Expanding polygenic risk scores to include gene-gene interactions

This manuscript (permalink) was automatically generated from lelaboratoire/rethink-prs-ms@76e826f on July 21, 2019.

Authors

- **Trang T. Le**☯
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Hoyt Gong**☯
  0000-0001-9339-4763 · hoytgong · GongHoyt
  Life Sciences and Management, Wharton School, University of Pennsylvania

- **Patryk Orzechowski**
  0000-0003-3578-9809
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Elisabetta Manduchi**
  0000-0002-4110-3714
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**†
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104 · Funded by National Institutes of Health Grant Nos. LM010098, LM012601, AI116794

☯ — These authors contributed equally to this work.

† — Direct correspondence to jhmoore@upenn.edu.

Abstract

This study expands the PRS to account for gene-gene interaction effects.

Introduction

As the field of traditional genomics rapidly expands its sequencing technologies and translational abilities, novel applications of genomic data are starting to arise in addressing disease burden. Beginning with the completion of the Human Genome Project in 2003, increased interest in cataloguing genomic data spurred the innovation of massively parallel, chip-based genotyping arrays. Leveraging these technologies, early researchers were able to characterize and categorize gene variants across millions of individuals internationally. In particular, the advent of projects such as the

International HapMap Project [1] and the 1000 Genomes Project sought to document haplotype [2] structure (i.e. gene variants) involved in specific diseases of the human genome. As such, the gross information of nucleotide polymorphisms within publicly available databases has rapidly increased in the beginning of the 21st century with the rise in omics sequencing capabilities. This genomic information, coupled with additional high resolution marks for other individual biological variants (e.g. transcripts, epigenetic marks, metabolites) has been touted to further drive precision medicine approaches using genetics.

Complementing the rapid growth in our understanding of gene variants in the human genome was the emergence of using statistical techniques, formalized as genome-wide association studies (GWAS), to identify gene variants associated with common human diseases. From a population perspective, GWA studies have sought to discern genetic connections to various phenotypes by studying genotypic variation at biallelic markers across the human genome [3,4,5]. Such non-candidate driven GWA studies consider gene variations (i.e. SNPs, deletions, intertions, CNVs) to resulting phenotype values to ultimately report allele frequency differences among a case and control group in the form of an odds ratio. This technical revolution in the field of genomic medicine fueled our progressing capabilities to map associations of gene variants with disease on an increasingly granular level to single nucleotide polymorphisms (SNPs).

In tandem with the movement towards precision medicine, the post-GWAS era strives to bring significant population-derived gene variants into individual level metrics actionable in health delivery settings. While GWA studies indeed capture gene variants associated with a phenotype of interest on a population level, translating such results to personalized individual metrics of risk requires additional granularity on aggregating contributions of many gene variants in the form of polygenic risk scores (PRS). Importantly, PRS provides an ability to explain inherited risk for disease in an individual by representing a weighted sum aggregate of risk alleles based on measured loci effect contributions derived from GWAS studies [6,7]. In quantifying the effect of particular combinations of genetic SNP variants towards risk prediction, PRS offers a probabilisitic susceptibility value of an individual to disease. Such genetic risk estimation scores are central to clinical decision-making, serving to reinforce individual health management in heritable disease detection and early prevention of various adult-onset conditions. The utility of PRS scores have been demonstrated in previous studies towards disease risk stratification across leading heritable causes of death in the developed world [10,11,8,9]

For each SNP i of an individual's genome of n possible SNPs for analysis, the PRS score is calculated via a summation across all significant SNPs as

$$PRS = \sum_{i=1}^{n} \beta_i \cdot SNP_i$$

where $/beta$ is the weighted risk contribution of the loci gene variant derived from risk score model parameters. Various approaches towards predicting risk of the same disease exist across PRS studies based on the above equation; models may vary according to the β weight according to the specific type of statistical model used to combine risk of individual variations, the n with respect to the specification of genetic variants considered, and the ability of the PRS to generalize to the entire population [6].

Historically, PRS models have previously characterized genomic architechture in a dichotimous division of Mendelian monogenic and polygenic inheritance, in which either one or many gene perturbations give rise to disease phenotypes in an individual, respectively [12,7]. Yet while such classification models arose from former sequencing technology and study design, a more realistic genetic archtechture of common adult-onset disease acknowledges dynamic interactions among a continuum of common low-risk to rare high-risk gene variants to cumultatively drive overall risk of an

individual [13]. When only considering rare (minor allele frequency, MAF < 0.5%), high-risk gene variants, such genomic variation only contributes approximately 1-10% towards adult-onset disease incidence [14,7]. Often, more relevant and complete sources of genetic risk is captured from complex smaller interactions of both common (MAF > 5%) and low-frequency (MAF > 0.5% and < 5%) genetic variants each contributing individual, appreciable effects [15,16]. Existing standard multivariate categorical data analysis approaches fall short in handling such enormous possible gene interaction combinations with both linear and nonlinear effects. In this context, more robust and efficient methods towards a polygeneic risk calculation are necessary in capturing the overlap between context-dependent effects of both rare and common alleles on human genetic disorder.

With respect to better understanding the epistasis across an individual's genome, various statistical models have been designed with the intent of capturing high dimensional gene-gene (GxG) interactions. The Multifactor Dimensionality Reduction (MDR) method is one such nonparametric, model-free framework that addresses these challenges and has been extensively applied to detect nonlinear complex GxG interactions associated with individual disease [17]. By isolating a specific pool of genetic factors from all polymorphism and cross-valiating prediction scores averaged across identified high risk multi-locus genotypes, the original MDR approach is able to categorize multi-loci genotypes into two groups of risk based on some threshold value. While created with the primary intention towards GxG interaction detection, the MDR model has additionally demonstrated applicability as a risk score calculation model in constructing PRS scores [18].

Modifications built on top of the MDR framework have been proposed in order to better capture multiple significant epistasis models and potential missed interactions owning to limiations of the original model in the higher dimensions. Model-Based Multifactor Dimensionality Reduction (MB-MDR) was formulated as a flexible GxG detection framework for dichotomous traits and unrelated individuals [19]. Rather than a direct comparison against a threshold level in the original MDR method, MB-MDR merges multi-locus genotypes exhibiting significant High or Low risk levels through association testing and adds an additional 'No evidence of risk' categorization.

In the present work, we aim to reformulate the PRS using the MB-MDR approach to better capture epistatic gene interactions of individual disease risk in a novel Multilocus Risk Score (MRS). In observing prediction accuracy results on an evidence-based simulated dataset from HIBACHI, we demonstrate the improved performance of our epistasis enriched MRS towards characterizing more granular disease etiology.

Methods

Multifactor Dimensionality Reduction (MDR) and model-based MDR (MB-MDR)

MDR is a nonparametric method that detects multiple genetic loci associated with a clinical outcome by reducing the dimension of a genotype dataset by pooling multilocus genotypes into high-risk and low-risk groups [17]. Extended from the original MDR algorithm, MB-MDR addresses existing limitations of MDR by increasing detectability of important interactions and decreasing bias by allowing O labels for individuals with no evidence for abberant risk. Several improvements have been made to MB-MDR since it was first introduced in 2009, and its current implementation efficiently and effectively detects multiple sets of significant gene-gene interactions in relation to a trait of interest while efficiently controlling type I error rates.

Besides the P values associated with each genotype combination, another important output of MB-MDR is the HLO matrices generated from the affected- and unaffected-subjects matrices (in the case of binary outcome). Briefly, for each genotype combination, an HLO matrix is a 3 x 3 matrix with each

cell containing H (high), L (low) or O (no evidence), indicating risk of an individual whose genotype pairs fall into that cell [20]. For an example binary outcome problem, a genotype combination SNP_1 and SNP_2 will have a χ^2 value, an associated P value and an HLO matrix that looks like

	$SNP_1 = 0$	$SNP_1 = 1$	$SNP_1 = 2$
$SNP_2 = 0$	O	O	O
$SNP_2 = 1$	O	H	L
$SNP_2 = 2$	O	L	H

We discuss in the following subsection how these values were utilized in the formulation of the Multilocus Risk Score (MRS).

[More on significance of SNP combination vs. significance of H/L/O here...]

Multilocus Risk Score (MRS)

We apply the MB-MDR software [20] v.4.4.1 to simulated datasets of n individuals, p SNPs to obtain the significance level of each combination of SNPs. We let k_d denote the number of significant combinations. In this study, no significance threshold is imposed at the SNP combination level and, thus, k_d reaches its maximum value of C_p^d.

For each subject i ($i = 1, 2, \cdot, n$), the d-way interaction risk score is calculated as

$$MRS_d(i) = \sum_{j=1}^{k_d} \chi_j^2 \times \mathrm{HLO}_j(X_{ij})$$

where χ_j^2 is the test statistic of each genotype combination j from a χ_j^2 test with one degree of freedom for the simulated binary trait, X_{ij} is the j^{th} genotype combinations of subject i and HLO_j represents the j^{th} recoded HLO matrix (1 = High, -1 = Low, 0 = No evidence). As an example, consider a pair $X_{*j} = (SNP_{j_1}, SNP_{j_2})$ with $\chi_j^2 = 8.3$ and corresponding HLO matrix of all O's except an L in the first cell. Then, all subjects' current risks would remain the same except the ones with $SNP_{j_1} = SNP_{j_2} = 0$ where their risks are subtracted by 8.3.

The final MRS score is the sum of all MRS_d for all d up to $\bar{d}$:

$$MRS(i) = \sum_{d=1}^{\bar{d}} MRS_d(i)$$

In this study, we consider 1-way and 2-way interactions, i.e. $\bar{d} = 2$, and hence, the combined risk is simply the total of the first two: $MRS = MRS_1 + MRS_2$.

Mutual information and information gain

We apply entropy-based methods to measure how much information about the phenotype is due to either marginal effects or the synergistic effects of the variants after subtracting the marginal effects. A dataset's main effect (i.e. marginal effect ME) can be measured as the total of mutual information between each genotype SNP_j and the phenotypic class y based on Shannon's entropy H [21]:

$$ME = \sum_j^k I(SNP_j; y) = \sum_j^k H(y) - H(y|SNP_j).$$

We measure the 2-way interaction information (i.e. degree of synergistic effects of genotypes on the phenotype) of each dataset by summing the pairwise information gain between all pairs of genetic attributes. Specifically, if we let X_j denote the j^{th} genotype combination (SNP_{j_1}, SNP_{j_2}), the total 2-way interaction gain (i.e. synergistic effects SE) is calculated as

$$SE = \sum_{j}^{k} IG(X_j; y) = \sum_{j}^{k} I(SNP_{j_1}, SNP_{j_2}; y) - I(SNP_{j_1}; y) - I(SNP_{j_2}; y),$$

where IG measures how much of the phenotypic class y can be explained by the 2-way epistatic interaction within the genotype combination X_j. We refer the reader to Ref. [22] for more details on the calculation of the entropy-based terms.

Simulated data

[Patryk...]

[objective: simulate a diverse collection of datasets]

For each simulated and real-world dataset, after randomly splitting the entire data in two smaller sets (80% training and 20% holdout), we built the MRS model on training data to obtain the χ^2 coefficients and calculated risk score for each individual in the holdout set. We assess the performance of the MRS by comparing the area under the Receiving Operator Characteristic curve (auROC) with that of the standard GRS method.

Manuscript drafting

This manuscript is collaboratively written using the Manubot software which supports open paper writing via GitHub with Markdown [23]. Manubot uses continuous integration to monitor changes and automatically update the manuscript. As a result, the latest version of this manuscript is always available at https://lelaboratoire.github.io/rethink-prs/.

Availability

Detailed simulation and analysis code needed to reproduce the results in this study is available at https://github.com/lelaboratoire/rethink-prs-ms/.

Results

MRS outperforms standard PRS

Figure 1: MRS produces improved auROC in the majority (335 green lines) of the 450 simulated datasets (each line represents a dataset). In many datasets, the original method performs poorly (auROC < 60%) while the new method yields auROC over 90%. This improvement in performance can be seen at the second peak (~50% auROC increase) in the density of the difference between two methods (right).

To assess whether this improvement in performance correlates with , in the following section, we untangled the two components in MRS and apply information theory to

Assess improvement in performance

As the amount of main effects increases (Fig. 2 left), MRS1 increasingly performs better than PRS, which is likely because encodings are inferred (top left). Meanwhile, MRS2's accuracy remain similar to that of PRS (middle left). On the other hand, when the amount of interaction effects increases (right), MRS1 performs mostly on par to PRS while MRS2 Combinging the gain from both MR1 and MRS2, MRS's performance progressively increases compared to the standard PRS.

Figure 2: Combining 1-way (MRS1) and 2-way (MRS2) risk scores, MRS shows increasing outperformance to standard PRS as dataset contains more main and interaction effects

References

1. **The International HapMap Project** *Nature* (2003-12) https://doi.org/dgd
DOI: 10.1038/nature02168 · PMID: 14685227

2. **An integrated map of genetic variation from 1,092 human genomes** *Nature* (2012-10-31) https://doi.org/f4k2v2
DOI: 10.1038/nature11632 · PMID: 23128226 · PMCID: PMC3498066

3. **Chapter 11: Genome-Wide Association Studies**
William S. Bush, Jason H. Moore
PLoS Computational Biology (2012-12-27) https://doi.org/gfr9pz
DOI: 10.1371/journal.pcbi.1002822 · PMID: 23300413 · PMCID: PMC3531285

4. **Genome-wide association studies for common diseases and complex traits**
Joel N. Hirschhorn, Mark J. Daly
Nature Reviews Genetics (2005-02) https://doi.org/bhcc36
DOI: 10.1038/nrg1521 · PMID: 15716906

5. **Genome-wide association studies: theoretical and practical concerns**
William Y. S. Wang, Bryan J. Barratt, David G. Clayton, John A. Todd
Nature Reviews Genetics (2005-02) https://doi.org/fcqz33
DOI: 10.1038/nrg1522 · PMID: 15716907

6. **What Are Polygenic Scores and Why Are They Important?**
Leo P. Sugrue, Rahul S. Desikan
JAMA (2019-05-14) https://doi.org/gfx8wg
DOI: 10.1001/jama.2019.3893 · PMID: 30958510

7. **The personal and clinical utility of polygenic risk scores**
Ali Torkamani, Nathan E. Wineinger, Eric J. Topol
Nature Reviews Genetics (2018-05-22) https://doi.org/gd46bh
DOI: 10.1038/s41576-018-0018-x · PMID: 29789686

8. **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder** *Nature* (2009-07-01) https://doi.org/cb5f2j
DOI: 10.1038/nature08185 · PMID: 19571811 · PMCID: PMC3912837

9. **Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting**
Pradeep Natarajan, Robin Young, Nathan O. Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F. Reilly, Adam Butterworth, … Sekar Kathiresan
Circulation (2017-05-30) https://doi.org/gbgqhb
DOI: 10.1161/circulationaha.116.024436 · PMID: 28223407 · PMCID: PMC5484076

10. **Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States**
Paige Maas, Myrto Barrdahl, Amit D. Joshi, Paul L. Auer, Mia M. Gaudet, Roger L. Milne, Fredrick R. Schumacher, William F. Anderson, David Check, Subham Chattopadhyay, … Nilanjan Chatterjee
JAMA Oncology (2016-10-01) https://doi.org/gf4w2z
DOI: 10.1001/jamaoncol.2016.1025 · PMID: 27228256 · PMCID: PMC5719876

11. **Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts**
Tyler M Seibert, Chun Chieh Fan, Yunpeng Wang, Verena Zuber, Roshan Karunamuni, J Kellogg Parsons, Rosalind A Eeles, Douglas F Easton, ZSofia Kote-Jarai, Ali Amin Al Olama, ...
BMJ (2018-01-10) https://doi.org/gcsrzs
DOI: 10.1136/bmj.j5757 · PMID: 29321194 · PMCID: PMC5759091

12. **Beyond Mendel: an evolving view of human genetic disease transmission**
Jose L. Badano, Nicholas Katsanis
Nature Reviews Genetics (2002-10) https://doi.org/cjfzf2
DOI: 10.1038/nrg910 · PMID: 12360236

13. **The continuum of causality in human genetic disorders**
Nicholas Katsanis
Genome Biology (2016-11-17) https://doi.org/f9fzd3
DOI: 10.1186/s13059-016-1107-9 · PMID: 27855690 · PMCID: PMC5114767

14. **Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits**
Zhihong Zhu, Andrew Bakshi, Anna A.E. Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, ... Jian Yang
The American Journal of Human Genetics (2015-03) https://doi.org/f64772
DOI: 10.1016/j.ajhg.2015.01.001 · PMID: 25683123 · PMCID: PMC4375616

15. **Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data**
Huwenbo Shi, Gleb Kichaev, Bogdan Pasaniuc
The American Journal of Human Genetics (2016-07) https://doi.org/f8xxkd
DOI: 10.1016/j.ajhg.2016.05.013 · PMID: 27346688 · PMCID: PMC5005444

16. **Common SNPs explain a large proportion of the heritability for human height**
Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, ... Peter M Visscher
Nature Genetics (2010-06-20) https://doi.org/fjjm4v
DOI: 10.1038/ng.608 · PMID: 20562875 · PMCID: PMC3232052

17. **Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer**
Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, Jason H. Moore
The American Journal of Human Genetics (2001-07) https://doi.org/bh3x75
DOI: 10.1086/321276 · PMID: 11404819 · PMCID: PMC1226028

18. **Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction**
Hongying Dai, Richard J Charnigo, Mara L Becker, J Steven Leeder, Alison A Motsinger-Reif
BioData Mining (2013-01-08) https://doi.org/gb5fmn
DOI: 10.1186/1756-0381-6-1 · PMID: 23294634 · PMCID: PMC3560267

19. **Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data**

Jestinah M Mahachie John, François Van Lishout, Kristel Van Steen
European Journal of Human Genetics (2011-03-16) https://doi.org/bndfnk
DOI: 10.1038/ejhg.2011.17 · PMID: 21407267 · PMCID: PMC3110049

20. **An efficient algorithm to perform multiple testing in epistasis screening**
François Van Lishout, Jestinah M Mahachie John, Elena S Gusareva, Victor Urrea, Isabelle Cleynen, Emilie Théâtre, Benoît Charloteaux, Malu Luz Calle, Louis Wehenkel, Kristel Van Steen
BMC Bioinformatics (2013-04-24) https://doi.org/f4v3n7
DOI: 10.1186/1471-2105-14-138 · PMID: 23617239 · PMCID: PMC3648350

21. **A Mathematical Theory of Communication**
C. E. Shannon
Bell System Technical Journal (1948-07) https://doi.org/b39t
DOI: 10.1002/j.1538-7305.1948.tb01338.x

22. **Epistasis Analysis Using Information Theory**
Jason H. Moore, Ting Hu
Methods in Molecular Biology (2014-11-03) https://doi.org/gf484q
DOI: 10.1007/978-1-4939-2155-3_13 · PMID: 25403536

23. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653