Expanding polygenic risk scores to include gene-gene interactions

This manuscript (permalink) was automatically generated from lelaboratoire/rethink-prs-ms@1d6dde4 on July 10, 2019.

Authors

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

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

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

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

- **Jason H. Moore**[†]
  iD 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

[1]

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 catalouging genomic data spurred the innovation of massively parallel, chip-based genotyping arrays. Leveraging these technologies, early researchers were able to characterize and catalogue gene variants across millions of individuals internationally. In particular, the advent of projects such as the International HapMap Project [2] and the 1000 Genomes Project sought to document haplotype [3] 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 complement 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 [4,5,6]. 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).

Nonetheless 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). In tamdem with the movement towards precision medicine, the post-GWAS era strives to bring significant population-derived gene variant into individual level metrics actionable in clinical delivery settings. Importantly, PRS have been one such approach developed to explain individual inherited risk for disease by **placing unique weights on a selection of SNPs from the GWAS**.

[put in PRS equation].

[This is just one way, the most basic. many have tried to reformulate the PRS in various ways.]

[cite common ones. Which ones are most common GRS scores?]

[In this study we aim to reformulate the PRS score with MDR reduction to better detect GxG interactions. MDR as a form of feature engineering the proper encoding for detecting GxG interactions]

[explain more technical details of MDR]

END OF INTRO -> rest is all methods & performing MDR

Methods

Multilocus Risk Score (MRS)

Compute risks from significant interactions i = 1…n subjects p SNPs j = 1…k significant combinations We apply the software… [7] to obtain the significance level of each combination of SNPs. allow for parallel computation The maximum value of k is C_p^d. For each subject i, the d-way interaction risk score is calculated as

$$R_d(i) = \sum_{j=1}^{k} \chi_j^2 \times HLO_j(X_{ij})$$

where χ_j^2 is the test statistic of each multi-locus combination j from a χ_j^2 test with one degree of freedom for the simulated binary trait, HLO_j is the j^{th} re-coded HLO-matrix and X_j is one of k combination of SNPs.

Simulated data

[Patryk…]

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 where

Results

Information gain

iPRS outperforms standard PRS

MM12 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).

References

1. **On the utilization of polygenic risk scores for therapeutic targeting**
Greg Gibson
PLOS Genetics (2019-04-25) https://doi.org/gf4gdx
DOI: 10.1371/journal.pgen.1008060 · PMID: 31022172 · PMCID: PMC6483161

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

3. **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

4. **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

5. **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

6. **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

7. **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