Weighted Polygenic Risk Scores and Non-Negative Matrix Factorization

Xavier Loffree

Weighted Polygenic Risk Scores

Project with Lai Jiang, Celia Greenwood

Goal

- Improve performance of polygenic risk scores by determining a weight for each single nucleotide polymorphism and then using those weights when calculating polygenic risk scores

Polygenic Risk Scores (PRS)

- Estimate risk of an individual developing an heritable disease
- GWAS use regression to determine association of each SNP with particular disease
- PRS sum the strength of the association at a set of associated SNPs:

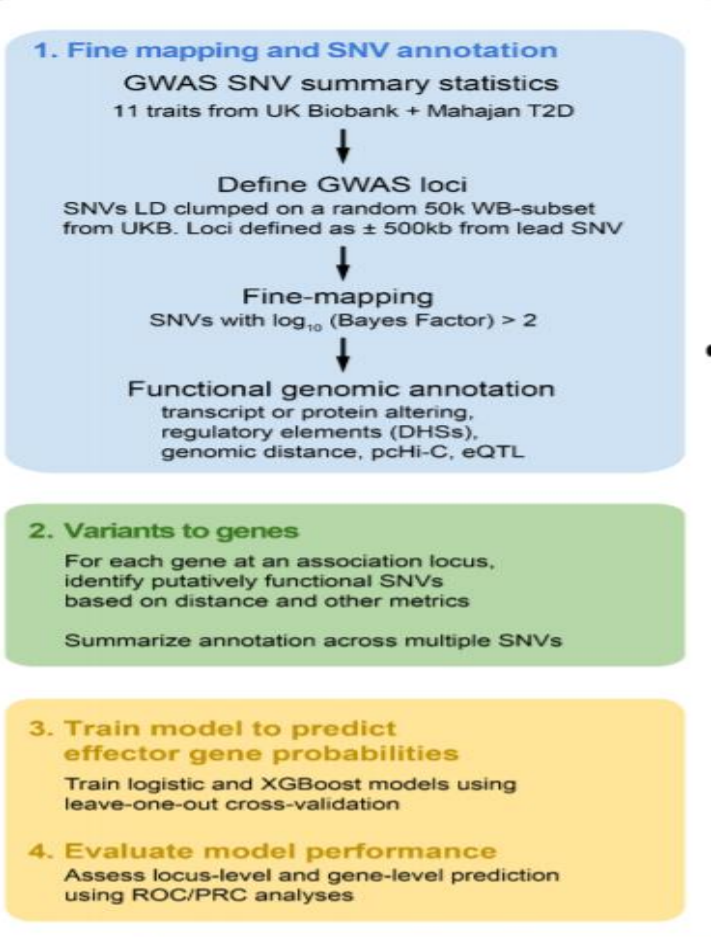
$$PRS_i = \sum_j^n \beta_j * quantity_{ij}$$

Effector Index (EI) Project

- Predict which is the most likely causal gene in a region showing GWAS associations at many SNPs and across several genes.
- Use various genomic features and annotations in a (boosted tree) model predicting gene most likely to be causal.

Forgetta, V., Jiang, L., Vulpescu, N. A., Hogan, M. S., Chen, S., Morris, J. A., Grinek, S., Benner, C., Jang, D.-K., Hoang, Q., Burtt, N., Flannick, J. A., McCarthy, M. I., Fauman, E., Greenwood, C. M. T., Maurano, M. T., & Richards, J. B. (2020). An effector index to Predict CAUSAL genes AT GWAS Loci. https://doi.org/10.1101/2020.06.28.171561

EI Project Flow Chart

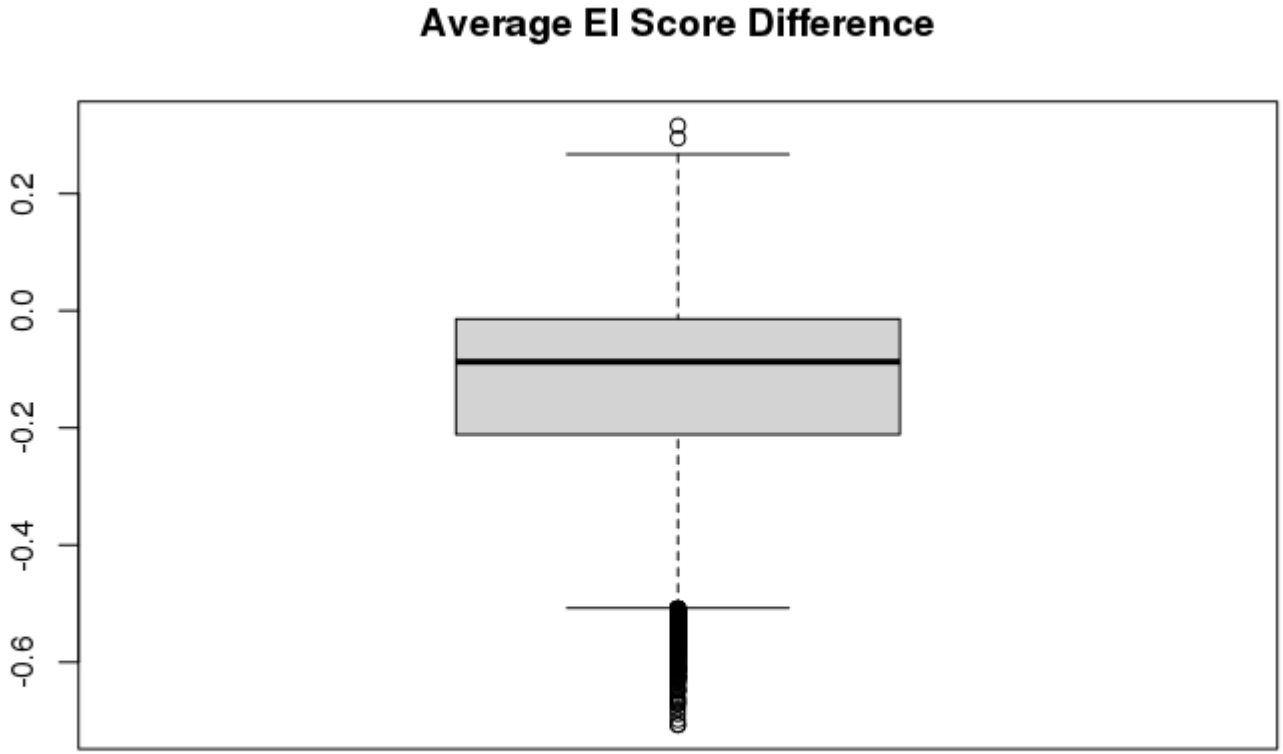


Our Project: Weighted PRS

- Incorporate the causality predictions from the EI project
- Use these results to determine weights for SNPs
- SNPs with high predictions for being important in the Ei algorithm are given greater weights, less important are given lower weights
- Once the weights are incorporated, this should result in a more accurate PRS

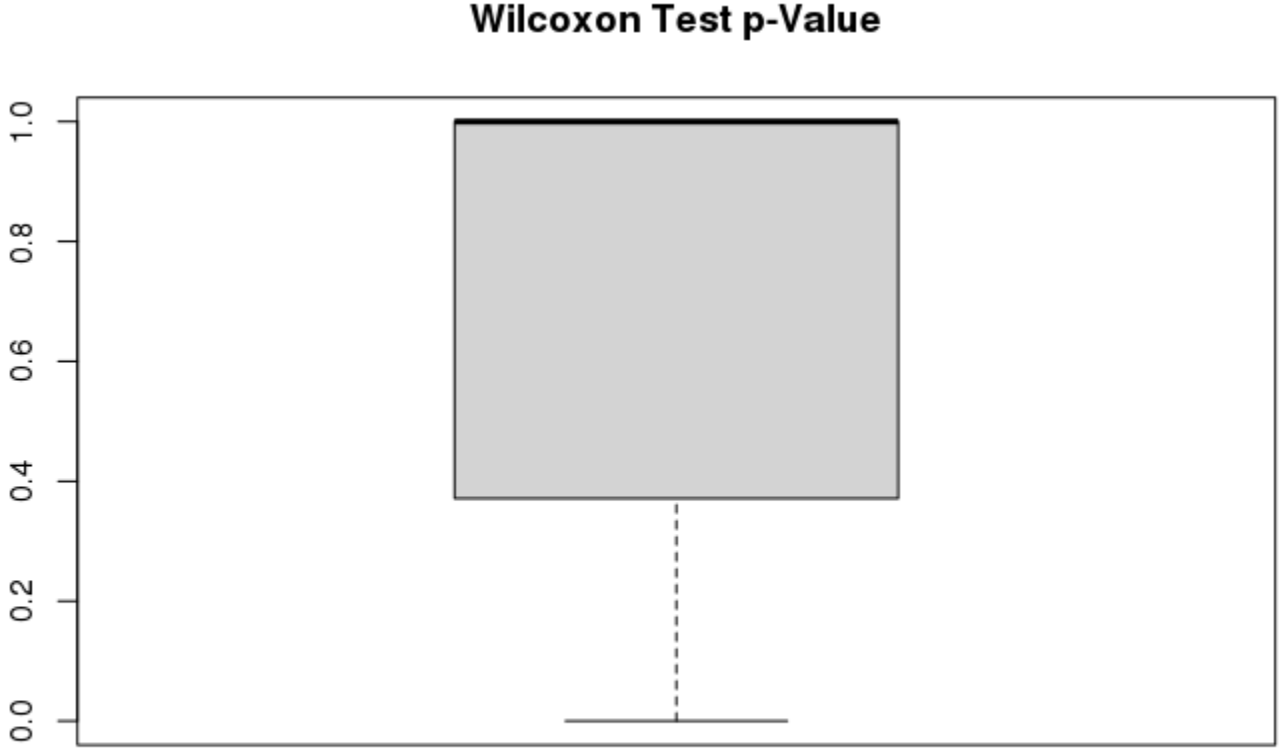
Converting Gene-Specific EI Scores to SNP-Specific EI Score

- Problem: EI predicts causal genes, but we are looking for causal SNPs
- Employed a leave-one-out algorithm to arrive at SNP-specific values from the EI algorithm
- Ran the algorithm with each SNP removed, then compared the EI values when the SNP was included in the algorithm and when it was not
- Efficiency was a challenge at first


Average EI Score Difference
0.2
0.0
-0.2
-0.4
-0.6
8

Analyzing the Change

- Converted EI scores into rankings within the locus for each SNP-gene combination
- Used a Wilcoxon test to compare the EI score rankings of each SNP-gene combination within a locus before and after the new EI scores were calculated
- Wilcoxon tests give us p-values that can be used to check if the ranking change is significant


Wilcoxon Test p-Value
0.0
0.2
0.4
0.6
0.8
1.0

Next Steps

- Conduct EI-weighted PRS score for each of 12 traits associated with T2D used in the EI project
- Use UKBB data and SNPs with p-value less than 0.05 from Wilcoxon test

Non-Negative Matrix Factorization

Project with Irene Zhang, Celia Greenwood

Introduction

- Udler et al. used non-negative matrix factorization (NMF) to explore soft clustering of the SNPs that are associated with T2D
- Also finds extent to which SNPs are associated with T2D
- Approximates a matrix Z with a lower rank approximation WH
- Columns of phenotypes
- Rows of SNPs
- Entries of beta values of association with T2D

$$X \approx WH$$

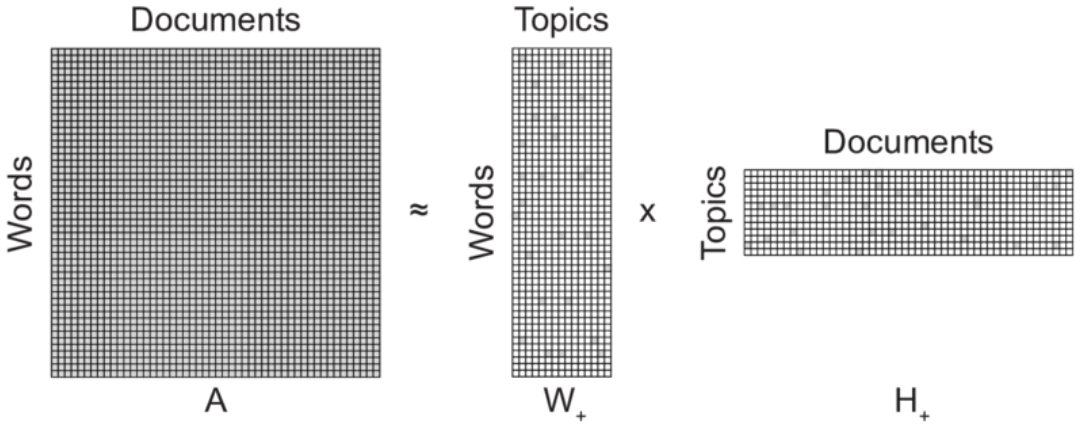
Goal

- Replicate Udler's study on a larger scale
- Study the clustering structure of the risk alleles for a particular disease
- Determine genes most associated with a particular disease
- Improve upon accuracy of a typical PRS

Crossover with EI Project

- Compare SNPs most associated with T2D found in EI project and in NMF project
- Similar results should validate both methods
- Also important to know if and why results differ

Uses of NMF



- Clustering
- Dimension reduction
- Imputation
- Recommendation algorithms for music, movies, etc.

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime topic modeling. Crime Science, 6(1). https://doi.org/10.1186/s40163-017-0074-0

Data Collection

- Searched in UKBB Showcase for traits relevant to T2D that were used in Udler's paper
- Included all SNPs with p-value less than 0.05 from previous studies by the Broad Institute
- Number of SNPs was too large (find number)
- Used a small subset of the total SNPs to test the NMF algorithm by only included the SNPs that were common across all 12 traits

Preparing the Test Matrix

- Created matrix with traits as the columns, snps as the rows, and beta coefficients as the entries
- NMF algorithm does not accept negative values
- For each trait, have a column filled with the absolute values of the negative entries (all positive entries replaced with zeroes) and another column with the positive entries (all previously negative entries replaced with zeros)
- Have yet to interpret results

Next Steps

- Conduct NMF on large scale with all relevant SNPs from Broad Institute with p-value less than 0.05
- Conduct NMF on a matrix with entries equal to the EI algorithm reweighted PRS coefficients
- See if this provides new findings

Thank You!

- Special thank you to Lai Jiang, Irene Zhang, Celia Greenwood, Tianyuan Lu, Kathleen Klein, Yixiao Zeng, Jeffrey Yu