Comparison of predictive abilities and model performance of biotx.ai's algorithmic pipeline with Polygenic Risk Scores

Abstract / Executive Summary

Polygenic Risk Scores (PRS) are the industry and research standard for predicting traits based on genetic data, when taking into account variation in multiple genetic variants. For each individual a score is calculated based on the variants associated with a trait and their respective impact on the trait. This paper compares PRS to biotx.ai, a novel approach that uses data-mining of contextual information to create polygenic disease models. We show that biotx.ai outperforms PRS in terms of ROC curve, explanatory value and requires considerably less data for analysis.

Comparison PRS / biotx.ai

Polygenic Risk Scores

In human genetics Polygenic Risk Scores are calculated by computing the sum of risk alleles corresponding to a phenotype of interest in each individual, weighted by the effect size estimate of the most powerful GWAS on the phenotype. Studies have shown that substantially greater predictive power can be achieved by using PRS rather than a small number of genome-wide significant SNPs (Choi et al., 2018). Even though PRS compound a large number of SNPs, the score is entirely additive that is each SNPs contribution to the score is considered separately, complex interactive patterns are not taken into account (see Box 1 below).

Box 1. Additive versus interactive modeling

The name 'Polygenic Risk Score' itself suggests that PRS models polygenic interactive effects. This is, however, not the case. In PRS, the separate effects of a large number of SNPs are added into one risk score. This does not capture any true interactions in which the effect of a SNP is contingent on the presence of one or more other variants, e.g. *rs_XXX (G/G)* leads to an increased risk for developing Diabetes, but only if *rs_YYY (A/C)* and *rs_ZZZ (T/T)* are present.

Biotx.ai

Biotx.ai's approach consists of two complementary modules that are in feedback with each other. The contextual module uses information mined from the scientific literature, pathway libraries and protein co-expression data and an evaluation module that estimates predictive power of a feature based on that contextual information with extreme computational efficiency When testing for monogenic effects the number of features in a GWAS equals the number of SNPs, but when testing for polygenic interactions, the number of features is the number of SNPs exponentiated with the complexity of the interaction one tests for. This can easily lead to a feature space that is so large that it becomes unmanageable. A combination of using the biological contextual information and computationally efficient algorithms (over 80,000 times as fast as logistic regression) allows for a massive reduction of the feature space and time required to scan it for association hypotheses.

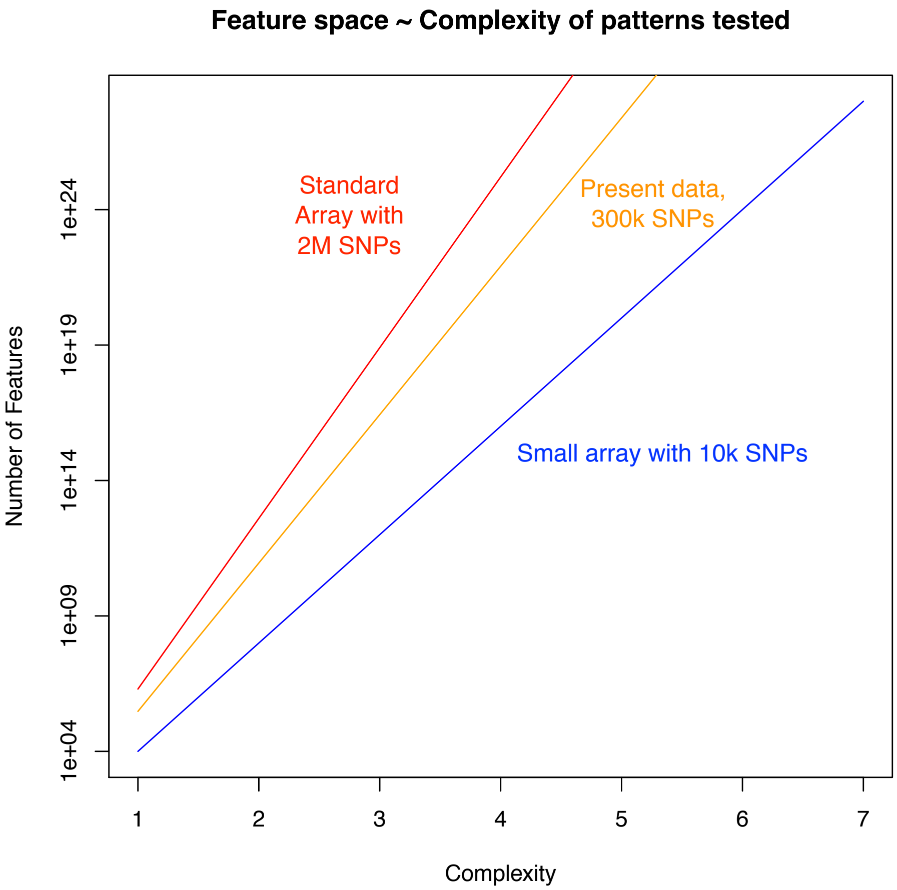

Figure 1. Logarithmic representation of the number of features as a function of array size (number of variants) and complexity of the interactions tested for. Testing for complex patterns with standard methods leads to an unmanageably large feature space even for small arrays.

The final result is a manageable set of polygenic hypotheses, each containing up to 7 SNPs that, in interaction, better predict a given phenotype. In contrast to PRS, which merely adds the effects of different SNPs, biotx.ai models true interactions (see Box 1).

Table 1. Key differences of Polygenic Risk Scores (PRS) and biotx.ai

Polygenic Risk Scores	**biotx.ai**
Effects of multiple SNPs are only accounted for only by addition	True modeling of interactions
Can use summaries from previous GWAS (which results in overfitting)	Incorporates contextual information and previous research via NLP
Uses p-value threshold for inclusion of SNPs	True predictive power calculated via bootstrapping over validation data
Tens of thousands of SNPs required for predictions, functional relations are opaque	Produces concise polygenic term with verifiable functional insights

Data

The Parkinson's Progression Marker Initiative, short PPMI, data (https://www.ppmi-info.org) was used for the comparison. The data set contains 471 subjects, 368 cases, 152 controls, for each subject genotyping information from two complementary chips (NeuroX and ImmunoChip) was collected. After careful quality control and harmonization we merged that information into a single dataset with 380,939 variants in total. An additional set of quality control steps were performed on variants and individuals (see Supplementary Methods) filtering for excessive missingness in genotyping across SNPs and individuals, keeping only autosomal SNPs, filtering SNPs by their Hardy-Weinberg equilibrium, removing individuals with abnormal heterozygosity rates and finally removing individuals that were related.The final quality controlled dataset contained 369,036 variants and 436 individuals passing the various filters. The data was then split into training, validation and test sets. The exact same sets were used for both methods, PRS and biotx.ai. For a more detailed description of the data preparation please see Supplementary Materials 1.

Method

For PRS results of a GWAS conducted with PLINK (Purcell et al., 2007) were used to calculate PRS for 4,195 different p-value thresholds for the subjects in the training, validation and test set. The PRS of the subjects in the training set were then used to train a separate logistic regression classifier for each p-value threshold. The validation data set was used to determine which p-value threshold produces the best classifier and this classifier was then used to predict the test set. This classifier is based on the PRS of 16,135 different SNPs.

For biotx.ai the training set was used to generate and filter hypotheses. After 75,000 iterations the most performant 500 polygenic hypothesis were obtained. The validation set was then used to further reduce this to less than 100 hypotheses. The remaining hypotheses were summarized in a term that was used to train a LASSO (Tibshirani, 1996, Hastie and Qian, 2014) regression model on the training data. This model, based on less than 50 SNPs in interaction, then predicted the test set.

Results

The Receiver Operating Characteristics (ROC) were used to evaluate the quality of prediction of both approaches. Biotx.ai Area under Curve is 26 perentage points or 46% larger than that of the PRS with Sensitivity and Specificity being 19 percentage points (31%) and 24 percentage points (42%) higher at the Youden Point.

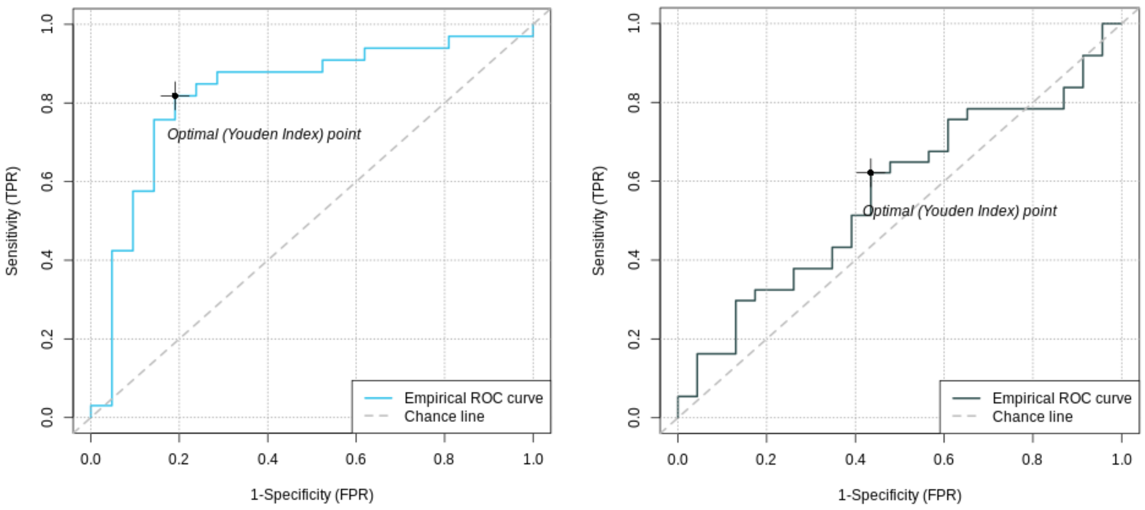

Figure 2. ROC Curve for biotx.ai (left) and PRS (right)

Table 2. Key Statistics for both methods at the optimal (Youden Index) point

Method	AUC	Accuracy	Misclassification	Sensitivity	Specificity
biotx.ai	0.82	0.81	0.19	0.81	0.80
PRS	0.56	0.6	0.40	0.62	0.56

Beyond prediction biotx.ai also provides insights about the genes associated with the disease. Due to its ability to detect true interactions between SNPs, 47 genes were associated with the disease that would not have shown any significance in a purely additive approach like PRS.

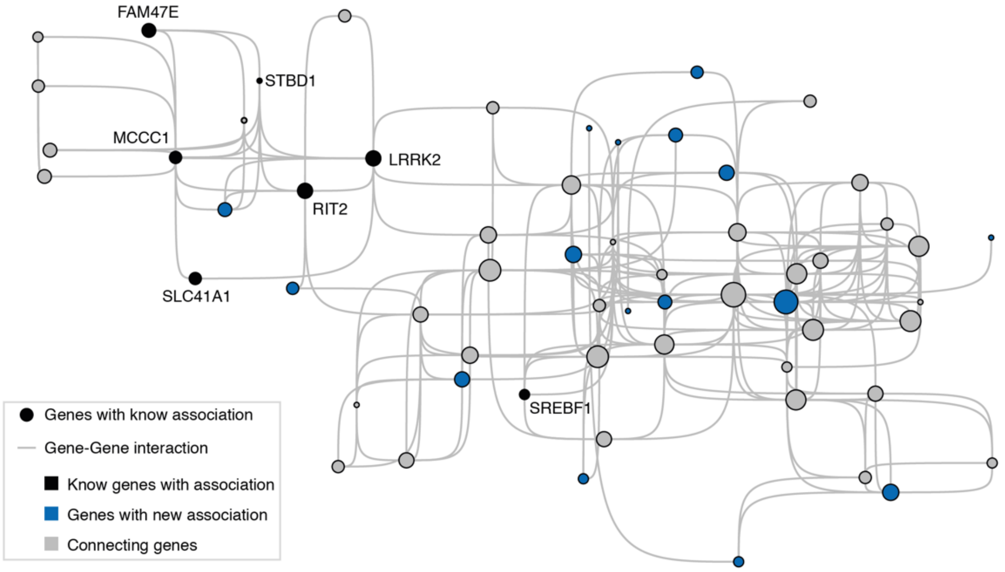
FAM47E
STBD1
MCCC1
LRRK2
RIT2
SLC41A1
SREBF1
● Genes with know association
— Gene-Gene interaction
■ Know genes with association
■ Genes with new association
■ Connecting genes

Figure 3. Contextualized information of genes with disease associations allows for meaningful biological interpretation of new discoveries. The graph shows a gene-gene interaction network where genes are represented as nodes and interactions between them as edges. Genes known to play a role in Parkinson's Disease are colored in black, Genes with novel association discovered by biotx.ai's algorithm are represented in blue. Intermediate

nodes in grey represent genes act as intermediary interactions. The context provided enables identify secondary genes that could play an important role in the disease.

Conclusion

Biotx.ai outperforms PRS by over 45%. Beyond that the approach is able to associate new variants with the diseases that would have not shown up under an additive approach such as PRS. PRS models also require a large set of SNPs which leads to overfitting and limits their use in clinical practice. Biotx.ai generates more parsimonious models which do not have such limitations – with less than 50 SNPs in interaction, biotx.ai was able to outperform the PRS model based on 16,135 SNPs. This advantage goes beyond genomics and applies to any area where small sample sizes meet large feature spaces.

References

Choi, S. W., Mak, T. S. H., & O'reilly, P. (2018). A guide to performing Polygenic Risk Score analyses. *BioRxiv*, 416545.

Hastie, T., & Qian, J. (2014). Glmnet vignette. *Retrieve from http://www web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed October*, *30*, 2019.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.