Identifying Gene Contribution to Neuropsychiatric Disease: PrediXcan X PheRS

Jessica E.H. Brown[1], Dan Zhou[1], Eric Gamazon[1,2], Haky Im[3], Nancy J. Cox[1,4]

[1]*Vanderbilt University*, [2]*University of Cambridge*, [3]*University of Chicago*, [4]*Vanderbilt Genetics Institute*

VANDERBILT UNIVERSITY

Background

The function of extracellular matrix-related genes is structural support and biochemical or biomechanical cues for cells or tissues. Based on results of studies of the phenotypic consequences, including attention deficit hyperactivity disorder and pervasive developmental delay, of loss of genes that traffic fibrillar collagens to enable them to perform their roles in extracellular matrix (Unlu et al., 2020 Nature Medicine), **we hypothesized that matrisome genes contribute to risk of neuropsychiatric disease in addition to other medical phenome**.

PredixVU

PredixVU is the output of data from applying PrediXcan to BioVU. PrediXcan is a method to predict gene expression based on cis-regulatory variation (figure 1). Vanderbilt Medical Center's EHR-linked biobank, BioVU, contains over 250,000 DNA samples, with over 120,000 samples which have had genome-wide genotyping and between 10,000 and 20,000 which have had either exome or whole genome sequencing.

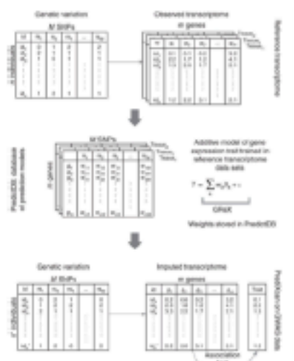

Figure 1. Workflow of PrediXcan from the input of the genetic variation of individuals through to an imputed transcriptome.

Neuropsychiatric Phenome Risk Scores (PheRS)

We identified the subset of matrisome genes with high quality prediction performance in at least one brain region (350 brain matrisome genes of 1,016 matrisome genes). Neuropsychiatric PheRS were calculated for ~10,000 European ancestry BioVU subjects for neuropsychiatric disease using all phecodes within the neuropsychiatric diagnoses. Each phecode was weighted based on the inverse prevalence of the code in BioVU. The weight for phenotype p is calculated as

$$w_p = log\frac{N}{n_p}$$

where N is number of individuals in the total population and n_p is the number of individuals with phenotype p. We applied this weighting so that individuals with a single or a few rare features of neuropsychiatric diseases have a higher PheRS than an individual with a single or a few more common features. For an individual i, the PheRS for a single disease defined by m phecodes was calculated by

$$PheRS_i = \sum_{p=1}^{m} w_p x_{i,p} \quad where\ x_{i,p} = \begin{cases} 1\ if\ individual_i\ has\ phenotype_p \\ 0\ otherwise \end{cases}$$

Individuals' values for neuropsychiatric PheRS range from 0 to 284. The distribution skews towards 0 (figure 2).

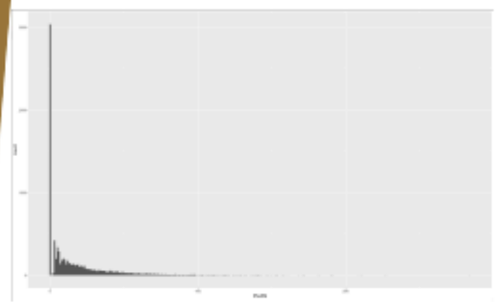

Figure 2. Distribution of neuropsychistric PheRS. The distribution is bimodal, showing two peaks, one of which is at 0.

Regression Modeling

We evaluated the association between PheRS and predicted gene expression – derived from PredixVU, the application of PrediXcan on Vanderbilt University's biobank BioVU – through the the following regression:

$$PheRS \sim rcs(predicted gene expression, 3) + gender + PC1 + PC2 + PC3$$

Using permutation analyses, we can set type I error thresholds depending on stringency requirements. Through these analyses, we observe the Bonferroni correction is too conservative (figure 3).

Forty matrisome genes show highly significant associations between neuropsychiatric PheRS and their predicted gene expression, including *ANGPT2* (p=6.3e-15), *FBN1* (p=2.5e-13), and *TNF* (p=2.1e-12).

Due to the skewed distribution of PheRS, we will be exploring the use of alternate regression models, including ordinal regression.

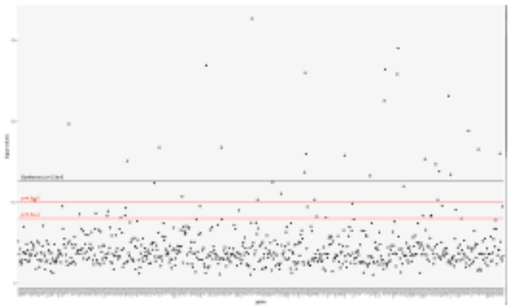

Figure 3. P-values showing significance of association between predicted gene expression and PheRS. Black line shows Bonferroni corrected p-value, while the red lines show alternative p-values determined through permutation tests.

Conclusions

Further refinement of targets will include ascertaining associations with multiple early onset phenotypes expected to be observable in zebrafish. Additionally, we are assessing the hypothesis that matrisome genes are more likely to be associated with neuropsychiatric phenotypes than random gene sets of the same size through gene set permutation tests. Through these analyses, specific genes or gene sets can be identified for further study in model organisms.