Genetic Risk Prediction with cis-eQTLs and Polygenic Risk Scores

Garvey Li
gjl001@ucsd.edu

Tiffany Amariuta-Bartell
tamariutabartell@ucsd.edu@ucsd.edu

Abstract

This paper was created with the intentions of utilizing cis-eQTLs to predict the the genetic risk of disease for an individual through what is known as a polygenic risk score (PRS). We conducted cis-eQTL analysis on various genes using the data of 489 individuals of European descent from the 1000 Genomes Project Dataset. The resulting cis-eQTL findings were used to construct PRS distributions for each gene. By examining these distributions, we identified correlations between genetic variations and gene expression levels, providing insights into an individual's susceptibility to genetic diseases.

Code: https://github.com/GarveyJLi/genetic_risk_prediction_with_prs

1 Introduction

Polygenic risk scores (PRS) can be thought of a weighted sum of an individuals genetic variation across their genome. They can be used to represent the probability that a phenotype or attribute (e.g. height, eye color, disease) will be expressed for that individual. PRS distributions are created from the analysis of eQTLs, which are single nucleotide polymorphisms (SNPs, a single variation in DNA when a nucleotide base is replaced by another). These distributions can be created using any number of SNPs across any number of genes relating to a attribute. For this study, however, we more specifically look at cis-eQTLs, or SNPs within a 1 Megabase window around a gene, as they typically show stronger associations with expression than SNPs outside the window (these SNPs are known as trans-eQTLs). To analyze the potential of these cis-eQTLs, their uses will be investigated in three parts: A genome wide analysis of cis-eQTLs, an analysis of PRS model performances on individual genes, and using Genome Wide Association Studies summary statistics to create a PRS model to predict a polygenic trait. Leveraging cis-eQTLs allows for a more precise assessment of the correlations between the genetic variations and expression, resulting in more precise PRS. Leveraging PRS can significantly enhance healthcare by enabling preventive measures. It would enable individuals to learn of any genetic predispositions to disease, potentially before any symptoms appear, thus facilitating earlier and more effective interventions.

2 Prior Work

Previous research on polygenic risk scores (PRS) and eQTL analysis has established a foundation for understanding the genetic basis of complex traits and disease risk. Studies such as Consortium (2009) and Khera et al. (2018) demonstrated the potential use cases of PRS in various polygenic diseases such as schizophrenia, type 2 diabetes, and breast cancer. These studies highlighted how aggregating genetic risk factors across the genome can provide insights into individual disease risk, even for complex conditions with multifactorial genetic and environmental influences. On a broader scale, Consortium (2020) analyzed RNA-seq samples and characterized genetic associations with both cis/trans eQTLs/sQTLs within the samples. Additionally, it was found that the associations with various QTLs and gene expression levels were influenced by various factors such as the type of cell or tissue the sample was from as well as the sex of the individual. Despite these advancements, challenges persist in PRS application. There is limited representation of diverse populations, which impacts the generalizability of PRS models. The data available (e.g. in GWAS, 1000 Genomes) is composed of a disproportionate amount of individuals of European ancestry in comparison to the global population (Martin et al. 2019). The non-representative sample has resulted in risk prediction accuracies to vary greatly between ethnicity groups, especially for those of African descent, limiting their predictive power for general populations. Furthermore, the results from a eQTL analysis and PRS can be easily misinterpreted by physicians and patients, which may cause harm to the patient depending on the condition in question and the patients demographics. Results from these diagnoses and the individual's data can also be used maliciously if it falls into the wrong hands (e.g. used to null

insurance policies, employment contracts, etc). With these issues in mind, our study seeks to build on this foundational work by focusing on cis-eQTL data specifically from the 1000 Genomes Project to generate PRS distributions at the gene level. By targeting gene-specific risk assessments, we aim to refine PRS accuracy for predicting individual disease susceptibility, addressing gaps in population diversity and moving towards a more personalized approach to genetic risk assessment.

3 Data

The genotype data used in our study primarily comes from the 1000 Genomes Project (Consortium et al. 2015), and consists entirely of male individuals with European ancestry. In total, there are 1,190,321 SNPs recorded for each of the 489 individuals sampled. These SNPs span all chromosomes in the human genome and have information pertaining to chromosome, unique identifiers, position (in base pairs), the reference allele (original nucleotide base), and the effect allele (variant nucleotide base). Metadata regarding the individuals contains unique identifiers, parental information, gender, and trait. Most importantly, the 1000 Genomes data contains the genotype for each individual and SNP, with values 0, 1, and 2 denoting 0 variant alleles (or base pairs), 1 variant allele, and 2 variant alleles (since there are 2 copies of every chromosome). Another dataset we will be using is a phenotype dataset containing 23,722 gene expression levels for most of the individuals in the 1000 Genomes data. By conducting cis-eQTL analysis on the SNP data, we are able to construct PRS for each individual and gene as predictions, and compare them to the gene expression levels in the phenotype dataset as our ground truth. In addition to genotype and phenotype data, we will be using GWAS summary statistic data that pertains to an individuals height, along with a single data sample for a European female individual, allowing us to make a basic evaluation of the resulting PRS model.

4 Methods

To calculate the effect sizes of each cis-eQTL for a gene, an eQTL analysis was run on the SNPs within a 1 Mb window around the center (center +/- 500 kb) of the gene. For each cis-eQTL found, a linear regression model of the following form is fit:

$$\alpha + genotype \times \beta = phenotype$$

where $genotype$ is the number of variant alleles for an individual, $phenotype$ is the level of gene expression for that individual, and β is the effect size of the SNP. This will be done across all cis-eQTLs present in the 1000 Genomes dataset, which will allow us to determine which SNPs contain unique information about genetic associations by using various methods. Two methods we will consider is p-value thresholding and linkage disequillibrium (LD) clumping. The thresholding is done by setting a p-value proportionate to the number

of SNP-gene pairs for normalization, and keeping only the SNPs whose effect size p-values are less than that p-value. As for LD clumping, the most significant SNP (usually the one with the smallest p-value) is compared to all other significant SNPs for a gene using the R^2 correlation coefficient. A R^2 threshold with range (0, 1] is then set, and the SNPs with R^2 value above the threshold are omitted, as we want to omit any cross-correlated SNPs.

To create a PRS with a set of analyzed SNPs, the collection of SNPs is must be pruned and thresholded using the p-value and LD clumping methods mentioned above. If this is not done, SNPs that are nearby each other may be highly associated with one another and introduce multicollinearity into the model. With the pruned and thresholded SNP effect sizes, the PRS for each individual can be calculated using the following equation:

$$PRS_j = \frac{\sum_i^N S_i \times G_{ij}}{P \times M_j}$$

Table 1: Descriptions of symbols used in the PRS calculation.

Symbol	**Description**
j	Individual
i	SNP
S_i	SNP effect size (β)
G_{ij}	Number of variant alleles observed
P	Ploidy of the sample (2 for humans)
M_j	Number of non-missing SNPs observed in the sample

To predict the phenotype of an individual using this PRS distribution, a PRS would need to be calculated for the individual using the effect sizes calculated in the cis-eQTL analysis. The score would then be compared to the PRS distribution, and can be categorized as high or low risk based on a threshold that maximizes prediction accuracy. Having thresholds for the PRS distribution and the known gene expression levels used in training to generate classes would allow us to directly measure the predictive power of the model. However, as is the case with the dataset acquired from the 1000 Genomes Project, the gene expression levels have no known threshold and cannot be split into distinct categories, so a different approach must be taken. The overall performance of a PRS distribution can be determined using a goodness of fit metric between the PRS values and gene expression levels for each individual, such as the R-squared correlation coefficient.

It should also be noted that using the entire dataset to create the PRS distribution and choose hyperparameters may result in overfitting, so a training, validation, and test set is used to better measure model performance. The data is split into 60% training, 30% validation and 10% testing. The cis-eQTL analysis to calculate effect sizes is done on the training data, and the validation sets are used for hyperparameter selection when creating the PRS distributions, using $R2$ to measure performance. Additionally, to demonstrate the performance of the PRS model in various potential use cases, the analysis will be limited to a subset of 10 genes known to be related to various polygenic traits such as cardiovascular

disease (CVD), type 1 diabetes (T1D), and type 2 diabetes (T2D).
To evaluate the results of the PRS models generated for each gene, a baseline or null model for each gene will be established to ensure that the performance metrics of the PRS models are significant. This will be done by shuffling the effect sizes from the cis-eQTL analyses before creating the PRS distributions, effectively removing the true genetic associations while preserving the structure of the data. In the case that there are so few cis-eQTLs (e.g. less than 10) that the shuffling the effect sizes does not remove these associations, additional cis-eQTLs are added to the cis-eQTL analysis to generate the randomized PRS distribution.
In the case of creating a PRS model from the GWAS summary statistics, the exact same approach is taken for training, but as the GWAS data does not have any phenotype or trait data, it can not be evaluated or be finetuned through hyperparameters. So, a PRS model will be created with hyperparameters $p-val = 0.001$ and LD $R^2 = 0.1$ on a GWAS analysis pertaining to an individual's height, and will be loosely evaluated using a data sample from a female European individual.

5 Results

5.1 Genome Wide cis-eQTL Analysis

The 1000 Genomes Project dataset contains 1,059,922 unique cis-eQTLs out of a total 1,190,321 SNPs across 22 chromosomes. Upon conducting a cis-eQTL analysis across all genes in all chromosomes, there are 9,234,074 pairs of genes and cis-eQTLs. However, by thresholding these cis-eQTLs using the following derived p-value, we can determine how many significant cis-eQTLs there are:

$$p-value = 0.05/\#\text{ of SNP-gene pairs}$$

This resulted in 27,358 significant SNPs at $p-value = 5.4147e-09$. Another way to analyze the significance of these SNPs is by using linkage disequillibrium clumping at the gene level. As this can only be done at the gene level, only a single gene, GGT5 (ENSG00000099998.12) on chromosome 22, will be considered for simplicity.

Table 2: Number of cis-eQTLs at different LD r^2 thresholds with the lead variant for ENSG00000099998.12.

LD r^2 Threshold	Number of cis-eQTLs
$r^2 > 0.8$	1
$r^2 > 0.5$	12
$r^2 > 0.1$	78
$r^2 > 0$	358

The results in Table 2 of the various clumping threshold implies that of the 358 cis-eQTLs

associated with the gene, 12 of them are highly correlated to the most significant SNP. This can is further reinforced by utilizing a LocusZoom plot to visualize both the significance of each SNP as well as their correlations to the most significant SNP, as was done in Figure 1. From this plot, it is apparent that the majority of the genes cis-eQTLs hold little information, and those that do convey stronger associations to gene expression often convey the same information. This demonstrates why pruning and thresholding cis-eQTLs for PRS model creation is necessary, as multi-collinearity may introduce high levels of variance an noise to the effect sizes required to calculate PRS values.

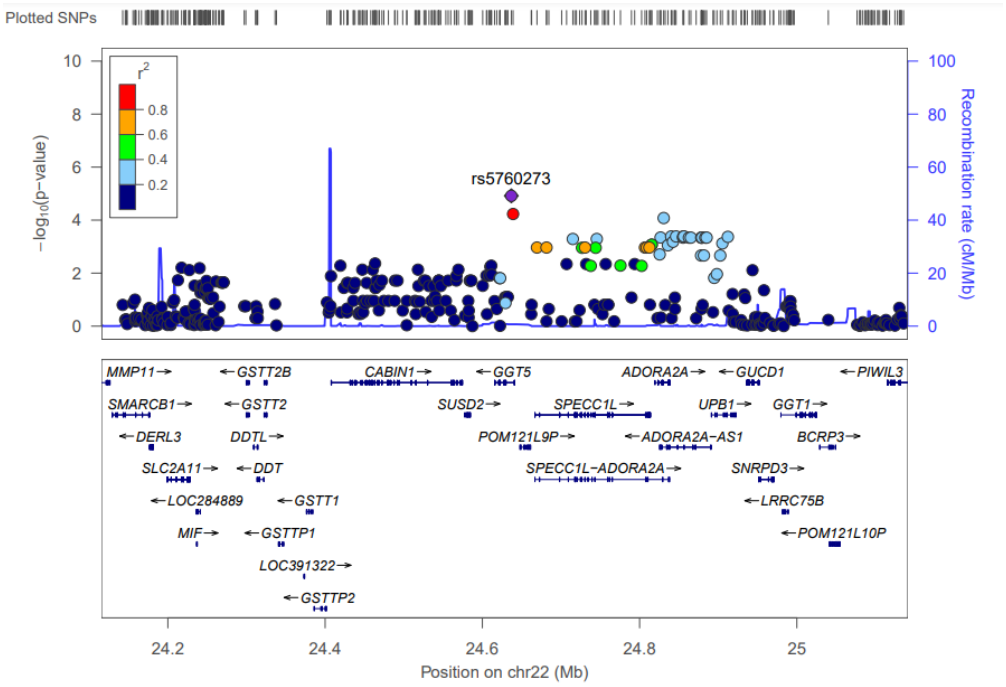


Figure 1: Locus Zoom plot for ENSG00000099998.12.

5.2 PRS Model for Individual Genes

Among the 10 genes considered in this analysis for cardiovascular diseases, type 1 diabetes, and type 2 diabetes, there are 1,783 unique cis-eQTLs corresponding to the cardiovascular disease genes, 248 unique cis-eQTLs for the type 1 diabetes genes, and 1,465 unique cis-eQTLs for the type 2 diabetes genes. The PRS models for each of these genes were trained using their respective cis-eQTLs, and the performance of each gene's null model and PRS model are show in Table 3 and Table 4, respectively.

5.2.1 Cardiovascular Disease PRS Model Results

With regards to the PRS models for cardiovascular disease genes, the results varied in comparison to their null counterparts. The model for the ENSG00000092054.11 gene was the

Table 3: Baseline Random PRS Model Performance

Gene ID	Train R^2	Valid R^2	Test R^2	Beta p-val	LD R^2	Trait
ENSG00000132170.15	0.112279	0.169664	0.022144	0.01	0.3	CVD
ENSG00000159640.8	0.017079	0.118548	0.094465	1.00	0.1	CVD
ENSG00000092054.11	0.002015	0.103439	0.007787	0.05	0.4	CVD
ENSG00000164867.5	0.019482	0.096251	0.000444	1.00	0.1	CVD
ENSG00000196126.6	0.000026	0.185011	0.019816	1.00	0.2	T1D
ENSG00000196735.6	0.194088	0.239903	0.045924	1.00	0.2	T1D
ENSG00000179344.11	0.129430	0.252629	0.261632	1.00	0.1	T1D
ENSG00000137309.14	0.018445	0.012736	0.003138	1.00	0.1	T2D
ENSG00000148737.11	0.002039	0.069765	0.044040	0.05	0.3	T2D
ENSG00000145996.7	0.002286	0.096535	0.091981	0.05	0.2	T2D

Table 4: PRS Model Performance

Gene ID	Train R^2	Valid R^2	Test R^2	Beta p-val	LD R^2	Trait
ENSG00000132170.15	0.134402	0.137487	0.235447	0.00100	0.3	CVD
ENSG00000159640.8	0.082613	0.057347	0.104214	1.00000	0.2	CVD
ENSG00000092054.11	0.201732	0.161565	0.075967	1.00000	0.1	CVD
ENSG00000164867.5	0.081935	0.218675	0.000002	1.00000	0.4	CVD
ENSG00000196126.6	0.342950	0.435084	0.400605	1.00000	0.2	T1D
ENSG00000196735.6	0.417692	0.461886	0.352035	0.00001	0.1	T1D
ENSG00000179344.11	0.517399	0.655920	0.687491	0.00001	0.4	T1D
ENSG00000137309.14	0.019076	0.011291	0.000730	1.00000	0.1	T2D
ENSG00000148737.11	0.135595	0.171143	0.117373	1.00000	0.3	T2D
ENSG00000145996.7	0.099658	0.129283	0.040357	1.00000	0.3	T2D

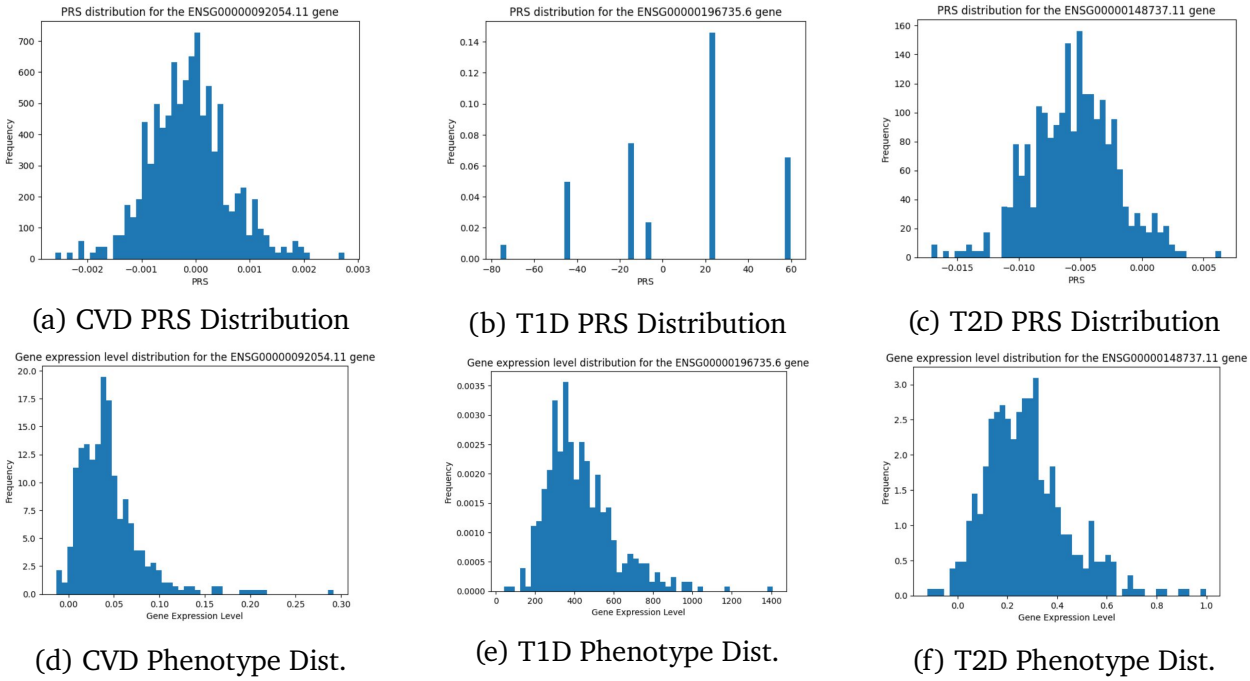
only model to significantly outperform the corresponding null model in the training, validation, and test sets. Of the remaining genes, the models for ENSG00000132170.15 and ENSG00000159640.8 had better training and test scores than the nulls. The last of cardiovascular models performed much worse across all metrics, indicating that there were no genetic association being captured by the model.

5.3 Type 1 Diabetes PRS Model Results

The PRS models for the type 1 diabetes genes performed remarkably well compared to all other models, with R^2 values ranging from 0.3 to 0.6 across all sets. These results suggest that the cis-eQTLs retained after pruning and thresholding are highly predictive of gene expression levels and are robust to a majority of the samples in the total dataset. However, it is also apparent that the correlations were so strong that even the null models showed relatively high R^2 values, which could be attributed to several factors. One explanation is the low number of SNPs (< 10) included in the PRS models for type 1 diabetes. With fewer

Table 5: Descriptions of columns in Tables 3 and 4.

Column	Description
Gene ID	Gene Ensembl ID
Train R^2	R^2 for the training set
Valid R^2	R^2 for the validation set
Test R^2	R^2 for the test set
Beta p-val	Hyperparameter: p-value threshold for SNP effect sizes
LD R^2	Hyperparameter: R^2 correlation coefficient threshold for clumping
Trait	Corresponding polygenic trait



(a) CVD PRS Distribution (b) T1D PRS Distribution (c) T2D PRS Distribution

(d) CVD Phenotype Dist. (e) T1D Phenotype Dist. (f) T2D Phenotype Dist.

Figure 2: PRS Distributions of a gene for each trait

SNPs, the randomization process in the null models may not completely disrupt the relationships between the PRS values and gene expression levels, thereby resulting in artificially inflated R^2 values for the null models.

5.3.1 Type 2 Diabetes PRS Model Results

Similarly to the cardiovascular models, the performances of the ENSG00000148737.11 and ENSG00000145996.7 models demonstrate significant potential. They show much stronger correlations between the PRS distributions and the gene expression levels compared to the baseline models, suggest that the cis-eQTLs used in these models are also capturing meaningful genetic associations that influence gene expression for these genes.

5.3.2 PRS Distributions

It seems that despite the shapes of the phenotype distributions, the PRS distributions are approximately normal except for the case of a gene associated with type 1 diabetes. Interestingly, Figure 2b displays a rather sparse distribution. It was found that this particular model comprised of just 2 cis-eQTLs. This model also happens to have one of the highest correlation values, indicating that the 2 SNPs contain a majority of the information on genetic association.

5.4 PRS Model from a GWAS

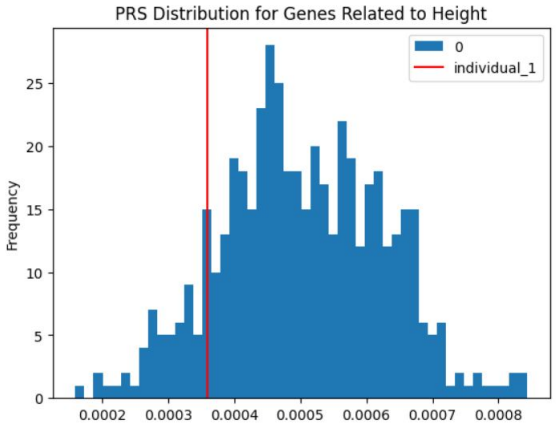


Figure 3: Sample of a female European individual compared to a PRS Distribution for the heights of male European individuals

As this PRS distribution was generated using the 1000 Genomes data containing male European samples, the peak of the distribution most likely correlates with the average European male height, which is 5 feet 11 inches (180 cm). Given that the height of the female individual is 5 feet and 6 inches (167.64 cm), which aligns reasonably well with the distribution when considering the sex-based difference in average heights. For European populations, the average height for females is typically about 13–15 cm shorter than males, meaning the observed result is consistent with the general patterns of genetic influences on height.

6 Discussion

The results indicate that a majority of the models were able to capture distinct relationships between their genes and corresponding phenotypes, as shown by the outperformance in relation to the null models generated for each gene. These results may not be deterministic on their own, but results can be further improved by creating a PRS distribution across multiple genes for a trait, using some weighted average of the gene expression levels as the

ground truth in cross validation and model evaluation. However, it should be noted that the results of these models come with severe limitations due to the nature of the data, and must be interpreted carefully. As mentioned before, this dataset is comprised entirely of samples from male European individuals, and may not accurately represent genomic associations in other populations with regards to sex and ethnicity. Another issue comes from the size of the dataset. Although there was a large number of SNPs to use in the generation of the PRS distributions, there were very few samples to train and evaluate the models on. Some of the inconsistencies between training, validation, and test scores can be attributed to the small sizes of each set. Most notably, the test sets consisted of only 40 samples (10% of the data), leading to high variations in the R^2 scores. Using a smaller validation set to allow for more testing data led to large variations in hyperparameter selection, resulting in overfitting and poor performance on the training and test sets. To account for these small validation and test sets, different methods of cross-validation can be used to more accurately evaluate each model (e.g. LOOCV, k-folds, stratified k-folds). As for the interpretations of PRS models, it is important to note that PRS models do not account for environmental factors that can significantly influence traits such as cardiovascular diseases, diabetes, and height. So although there may be strong genetic associations, it is not necessarily the case that they are the main cause. Nevertheless, these initial findings demonstrate that genetic risk prediction using PRS models is feasible and can be made more reliable using robust methodologies.

7 Conclusion

The findings from this analysis highlight the potential of PRS models in uncovering genetic associations for polygenic traits. While many models effectively captured the relationships between gene expression levels and their respective phenotypes, there is substantial room for improvement. Future work should prioritize expanding the dataset to include a more diverse range of individuals and larger sample sizes to improve generalizability. Incorporating more robust methods for hyperparameter selection and leveraging additional cross-validation strategies can also mitigate overfitting and variability. Despite these challenges, the results from this study provide a promising foundation for the application of PRS models in studying complex genetic traits. These models could pave the way for advancements in personalized medicine, allowing for more precise genetic risk predictions and tailored healthcare solutions. Future research should focus on addressing the current limitations while exploring ways to scale the models to larger, more diverse datasets, ultimately leading to more accurate and broadly applicable predictions.

References

Consortium, Genomes Project, A Auton, LD Brooks, RM Durbin, EP Garrison, and HM Kang. 2015. "A global reference for human genetic variation." *Nature* 526 (7571): 68–74

Consortium, GTEx. 2020. "The GTEx Consortium atlas of genetic regulatory effects across human tissues." *Science* 369 (6509): 1318–1330

Consortium, International Schizophrenia. 2009. "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." *Nature* 460 (7256): 748–752

Khera, Amit V, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature genetics* 50 (9): 1219–1224

Martin, Alicia R, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. 2019. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51 (4): 584–591