Identification of cis-eQTLs in 1000 Genomes LCLs and Predicting Genetic Risk for Multiple Diseases with Polygenic Risk Scores

Pallavi Prabhu
pprabhu@ucsd.edu

Tiffany Amariuta
tamariutabartell@ucsd.edu

Abstract

While humans appear to be very genetically different, 99.9% of the human genome is actually identical and genetic variation only accounts for 0.1%. This 0.1% consists of SNPs, or single nucleotide polymorphisms, which are essentially places in the genome where individuals may have a different nitrogenous base (A,T,C, or G) lending to different alleles (gene variants) and plays an instrumental role in genetic variation. Leveraging SNPs, researchers can identify which ones are statistically significant in impacting gene expression in individuals, using this information in order to construct polygenic risk scores, PRS, which are scores assigned to individuals identifying their risk for a certain disease. Polygenic risk scores using statistically significant and independent SNPs, offer a new approach to medicine and healthcare, providing individuals with an in-depth understanding of their predisposition to disease and informing potential preventative measures. While prior studies have performed cis-eQTL -loci near a gene that impact gene expression- analyses on the entire genome and created polygenic risk scores based on GWAS data (genome wide association study) for common diseases, these studies have been based on a primarily European ancestry which may not be as accurate for populations of other backgrounds. Recreating a cis-eQTL analysis, where cis is defined as SNPs located within 500KB of the gene, this project finds cis-eQTLs in 1000 Genome LCLs across all protein-coding genes genome-wide. Using this information, the project predicts the genetic risk of these individuals, merging this data with an individual's 23andme data to illustrate where they fall within the distribution of risk for multiple diseases.

Code: https://github.com/pallavisprabhu/dsc180a-cis-eQTL-and-polygenic-risk-scores

1 Introduction

SNPs, or single nucleotide polymorphisms are locations in the genome in which some people might have a different base and thus allele for a gene. And while 99.9% of the human genome is identical across humans, that 0.1% is what gives rise to human genetic variation and diversity (in addition to environmental influences) (Learn. Genetics 2199). An eQTL, expressive quantitative trait loci, are regions in the genome that influence the expression of genes in the population. A *cis*-eQTL are regions near the gene such as 500 kilo bases (Kb) upstream and downstream the start and end of the gene that influence gene expression. Studying the genome in this manner, identifying these SNPs and their impact on gene expression gives a means to create polygenic risk scores, ie, identifying the risk an individual has for a disease based on their collection of SNPs (GTEx Consortium 2020). This has the potential to be very informative in one's individual health risk since most diseases are not monogenic, i.e most diseases are not Mendelian and do not express based on the presence of a single allele. Rather, most diseases are polygenic, little accumulations or the combination of certain SNPs (Khera et al. 2018). Thus, polygenic risk scores can be instrumental in understanding an individual's health.

Polygenic risk scores aim to give individuals an understanding of their predisposition to certain diseases, but have not yet been deployed to the public for a multitude of reasons including that prior reference data of which polygenic risk scores were based on have historically been on a population of European ancestry which can give incomplete and inaccurate information of the individual is of a different background (Amariuta et al. 2020). For example some associations in certain genes with Type II diabetes that are present in Latino populations are rare in this European-centric data (Martin et al. 2019). This makes using polygenic risk scores on non-European populations far less accurate. Another reason they haven't been deployed in the public is the role of the environment. Phenotype not only relies on genotype, but the combination of both genotype and environment. Thus, the data at the foundation of polygenic risk scores may not accurately describe a certain population based on their differing environments. Using this data taken from the global scale, while it works by considering a lot of varying environments, it fails to take into account the unique environment and circumstances that the particular individual may be a part of (Martin et al. 2019).

This project uses the GWAS 1000 Genome Project which was published in 2015. The data comes in the form of `.bed`, `.bim`, and `.fam` files for each chromosome. The `.bed` file contains genotype data for all individuals in the sample; the `.bim` file contains the SNP information; and the `.fam` file contains family information for all the individuals in the sample. A public project created as a source of genetic variation, the data covers over 1000 samples from 14 global populations including but not limited to African ancestry in the United States, Bengali in Bangladesh, Han Chinese in China, British from England and Scotland, and Puerto Rican from Puerto Rico (Coriell Institute for Medical Research 2024). Generating polygenic risk scores for each individual from the 1000 project for a certain disease not only allows us to understand the distribution of risk on a global scale — actually considering individuals of different ethnic backgrounds— but also allows researchers to identify

the risk a specific individual has in comparison to the rest of this global population using their 23andMe data. This work is instrumental in helping individuals to make informed health decisions — knowing one's genetic predisposed risk to certain diseases can inform any potential diet, lifestyle, or family planning, contributing greatly to preventive health care (Martin et al. 2019).

2 Methods

2.1 Part 1: Identifying cis-eQTLs in 1000 Genomes LCLs

First, a Python `py` file was created to host the project, importing all the necessary Python packages and installing Plink and Plink2 by downloading the software online in order to actually do analyses with the 1000 Genomes data which contains `.bed`, `.bim`, and `.fam` files.

The `cis-eQTL` function takes in the chromosome number as an integer. The gene expression data was filtered to only include protein coding genes and genes on the specified chromosome. For each of these genes, a coordinate file was created to define the *cis* region as 500Kb upstream and downstream the beginning and end of the gene. This cannot be done for all the genes on the chromosome due to some genes being near the start or end of the chromosome, so the successes were tracked. For each of these successful genes, Plink files were created extracting the SNPs in the specified loci. In order to perform linear regression on the genetic data in Python, the genotype data was then converted to `raw` format. Finally, for each gene, the common samples between the genotype and gene expression were found and a feature matrix for the genotype and target vector of gene expression was defined. For each SNP in the gene, linear regression was performed and summary statistics were saved. Finally, the function outputs the summary statistics including the gene, SNP, β_0 (intercept), β_1 (slope), R^2 (coefficient of determination) and P-value, as a DataFrame in addition to saving it as a `txt` file in the directory `./ciseqtls_sumstats/sumstats_{chromosome}` where chromosome is the number chromosome as given as the argument of the function.

The function `cis-plot` has two parameters, the file path to the summary statistics file and a gene on that chromosome, and outputs a locus zoom plot, plotting the $\log_{10}$ of the P-value, in order to better visualize the scaling across the SNPs for the gene, marking which SNPs are above the threshold for significance at $\log_{10}(0.05)$ as noted by a red line. This function also prints the SNP Ids for all the statistically significant SNPs for the gene. With both of these functions, one can find the cis-eQTLs for each protein coding gene on each chromosome.

2.2 Part 2: Generating Polygenic Risk Scores

Tiffany's 23andMe text data was first converted to vcf using the GRCh37 reference genome and was then converted to Plink file format. GWAS data for Diabetes Mellitus, Heart Disease, Parkinson's Disease, Skin Cancer, and Arthritis were downloaded from the GWAS Cat-

alog. First the GWAS data was cleaned to extract the effect alleles, remove duplicate SNPs, drop unnecessary columns, and apply the natural logarithm to the odds ratio column in order to be used.

Since the 1000 Genomes data were separated by chromosome and Tiffany's 23andMe data had her own file for all her genetic data, Plink was used to merge all the data into a singular set of bed, bim, and fam files.

For Part 2, a second `py` file was created for organization. The function `prs` was created with two parameters, the file path to the GWAS data for a disease, and the name of the disease as a string. First clumping and thresholding were performed on the merged 1000 Genomes and Tiffany's data (Choi 2019). Thresholding is the process of keeping SNPs with a P-value below the specified threshold (0.0001 in this case with an R^2 threshold of 0.2) whereas clumping is the process of keeping independent SNPs, accounting for linkage disequilibrium. SNPs were clumped based on 500Kb. Linkage disequilibrium describes how genes that are located close to one another on a genome are more likely to be inherited together. Thus, these SNPs located close together might all show a significant impact on expression when it is actually just one SNP and the other SNPs are just located nearby.

The SNPs were then extracted from the clumped output, and Plink files were created with this data. Reading these SNPs, a score file was created using the extracted SNPs and adding their associated effect size and effect allele from the GWAS. Polygenic risk scores were calculated from the score file using Plink. The resulting `.sscore` file contained the PRS for each individual in the sample. Tiffany's score was extracted to mark her place in the distribution. The 1000 Genomes PRS scores were plotted on a histogram and Tiffany's individual data was marked with a line to illustrate her placement in the distribution of risk. This same process was repeated for all five diseases to illustrate her risk in comparison to the population.

2.3 Running the Project

Both `py` files were imported in a Jupyter Notebook to run and project with different chromosomes and genes for Part 1, and for all the different diseases for Part 2 by calling the respective functions.

3 Results

The summary statistics computed by the cis-eQTL comes in the form of a tab separated file for each chromosome with the columns Chr (Chromosome), Gene, SNP, beta_0, beta_1, R_sq (R^2) Standard Error, and P-value (see Table 1). While the summary statistics are by chromosome, to actually visualize the cis-eQTLs, you would plot the locus plot for each gene individually since one would be interested in seeing which SNPs are statistically significant in impacting a single gene's expression.

Table 1: Summary Statistics of cis-eQTL on Chromosome 1

Chr	Gene	SNP	beta_0	beta_1	R_sq	Standard Error	P-value
1	SAMD11	rs3094315_A	3.289710	0.082625	0.002426	0.090597	0.362409
1	SAMD11	rs3131972_G	3.291351	0.081776	0.002384	0.090448	0.366567
1	SAMD11	rs3131969_G	3.276884	0.086575	0.002388	0.095694	0.366258
...	...	...	...	...	...	...	...
1	PGBD2	rs34013644_T	1.730611	-0.011718	0.000130	0.055484	0.832862
1	PGBD2	rs12746903_T	1.905123	-0.099312	0.001828	0.125485	0.429244
1	PGBD2	rs12726733_C	1.726229	-0.009268	0.000056	0.066805	0.889742

The graph of cis-eQTL analysis for different genes (see Figure 1 and Figure 2) illustrates that regardless of the gene, there are usually a couple hundred SNPs associated with that gene but only a handful of SNPs with significant P-values.

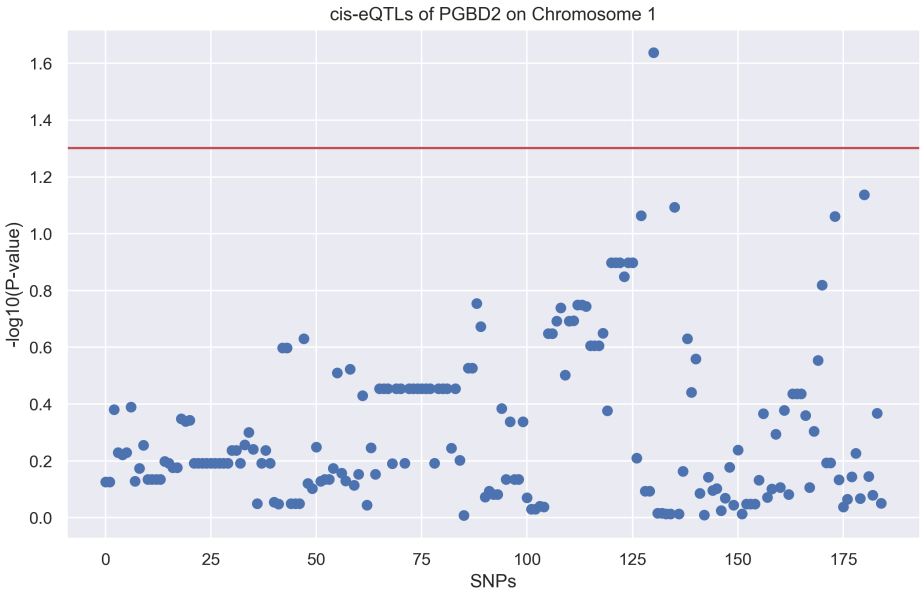

Figure 1: cis-eQTL of the PGBD2 gene on Chromosome 1

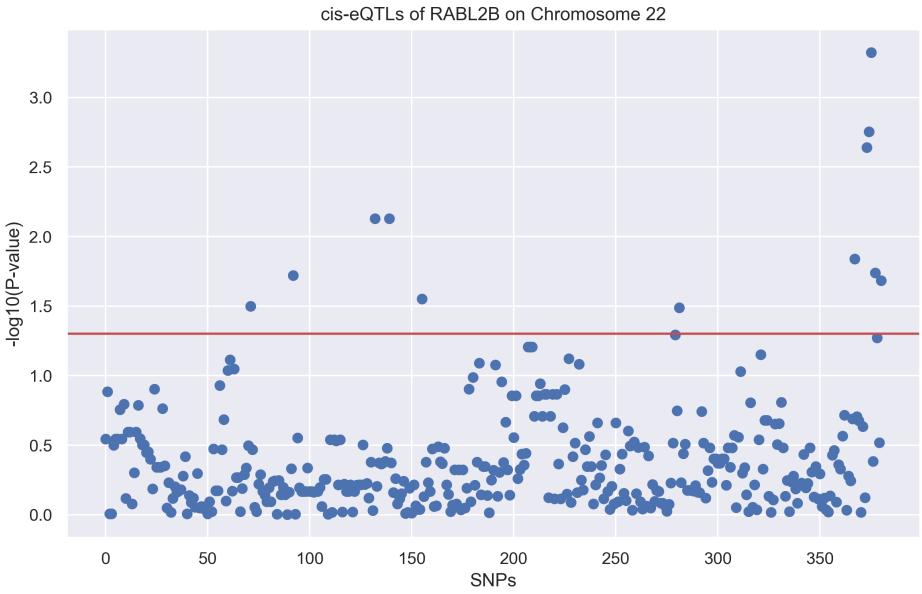

Figure 2: cis-eQTL of the RABL2B gene on Chromosome 22

Moreover, the PRS distribution for all five diseases resembles a normal distribution. For diabetes mellitus, Tiffany's risk appears to be around the middle where a majority of the population's scores are, but slightly towards the right. For Heart Disease, Tiffany appears to be more towards the right, more in between the middle of the distribution and the right end (see Figure 3).

For Parkinson's, like Heart Disease, Tiffany appears to be more towards the right, in between the middle of the distribution and the right end. For both skin cancer and arthritis, Tiffany appears to fall right in the middle of the distribution (see Figure 4 and Figure 5).

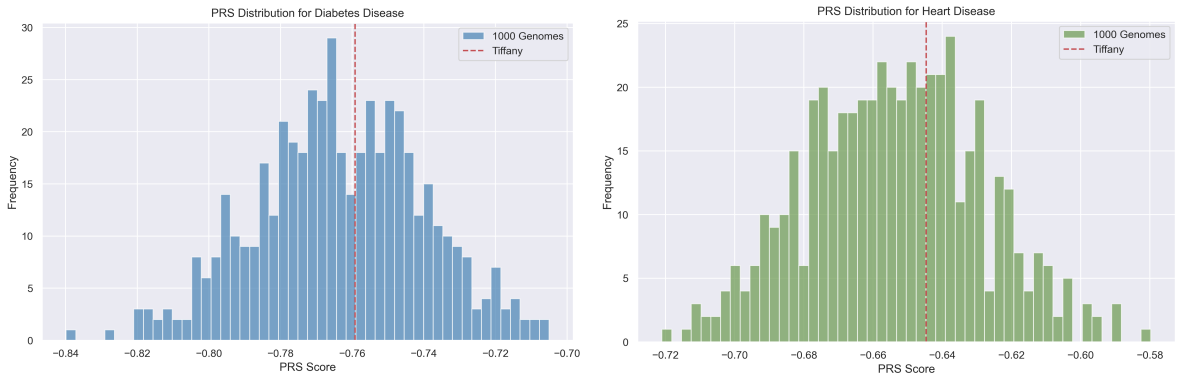

Figure 3: The PRS Distribution for Diabetes Mellitus (left) and Heart Disease (right)

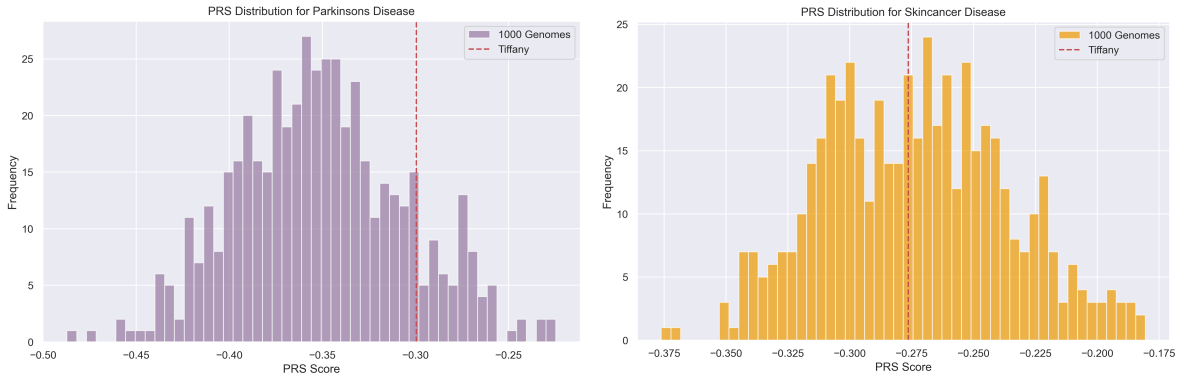

Figure 4: The PRS Distribution for Parkinson's Disease (left) and Skin Cancer (right)

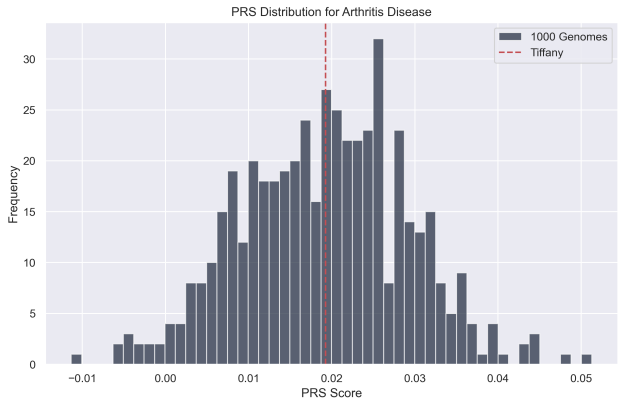

Figure 5: The PRS Distribution for Arthritis

4 Discussion

4.1 Part One: Identifying cis-eQTLs in 1000 Genomes

In Part 1 of the project, there are a handful of SNPs that are below the threshold for significance for each gene. Because the scaling of the y-axis is $\log_{10}$(P-value), lower p-values have higher $\log_{10}$(P-value), meaning data points above the red line are significant.

Moreover, for each SNP that is significant, there are SNPs close by that are also above the threshold or nearly do creating this peaks of significance. This is due to gene linkage. During crossover and recombination when creating gametes (a reproductive cell), genes that are located close to each other on the chromosome tend to be inherited together since they have less chance of crossing over and recombining. Thus, these SNPs that are located close to each other on the genome are likely to be inherited together. Therefore, if one of these SNPs is actually significant in influencing gene expression, it may appear like all its nearby SNPs are as well since they have been inherited together, thus creating this swarm pattern where nearby SNPs may have similar p-values.

This aligns to previous work exploring cis-eQTLs in the genome, as linkage disequilibrium creates these peaks of associations (GTEx Consortium 2020). Linkage disequilbrium, this phenomenon of associations of alleles due to their close proximity to each other makes it challenging to identify the the influence of casual variants or SNPs of a gene. Clumping, as explored in Part 2, attempts to remove these correlated SNPs, only keeping both weakly correlated SNPs and significant SNPs (Choi 2019).

4.2 Part Two: Generating Polygenic Risk Scores

In Part 2 of the project, PRS scores were created and plotted for diabetes mellitus, heart disease, Parkinson's disease, skin cancer, and arthritis, with Tiffany's individual score being marked in red to illustrate her risk in comparison to the population.

For Diabetes Mellitus, she appears to have slightly more than average risk in comparison to the population since very slightly towards the right of the middle of the distribution For Heart Disease, she appears to have slightly more than the average risk, since she is again a bit right of the middle of the distribution. For Parkinson's, she appears to have more than average risk compared to the population since she is in the middle of the middle and right end of the distribution. For both Skin Cancer and Arthritis, she appears to be right in the middle of the distribution, indicating average risk for both diseases.

While it is difficult to compare the accuracy of these data in comparison to prior approaches that have relied in Euro-centric data since these are all potential disease risk rather than a definitive telling of whether the individual will get a disease or not, this approach aligns with previous polygenic risk score work (Amariuta et al. 2020).

However, what this approach fails to consider, like prior studies, is the role environment and epigentics plays in gene expression (Martin et al. 2019). Phenotype is the combination of genotype and environment, and while using a diverse dataset like 1000 Genomes attempts to include the potential role environment may play in gene expression, it can not fully capture the nuances in impact different environments may have in gene expression.

5 Conclusion

Ultimately, the cis-eQTL analysis identifies genetic variants, SNPs, that influence gene expression of nearby genes. Finding these associations lends insight on how genetic variation impacts gene expression and regulation. Actually identifying which parts of the genome are impacting gene expression has profound impact in understanding which genetic variants can contribute to diseases like diabetes or cancer. Moreover it also paves the way for precision medicine and personalized health treatments, allowing personalized medicine and healthcare to target genetic variants that may be causing illness.

Furthermore, quantifying one's risk for a disease using polygenic risk scores has does intrumental work for preventative healthcare. Analyzing the genome and identifying one's potential genetic risk for a number of diseases arms individual's with the power and knowl-

edge to make smart and informed decisions about their lifestyles and health, noting any predisposition for disease and giving individuals the opportunity to prevent like making diet and lifestyle changes. Moreover, they can also be used in family planning, informing potential parents about diseases they may be carriers for and may pass along to any potential offspring, and getting the jump on testing for conditions one may be susceptible to. Essentially, by knowing the conditions/diseases one might be predisposed to, the individual can make more informed decisions about their health to prevent actually getting it whether that be taking medications, making changes to lifestyle, or getting regularly tested.

Polygenic risk scores have yet to be deployed in the clinic because most of the data underlying polygenic risk scores is primarily based on European data and is thus far more accurate for individuals of European descent (Martin et al. 2019). Using the 1000 Genomes data, this project works to dissolve this bridge between polygenic risk scores and practical use, building a more well-rounded model for polygenic risk score by using data based on varying ancestories. However, where this project still falls short is its limitation in including the role of environment in influencing gene expression.

References

Amariuta, Tiffany, Kazuyoshi Ishigaki, Tazro Ohta Hiroki Sugishita, Masaru Koido, Kushal K. Dey, Koichi Matsuda, Yoshinori Murakami, Alkes L. Price, Eiryo Kawakami, Chikashi Terao, and Soumya Raychaudhuri. 2020. "Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements." In *Nature Genetics*. [Link]

Choi, Shing Wan. 2019. "Basic Tutorial for Polygenic Risk Scores." [Link]

Coriell Institute for Medical Research. 2024. "1000 Genomes Project." [Link]

GTEx Consortium. 2020. "The GTEx Consortium atlas of genetic regulatory effects across human tissues." In *Science*. [Link]

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, and Sekar Kathiresan. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." In *Nature Genetics*. [Link]

Learn. Genetics., "Making SNPs Make Sense."

Martin, Alicia R, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Current clinical use of polygenic scores will risk exacerbating health disparities." In *Nature Genetics*. [Link]

Appendices

A.1 Training Details

A.1.1 cis-eQTL Calculations

cis-eQTLs are identified by performing linear regression on each SNP where X is the matrix of genotypes where each each row is the person:

$$\hat{y} = \beta_0 + \beta_1 X + \epsilon$$

A.1.2 Polygenic Risk Score Calculations

Plink calculates polygenic risk scores from the score file using `–score`. This is done by weighting the genotypes by the effect size of the SNP.

$$\text{PRS} = \sum(\text{Genotype} \times \text{Effect Size})$$

This in practice looks like where X is the feature matrix, i represents an individual in the sample, and j represents the SNP:

$$\text{Y}_j = \sum_{j=0}^{n} X_{ij}\beta_j + \epsilon_i$$

Refer to the GitHub repository linked in the beginning for more specific details of the code used to run these analyses.

A.2 Additional Figures

cis-eQTLs of different genes on chromosome 14 to better visualize how different genes on the same chromosome have different collection of SNPs and different patterns of significant SNPs:

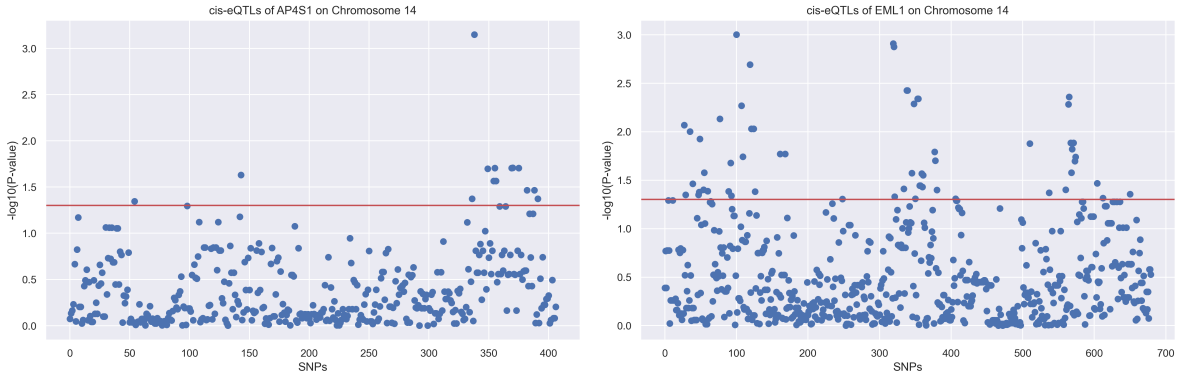

Figure A 1: cis-eQTLs of the AP4S1 gene (left) EML1 gene (right) on Chromosome 14

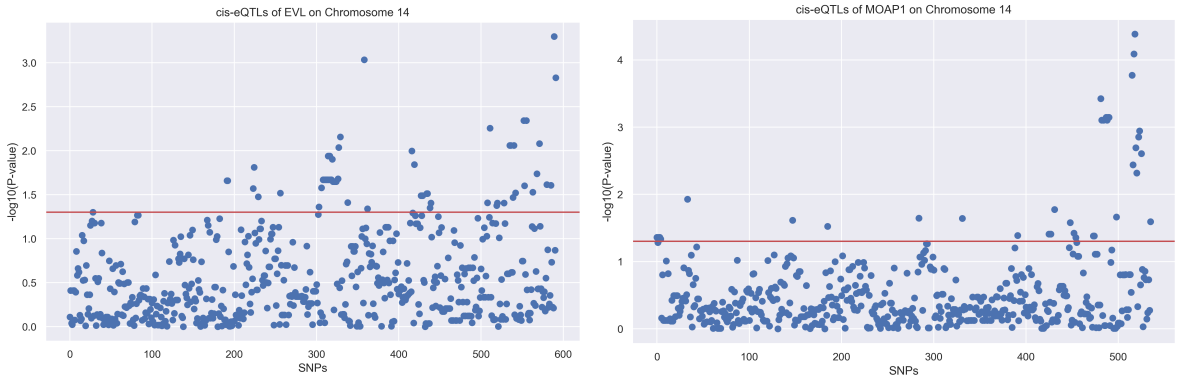

Figure A 2: cis-eQTLs of the EVL gene (left) MOAP1 gene (right) on Chromosome 14

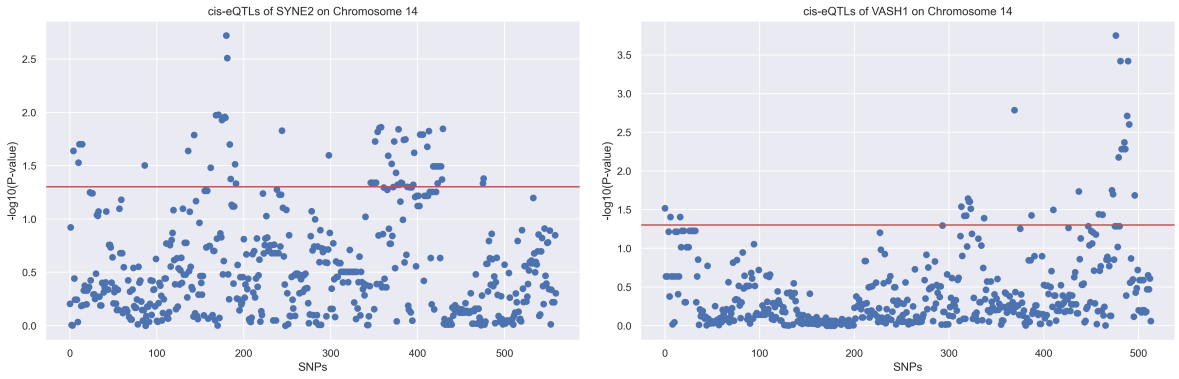

Figure A 3: cis-eQTLs of the SYNE2 gene (left) VASH1 gene (right) on Chromosome 14

A.3 Additional Tables

Table A 1: Condensed Summary Statistics of cis-eQTL on Chromosome 14. Position refers to the SNP's relative position to one another on the genome, denoting the chronological order of the SNPs.

Chr	Gene	SNP	P-value	Position
14	EML1	rs11627868_A	0.407210	0
14	EML1	rs12897121_T	0.169447	1
14	EML1	rs11623713_G	0.407210	2
...	...	...	...	...
14	EML1	rs1191036_C	0.263779	677
14	EML1	rs1191035_T	0.263779	678
14	EML1	rs1191030_G	0.296223	679

Table A 2: Condensed Summary Statistics of statitisically significant cis-eQTLs on Chromosome 14. Note how the statistically significant SNPs are located close to one another due to gene linkage.

Chr	Gene	SNP	P-value	Position
14	EML1	rs2400461_C	0.008526	27
14	EML1	rs7148549_A	0.044834	29
14	EML1	rs1955903_C	0.009955	35
...	...	...	...	...
14	EML1	rs1152792_G	0.035855	357
14	EML1	rs1152795_G	0.026979	359
14	EML1	rs7141044_C	0.037227	360
...	...	...	...	...
14	EML1	rs1957515_T	0.039768	560
14	EML1	rs12232161_G	0.005206	564
14	EML1	rs4905891_A	0.004370	565
..	...	...	...	...

Table A 3: Tiffany's PRS as Percentiles

	Percentile
Diabetes Mellitus	56.122449
Heart Disease	63.469388
Parkinson's Disease	86.530612
Skin Cancer	47.55102
Arthritis	49.183673