2019

Evaluating PrediXcan's Ability to Predict Differential Expression Between Alcoholics and Non-Alcoholics

John E. Drake Jr
Virginia Commonwealth University

Evaluating PrediXcan's Ability to Predict Differential Expression Between Alcoholics and Non-Alcoholics

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics.

by

John Edward Drake Jr

Bachelor of Science in Biology, James Madison University, 2009

Advised by Vladimir Vladimirov, MD PhD
Dept. of Psychiatry,
Virginia Institute for Psychiatric and Behavioral Genetics,
Dept. of Physiology and Biophysics
Center for Biomarker Research and Precision Medicine,
School of Pharmacy
Virginia Commonwealth University,
Lieber Institute for Brain Development,
Johns Hopkins University

Virginia Commonwealth University
Richmond, Virginia
April 2019

Acknowledgements

The author would like to thank several people who helped him along his journey. First, I would like to thank my family for their support and encouragement. I would like to thank Dr. Allison Johnson for giving me the opportunity to join the Bioinformatics program. As a nontraditional student, I understand the faith she placed in my application and I am grateful for the opportunity she gave me. I would like to thank my committee members, Dr. Maria Rivera and Dr. Mikhail Dozmorov, for their feedback and time. And lastly, but not least, I would like to thank Dr. Vladimir Vladimirov, who has been an inspirational mentor and advisor throughout my graduate career.

Abstract

PrediXcan is a recent software for the imputation of gene expression from genotype data alone. Using an overlapping set of transcriptome datasets from postmortem brain tissues of donors with alcohol use disorder and neurotypical controls, which were generated by two different platforms (e.g., Arraystar and Affymetrix), and an additional unrelated transcriptome dataset from lung tissue, we sought to evaluate PrediXcan's ability to impute gene expression and identify differentially expressed genes. From the Arraystar platform, 1.3% of matched genes between the measured and imputed expression had a Pearson correlation ≥ 0.5. Our attempt to replicate this finding using the expression data from the Affymetrix platform also lead to a similarly poor outcome (2.7%). Our third attempt using the transcriptome data from lung tissue produced similar results (1.1%) but performance improved markedly after filtering out genes with a low predicted R^2, which was a model metric provided by the PrediXcan authors. For example, filtering out genes with a predicted R^2 below 0.6 led to 16 genes remaining and a Pearson correlation of 0.365 between the measured and imputed expression. We were unable to reproduce similar performance gains with filtering the Arraystar or Affymetrix alcohol use disorder datasets. Given that PrediXcan can impute a narrow portion of the transcriptome, which is further reduced significantly by filtering, we believe caution is warranted with the interpretation of results derived from PrediXcan.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Assessing gene expression via high-throughput expression platforms is becoming an increasingly ubiquitous approach to study the molecular underpinnings underlying complex phenotype differences between clinical samples. In hybridization (i.e. microarray) experiments, RNA is reverse transcribed to copy DNA (cDNA), labeled with a fluorescent tag, and allowed to hybridize to complementary probes on a microarray platform; thus, the expression of a gene can be assessed by detecting the presence of the label at specific probe sites. As such platforms contain numerous probes, they are able to quantify the expression of tens of thousands of genes per sample, which make these platforms attractive for an unbiased assessment of gene expression patterns. However, microarray platforms can only detect the presence of known transcripts and are known for being sensitive to technical artefacts. For instance, one study found that between 50.5% to 82.8% of all probes, using either an Affymetrix or Agilent platform, correlated significantly with array batches (Leek, 2010). This sensitivity can be explained by seemingly minuscule factors, such as ozone levels that meet the Environmental Protection Agency's standard for "good" air quality, having been shown to effect expression measurements (Byerly, 2009).

RNA-sequencing is another popular method to assess expression levels, which isn't as sensitive to background noise (Su, 2014). By reverse transcribing the RNA sample to DNA, amplification via polymerase chain reaction, and then sequencing, RNA-sequencing is able to provide a more comprehensive measurement of the transcriptome, which also includes detection of alternative and unknown transcripts. Unfortunately, measuring differential expression of various phenotypes is hampered by small effect sizes caused by the reliance on clinical assessments, high genetic heterogeneity, and co-regulation of genes unrelated to the phenotype of interest. Additionally, RNA is

quite prone to degradation, which frequently leads to differential degradation among species of RNA (Garneau 2007; Ferrer 2008). Consequently, these factors add noise, which researchers often overcome with larger sample sizes. However, for psychiatric illnesses, where researchers are limited to post-mortem samples, as the disease organ is the brain, procuring sufficient samples is fraught with difficulty and high costs.

One such psychiatric ailment is alcohol use disorder (AUD). As defined by the Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5; American Psychiatric Association, 2013), AUD is measured by its severity (mild, moderate, or severe) based on the cumulative sum of eleven different symptoms; such symptoms include alcohol interfering with one's job or family, being fixated on drinking alcohol, and drinking more in order to reach a desired effect. For Americans over the age of 18, the average rate of AUD was high for twelve-month (13.9%) and lifetime (29.1%) prevalence (Grant, 2015); when one considers white men, the rates for twelve-month (17.6%) and lifetime (36.0%) are even more despairing. Additionally, two meta-analyses of global epidemiology studies have estimated AUD relative risk for all-cause mortality for both sexes to be 3.45 (2.96-4.02) and 3.55 (2.48-5.09) (Laramee, 2015; Roerecke, 2013). Much of this can be explained by the numerous physical ailments strongly associated with alcoholism (e.g. liver cirrhosis, heart disease, etc). But despite the significant health effects, it is estimated that only 24.1% of those with alcohol dependency will receive treatment in their lifetime (Hasin, 2007).

Unfortunately, AUD treatment is limited as there are only 3 unique Food and Drug Administration approved agents for the treatment of alcohol abuse, with the last one being approved in 2004 (Ray, 2018). Combinational treatments (therapy and pharmacological intervention) have shown efficacious results, but such treatments have failed a large portion of participants and the longevity of abstinence is questionable; for instance, one combinational trial (naltrexone, acamprosate, and cognitive behavioral therapy) reached an abstinence rate of 67% but only spanned 84 days (Feeney,

2006). Thus, when one considers alcohol abuse's prevalence, mortality risk, and lack of pharmaceutical options, the necessity to find additional treatments becomes apparent. But because of the challenges face with differential expression, elucidating these complex psychiatric phenotypes has been challenging.

1.1 PrediXcan

PrediXcan is a recent software developed for predicting expression levels from genotype data alone (Gamazon, 2015). It accomplishes this by building upon the idea of an expression quantitative trait loci (eQTL) analysis, which finds associations between expression levels and single nucleotide polymorphisms (SNPs), which represents sites of genomic variability. But instead of using an ordinary least square regression to find such associations, PrediXcan utilizes an elastic net, which combines two penalizing regression methods. The first is ridge regression, which down weights correlated predictors and hence is able to account for linkage disequilibrium, which is the non-random association of alleles, without pruning/clumping highly correlated SNPs (Malo, 2008). The second is lasso regression, which uses shrinkage i.e., the values of nonmeaningful variables are shrunk towards a central point (usually the mean) and eventually eliminated from the model, creating a more parsimonious model (Waldmann, 2013). But as lasso regression performs poorly in certain situations, such as with highly correlated variables, it does not necessarily perform better than ridge regression in performance (Zou, 2005). Thus, the elastic net model essentially amounts to a compromise between ridge regression and lasso regression.

By utilizing elastic net methodology and a transcriptome dataset, which contain both genotype and expression data, the authors of PrediXcan have been able to ascribe weights to the genotype

components (i.e. SNPs), which corresponds to the expression of select genes. These weights have been made available by the authors, in the form of an SQLite file, to be used by PrediXcan. When a user runs PrediXcan with unrelated genotype data, PrediXcan will match SNPs in the unrelated genotype data to the weights in the SQLite file. And when matches are found, those weights will be used to add/subtract from the expression value of the corresponding gene of a given individual. In doing so, PrediXcan is able to impute gene expression from unrelated genotype data at the individual level.

One of the most popular and comprehensive transcriptome datasets is the Genotype-Tissue Expression (GTEx) project, which aims to include the genotype and tissue specific expression of over 900 deceased donors (The GTEx Consortium, 2013). Currently, version 7 is the most recent release of GTEx and incorporates the genotype and expression data from 635 donors. The authors of PrediXcan have taken this data, adjusted it to account for sex and experimental/population confounders, used the elastic net methodology to calculate the expression weights for the genotypic components, and made the SQLite weights files available on predictdb.org.

Additionally, the authors utilized 10-fold cross-validation to derive quality measure for each gene model. With this procedure, a subset of expression data was extracted out, the elastic net was conducted, and the model's imputed results was compared against the extracted expression. This step was repeated 9 more times and thus allowed the model to be evaluated using the same transcriptome dataset. From the cross-validation, several quality measures were calculated and provided in the GTEx version 7 SQLite weights files. One quality measure is "rho_avg", which is the average Pearson correlation between the imputed and measured expression for each of the hold out folds during cross validation; this measure will be referred to as **predicted correlation**. The square value of this measure is labeled "pred.perf.R2" and will be referred to as **predicted R^2.** Another quality measure is "test_R2_avg", which is the average coefficient of determination for each of the cross-validated models and will be referred to as **model R^2**.

Considering the stability and "immutability" of DNA i.e., DNA does not change from one tissue to another, means one can obtain genetic information from peripheral tissues like whole blood or buccal tissue. In contrast, gene expression various among tissues and thus may be more burdensome to obtain, especially from internal organs. And with the GTEx transcriptome dataset, the authors of PrediXcan have been able to create SQLite weights files for 53 different tissues, with 13 originating within the brain. Thus, PrediXcan offers an alternative to circumvent the difficulties associated with generating gene expression data from difficult to obtain organs (such as brain) and thus may aid in the study of psychiatric ailments.

1.2 PrediXcan in Recent Literature

As of March 25, 2019, google scholar reported 374 citations for the nature genetics' publication of PrediXcan. Many of these studies referenced the increase power of PrediXcan as the inspiration for their study. As users are testing each gene once, with any number of SNPs for a given gene, the multiple testing burden is significantly reduced when compared to a single variant eQTL analysis. This count potentially reduces the number of tests from the millions to the thousands, depending on the number of SNPs in the genotype data and the genes included in the SQLite weights file (Li, 2018). Thus, for many studies, PrediXcan is an exploratory tool that was used to confirm measured differential expression and find newly differentially expressed genes, which past studies were underpowered to find.

Many studies take a critical examination of PrediXcan, which is reassuring considering how new this software is. One such study examined individuals with amyloid deposition in the brain, which is a risk factor for Alzheimer's Disease, against the GTEx cohort (Hohman, 2018). As this study required a more stringent prediction R^2 of ≥ 0.15, the authors show restraint in utilizing PrediXcan. More importantly, the

authors replicate their findings in another cohort and examined how the results conformed to past findings. Lastly, the discussion highlighted many of the limitations with the GTEx transcriptome dataset and reiterated the need for the newly identified differentially expressed genes to be validated further.

Other studies are far more ambitious in their use of PrediXcan. One study conducted a trans-ethnic meta-analysis on SNPs associated with breast cancer risk despite that the GTEx transcriptome data largely originated from Caucasian donors (Hoffman, 2017). Additionally, they reported significantly differentiated genes regardless of the model's strength; one such gene was DHODH, whose gene model had a predicted R^2 value of 0.026. But the most concerning element of these studies is that their results may be referenced in other papers without sufficient attention to the imputation methodology. One such study found that lower expression of RPRD1B was associated with cisplatin-induced peripheral neuropathy using PrediXcan (Dolan, 2017). A systematic review of cisplatin toxicity reported the study's findings on RPRD1B with no mention of PrediXcan, imputation, or the elastic net methodology (Chovanec, 2017). Thus, there is the possibility that that the results will be given similar weight and attention to non-imputation results. If PrediXcan accurately imputes gene expression, such interpretations aren't problematic. But if PrediXcan's accuracy is low or limited, such interpretations could incorrectly define the narrative and hinder scientific inquiry.

1.3 Concerns about PrediXcan Methodology

Considering PrediXcan ability to overcome limitations with differential expression analysis, it's use has grown overtime. But with any new methodology, it's important that alongside its benefits, its limitations are also considered. Here, 4 such concerns have been identified.

I. For the GTEx transcriptome dataset, PrediXcan only includes cis-SNPs (within 1Mb of gene start or end) within the model. While testing for cis-SNPs is typical with eQTL analysis, as a consequent of the multiple-testing burden (Nica, 2013), large twin studies have asserted that trans-SNPs may play a larger role. One study found that only 10% of the genetic variance for skin tissue was explained by cis-SNPs (Grundberg, 2012). Additionally, during the model-building procedure, ambiguous SNPs (e.g. A/T) were removed so that the software could identify and correct genotype dosages given by the user. Thus, a considerable amount of genotype information had been eliminated prior to deriving each model, which could affect PrediXcan's performance.

II. For GTEx v7, the authors required each gene model to have a predicted correlation above 0.10. Additionally, the p-value for that correlation must be below 0.05. Hence, these quality-control measures reduced the number of genes within each tissue-specific SQLite weights file. For GTEx v7 whole blood, the tissue-specific weight file with the highest average predicted R^2, only 6,297 genes were included.

III. Only the top 9 GTEx tissues, which had the largest sample sizes, were evaluated in the original manuscript. Both the predicted R^2 and model R^2 were low for these models. For instance, the average predicted R^2 for all gene models for whole blood was 0.118 (median 0.074). Additionally, the average model R^2 for all the gene models for whole blood was 0.096 (median 0.050).

IV. While the GTEx cohort is fairly inclusive with only a handful of exclusionary criteria, like metastatic cancer (GTEx, 2013), the cohort is largely European, skews towards older adults, and does not include expression measurements for all donors. For PrediXcan's tissue-specific weight files, derived from GTEx v7, only European donors were selected and the average number of donors per weight file was around 185. Thus, there is a question about

how well PrediXcan's methodology can capture the expression differences found in less common ailments and populations, which may have not been sufficiently represented in the GTEx cohort or tissue-specific samples. Additionally, the general heterogeneity of the cohort may lead to higher expression variance and thus limit the power of any models derived from it.

1.4 An Opportunity to Evaluate PrediXcan:

A transcriptome dataset has been made available from deceased donors diagnosed with alcohol dependence (n=32) and neurotypical controls (n=31). The tissue was obtained by the Australian Brain Donor Program and include Caucasian male (n=54) and females (n=9) that met the inclusion criteria, which excluded individuals with infectious disease, significant head trauma, long agonal, and long delays in procuring the tissue (McMichael, 2019, *pre-print: https://doi.org/10.1101/583203, under review*). The expression data was obtained from the nucleus accumbens of these donors, which animal models have strongly implicated in pavlovian reward-related learning (Day, 2011) and addictive drug behaviors (Scofield, 2016). Additionally, eQTL analysis of the nucleus accumbens have shown differential expression for those with alcoholism (Mamdani, 2015), including for this particular dataset (McMichael, 2019). Thus, not only is the organ appropriate for finding differential expression of alcoholics, such findings have been demonstrated for hundreds of genes.

As the PrediXcan authors have made the SQLite weights file available for the nucleus accumbens, which was derived from the GTEx version 7 transcriptome dataset, there is an opportunity to assess PrediXcan's methodology. Twin studies have estimated the heritability of alcohol use disorder to range from 48-58% (Prescott, 1999). Given this is a disease with a large genetic component, which PrediXcan's

methodology models, the use of PrediXcan for this phenotype is appropriate. Furthermore, studies evaluating PrediXcan for assessing psychiatric phenotypes are rare and, to the best of my knowledge, have not focused on alcoholism.

As the nucleus accumbens was not evaluated in the original PrediXcan manuscript and considering the general concerns/performance surrounding GTEx weights files, we hypothesize that PrediXcan will impute the expression for most genes, which are associated with alcohol use disorder, poorly. That being said, the PrediXcan authors calculated the heritability of their Depression Gene Networks model, which is based on whole blood samples, and found that the predicted R^2 increased with higher heritability (Gamazon, 2015). This observation isn't unusual as the effect sized of cis-eQTLs increase with a gene's heritability (Grundberg, 2012). Additionally, it has been shown that the heritability of gene expression is primarily explained by cis loci ---not trans--- in tissues with a heterogenous mixture of cell types, as is common with the GTEx datasets (Price, 2011). Thus, it seems reasonable that PrediXcan, which utilized models built on cis associations, may be able to accurately impute expression values for highly heritable gene. And while heritability estimates for GTEx are unavailable, the fact that heritability and predicted R^2 are correlated means that predicted R^2 can be used as a surrogate. Hence, it is also hypothesized that the accuracy of PrediXcan imputation will increase by selecting genes with a high predicted R^2 model.

Chapter 2

Measuring Differential Expression

To gauge PrediXcan's accuracy, one must first measure the differential expression between donors diagnoses with alcohol use disorder and neurotypical controls. The expression measurements of the 63 donors were obtained using Arraystar Human LncRNA Array v3.0 (Rockville, MD, USA). This platform detects 61,464 transcripts encompassing both protein coding genes and long noncoding RNAs. As Arraystar limits information concerning their proprietary probe designs, annotation information was available for only 11,810 of those genes (19%) that were found to be differentially expressed (t-value ≥ |0.5|) as determined by the methodology of the original manuscript for this transcriptome dataset (McMichael, 2019). Additionally, those differential expression results were made available, were labeled **manuscript results** within this work, and were used as a comparison to this work's methodology.

Differential expression analysis was conducted with R version 3.5.1 with the tidyverse package, which was used for data wrangling and visualization, and the Linear Models of Microarray Data (Limma) package, which handled the statistical elements (Smyth, 2004). Essentially, limma follows the traditional approach of determining group differences by building a linear regression model for each probe. But instead of using a simple t-test, Limma employs empirical Bayes methodology to "borrow strength" from the other probes (Ritchie, 2015). In other words, limma utilizes all probes within the dataset to calculate the variance of each individual probe, which ultimately increases the power of the model. But before using limma's toptable function, which produced the results, several considerations were taken into account.

First, the raw data was log2-transformed, background corrected using normexp, and had controls probes removed from further consideration. Second, each array was examined for irregularities utilizing a MA-plot. No irregularities were found, which was likely a result of the transcriptome dataset including only arrays used in the final analysis of the original manuscript (i.e. the dataset had problematic arrays filtered out). Next, normalization was applied, which reduces the variances found between arrays so that they can be compared (Quackenbush, 2002). It should be noted that there is considerable debate concerning the best normalization procedure. Quantile normalization, which is limma's default procedure for single-channel arrays, forces each array to have the same distribution. As there is no biological reason to expect that the bulk expressed transcriptome will show stark expression differences between alcoholics and neurotypical controls, quantile normalization seemed appropriate and was selected.

Demographic information (age of death, sex, brain weight, hemisphere of sample, neuropathology, and smoking status), experimental conditions (Batch, pH of sample, postmortem interval from procuring sample, and RNA integrity number) were collected for all samples and checked for outliers; no outliers were found, which was expected as the transcriptome dataset has been pre-screened. As these elements may influence differential expression or measurements of differential expression, it is prudent to account for them within the model. But while correcting for confounding variables can lead to increased power and more accurate results, over adjusting a model can lead to lower power (Kahan, 2014). Thus, a principal component analysis (PCA) was conducted, which captures the variance of potential demographic/experimental confounders as principal components (PCs).

R's prcomp function was used to conduct the PCA. A scree plot was used to identify the inflection point of the PCs' eigenvalues. The 3 PCs to the left of the inflection point were retained as those PCs explained a substantial amount of variance in the demographic/experimental data. R's varimax function was used on the loadings of each PC in order to aid in interpreting and verification of

the PCA results. The correlation between each of the potential confounding variables and the retained PCs with varimax rotation are displayed in Table 1.

Variable	PC1	PC2	PC3
Batch	-0.661	0.240	0.382
Age of Death	-0.362	-0.153	0.203
Sex	-0.828	0.164	-0.097
Brain Weight	0.620	0.310	0.108
pH	0.147	-0.165	-0.796
Postmortem Interval	-0.053	-0.587	0.112
Hemisphere	0.166	-0.191	0.676
Neuropathology	-0.135	-0.683	0.200
Smoking Status	-0.116	0.591	0.225
RNA Integrity	0.464	0.214	-0.609

Table 1: Correlation between potential confounding variables and retained principal components, which were produced by the prcomp function and manual varimax rotation.

The differential expression of each gene was tested using a robust linear regression, which was adjusted with the top three unrotated PCs. t-values were selected to make comparisons between limma's results and the manuscript results. The histograms of both approaches can be seen in figure 1. The Pearson correlation between the t-values of both approaches was 0.92 and the corresponding scatter plot can be viewed in figure 2. Note: whitespace found in the manuscript's histogram and scatter plot were caused by the incomplete annotation.

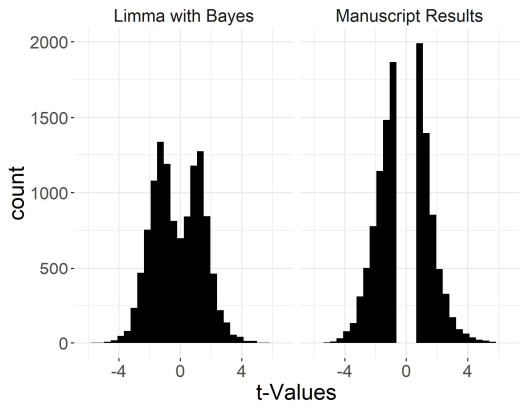

Figure 1. Histogram of t-values calculated by Limma, using empirical Bayes methodology, and the manuscript's methodology.

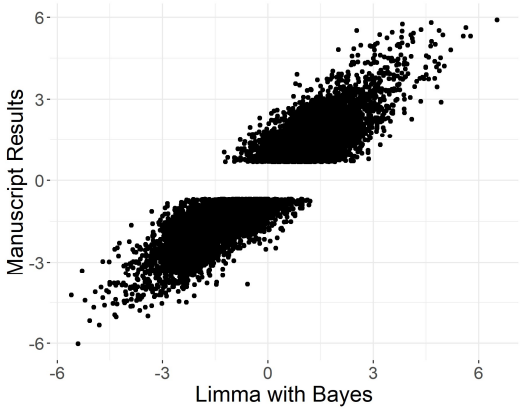

Figure 2. Scatter plot of comparable t-values calculated by Limma, using empirical Bayes methodology, and the manuscript's methodology.

A Venn diagram was provided in Figure 3, which compared t-values ≥ |3.25| from Limma's empirical Bayes methodology and the manuscript's methodology; this roughly corresponds to a False Discovery Rate p-value of 0.05. As there was a strong Pearson correlation between both techniques, these initial results suggest that the empirical Bayes test is slightly inflating the t-values. That being said, it's imperative to test this inclination.

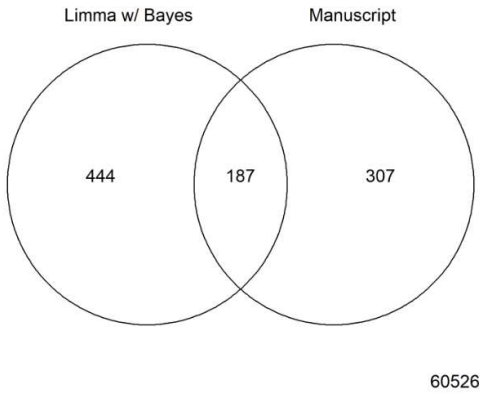

Figure 3. Venn diagram of t-values ≥ |3.25| with Limma, using the empirical Bayes methodology, and the manuscript's methodology.

2.1 Limma without Empirical Bayes

In order to perform differential expression, Limma interacts with an expression object, which is modified at each step. The toptable function, which outputs the results, is dependent on the empirical Bayes function to calculate and apply certain test-statistics (t-values, p-values, and log-odds) onto the expression object. Thus, Limma does not offer a way to disable the empirical Bayes methodology. But by manually calculating and applying those test-statistics to the expression object, one can circumvent the Bayes methodology with another. Thus, the same methodology was followed except a simple t-test was performed and the necessary test-statistics were manually added to the expression object prior to using the toptable function.

A histogram (Figure 4), scatter plot (figure 5) and Venn diagram (figure 6) of Limma without the empirical Bayes methodology are provided below. The Pearson correlation between comparable t-values was 0.92. Thus, the differences between the two methods were minor, suggesting that the empirical Bayes methodology was not responsible for the inflation of t-values. In fact, there was a slight increase of t-values $\geq |3.25|$ with the simple t-test, suggesting that the empirical Bayes methodology was deflating ---not inflating--- the t-values. This could be caused by a number of probes having an unusually low variances, which Limma was correcting.

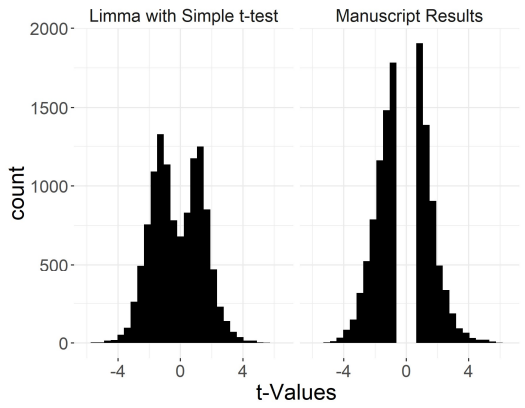

Figure 4. Histogram of t-values calculated by Limma, using a simple t-test, and the manuscript's methodology.

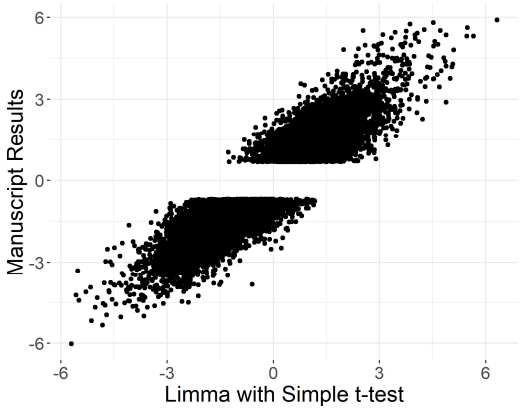

Figure 5. Scatter plot of comparable T-values calculated by Limma, using a simple t-test, and the manuscript's methodology.

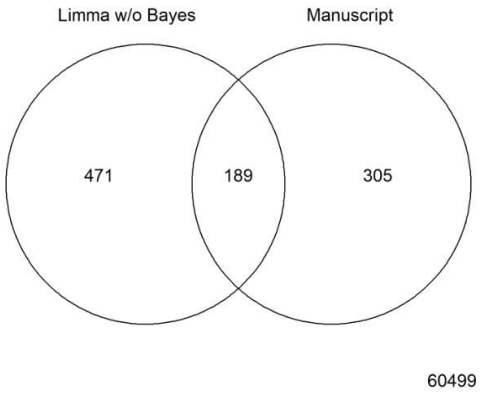

Figure 6. Venn diagram of t-values ≥ |3.25| with Limma, using a simple t-test, and the manuscript's methodology.

2.2 Limma using NCSS Derived Principal Component Analysis.

The manuscript methodology differed from this work in several ways which included its handling of normalization, accounting for batch effect, and removal of probes. That being said, batch effect tends to be one of the most prominent confounding elements in microarray platforms and the prcomp function (Table 1) did not appear to account for much of the batch variance. Thus, there were concerns that the original PCA did not properly account for confounding variables. Following the manuscript as a

rough guideline, NCSS version 12 was used to perform a robust PCA with a varimax rotation. The top 3 PCs were still retained. As the NCSS software utilizes a version of the modern QL algorithm, it's was expected to produce different loadings from the prcomp function. The correlation between the variables and the 3 rotated PCs are displayed in Table 2. Additionally, a histogram (Figure 7), scatter plot (figure 8) and Venn diagram (figure 9) of Limma with the empirical Bayes methodology with NCSS PCA were also provided.

Variable	**PC1**	**PC2**	**PC3**
Batch	0.755	-0.051	-0.300
Age of Death	0.490	0.181	0.048
Sex	0.791	-0.037	0.024
Brain Weight	-0.348	-0.390	0.044
pH	-0.204	0.002	0.836
Postmortem Interval	0.001	0.636	0.096
Hemisphere	-0.189	0.264	-0.642
Neuropathology	-0.051	0.727	-0.141
Smoking Status	0.168	-0.458	-0.437
RNA Integrity	-0.501	-0.426	0.435

Table 2: Correlation between potential confounding variables and retained principal components, which were produced by NCSS's robust PCA with varimax rotation.

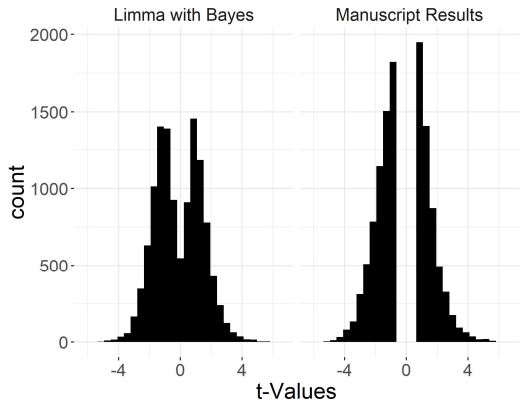

Figure 7. Histogram of t-values calculated by Limma (left) and manuscript's methodology (right).

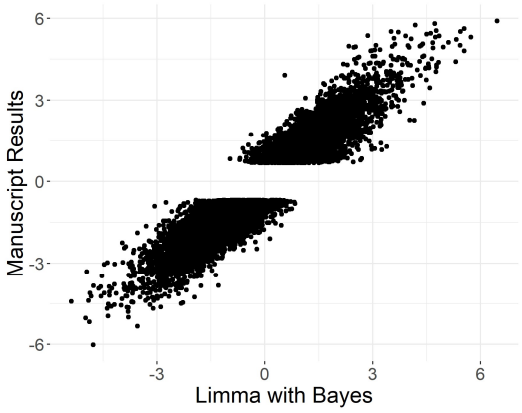

Figure 8. Scatter plot of comparable t-values calculated by Limma and the manuscript's methodology.

Note: Limma utilized empirical Bayes methodology. The model was adjusted utilizing PCs derived from NCSS.

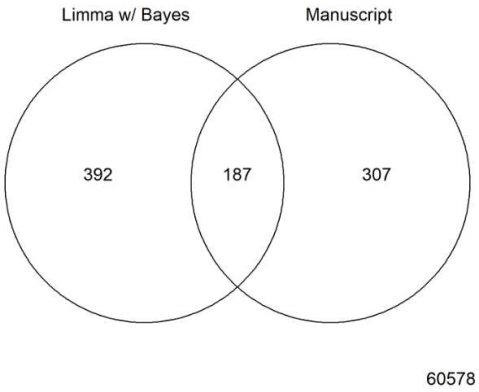

Figure 9. Venn diagram of t-values ≥ |3.25| with Limma and the manuscript's methodology. Limma utilized empirical Bayes methodology. The model was adjusted utilizing PCs derived from NCSS.

The Pearson correlation slightly improved to 0.94 and the number of probes, with a t-value ≥ |3.25|, decreased. That being said, the differences between prcomp and NCSS were not substantial, highlighting that the original analysis (Limma, empirical Bayes, prcomp) had an adequate PCA. Additionally, the original analysis produced a Pearson correlation of 0.92 and a non-significant Wilcoxon rank sum test (p-value = 0.97). Thus, while there may be substantial nonoverlapping in the Venn diagram, that result was caused by slight variations ---not stark differences--- in the t-values. Considering those factors, the original analysis seemed appropriate and was used as a base comparison for PrediXcan.

Chapter 3

Evaluating Imputed Expression

The genotype data of the transcriptome dataset, which included donors with alcohol use disorder, originated from the Affymetrix Genome-Wide Human SNP Array 6.0 platform. The data had been pre-imputed and filtered to exclude SNPs with a low (≤ 0.01) minor allele frequency (MAF). In total, 3,703,165 autosomal SNPS were provided with annotations. The GTEx version 7 nucleus accumbens SQLite weight file was downloaded from predictdb.org, which was built using the genotypic and expression data from 114 donors. This file contains weights for 113,923 SNPs that were associated with 3,633 genes; 2,641 were protein-coding (73%), 485 were long non-coding RNAs (13%), and 507 were pseudogenes (14%).

Python 3.7 was used to transform the genotype data into a PrediXcan dosage files, which are chromosome-specific compressed files containing the genotype information and SNP annotations. Because PrediXcan cannot handle NA values, which relate to unknown calls, SNPs with any NAs were removed from further consideration. This reduced the initial number of SNPs to 2,217,199 (60% of the original dataset). Finally, a sample file, which helps PrediXcan identify samples, and a phenotype file, which is used during PrediXcan's association test, were created.

Several quality checks were conducted before running PrediXcan. First, GTEx used the human genome reference GRCh37; thus, the genotype annotation was compared to ensure both the SNP position and allele information matched Ensembl's GRCh37 build. R version 3.5.1 with the biomaRt package (Durinck, 2011) was used to retrieve and compare annotations. No discrepancies were found. Secondly, the SQL weight files do not follow the alternative/reference convention due to annotation discrepancies between different databases. Instead, the weight file reference the "effect allele" and only include non-complementary SNPs, which allows PrediXcan to correct dosages if the other allele was

coded for. Still, the transformed genotype files were checked against the SQLite weight file for coding discrepancies; none were found.

Python 2.7 on Ubuntu 18.04.1 LTS was used to run PrediXcan, which produced imputed expression levels for each individual. Several steps were taken so that the imputed expression could be compared against the measured expression. First, the PrediXcan output included Ensembl gene IDs, which were not available in the Arraystar annotation. Thus, Python 3.7 with the MyGene package (Xin, 2016) was used to convert the IDs into gene symbols. Secondly, a handful of genes, which were found in the measured expression, had duplicate probes (n=11). The median values of those probes were taken. Finally, PrediXcan works in an additive fashion where all genes start with an initial value of zero. If the output of a gene includes zeros for all imputed individuals, that indicates that the genotypic data lacked matching SNPs in the weights file. Thus, these genes were removed (n=411) as they weren't informative.

As the gene models were trained on log2 inverse-rank normalized data, the measured expression data was transformed in the same manner. Thus, the imputed expression results could be negative and would indicate that the predicted expression was lower than the average expression of the GTEx cohort, which was used to create the model. For genes in both datasets, a Pearson correlation was used to compare measured expression values and imputed expression values (Table 3). Additionally, the genes with the highest Pearson correlation are highlighted in Figure 10.

		ALL	$PR^2 \geq 0.1$	$PR^2 \geq 0.2$	$PR^2 \geq 0.3$	$PR^2 \geq 0.4$	$PR^2 \geq 0.5$	$PR^2 \geq 0.6$
Number of Genes:		975	546	262	142	69	25	10
First 5 with AUD:								
	1	-0.042	-0.065	-0.039	0.031	-0.116	-0.148	0.195
	2	-0.007	-0.030	-0.098	-0.069	-0.087	0.095	0.131
	3	0.004	0.049	0.019	-0.027	0.011	0.037	-0.234
	4	-0.017	-0.051	-0.104	-0.114	-0.097	-0.091	-0.280
	5	0.014	0.017	0.114	0.137	0.177	0.387	0.751
First 5 Controls:								
	1	0.011	0.017	0.001	-0.031	0.006	0.022	0.076
	2	-0.002	-0.026	-0.078	-0.104	-0.051	0.169	0.826
	3	0.013	0.044	0.100	0.081	-0.016	0.128	-0.019
	4	-0.027	-0.035	-0.038	-0.133	-0.032	-0.272	-0.504
	5	-0.005	0.022	0.016	-0.010	-0.117	-0.114	-0.251
Mean (All Donors)		0.021	0.012	0.016	0.022	0.003	0.054	0.043
SD (All Donors)		0.193	0.219	0.256	0.286	0.323	0.320	0.317

Table 3. Pearson correlation of matching genes between measured expression and imputed expression for a given individual. The first 5 donors with alcohol use disorder and neurotypical controls were highlighted. Pearson correlation of matching genes were also calculated at certain predicted R^2 thresholds (≥0.1, ≥0.2, ≥0.3, ≥0.4, ≥0.5, or ≥0.6).

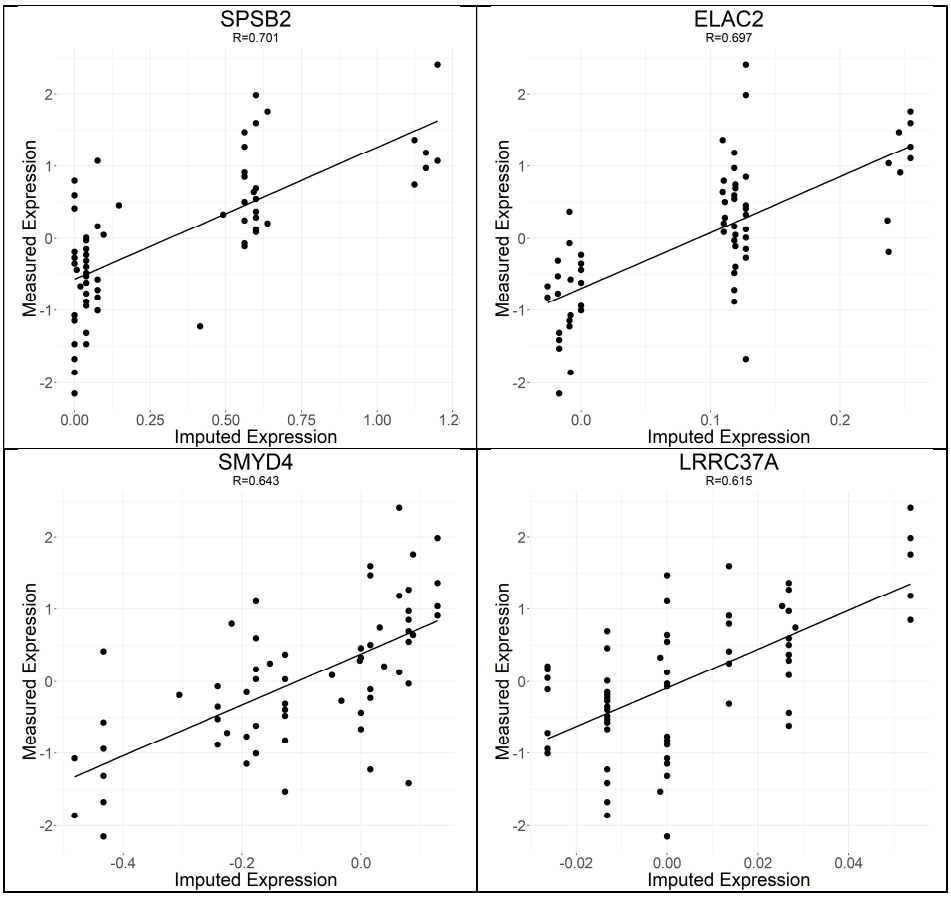

Figure 10: Examples of well predicted genes. The predicted R^2 for SPSB2, ELAC2, SMYD4, and LRRC37A was 0.418, 0.323, 0.139, 0.531.

The results show that only 975 genes were matched between the imputed expression and measured expression. Overall, the correlation was poor and did not significantly improve by selecting genes with a specified predicted R^2. There were a handful of well correlated genes (Figure 10) but only 13 genes had a Pearson correlation ≥ 0.5 (1.3% of all matched genes). And while most of those genes had a predicted R^2 above the average gene model, which was 0.16 for the nucleus accumbens, none of those genes had a strong predicted R^2.

Additionally, some of the well correlated genes had defined striations, which suggests that much of the imputed expression was defined by a single SNP. For example, the most likely explanation for ELAC2's results would be a single SNP, with a high weight coefficient, and a somewhat even-

distribution of the three genotypes (homozygous major, heterozygous, and homozygous minor). The jittering of those lines was likely the result of other SNPs with a lower weight coefficient. This observation is further supported by the fact that ELAC2 has only 7 SNPs in its gene model (on average, nucleus accumbens' gene models contained 31 SNPs).

3.1 Selecting Genes with Model R^2

In the SQLite weight files for GTEx version 7, the model R^2 quality metric was added. This metric is the average coefficient of determination for each of the cross-validated models. As such, it expresses how much variance a particular model is able to account for. Thus, the metric may be a stronger indicator of a model's accuracy and was used to select genes (Table 4). Overall, the correlations did not improve by selecting genes with a higher model R^2. The differences between the model R^2 and the predicted R^2 were not substantial.

		ALL	$MR^2 \geq 0.1$	$MR^2 \geq 0.2$	$MR^2 \geq 0.3$	$MR^2 \geq 0.4$	$MR^2 \geq 0.5$	$MR^2 \geq 0.6$
Number of Genes:		975	343	188	98	46	19	6
First 5 with AUD:								
	1	-0.042	-0.060	0.045	-0.077	-0.093	-0.160	0.168
	2	-0.007	-0.071	-0.062	-0.129	0.082	0.188	0.041
	3	0.004	0.030	-0.023	-0.023	-0.113	0.032	-0.328
	4	-0.017	-0.089	-0.090	-0.092	-0.102	-0.148	-0.229
	5	0.014	0.045	0.167	0.173	0.123	0.409	0.935
First 5 Controls:								
	1	0.011	0.000	0.009	-0.080	-0.022	0.024	0.284
	2	-0.002	-0.026	-0.092	-0.107	-0.012	0.240	0.933
	3	0.013	0.078	0.101	0.065	-0.025	0.285	-0.021
	4	-0.027	-0.033	-0.106	-0.122	-0.179	-0.325	-0.450
	5	-0.005	0.002	0.024	-0.020	-0.074	-0.162	-0.324
Mean (All Donors):		0.021	0.022	0.031	0.002	-0.021	0.026	-0.012
SD (All Donors):		0.193	0.237	0.271	0.298	0.323	0.294	0.383

Table 4. Pearson correlation of matching genes between measured expression and imputed expression for a given individual. The first 5 donors with alcohol use disorder and neurotypical controls were highlighted. Pearson correlation of matching genes were also calculated at certain model R^2 thresholds (≥0.1, ≥0.2, ≥0.3, ≥0.4, ≥0.5, or ≥0.6).

3.2 Recoding Missing SNPs

As PrediXcan cannot handle missing values, SNPs with any NAs were removed from further consideration. Thus, a single NA would cause all genotypic data for that SNP to be disregarded, which is reflected in the fact that 40% of the genotypic data was removed. As PrediXcan depends on matching SNPs to the SQLite weight file, there was a strong possibility that the filtering step eliminated predictive information and hampered PrediXcan's accuracy. Thus, recoding missing values was attempted by assigning them the value [2 X Population's Effect Allele]. This step ascribes a neutral estimator to each missing SNP by assuming the sample's genotype matched the allele frequency in its population.

The National Center for Biotechnology Information (NCBI) provides comprehensive information about each SNP, including allele frequencies for different populations. Additionally, the URL of a given SNP is predictive, allowing one to automate the retrieval of relevant information. Thus, a virtual private

server running ubuntu 18.04.2 LTS executed a bash script, which visited and extracted allele frequencies for 1,485,966 SNPs. Considering that the donors were Caucasian Australians, the European population was chosen. Information relating to 973,016 SNPs had allele frequencies for the European population. That information was checked against the annotations for mismatch alleles and nonsensical results (i.e. same nucleic acid for the reference and alternative allele); 7,931 results were removed due to quality-control measures. Thus, 965,085 SNPs were able to be recoded, which increased the total number of SNPs available for PrediXcan to 3,182,284 (86% of the original dataset).

The results of the recoding are highlighted in table 4. As more SNPs were included, there were less genes removed for having all imputed zeros (n=333) and slightly more matching genes (n=999). And while filtering by predicted R^2 brought a slight improvement in the mean correlation, PrediXcan accuracy remain poor and only 14 genes had a Pearson correlation ≥ 0.5 (1.4% of all matched genes). Additionally, the genes with the highest correlation did not change. Thus, recoding missing values brought minor improvements.

		ALL	$PR^2 \geq 0.1$	$PR^2 \geq 0.2$	$PR^2 \geq 0.3$	$PR^2 \geq 0.4$	$PR^2 \geq 0.5$	$PR^2 \geq 0.6$
Number of Genes:		999	558	265	144	70	25	10
First 5 with AUD:								
	1	-0.041	-0.052	-0.037	0.000	-0.153	-0.233	0.249
	2	0.031	0.009	-0.042	-0.036	-0.084	0.088	0.181
	3	0.020	0.074	0.068	0.031	0.086	0.101	-0.083
	4	-0.035	-0.045	-0.098	-0.114	-0.132	-0.087	-0.132
	5	0.032	0.039	0.098	0.101	0.153	0.278	0.722
First 5 Controls:								
	1	0.020	0.010	-0.008	-0.035	-0.030	0.070	0.117
	2	-0.009	-0.042	-0.121	-0.186	-0.207	-0.086	0.580
	3	0.014	0.027	0.064	0.069	-0.032	0.142	0.058
	4	-0.006	-0.009	-0.016	-0.102	- 0.047	-0.262	-0.595
	5	-0.017	0.017	0.015	0.006	-0.066	-0.055	-0.175
Mean (All Donors):		0.025	0.020	0.024	0.031	0.025	0.060	0.050
SD (All Donors):		0.194	0.219	0.253	0.283	0.318	0.323	0.309

Table 5. Pearson correlation of matching genes between measured expression and imputed expression for a given individual. Missing dosages in genotype data was recoded prior to using PrediXcan. The first 5 donors with alcohol use disorder and neurotypical controls were highlighted. Pearson correlation of matching genes were also calculated at certain predicted R^2 thresholds (≥0.1, ≥0.2, ≥0.3, ≥0.4, ≥0.5, or ≥0.6).

3.3 Association Test

If given a phenotype file, to distinguish what group each sample belonged to, PrediXcan can run an association test. Thus, a linear association test was performed using the data that had missing SNPs recoded. The Benjamani and Hochberg adjusted p-values were calculated with the p.adjust function in R version 3.5.1. None of the genes reached significance (≤0.05).

For exploratory reasons, genes with significant (≤0.05) unadjusted p-values were examined. In total, 164 genes were identified and 49 of those genes were represented in the Arraystar annotation. Of those 49 genes, only 2 (NOC3L and IFITM4P) were significant in the differential expression analysis of the measured expression. Additionally, those genes' Pearson correlations, which compared the imputed and measured expression, are shown in Figure 11. Both of those genes had a modest Pearson

correlation and poor predicted R^2. Thus, the association test performed poorly, which is understandable given the poor imputation of expression.

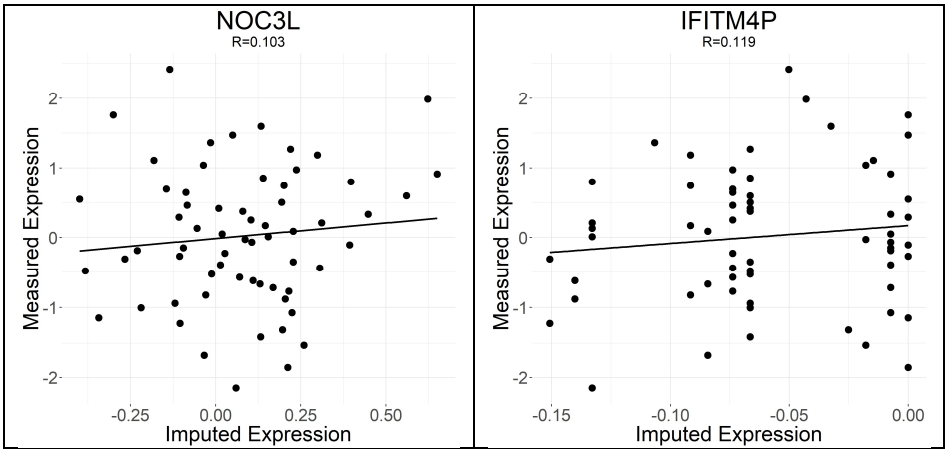

Figure 11. Pearson correlation of two genes, which were significant in the differential expression analysis (p-value adjusted) and PrediXcan association test (p-value unadjusted). The Pearson correlation and predicted R^2 for NOC3L (0.057) and IFITM4P(0.050) were low.

Chapter 4

Affymetrix Expression

As part of another study on the same population, expression data was collected utilizing the Affymetrix GeneChip Human Genome U133A 2.0 platform, which measured the expression of 18,400 human transcripts and variants. Only a subset of the neurotypical controls (n=18) and donors with alcohol use disorder (n=18). While there are numerous differences between the two platforms, including RNA preparation and scanner, one impactful difference is the size of the probes. Affymetrix utilizes 25-mer oligonucleotide probes whereas the Arraystar platform are more than double in length (60 nt). As shorter probes tend to be more sensitive and longer probes have a higher rate of cross-hybridization (Chou, 2004), discrepancies may exist between different platforms.

Thus, PrediXcan was evaluated against the Affymetrix array using the same procedure as described in chapter 3 with several exceptions. First, the data was not log2 transformed, background corrected, or normalized, as it was already done as described in the data's publication (Mamdani, 2015). Secondly, the annotations for the platform were created using Affymetrix's batch query web app (https://www.affymetrix.com/analysis/netaffx/batch_query.affx). Additionally, the genotypic data with recoding of missing SNPs was used. The results are highlighted in table 6.

		ALL	$PR^2 \geq 0.1$	$PR^2 \geq 0.2$	$PR^2 \geq 0.3$	$PR^2 \geq 0.4$	$PR^2 \geq 0.5$	$PR^2 \geq 0.6$
Number of Genes:		1500	832	390	193	91	36	9
First 5 with AUD:								
	1	0.049	0.057	0.050	0.137	0.189	0.301	0.442
	2	0.084	0.104	0.128	0.108	0.171	0.049	0.036
	3	0.022	0.031	0.107	0.121	0.105	0.261	0.403
	4	0.026	0.029	0.060	0.078	-0.019	0.164	0.050
	5	-0.005	-0.039	0.006	0.029	-0.007	0.092	0.098
First 5 Controls:								
	1	0.006	0.012	0.027	0.054	0.014	-0.198	-0.010
	2	-0.004	0.020	0.008	0.014	-0.026	0.074	0.328
	3	0.045	0.061	0.090	0.088	0.076	-0.077	-0.229
	4	0.054	0.095	0.072	0.053	0.075	0.213	0.576
	5	-0.027	0.007	0.003	0.011	0.006	0.119	0.192
Mean (All Donors):		0.025	0.033	0.046	0.058	0.054	0.026	0.099
SD (All Donors):		0.239	0.264	0.301	0.336	0.374	0.423	0.415

Table 6. Pearson correlation of matching genes between the Affymetrix measured expression and imputed expression for a given individual. Missing dosages in genotype data was estimated prior to using PrediXcan. The first 5 donors with alcohol use disorder and neurotypical controls were highlighted. Pearson correlation of matching genes were also calculated at certain predicted R^2 thresholds (≥0.1, ≥0.2, ≥0.3, ≥0.4, ≥0.5, or ≥0.6).

The fact that only 1,500 genes matched between the Affymetrix platform and the PrediXcan results were surprising for several reasons. First, the nucleus accumbens weight file included 2,641 protein coding genes. Additionally, the Affymetrix platform attempts to capture the entire protein coding transcriptome, which may only total 19,000 genes for humans (Ezkurdia, 2014), and more importantly, makes the full annotations available. Thus, the protein coding genes within the weight file were examined further; numerous protein coding genes included read-throughs, putative and variant genes were discovered, which explained the lower than expected matching.

The correlation between Imputed expression and Affymetrix expression are show in Figure 12 and show the top correlated genes all had a predicted R^2 above the average (i.e. 0.16). But overall, the

accuracy of PrediXcan was poor and increasing the predicted R^2 did not bring a substantial gain in performance. Only 41 genes (2.7% of all matched genes) had a Pearson correlation ≥ 0.5.

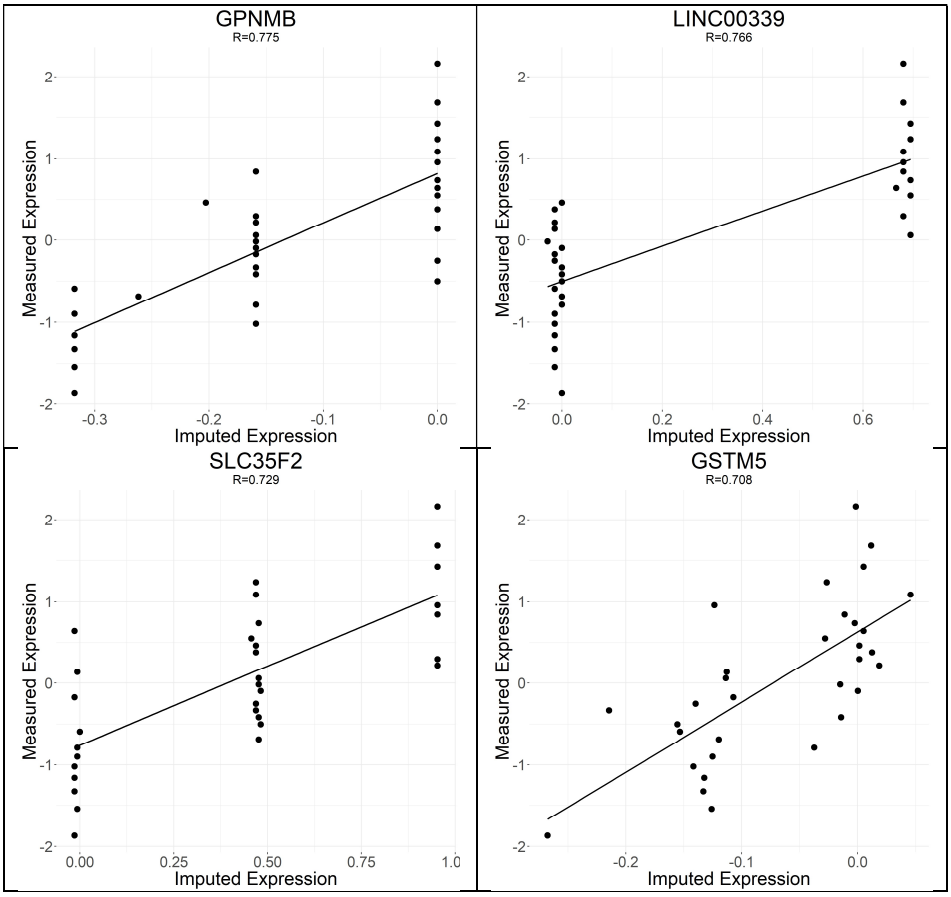

Figure 12: Examples of well predicted genes. GPNMB, LINC00339, SLC35F2, and GSTM5 had a predicted R^2 of 0.308, 0.521, 0.390, and 0.494.

Chapter 5

Lung cohort

While PrediXcan's performance was evaluated on two different platforms, the poor correlation may have been the consequence of any number of coding errors (i.e. incorrect annotations, incorrect dosage file format, etc) rather than a true measure of PrediXcan's accuracy. Unfortunately, the GTEx transcriptome dataset is not readily available, which prevented us from reproducing the findings of the original PrediXcan paper. Thus, to evaluate the previous chapters' methodology, the NCBI's Gene Expression Omnibus (GEO) was used to find a transcriptome dataset that offered the greatest likelihood of validating PrediXcan's imputation. This meant selecting a dataset that would be used alongside one of PrediXcan's more robust models (i.e. included a higher number of samples in the model), one that matched the donors used in PrediXcan's models (i.e. Europeans), and one that included healthy tissue.

The genotypic and small airway epithelium expression data from 15 healthy non-smoking European donors were selected from a larger publicly available transcriptome dataset (Butler, 2011). This dataset was downloaded from NCBI's GEO site on February 7th, 2019 (GEO ID: GSE22047). Additionally, this dataset contained similar platforms for the genotypic data (Affymetrix Genome-Wide Human SNP 5.0 Array) and expression data (Affymetrix Human Genome U133 Plus 2.0 Array). In total, preprocessed genotypic data pertaining to 443,816 SNPs and expression data pertaining to around 47,000 transcripts was provided. The lung SQLite weights file was selected, which utilized the GTEx v7 transcriptome dataset and incorporated information from 333 donors. The file included weights for 209,035 SNPs, which were associated with 7,968 genes; 6,145 were protein-coding (77%), 940 were long non-coding RNAs (12%), and 883 were pseudogenes (11%). Compared to the nucleus accumbens'

weights file, there were substantially more GTEx donors incorporated in the gene models (192%) and more genes (119%).

The methodology for this chapter closely follows chapter 3 with a few exceptions. First, the data was not background corrected or normalized as those steps were conducted as specified in the dataset's manuscript (Butler, 2011). Secondly, the expression annotation provided on the dataset's GEO page was used. As this annotation included the Ensembl stable IDs, those IDs were used instead of gene symbols for comparisons between the imputed and measured expression.

Prior to comparing the results, the imputed expression of 1,374 genes were removed due to all individuals having a value of zero (i.e. non-informative imputation). In total, 3,735 genes matched, which is likely the consequence of many more genes modeled in the SQLite weights file. The results are highlighted in table 7 and show a consistent improvement in the correlation by using predicted R^2 for filtering. The top 4 genes, with the highest Pearson correlation, are shown in figure 13; the gene models for these genes all had a substantially higher predicted R^2 than the average predicted R^2 in the lung SQLite weights file, which was 0.11. Only 41 genes (1.09% of matched genes) had a Pearson correlation ≥ 0.5.

Overall, the results highlight PrediXcan's ability to predict some genes well and validate the notion of using predicted R^2 as a metric to increase imputation quality. It should be noted that the GTEx expression data comes from tissue found 1 cm below the pleural surface of the lung, which is a heterogenous mixture of different cell types. As the expression data originated from a particular cell type (small airway epithelium), the results also give support to PrediXcan's methodology, which has been used to model genes within heterogenous tissue.

	ALL	$PR^2 \geq 0.1$	$PR^2 \geq 0.2$	$PR^2 \geq 0.3$	$PR^2 \geq 0.4$	$PR^2 \geq 0.5$	$PR^2 \geq 0.6$
Number of Genes:	3735	1450	626	320	158	69	16
First 5 Individuals:							
1	0.049	0.084	0.102	0.149	0.218	0.169	0.200
2	0.035	0.091	0.126	0.217	0.220	0.309	0.502
3	0.007	0.033	0.082	0.131	0.079	0.146	0.300
4	0.023	0.018	0.095	0.101	0.208	0.176	0.348
5	0.054	0.063	0.119	0.173	0.163	0.103	0.344
Mean (All Donors):	0.049	0.088	0.124	0.181	0.213	0.268	0.365
SD (All Donors):	0.275	0.301	0.318	0.313	0.314	0.320	0.378

Table 7. Pearson correlation of matching genes between the measured expression and imputed expression of lung tissue for a given individual. The first 5 donors were highlighted. Pearson correlation of matching genes were also calculated at certain predicted R^2 thresholds (≥0.1, ≥0.2, ≥0.3, ≥0.4, ≥0.5, or ≥0.6).

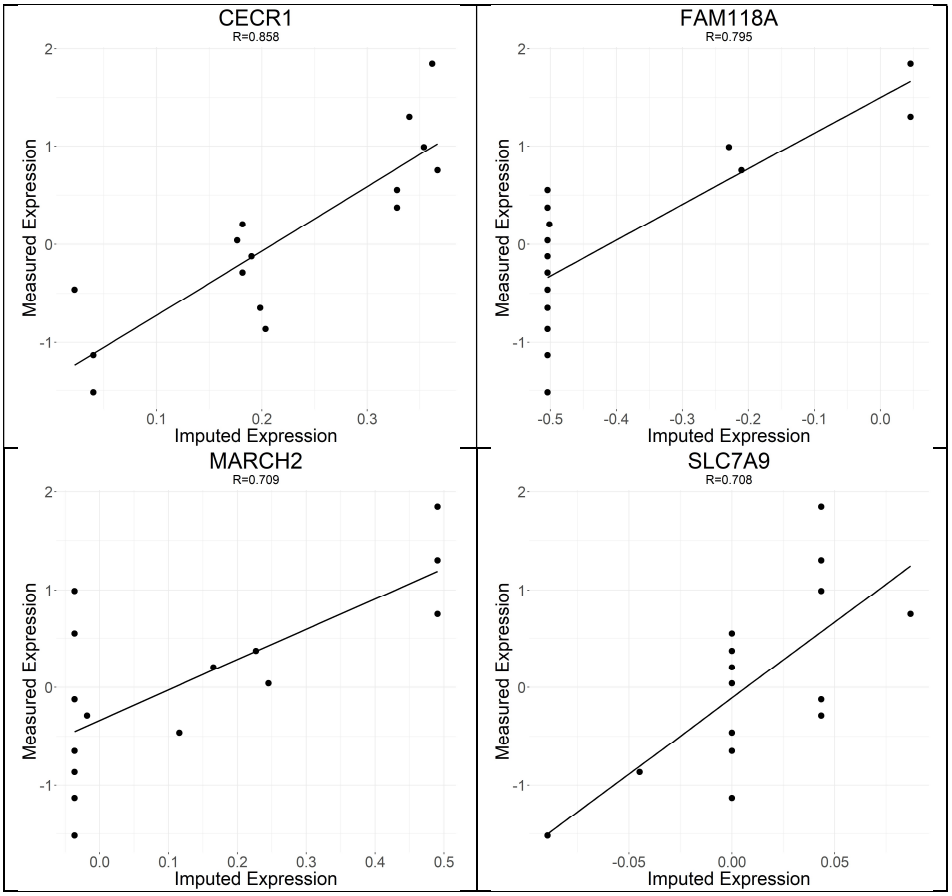

Figure 13: Examples of well predicted genes. CECR1, FAM118A, MARCH2, and SLC7A9 had a predicted R^2 of 0.306, 0.703, 0.365, and 0.267.

Chapter 6

Discussion

From the onset of this study, there had been two main goals. Firstly, evaluating PrediXcan ability to identify differential expression between donors with alcohol use disorder and neurotypical controls and secondly, assessing whether the predicted R^2 metric can be used to improve imputation quality. On both accounts the performance of PrediXcan was poor regardless of phenotype and tissue origins. However, the results did not invalidate PrediXcan's methodology altogether. PrediXcan was able to predict a handful of genes well and those genes largely had a predicted R^2 metric above the average gene model, which indicates that the well-predicted genes are not a factor of chance alone. Additionally, the lung dataset demonstrated that the predicted R^2 metric could be used to filter out poorly performing gene models and improve imputation. Considering that predicted R^2 is a surrogate for heritability, the results show that the strong performance of select genes is a consequence of high heritability.

Despite some of the propitious results of this study, there are apparent limitations to PrediXcan that urge a cautionary approach. Most apparent is the fact that the majority of the genes are not represented within each of the PrediXcan's tissue specific weights file. Additionally, the overall imputation quality with the alcohol use disorder dataset and lung dataset was poor. And if one utilizes a quality metric to limit genes with a high predicted R^2 or model R^2, the number of genes quickly is reduced. For instance, only 36 genes in the lung dataset could be compared when selecting gene models with a predicted $R^2 \geq 0.5$. On this basis alone, PrediXcan cannot replace traditional differential expression studies.

But even if one accepted those limitations, the well predicted genes highlight another concern. Despite these genes having predicted R^2 metrics above the average gene model in their weights file, these genes often did not have a strong predicted R^2. Put in another way, genes with higher quality metrics were not imputed with greater accuracy. This shortcoming can also be seen in the lung dataset, where the improved correlations markedly fell short of the predicted R^2 quality metric. As that metric is itself a correlation^2, one would expect the opposite. Additionally, it is worth mentioning that only a small percentage of compared genes (1.1%-2.7%) had a Pearson correlation ≥ 0.5 between the measured and imputed expression. Thus for the lung tissue, where the average correlation improved by filtering out genes with a low predicted R^2, that average was largely being driven by weak to modest correlations. Together, these factors not only highlight the rarity of strong models but an inability to discern strong models from weak models, which may lead to spurious results and mislead researchers.

While the notion of imputing gene expression from genotypic data alone may seem unreliable, especially considering the dynamic nature of gene expression, we did observe some limited success in the results. Thus while PrediXcan will likely not replace traditional approaches to measure gene expression, the authors nonetheless have laid a foundation for future development. For example, EpiXcan is another recent software, which builds upon that concept by incorporating epigenomic data into its models and reportedly outperforms PrediXcan (Zhang, 2019, *pre-print: https://dx.doi.org/10.1101/532929*). However, we have not tested EpiXcan here.

Given the merit in PrediXcan's methodology, it is our intention to continue to understand and explore ways to improve PrediXcan's imputation. Recently, we've been given access to genotypic and expression data from 326 Australian twins (Powell, 2012). As this cohort is from the same population as those in the alcohol use disorder dataset, it will allow us to also calculate narrow sense heritability of gene expression and thus circumvent the need for a surrogate quality metric. Additionally, we've applied for and have obtained access to the GTEx v7 transcriptome dataset. With this data, we will be

able to rebuild PrediXcan models, enact different quality control metrics, and attempt to improve the results of PrediXcan. Also, our access to the GTEx dataset may allow us to understand why the nucleus accumbens tissue performed so poorly. While the lower number of donors included in the elastic net methodology are a prime suspect, this can be tested by creating a lung model with the same number of donors and observing the outcome.

Bibliography

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders 5th edn (American Psychiatric Association, 2013).

Butler, M. et al. Modulation of Cystatin A Expression in Human Airway Epithelium Related to Genotype, Smoking, COPD and Lung Cancer. Cancer Res 71, 2572–2581 (2011).

Byerly, S. et al. Effects of ozone exposure during microarray posthybridization washes and scanning. J. Mol. Diagnostics 11, 590–597 (2009).

Chou, C. C., Chen, C. H., Lee, T. T. & Peck, K. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Res. 32, (2004).

Chovanec, M. et al. Long-term toxicity of cisplatin in germ-cell tumor survivors. Ann. Oncol. 28, 2670–2679 (2017).

Day, J. & Carelli, R. The Nucleus Accumbens and Pavlovian Reward Learning. Neuroscientist. 13, 148–159 (2007).

Dolan, M. E. et al. Clinical and Genome Wide Analysis of Cisplatin-Induced Peripheral Neuropathy in Survivors of Adult-Onset Cancer. Clin Cancer Res. 23, 5757–5768 (2017).

Durinck, S., Spellman, P., Birney, E. & Huber, W. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nat Protoc 4, 1184–1191 (2011).

Ezkurdia, I. et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Hum. Mol. Genet. 23, 5866–5878 (2014).

Feeney, G. F. X., Connor, J. P., Young, R. M. D., Tucker, J. & McPherson, A. Combined acamprosate and naltrexone, with cognitive behavioural therapy is superior to either medication alone for alcohol abstinence: A single centres' experience with pharmacotherapy. Alcohol Alcohol. 41, 321–327 (2006).

Ferrer, I., Martinez, A., Boluda, S., Parchi, P. & Barrachina, M. Brain banks: Benefits, limitations and cautions concerning the use of post-mortem brain tissue for molecular studies. Cell Tissue Bank. 9, 181–194 (2008).

Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47, 1091–1098 (2015).

Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. Nat. Rev. Mol. Cell Biol. 8, 113–126 (2007).

Gowon O. McMichael, John Drake, Eric Sean Vornholt, Kellen Cresswell, Vernell Williamson, Chris Chatzinakos, Mohammed Mamdani, Siddharth Hariharan, Kenneth S. Kendler, Michael F. Miles, Gursharan Kalsi, Brien P. Riley, Mikhail Dozmorov, Silviu-Alin Bacanu, V. I. V. Assessing the role of long-noncoding RNA in nucleus accumbens in subjects with alcohol dependence. pre-print bioRxiv (2019). doi:https://doi.org/10.1101/583203

Grant, B. F. et al. Epidemiology of DSM-5 Alcohol Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 72, 757–766 (2015).

Grundberg, E. et al. Mapping cis - and trans -regulatory effects across multiple tissues in twins. Nat Genet. 44, 1084–1089 (2012).

Hasin, D., Stinson, F., Ogburn, E. & Grant, B. Prevalence, Correlates, Disability, and Comorbidity of DSM-IV Alcohol Abuse and Dependence in the United States. Arch. Gen. Psychiatry 64, 830–842 (2008).

Hoffman, J. D. et al. Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet. 13, 1–19 (2017).

Hohman, T. J., Dumitrescu, L., Cox, N. J. & Jefferson, A. L. Genetic resilience to amyloid related cognitive decline. Brain Imaging Behav. 11, 401–409 (2017).

Kahan, B. C., Jairath, V., Doré, C. J. & Morris, T. P. The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. Trials 15, 1–7 (2014).

Laramée, P. et al. Risk of All-Cause Mortality in Alcohol-Dependent Individuals: A Systematic Literature Review and Meta-Analysis. EBioMedicine 2, 1394–1404 (2015).

Leek, J. T. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 1–15 (2010).

Li, B. et al. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. Pac Symp Biocomput 23, 448–459 (2018).

Malo, N., Libiger, O. & Schork, N. J. Accommodating Linkage Disequilibrium in Genetic-Association Analyses via Ridge Regression. Am. J. Hum. Genet. 82, 375–385 (2008).

Mamdani, M. et al. Integrating mRNA and miRNA weighted gene co-expression networks with eQTLs in the nucleus accumbens of subjects with alcohol dependence. PLoS One 10, 1–28 (2015).

Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. B-Biological Sci. 368, (2013).

Powell, J. E. et al. Genetic control of gene expression in whole blood and lymphoblastoid cell. Genome Res. 22, 456–466 (2012).

Prescott, C. A. & Kendler, K. S. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. Am. J. Psychiatry 156, 34–40 (1999).

Price, A. L. et al. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. 7, (2011).

Quackenbush, J. Microarray data normalization and transformation. Nat. Genet. 32, 496–501 (2002).

Ray, L. A., Bujarski, S., Roche, D. J. O. & Magill, M. Overcoming the 'Valley of Death' in Medications Development for Alcohol Use Disorder. Alcohol. Clin. Exp. Res. 42, 1612–1622 (2018).

Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).

Roerecke, M. & Rehm, J. Alcohol use disorders and mortality: A systematic review and meta-analysis. Addiction 108, 1562–1578 (2013).

Scofield, M. D. et al. The Nucleus Accumbens: Mechanisms of Addiction across Drug Classes Reflect the Importance of Glutamate Homeostasis. Pharmacol. Rev. 68, 816–871 (2016).

Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Stat. Appl. Genet. Mol. Biol. 3, 1–25 (2004).

Su, Z. et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 32, 903–914 (2014).

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013).

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. & Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. Front. Genet. 4, 1–11 (2013).

Xin, J. et al. High-performance web services for querying gene and variant annotation. Genome Biol. 17, 1–7 (2016).

Zhang, W. et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. bioRxiv Prepr. 1–24 (2019). doi:https://dx.doi.org/10.1101/532929

Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320 (2005).

Appendix

Arraystar AUD Cohort:	Gene	SNPS in Model	Predicted R^2	Pearson Correlation
	ELAC2	7	0.323	0.697
	SPSB2	22	0.418	0.691
	SMYD4	38	0.139	0.643
	LRRC37A	38	0.531	0.615
	CNGA1	33	0.403	0.597
	XRCC1	16	0.079	0.576
	SNX31	23	0.287	0.568
	ZNF626	23	0.056	0.559
	ZNF880	50	0.712	0.554
	HBS1L	65	0.146	0.528
	LINC00667	19	0.339	0.527
	DUS3L	16	0.328	0.522
	POMZP3	60	0.581	0.520
	L3HYPDH	48	0.385	0.502

Table 8. Well predicted genes (Pearson correlation ≥ 0.50 between imputed and measured expression) for Arraystar alcohol use disorder cohort with recoding of missing SNPs.

Affymetrix AUD Cohort:	Gene	SNPS in Model	Predicted R^2	Pearson Correlation
	GPNMB	13	0.308	0.775
	LINC00339	21	0.521	0.766
	SLC35F2	7	0.390	0.729
	APIP	103	0.660	0.708
	GSTM5	24	0.494	0.708
	PILRB	56	0.591	0.677
	NMRK1	54	0.497	0.674
	KNOP1	49	0.498	0.656
	ALDH8A1	71	0.330	0.646
	TMEM80	33	0.328	0.629
	GSTM3	49	0.241	0.625
	SLC27A2	11	0.300	0.613
	ERMAP	7	0.285	0.607
	STAG3L4	24	0.495	0.604
	TYW1	41	0.415	0.590
	MTMR3	41	0.195	0.585
	DHRS11	47	0.237	0.579
	CHAF1A	42	0.278	0.573
	HBS1L	65	0.146	0.573
	C1GALT1	47	0.303	0.571
	HPR	14	0.587	0.571
	LIN7C	12	0.053	0.568
	LXN	46	0.132	0.566
	SNRPC	66	0.367	0.566
	SPATA20	8	0.611	0.556
	TCFL5	191	0.205	0.544
	MED17	13	0.048	0.542
	LARS2	41	0.298	0.539
	RPS18	7	0.047	0.539
	MARCH2	35	0.417	0.532
	MTCH2	24	0.226	0.527
	POLI	137	0.454	0.526
	UROS	45	0.579	0.517
	TRAPPC4	10	0.493	0.516
	MFF	64	0.256	0.515
	PMS2P3	40	0.160	0.513
	LDAH	6	0.400	0.512
	ST7L	17	0.053	0.512
	PIGZ	81	0.392	0.510
	XRCC1	16	0.079	0.508
	TOR1B	18	0.064	0.504

Table 9. Well predicted genes (Pearson correlation ≥ 0.50 between imputed and measured expression) for Affymetrix alcohol use disorder cohort.

Affymetrix Lung Cohort:	Gene	SNPS in Model	Predicted R^2	Pearson Correlation
	CECR1	38	0.306	0.858
	FAM118A	30	0.703	0.795
	MARCH2	34	0.365	0.709
	SLC7A9	45	0.267	0.708
	PLD1	45	0.172	0.704
	GSTT2	42	0.608	0.687
	SNAPC1	13	0.108	0.685
	TRIOBP	21	0.051	0.676
	SAR1A	29	0.289	0.652
	DHX8	44	0.019	0.651
	ARSA	7	0.311	0.649
	TFRC	42	0.036	0.636
	POMT2	33	0.093	0.632
	TMEM230	63	0.216	0.619
	TIE1	22	0.111	0.614
	CATSPERG	39	0.243	0.614
	MOV10L1	22	0.016	0.605
	CDIP1	17	0.084	0.600
	TFIP11	68	0.015	0.598
	GDPD3	20	0.074	0.591
	PPIL2	34	0.276	0.585
	MTMR3	27	0.165	0.583
	MKRN2	61	0.088	0.577
	CHMP2B	41	0.010	0.572
	CPNE3	49	0.107	0.571
	OXT	30	0.025	0.570
	IL27RA	16	0.049	0.568
	HSD17B7P2	20	0.102	0.564
	SIRT4	21	0.079	0.557
	ASB4	119	0.145	0.554
	WAC	13	0.055	0.533
	CECR5	32	0.015	0.532
	TUBB4A	10	0.043	0.525
	DAPP1	33	0.060	0.524
	RBM6	52	0.439	0.523
	EIF4B	17	0.111	0.522
	CACNA1I	9	0.016	0.520
	LIN7B	10	0.081	0.516
	TTC19	34	0.149	0.514
	GRAMD1A	28	0.120	0.502
	CD22	36	0.251	0.501

Table 10. Well predicted genes (Pearson correlation ≥ 0.50 between imputed and measured expression) for Affymetrix lung cohort.