Analysing the effects of alcoholism on gene expression

JAKUB BIAŁECKI

June 27, 2022

1. Abstract

Alcohol dependece (or alcoholism) is a condition which affects many individuals and their close ones. It is of great importance to understand the biological mechanisims which come into play regarding this condition. In this project we tried to explore the relationships between gene expression and alcoholism from a dataset of expression levels in brain tissiue of 35 men, divided into control and disease. We performed simple dimension reduction using PCA in hopes of finding some clues and managed to identify 10 potential genes of interest, which could be serve as our foothold into further analysis.

2. Introduction

It is well known that prolonged alcohol consumption leads to a multitude of problems for the individual. Chronic alcoholics will in time find themselves battling against a damaged liver, irregular heartbeat, signs of demetia or all of them combined. Additionally, the condition not only impacts the individual themsleves, but also family members, close friends and associates due to the high possibility of belligirent outbursts of the patient. The fact that ethanol is a relatively cheap substance to produce only contributes to the issue and makes it more commonplace that people would like to admit. Therefore, it is imperative to study the nature of alcohol addiction so we can devise better treatments for those affected, as well as make predictions who bears the highest risk of falling into addiction.

The condition itself has two componets: genetic and enviromental, with the genetic component contributing 40-60% to the risk of becoming addicted [1]. In this study we will focus on the genetic foundation and analyse the relations between gene expression in alcoholics versus control, with the view to finding interesting patterns, which could help identify genes responsible for (or the most affected by) alcoholism. The data we used came from Gene Expression Omnibus database and the query ID was 'GSE161986' [2]. It contained array based profiling of expression data from postmortem brain tissue samples. The samples underwent pre-selection based on (1) history of infectious disease, (2) circumstances surrounding death, (3) substantial brain damage, and (4) post-mortem interval $>$ 48 hours. In total there were 35 samples, 18 of which were from cases of alcohol dependance, while the rest constituted the control group. All of the samples came from men aged from 39 to 82, while most of them falling between 45 and 65.

In our analysis we first performed Principal Component Analysis to get a quick look at the data and see if there are any immediate patterns that pop up. Then we looked at the correlations between expression in control versus affected samples. Lastly, we calculated the difference between mean expressions of genes in control versus affected samples and selected 10 genes where the difference was the largest.

3. Results

In order to delve into the topic of relationships between gene expression and alcohol dependence we selected a fairly recent dataset (Nov 24, 2020) from the postmortem brain tissue, as the brain is one of the most affected organs by alcoholism. The expression data consisted of a 22215 by 35 table (22215 genes and 35 samples) and supplementary metadata, of which we focused on the type of sample (control/affected) and age of the men. The data in the table was provided in the log-trasformed form.

The first thing we decided to do was PCA, in order to see if we could quickly determine whether or not a pattern would emerge after dimensional reduction. We set 10 as our cutoff point and exluded all genes which had a mean expression less than the cutoff. Afterwards, we scaled the data around row-wise using the R 'scale' function. With the data prepared, we performed singular value decomposition and extracted the right singular vectors which corresponded to principal components. As a result, we obtained 35 PC's, the top 5 of which explained more than 80% of the variance. (Fig. 1) We then visualised the combinations of top 5 PC's (one PC on the X axis, the other PC on the y axis) and colored them by type of sample. The results were inconslusive (Fig. 2) so we moved on with our analysis. We also tried to group the points by the respective age group (Fig. 3) but from here we also couldn't draw any conclusions.

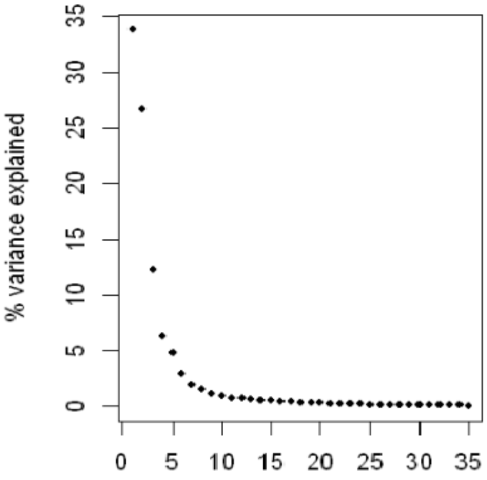


Figure 1: The percentage of the variance in the data explained by each principal component (PC)

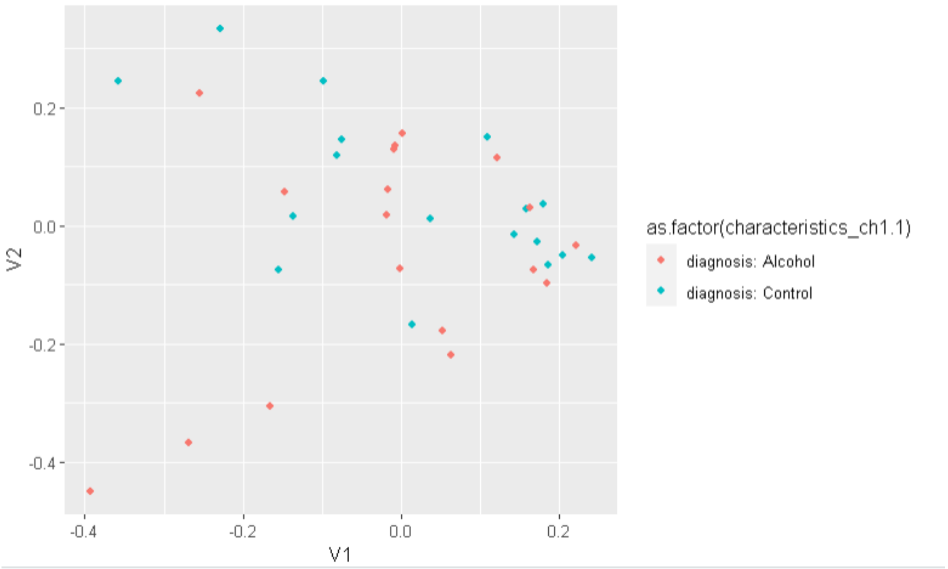


Figure 2: The first two principal components, denoted by V1 and V2 and colored by age group. The rest of the combinations ie. V1 and V3, V2 and V3 and so on can be found in the R notebook on github

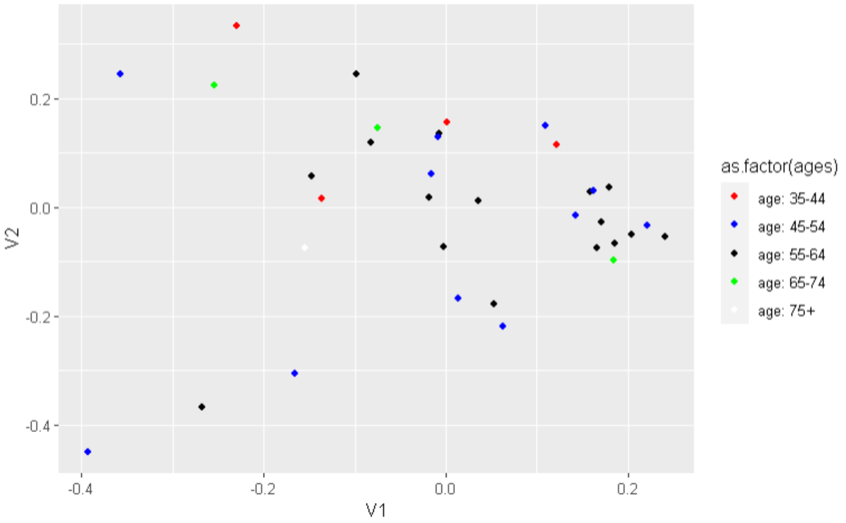
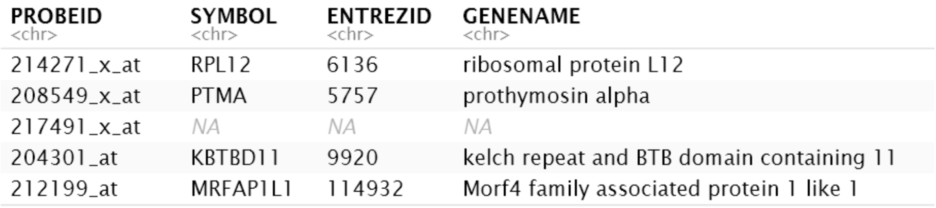


Figure 3: The first two principal components, denoted by V1 and V2 and colored by sample type. The rest of the combinations ie. V1 and V3, V2 and V3 and so on can be found in the R notebook on github

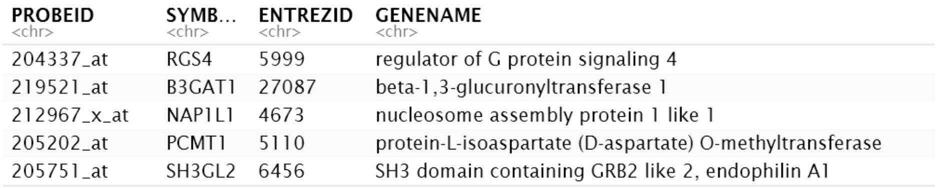
After performing the Principal Component Analysis, we considered another approach. We decided to search for particular genes, those whose expression greatly differed between alcohol and control samples. To do that, we calculated the means of all genes for both sample types and measured the difference between them. We selected the top 5 genes which were the most over-expressed relative to control, as well as the top 5 which were the most under-expressed. Because the original data did not come with gene ID's, but rather with probe ID's, we had to convert the probes into genes.

We did that using the 'hgu133plus2' annotation package [3]. The tables with overexpressed genes are shown on (Fig. 4) and underexpressed genes on (Fig. 5). The probe '217491_x_at' had to be annotated manually and corresponds to the gene COX7C - cytochrome c oxidase subunit 7C (Homo Sapiens). It is important to note that while the lists both contain the top 5 genes from both ends of the spectrum, the expression levels have diffrent ranges. Overexpressed genes had larger absolute diffrence in mean expression ranging from 0.75 to 1.43 while for underexpressed genes it was 0.56 to 0.67. By far the largest difference (1.43) was found in the 'Morf4 family associated protein 1 like 1' gene.



PROBEID <chr>	SYMBOL <chr>	ENTREZID <chr>	GENENAME <chr>
214271_x_at	RPL12	6136	ribosomal protein L12
208549_x_at	PTMA	5757	prothymosin alpha
217491_x_at	NA	NA	NA
204301_at	KBTBD11	9920	kelch repeat and BTB domain containing 11
212199_at	MRFAP1L1	114932	Morf4 family associated protein 1 like 1

Figure 4: List of the most overexpressed genes relative to control. Probe 217491_x_at corresponds to gene COX7C, ID = 1350



PROBEID <chr>	SYMB... <chr>	ENTREZID <chr>	GENENAME <chr>
204337_at	RGS4	5999	regulator of G protein signaling 4
219521_at	B3GAT1	27087	beta-1,3-glucuronyltransferase 1
212967_x_at	NAP1L1	4673	nucleosome assembly protein 1 like 1
205202_at	PCMT1	5110	protein-L-isoaspartate (D-aspartate) O-methyltransferase
205751_at	SH3GL2	6456	SH3 domain containing GRB2 like 2, endophilin A1

Figure 5: List of the most underexpressed genes

4. Discussion

Alcohol dependence continues to be a growing health concern among the general populace. It would therefore be benefitial to understand the connections between genetic factors and the disease itself. Our findings present a list of 10 genes of interest, which could serve as a starting point for further research. It is especially important to analyse the 'Morf4 family associated protein 1 like 1' gene, as it presented the largest difference in expression. However, we shouldn't overlook underexpressed genes such as the 'regulator of G protein signaling 4' gene, where although the difference wasn't as big, it was still significant.

5. Methods and materials

The data for this project was fetched from Gene Expression Omnibus (GEO) with the query ID = 'GSE161986'. All analysis was done in RStudio version 4.1.3 and used the following packages:

- Biobase

- GEOquery
- RColorBrewer
- gplots
- data.table
- ggplot2
- annotate
- hgu133plus2.db
- patchwork

We performed the Singular Value Decomposition (SVD) to find the principal components and ggplot2 to visualise them. Aside from packages mentioned we used generic R functions ex. when calculating the means of expressions. The code can be found at https://github.com/Kubinho1/CBS/tree/main/project.

References

[1] H. J. Edenberg and T. Foroud, "Review: The genetics of alcoholism: identifying specific genes through family studies," *Addiction Biology*, vol. 11, no. 3-4, pp. 386–396, 2006.

[2] E. Vornholt, J. Drake, M. Mamdani, G. McMichael, Z. N. Taylor, S.-A. Bacanu, M. F. Miles, and V. I. Vladimirov, "Network preservation reveals shared and unique biological processes associated with chronic alcohol abuse in nac and pfc," *PLOS ONE*, vol. 15, pp. 1–19, 12 2020.

[3] M. Carlson, "hgu133plus2.db: Affymetrix human genome u133 plus 2.0 array annotation data," R package version 3.13.0.