5003 FINAL ANALYSIS

2023-12-14

Gene expression analysis of Bipolar Disorder

A differential gene expression analysis is conducted amongst healthy control groups and those with Bipolar Disorder (BPD). The purpose of this assignment is to visualize differentiated gene expression profiles using limma (as the initial source used ANOVA), while also visualizing patterns with different phenotypic variables to determine the most prominent risk factors and diagnostic markers for this condition.

Dataset being used

Dataset Citation: Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT et al. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatry 2006 Oct;11(10):965-78. PMID: 16894394

Importing necessary libraries

Accessing the dataset

The structure and features/phenotypes of the dataset is assessed.

```
#Accessing data
data <- getGEO("GSE5389", GSEMatrix = TRUE, getGPL = TRUE)
```

```
## Found 1 file(s)
```

```
## GSE5389_series_matrix.txt.gz
```

```
data <- data[[1]] #simplify data structure and access primary information

#Getting a glimpse of data
exprs(data)[1:5, 1:5] # checking rows and columns
```

```
##            GSM123243 GSM123244 GSM123245 GSM123246 GSM123247
## 1007_s_at 1023.09929 591.85519 611.75265 491.68734 500.78403
## 1053_at     37.94475  43.60739  39.30148  37.18039  38.74891
## 117_at      50.02898  55.44346  36.80570  49.05599  49.12265
## 121_at     372.89125 317.76679 336.86170 352.93190 381.20518
## 1255_g_at   16.38924  18.82816  15.18879  17.86694  19.09526
```

```
data # checking summary
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22283 features, 21 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM123243 GSM123244 ... GSM123263 (21 total)
##   varLabels: title geo_accession ... Valproate treatment:ch1 (60 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 1007_s_at 1053_at ... AFFX-TrpnX-M_at (22283 total)
##   fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16 total)
##   fvarMetadata: Column Description labelDescription
## experimentData: use 'experimentData(object)'
##   pubMedIds: 16894394
## Annotation: GPL96
```

```
# str(data) #checking data structure
class(data) # checking data type, to confirm expression data
```

```
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

```
# head(fData(data)) # checking feature data (gene information)
# head(pData(data)) # checking phenotypic data (sample conditions, etc)

#understanding/exploring dimensions
dim(exprs(data))
```

```
## [1] 22283    21
```

```
dim(pData(data))
```

```
## [1] 21 60
```

```
dim(fData(data))
```

```
## [1] 22283    16
```

Dataset information

The dataset has 22,283 rows of genes/features and 21 columns of samples and phenotypes. Most of the data elements being a character (string), or integer.

Phenotypes

The phenotypes present in our data include the basic information on the actual data, such as the sample name, accession (ID) number, public status, and source data (species, tissues source, etc)

It also contains more important information for our analysis - Ch1: Disease status - 1.1: Subject age - 1.2: Gender - 1.3: Disease age of onset - 1.4: Duration of ilness - 1.5: Brain pH - 1.6: Time after death where sample was taken - 1.7: Side of brain sample was taken from - 1.8: Dosage of treatment (Fluphenazine) present within the sample - 1.9: Lithium treatment performed on patient - 1.10: Valproate treatment performed on patient - 1.11: Electroconvulsive therapy performed on patient - 1.12: Did the subject die from suicide - 1.13: Drug Abuse on a 1-5 scale - 1.14: Alcohol abuse

My hypothesis

As indicated previously, the purpose of this analysis is to find differentiated genes amongst those with BPD, and find the importance of these phenotypic characteristics and find any potential risk factors to BPD.

I predict the differentiation of a plethora of genes. Not only this, but I specifically suspect that such factors include as drug abuse and alcohol abuse, as they may have some genetic underpinnings to causing BPD.

Exploring and Processing the data

Presence of NA values are checked in the expression and the phenotype values, and also look further into the fData and pData to see if there are any unecessary/troublesome information. These are all commented out as it isn't relevant to the reader, but code is given to show exploration process.

```
#checking for na values
table(na.omit(exprs(data)) == TRUE) # this printed all FALSES meaning there isn't any NA expression val
```

```
##
##  FALSE
## 467943
```

```
#checking features and phenotypes
# head(fData(data))
# head(pData(data)) #repetitive columns found

##checking characteristics values (to explore anything odd in each data)
filtered_colnames <- grep("^characteristics_ch1", colnames(pData(data)), value = TRUE)

# Loop through each filtered column name and apply table function (commented out, but shown to display
# for (col in filtered_colnames) {
#   cat("Table for", col, ":\n")
#   print(table(pData(data)[[col]]))
#   cat("\n")
# } #all of them seem good, only NA values are control
```

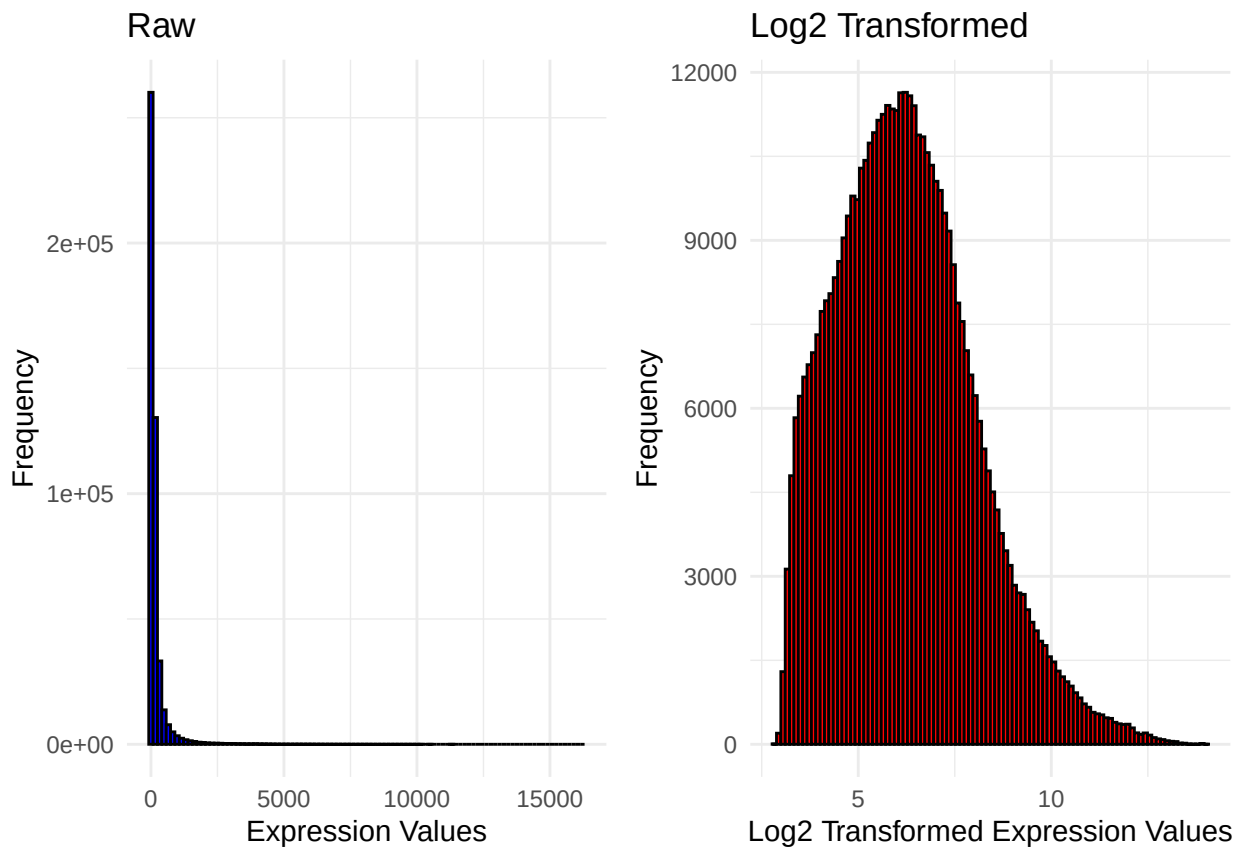
There are many miscellaneous information, including the presence of multiples of columns, as well logistical experiment information such as the biomaterial provider and protocol. These make up for most of our columns. Therefore, these columns will be filtered out.

Besides that, this dataset is relatively clean, therefore wrangling was not required.

```r
#subsetting first 24 columns
pData(data) <- pData(data)[, 1:24]
# pData(data) #check
```

Lets also take a look at the distribution of the data, and see if we can transform it to meet a normal distribution

```r
# Assuming 'exprs(data)' returns a matrix where columns are samples
# Convert expression values to a data frame for ggplot2
expression_values <- as.vector(exprs(data))
data_df <- data.frame(expression_values = expression_values)

# Histogram before transformation
raw_data <- ggplot(data_df, aes(x = expression_values)) +
  geom_histogram(bins = 100, fill = "blue", color = "black") +
  labs(title = "Raw",
       x = "Expression Values",
       y = "Frequency") + theme_minimal()

# Log2 transformation
log2_expression_values <- log2(expression_values + 1)  # Adding 1 to avoid log2(0)
data_log2_df <- data.frame(log2_expression_values = log2_expression_values)

# Histogram after log2 transformation
transformed_data <- ggplot(data_log2_df, aes(x = log2_expression_values)) +
  geom_histogram(bins = 100, fill = "red", color = "black") +
  labs(title = "Log2 Transformed",
       x = "Log2 Transformed Expression Values",
       y = "Frequency") +
  theme_minimal()


# Combine plots side-by-side
plot_grid(raw_data, transformed_data, ncol = 2, rel_widths = c(0.5, 0.5))
```




```
#Transforming data
exprs(data) <- log2(exprs(data))
```

Figure 1A and 1B: Distribution Histogram. The initial distribution of the dataset appeared notably skewed (Figure 1A). A log2 transformation was performed to mitigate this skewness and render the data more evenly distributed and representative (Figure 1B), resulting in a more Gaussian appearance and averting skewed outcomes.

PCA Analysis (along with some exploration)

Although this isn't particularly necessary for a DGE analysis, a PCA was conducted for initial exploratory analysis purposes. This was to find initial patterns, and determine which phenotypic traits seemed to express the most variance.

```
#generate pca
pca <- prcomp(t(na.omit(exprs(data))))
head(pca$x)[, 1:5] #check
```

```
##                   PC1        PC2         PC3        PC4        PC5
## GSM123243 -76.5142116   5.175249   1.7785242  -1.686203 -5.5191986
## GSM123244 -11.1898881 -45.174866  -1.7477647 -14.498252  0.5237915
## GSM123245  -2.4476624  12.560599   0.1810471  10.751140 -4.9506697
## GSM123246  10.4755950  -9.677050   1.8699374  -3.701150 -0.5335590
```

```
## GSM123247  -0.4495688 -32.610942   5.7513669  -5.838328 -4.6211775
## GSM123248   5.7427304   3.006894 -13.8414604  25.470970 -0.0840455
```

```
#plotting
pcaPlot <- data.frame(pca$x) #converting to dataframe for plotting
p1 <- ggplot(data = data.frame(pca$x), aes(x = PC1, y = PC2, colour = factor(data$characteristics_ch1.1
  theme(plot.caption = element_text(size = 7))
p2 <- ggplot(data = data.frame(pca$x), aes(x = PC1, y = PC2, colour = factor(data$characteristics_ch1.2
  theme(plot.caption = element_text(size = 7))

plot_grid(p1, p2, nrow = 2)
```

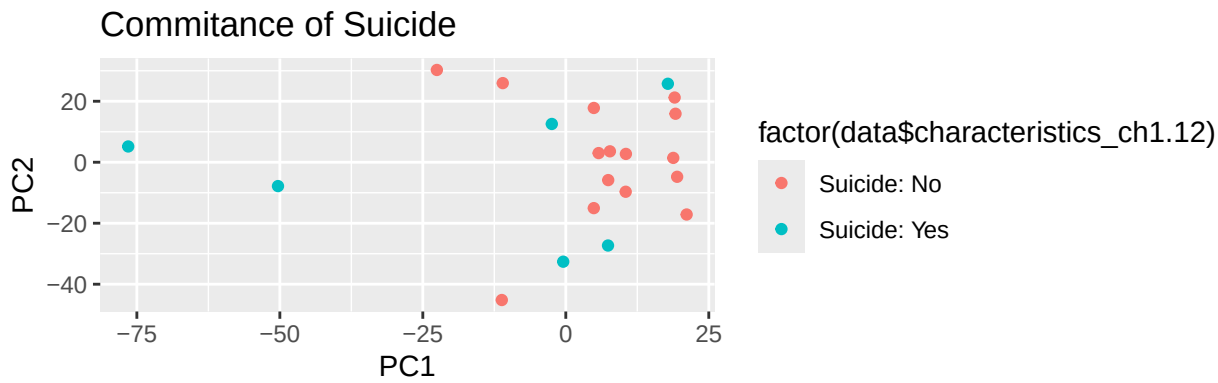


Figure 2A: Suicide is the driver of PC1

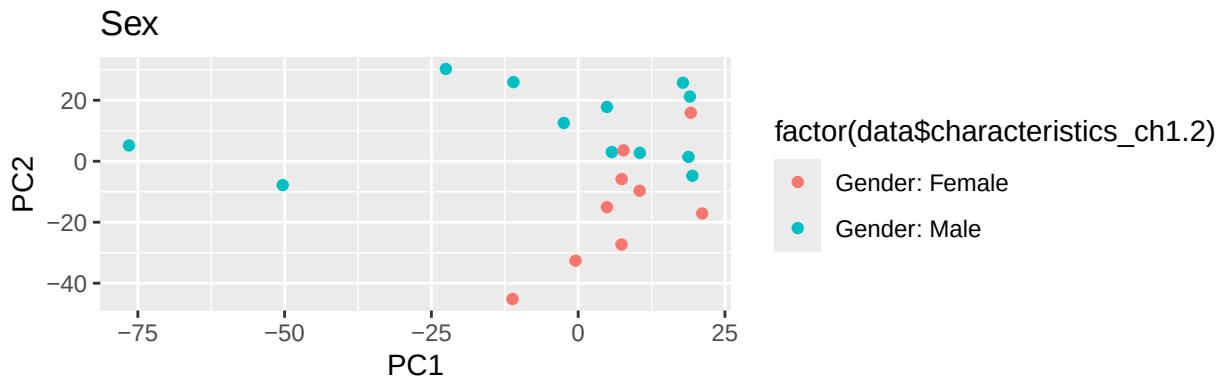
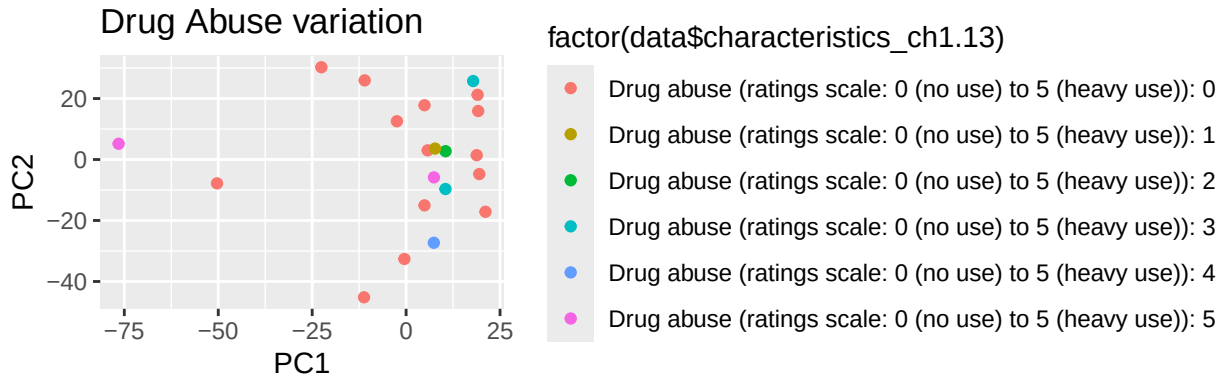


Figure 2B: Sex is the driver of PC1

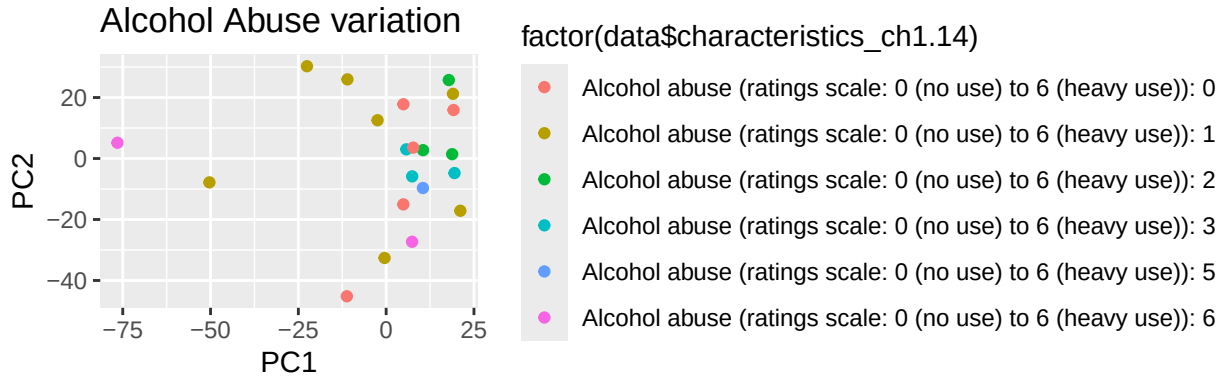
Figure 2A and 2B: Suicide commitance and Sex are the drivers of variation within the PCA, as they are differentiated along the "zero" value.

PC1 seems to be separated by suicide factor, while PC2 is separated by sex. However, there seemed to be some reasonable overlap for PC1, so there may be more factors at play here.

```
#Let's also plot drug and alcohol abuse to find any quick initial patterns
p3 <- ggplot(data = data.frame(pca$x), aes(x = PC1, y = PC2, colour = factor(data$characteristics_ch1.1
  theme(plot.caption = element_text(size = 7))
p4 <- ggplot(data = data.frame(pca$x), aes(x = PC1, y = PC2, colour = factor(data$characteristics_ch1.1
  labs(title = "Alcohol Abuse variation", caption = "Figure 2D: Alcohol Abuse variation: No significant
  theme(plot.caption = element_text(size = 7))

plot_grid(p3, p4, nrow = 2)
```



2C: Drug abuse variation: Some significant clustering is found



: Alcohol Abuse variation: No significant clustering is found

Figure 2C and 2D: Drug and Alcohol Abuse variation within PCA. Initial visualizations show little clustering and correlations.

There does seems to be some differentiation among PCs for drug abuse. This will be further explored.

```
##plot all other columns for exploration
#filtered_colnames <- grep("^characteristics_ch1", colnames(data), value = TRUE)

# Loop through each filtered column name and apply table function (commented out to prevent output over
# for (col in filtered_colnames) {
#   cat("Table for", col, ":\n")
#   print(ggplot(data = data.frame(pca$x), aes(x = PC1, y = PC2, colour = factor(data[[col]]))) + geom_
#   cat("\n")
# }
```

From plotting every other column, interesting or eye-catching information was not found. This is probably due to the low sample size.

Conducting differential gene expression analysis

Differentially expressed are assessed between the two groups. This is initiated by establishing contrasts between the control and disease group.

```
#define the two groups (control and disease)
design_bpd <- model.matrix(~0 + data$characteristics_ch1)
```

```r
# head(design_bpd) #check

#rename two group columns
colnames(design_bpd) <- c("BPD", "HC")
# head(design_bpd) #check

#establish contrast between groups
cont_matrix_bpd <- makeContrasts(BPD-HC, levels = design_bpd)
cont_matrix_bpd
```

```
##       Contrasts
## Levels BPD - HC
##    BPD         1
##    HC         -1
```

Table 1: Contrast Matrix. A contrast is established between the control and disease group. Groups are categorized to assess the gene expression differences between them.

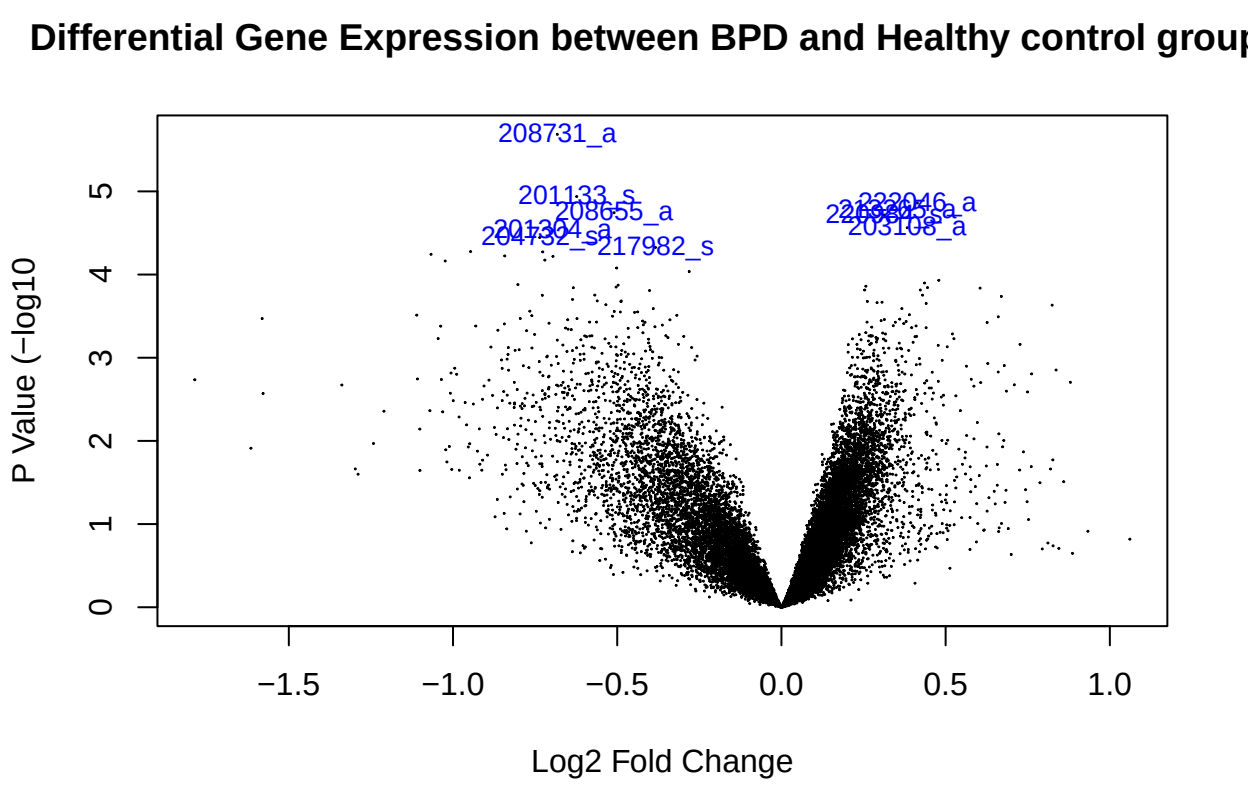
Differential Gene Expression Analysis is performed, via linear model systems limma.

```r
# fit linear model for each gene given the design matrix
fit_bpd <- lmFit(data, design_bpd)

# compute contrasts from linear model
fit2_bpd <- contrasts.fit(fit_bpd, cont_matrix_bpd)

# compute statistics based on lm
fit2_bpd <- eBayes(fit2_bpd, 0.01)

class(fit2_bpd) #check
```

```
## [1] "MArrayLM"
## attr(,"package")
## [1] "limma"
```

Once the model is created and fitted, expression values are visualized.

```r
#create volcano plot

volcanoplot(fit2_bpd,
            main = "Differential Gene Expression between BPD and Healthy control groups)",
            highlight = 10, hl.col = "blue", pch = 20, cex = 0.1,
            xlab = "Log2 Fold Change", ylab = "P Value (-log10")
```




```
#There seems to be no significantly expressed genes
```

Figure 3: Volcano Plot. Noticeable differentiation is observed among various genes, such as 208731_a, 201133_s, 222946_a, etc. Among these genes, 208731_a exhibited the most significant differentiation in terms of p-value magnitude. However, none reached the level of significance for coloration.

Most differentially expressed genes were stored in a list, then viewed.

```
# adjust for multiple testing (control false discovery rate)
top_bpd <- topTable(fit2_bpd, adjust.method = "fdr", number = 25)
#head(data.frame(top_bpd))

#find which columns need to be expressed to derive necessary information of the top 25
colnames(top_bpd)
```

```
##  [1] "ID"                                 "GB_ACC"
##  [3] "SPOT_ID"                            "Species.Scientific.Name"
##  [5] "Annotation.Date"                    "Sequence.Type"
##  [7] "Sequence.Source"                    "Target.Description"
##  [9] "Representative.Public.ID"           "Gene.Title"
## [11] "Gene.Symbol"                        "ENTREZ_GENE_ID"
## [13] "RefSeq.Transcript.ID"               "Gene.Ontology.Biological.Process"
## [15] "Gene.Ontology.Cellular.Component"   "Gene.Ontology.Molecular.Function"
## [17] "logFC"                              "AveExpr"
## [19] "t"                                  "P.Value"
## [21] "adj.P.Val"                          "B"
```

```r
#display table with necessary data
head(top_bpd[, c(1, 10, 11, 17:22)], 1) #showing output
```

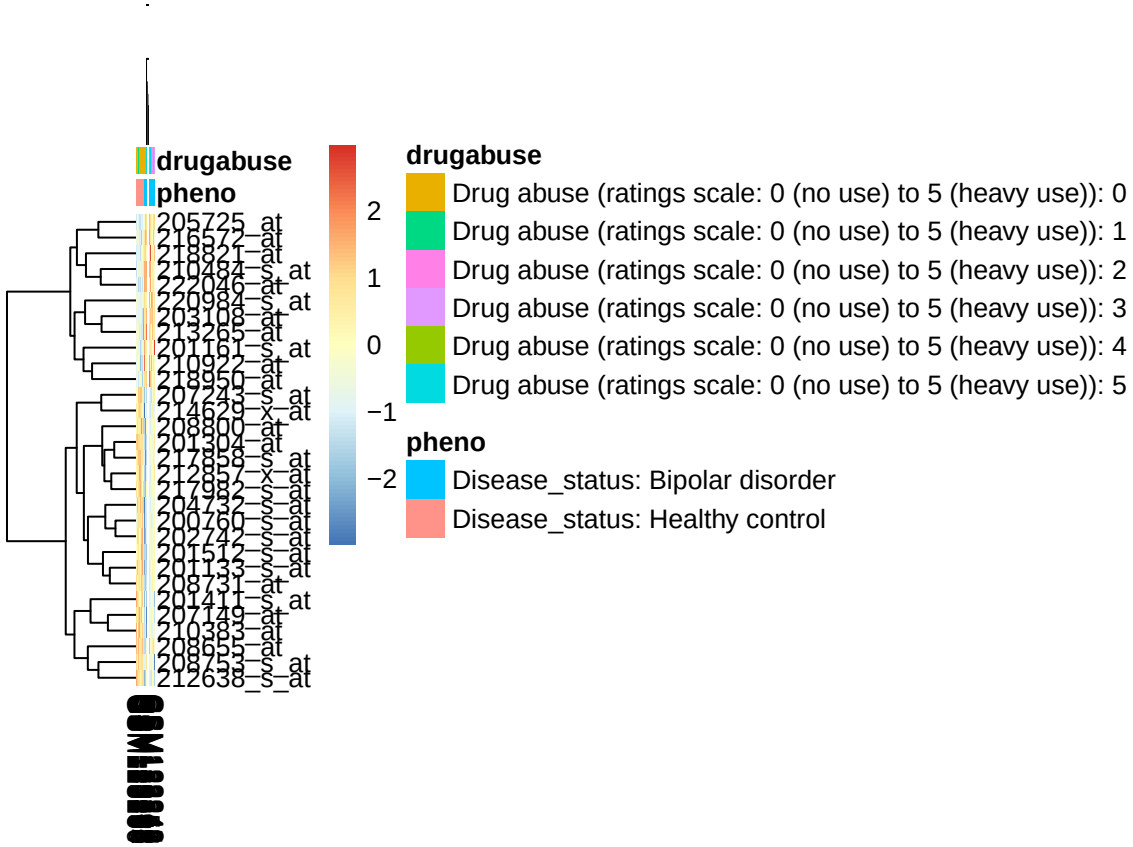
```
##                  ID                         Gene.Title Gene.Symbol      logFC
## 208731_at 208731_at RAB2A, member RAS oncogene family       RAB2A -0.6822309
##           AveExpr         t      P.Value  adj.P.Val        B
## 208731_at 9.65676 -6.328329 2.062249e-06 0.04595309 4.605781
```

Table 2: Top Differentiated Genes. Upon examination of the top_bpd list, it is evident that no genes exhibit significant differential expression, as indicated by adjusted p-values considerably greater than 0.01. Only one gene, 208731 (RAB2A), displays a p-value less than 0.05. Furthermore, none of the genes exhibit a logFC greater than 1.5, indicating a lack of substantial fold change in expression levels.

Generating a heatmap

A heatmap is generated to finally compare the differentiated values between different groups and phenotypes. This is done by creating different phenotypes for each phenotypic factor to see if there are any potential associations between phenotypes and certain gene expressions. Since our hypothesis states alcohol and drug abuse are the most prominent risk factors, those phenotypes will be plotted on the heatmap.

```r
# top 30 genes
tt_data <- topTable(fit2_bpd, adjust.method = "fdr", number = 30)

# check significance cut off, annotate which rows/genes are differentially expressed
tt_data$diff <- tt_data$adj.P.Val < 0.05 & abs(tt_data$logFC) > 0.5

# create data frame (including sample, diagnosis, and batch) # to colour by (for drug abuse)
anno_colour <- data.frame(row.names = colnames(data),
                          pheno = as.factor(pData(data)$characteristics_ch1),
                          drugabuse = as.factor(pData(data)$characteristics_ch1.13))

phm_data <- data[rownames(data) %in% rownames(tt_data),]
drug_heatmap <- pheatmap(phm_data, scale = "row", annotation_col = anno_colour, cutree_cols = 3)
```




```
# creating a heatmap for alcohol abuse
anno_colour <- data.frame(row.names = colnames(data),
                          pheno = as.factor(pData(data)$characteristics_ch1),
                          alcoholabuse = as.factor(pData(data)$characteristics_ch1.14))

phm_data <- data[rownames(data) %in% rownames(tt_data),]
alc_heatmap <- pheatmap(phm_data, scale = "row", annotation_col = anno_colour, cutree_cols = 3)

drug_heatmap
alc_heatmap
```

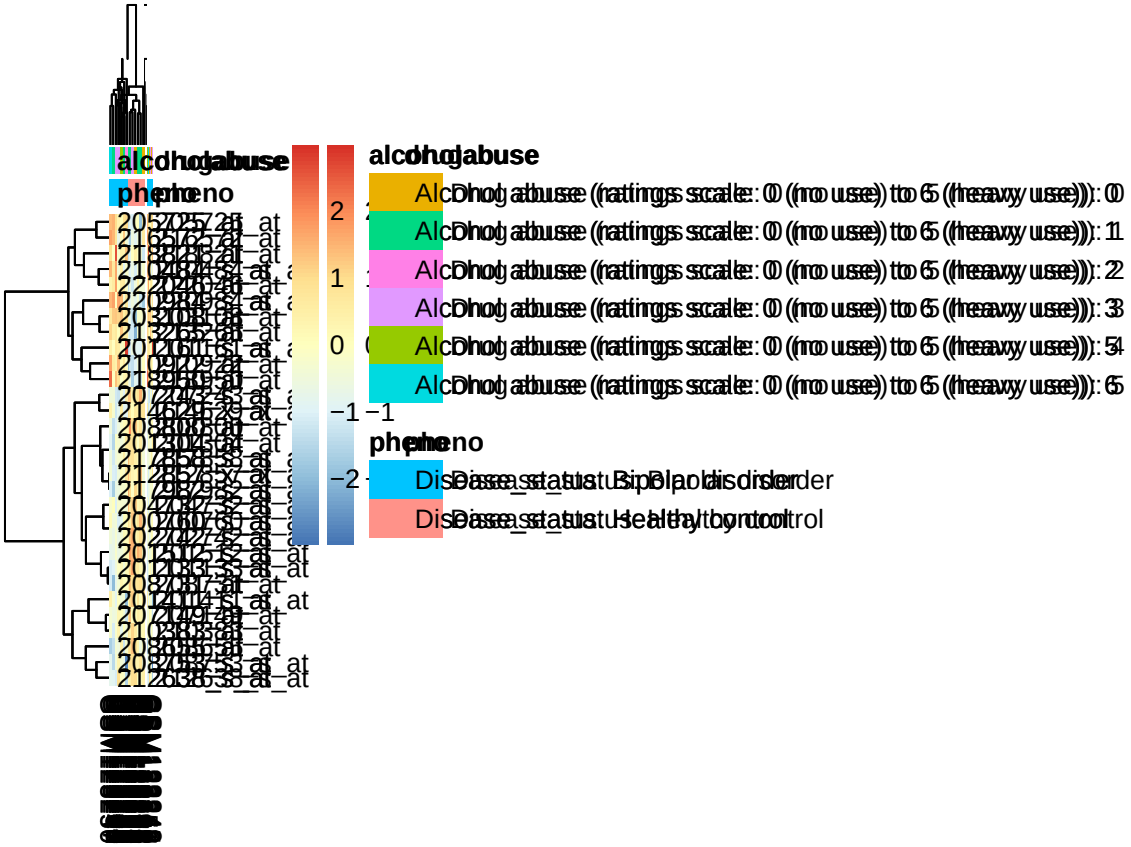


Figure 4A (Drug Abuse) and 4B (Alcohol Abuse): This heatmap illustrates distinct color variations between control and Major Depressive Disorder (MDD) groups, indicating differential gene expression patterns (where underexpressed genes in the control are overexpressed in MDD, and vice versa). Notably, despite substantial differentiation between these groups, no distinct differences in gene expression were observed among drug abuse or alcohol abuse categories, as gene expression levels were dispersed across the heatmap.

That being said however, there does seem to be one gene that is showing some hints of a pattern. The gene labelled as "213265_at" seems to have slighter darker red regions in regions where the alcohol abuse categories are at it's highest. This is only displayed within the MDD groups as well, which might provide some insights.

```
tt_data["213265_at",] #viewing information on gene
```

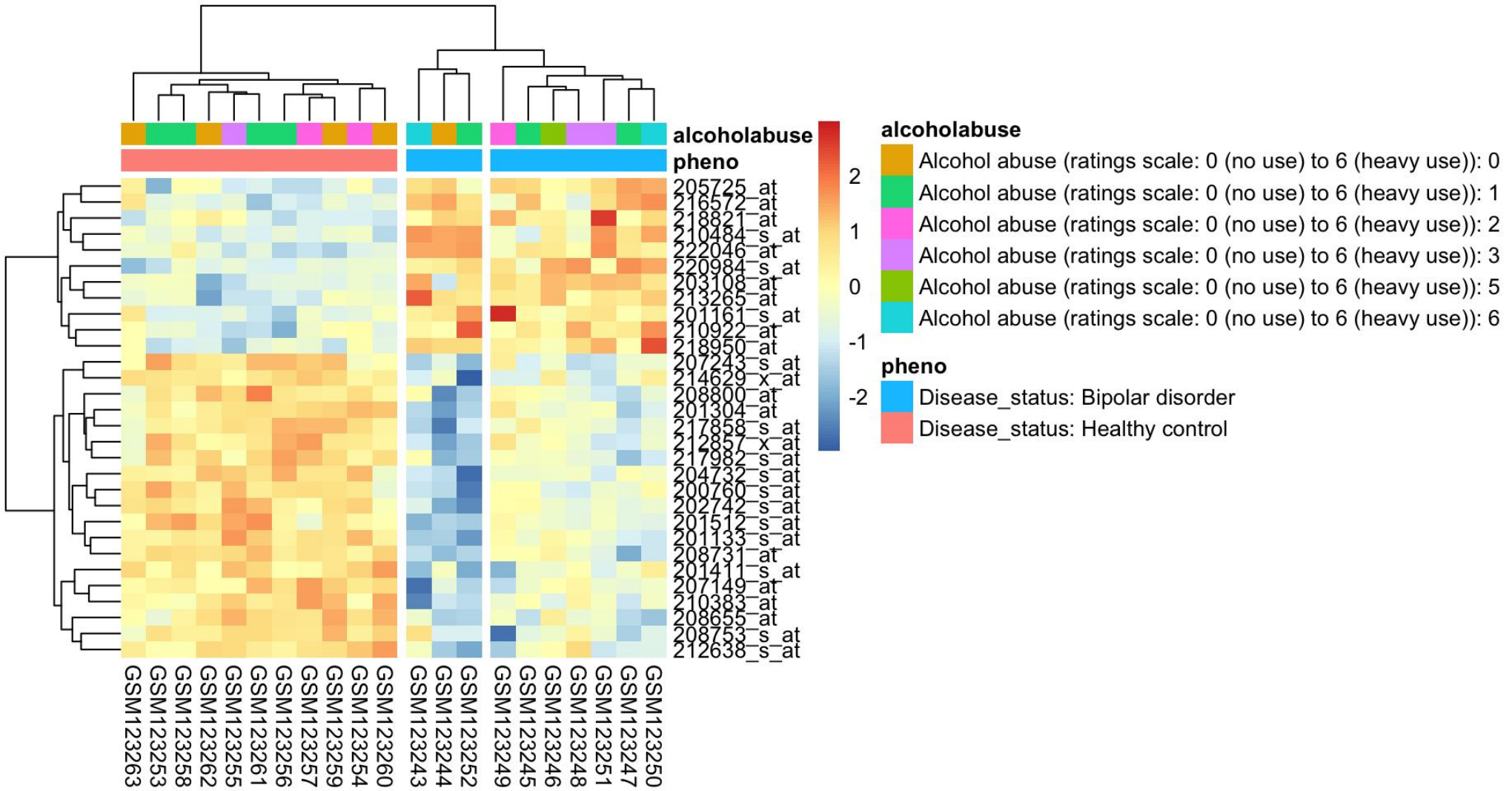
```
##                   ID   GB_ACC SPOT_ID Species.Scientific.Name Annotation.Date
## 213265_at 213265_at AI570199                     Homo sapiens     Oct 6, 2014
##                Sequence.Type Sequence.Source
## 213265_at Consensus sequence         GenBank
##
## 213265_at Consensus includes gb:AI570199 /FEA=EST /DB_XREF=gi:4533573 /DB_XREF=est:to76b07.x1 /CLONE=
##           Representative.Public.ID
## 213265_at                 AI570199
##
## 213265_at pepsin A-3-like /// pepsinogen 3, group I (pepsinogen A) /// pepsinogen 4, group I (pepsino
##                                     Gene.Symbol
## 213265_at LOC101929842 /// PGA3 /// PGA4 /// PGA5
##                                    ENTREZ_GENE_ID
```

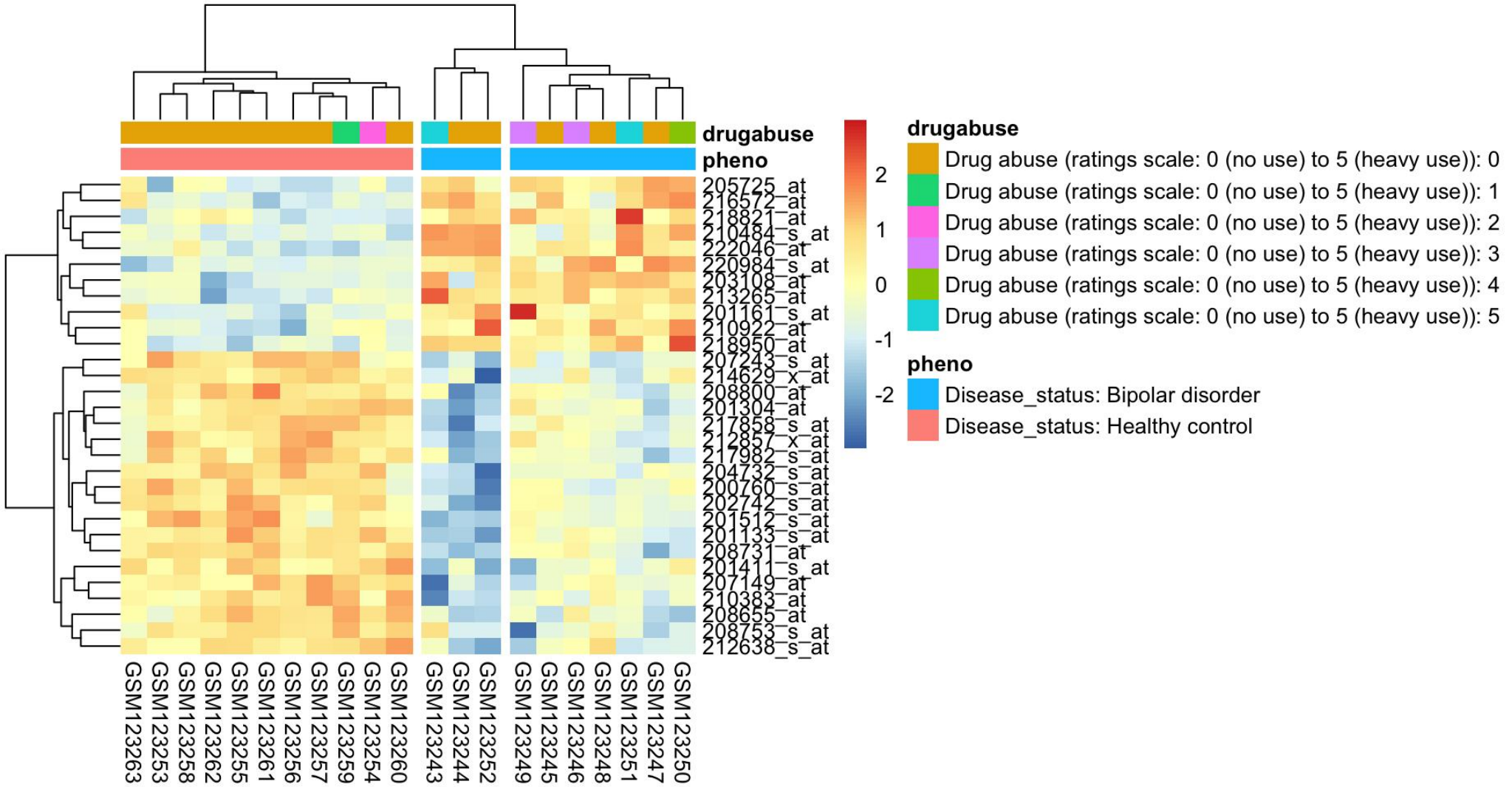
```
## 213265_at 5222 /// 643834 /// 643847 /// 101929842
##                                                    RefSeq.Transcript.ID
## 213265_at NM_001079807 /// NM_001079808 /// NM_014224 /// XM_005276404
##
## 213265_at 0006508 // proteolysis // inferred from electronic annotation /// 0007586 // digestion // i
##                                                 Gene.Ontology.Cellular.Component
## 213265_at 0005576 // extracellular region // inferred from electronic annotation
##
## 213265_at 0004190 // aspartic-type endopeptidase activity // inferred from electronic annotation ///
##               logFC  AveExpr       t      P.Value  adj.P.Val        B  diff
## 213265_at 0.3528145 6.224239 5.44818 1.666378e-05 0.07151453 2.856851 FALSE
```

Table 3: PGA3/4/5 information. All information regarding the pepsinogen gene, including the gene name, symbol, region, annotations and the significant differentiation expressed among both groups.

The gene in question is a Pepsinogen, a precursor to the enzyme pepsin, which is responsible for breaking down protein in the gastrointestinal system.

Ultimately, we conducted a differential gene expression analysis to assess whether or not drug and alcohol abuse can be characterized as a marker for Bipolar and Major Depressive Disorder, as having large subsets of differentiated genes between control and condition group can confirm genetic links to the addiction and conditions. By doing so, we found one particular gene of interest that is extremely differentiated across control and experimental groups: PGA, a pepsinogen involved in protein breakdown in digestion. Potential insights from this will be researched via literature reviews to find any other similar findings and explain this result. This will be discussed in the final paper.


alcoholabuse
pheno
205725_at
216572_at
218821_at
210484_s_at
222046_at
220984_s_at
203108_at
213265_at
201161_s_at
210922_at
218950_at
207243_s_at
214629_x_at
208800_at
201304_at
217858_s_at
212857_x_at
217982_s_at
204732_s_at
200760_s_at
202742_s_at
201512_s_at
201133_s_at
208731_at
201411_s_at
207149_at
210383_at
208655_at
208753_s_at
212638_s_at
GSM123263
GSM123253
GSM123258
GSM123262
GSM123255
GSM123261
GSM123256
GSM123257
GSM123259
GSM123254
GSM123260
GSM123243
GSM123244
GSM123252
GSM123249
GSM123245
GSM123246
GSM123248
GSM123251
GSM123247
GSM123250
2
1
0
-1
-2
alcoholabuse
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 0
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 1
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 2
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 3
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 5
Alcohol abuse (ratings scale: 0 (no use) to 6 (heavy use)): 6
pheno
Disease_status: Bipolar disorder
Disease_status: Healthy control


drugabuse
pheno
205725_at
216572_at
218821_at
210484_s_at
222046_at
220984_s_at
203108_at
213265_at
201161_s_at
210922_at
218950_at
207243_s_at
214629_x_at
208800_at
201304_at
217858_s_at
212857_x_at
217982_s_at
204732_s_at
200760_s_at
202742_s_at
201512_s_at
201133_s_at
208731_at
201411_s_at
207149_at
210383_at
208655_at
208753_s_at
212638_s_at
GSM123263
GSM123253
GSM123258
GSM123262
GSM123255
GSM123261
GSM123256
GSM123257
GSM123259
GSM123254
GSM123260
GSM123243
GSM123244
GSM123252
GSM123249
GSM123245
GSM123246
GSM123248
GSM123251
GSM123247
GSM123250
2
1
0
-1
-2
drugabuse
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 0
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 1
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 2
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 3
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 4
Drug abuse (ratings scale: 0 (no use) to 5 (heavy use)): 5
pheno
Disease_status: Bipolar disorder
Disease_status: Healthy control