Differential Gene Expression Analysis in Mouse Models with Varying Methamphetamine Consumption

Frida Galán

2025-02-09

Contents

Introduction

Drug addition, also called substance use disorder, is a disease that affects a person's brain and behavior and leads to an inability to control the use of a legal or illegal drug or medicine.(Mayo Clinic 2023) Substance use disorders pose a significant public health challenge, with far-reaching consequences for individuals, families, and society. The biological mechanisms underlying addiction are multifaceted, involving genetic, epigenetic, and environmental factors that interact to influence vulnerability and progression of the disorder. A critical aspect of addiction research is understanding the molecular and neurobiological changes that occur in the brain's reward circuitry. For addiction research, in-depth transcriptome analysis entails study of multiple brain regions comprising the addiction neurocircuitry, as well as the temporal patterns of drug intake, withdrawal, and relapse.(Wang et al. 2019)

Modern large-scale molecular methods, including RNA-sequencing (RNA-Seq), have been extensively applied to alcohol-related disease traits, but rarely to risk for methamphetamine (MA) addiction.(Hitzemann et al. 2019) Methamphetamine (N-methylamphetamine) is a potent central nervous system (CNS) stimulant that

is mainly used as a recreational or performance-enhancing drug and less commonly as a second-line treatment for attention deficit hyperactivity disorder (ADHD). METH has powerful addictive properties and, therefore, has devastating effects on health and other aspects of life of the people who abuse it. (Moszczynska and Callan 2017)

The data used in this analysis originates from a study by Hitzemann et al. (Hitzemann et al. 2019), which employed RNA-sequencing to investigate gene expression profiles in mice selectively bred for high and low voluntary MA intake. The study focused on three key brain regions—the NAc, PFC, and VMB—providing valuable insights into the transcriptional changes associated with MA consumption. By leveraging this dataset, our analysis aimed to identify differentially expressed genes in mice bred for high and low methamphetamine consumption.

Metodology

Section 1: Data loading and pre-processing

In this part, since we're usign recount3 to get the data from the study, we will load the library and get the data from the study. To do the rest of the analysis, we will need to create a RangedSummarizedExperiment (RSE) object, which is a Bioconductor class that stores genomic data in a convenient format. This object will contain the raw counts for each gene in the dataset, as well as metadata about the samples and genes.

```
## Load the library
library(recount3)

## Explore available mouse datasets in recount3
mouse_projects <- available_projects("mouse")

## Get the project of interest (in this case SRP193734 )
project_info <- subset(
  mouse_projects,
  project == "SRP193734" & project_type == "data_sources"
)

## Create a RangedSummarizedExperiment (RSE) object
rse_gene_SRP193734 <- create_rse(project_info)

## Convert raw counts to read counts
assay(rse_gene_SRP193734, "counts") <- compute_read_counts(rse_gene_SRP193734)
```

The next step is to explore the data to see what we're working with

```
## Expand the attributes of the samples
rse_gene_SRP193734 <- expand_sra_attributes(rse_gene_SRP193734)

colData(rse_gene_SRP193734)[
  ,
  grepl("^sra_attribute", colnames(colData(rse_gene_SRP193734)))
]

### DataFrame with 143 rows and 8 columns
###           sra_attribute.selected_line sra_attribute.Sex
###                              <factor>       <character>
```

```
### SRR8949519                High Drinker              male
### SRR8949520                High Drinker              male
### SRR8949521                High Drinker              male
### ...
### SRR8949659                 Low Drinker              male
### SRR8949660                 Low Drinker              male
###           sra_attribute.source_name
###                         <character>
### SRR8949519    High Drinker_Nucleus..
### SRR8949520    High Drinker_Nucleus..
### ...
### SRR8949657    Low Drinker_Pre-Fron..
### SRR8949658    Low Drinker_Pre-Fron..
### ...
```

```
## Inspect the attributes of the samples
rse_gene_SRP193734$sra.sample_attributes
```

```
### [1] "selected line;;High Drinker|Sex;;male|source_name;;High Drinker_Nucleus
### Accumbens|tissue;;Dissected Tissue (Brain) - Nucleus Accumbens"
### [2] "selected line;;High Drinker|Sex;;male|source_name;;High Drinker_Nucleus
### Accumbens|tissue;;Dissected Tissue (Brain) - Nucleus Accumbens"
### ...
### [142] "selected line;;Low Drinker|Sex;;male|source_name;;Low Drinker_Pre-
### Frontal Cortex|tissue;;Dissected Tissue (Brain) - Pre-Frontal Cortex"
### [143] "selected line;;Low Drinker|Sex;;male|source_name;;Low Drinker_Pre-
### Frontal Cortex|tissue;;Dissected Tissue (Brain) - Pre-Frontal Cortex"
```

```
## Get the data in the columns
names(colData(rse_gene_SRP193734))
```

```
### [1] "rail_id"
### [2] "external_id"
### [3] "study"
### [4] "sra.sample_acc.x"
### ...
### [183] "sra_attribute.selected_line"
### [184] "sra_attribute.Sex"
### [185] "sra_attribute.source_name"
### [186] "sra_attribute.tissue"
```

Finally, we are going to manipulate the data to make it more convenient to work with

```
## Casting the data of interest (for a better manipulation), since all mouse are male, the sex isn't re
rse_gene_SRP193734$sra_attribute.selected_line <- factor(rse_gene_SRP193734$sra_attribute.selected_line)
rse_gene_SRP193734$sra_attribute.tissue <- factor(rse_gene_SRP193734$sra_attribute.tissue)

## Check more information about our data
summary(rse_gene_SRP193734$sra_attribute.selected_line)
```

```
## High Drinker  Low Drinker
```

```
##          72          71
```

```
summary(rse_gene_SRP193734$sra_attribute.tissue)
```

```
##  Dissected Tissue (Brain) - Nucleus Accumbens
##                                            47
## Dissected Tissue (Brain) - Pre-Frontal Cortex
##                                            48
##   Dissected Tissue (Brain) - Ventral Midbrain
##                                            48
```

```
## Resume of our variable of interest
summary(as.data.frame(colData(rse_gene_SRP193734)[
  , grepl("^sra_attribute.*(selected_line|tissue)", colnames(colData(rse_gene_SRP193734)))
]))
```

```
##  sra_attribute.selected_line
##  High Drinker:72
##  Low Drinker :71
##
##                                     sra_attribute.tissue
##  Dissected Tissue (Brain) - Nucleus Accumbens :47
##  Dissected Tissue (Brain) - Pre-Frontal Cortex:48
##  Dissected Tissue (Brain) - Ventral Midbrain  :48
```

Section 2: Quality control and data cleaning

For this step, firts we're going to calculate the proportion of assigned genes in the dataset. This metric provides an indication of the quality of the sequencing data and the efficiency of the read alignment process. A high proportion of assigned genes suggests that a large fraction of the reads were successfully mapped to known genes, which is essential for downstream analyses such as differential gene expression analysis.

```
## Save a copy of the data before the quality control
rse_gene_SRP193734_unfiltred <- rse_gene_SRP193734

## Calculate the proportion of assigned genes
rse_gene_SRP193734$assigned_gene_prop <- rse_gene_SRP193734$recount_qc.gene_fc_count_all.assigned /
  rse_gene_SRP193734$recount_qc.gene_fc_count_all.total
#Get a summary of the assigned gene proportion
summary(rse_gene_SRP193734$assigned_gene_prop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7589  0.7971  0.8011  0.8006  0.8047  0.8175
```

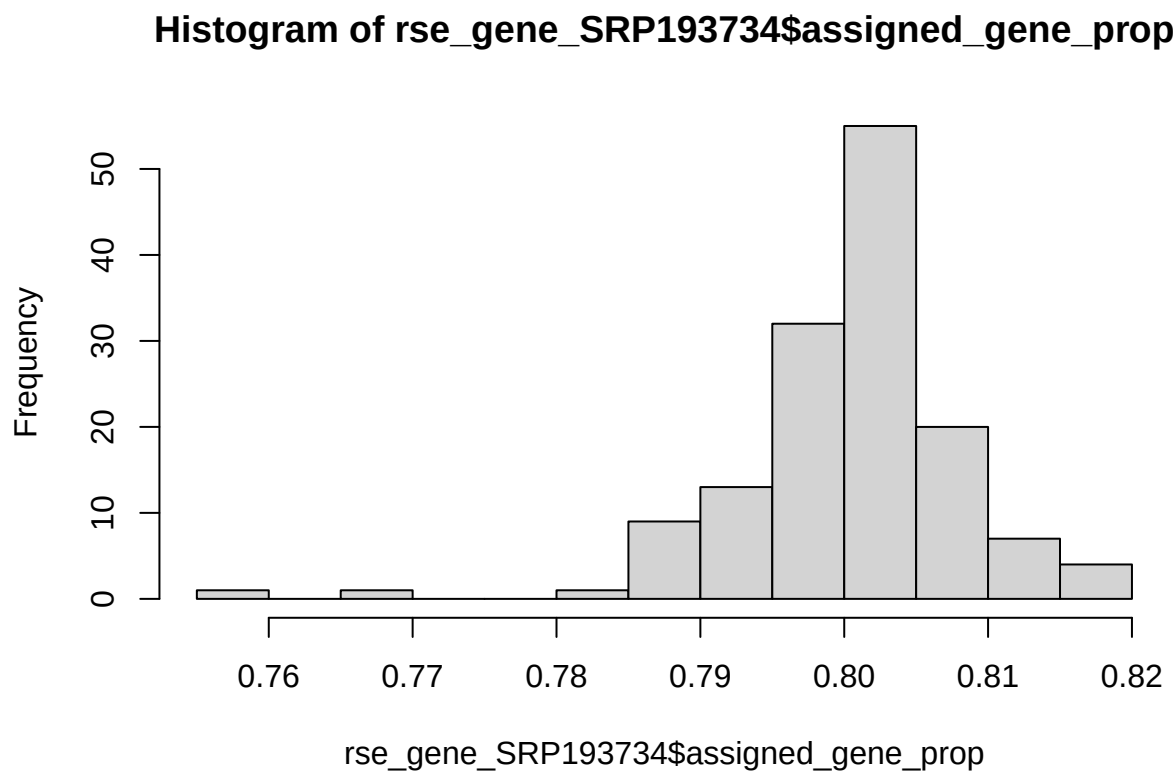
Secondly, we're going to check if there are significant differences

```
## Check if there is a difference between the groups
with(colData(rse_gene_SRP193734), aggregate(assigned_gene_prop,
                                            by = list(sra_attribute.selected_line, sra_attribute.tissue
                                            FUN = summary)
     )
```

```
##        Group.1                                       Group.2    x.Min.
## 1 High Drinker  Dissected Tissue (Brain) - Nucleus Accumbens 0.7589296
## 2  Low Drinker  Dissected Tissue (Brain) - Nucleus Accumbens 0.7881851
## 3 High Drinker Dissected Tissue (Brain) - Pre-Frontal Cortex 0.7940395
## 4  Low Drinker Dissected Tissue (Brain) - Pre-Frontal Cortex 0.7963435
## 5 High Drinker   Dissected Tissue (Brain) - Ventral Midbrain 0.7906245
## 6  Low Drinker   Dissected Tissue (Brain) - Ventral Midbrain 0.7853710
##   x.1st Qu.  x.Median    x.Mean x.3rd Qu.    x.Max.
## 1 0.7920566 0.7985304 0.7944864 0.8015310 0.8050681
## 2 0.7931817 0.7977923 0.7967840 0.8004943 0.8076634
## 3 0.8027410 0.8068420 0.8069616 0.8107848 0.8175259
## 4 0.8039593 0.8054463 0.8060886 0.8082870 0.8146139
## 5 0.7977708 0.8000756 0.7993344 0.8019944 0.8046663
## 6 0.7976182 0.8004639 0.7996494 0.8028321 0.8055571
```

Now, we're going to filter the data to remove low-quality samples. We're going to eliminate any sample with an assigned gene proportion below 0.3, as these samples may have poor sequencing quality and could introduce noise into the analysis.

```
hist(rse_gene_SRP193734$assigned_gene_prop)
```




```
table(rse_gene_SRP193734$assigned_gene_prop < 0.3)
```

```
##
## FALSE
##   143
```

```
rse_gene_SRP193734 <- rse_gene_SRP193734[, rse_gene_SRP193734$assigned_gene_prop > 0.3]
```

In this step, we're going to calculate the expression levels of the genes by using the expression count. By using the function 'rowMeans' helps us to calculate the average expression level of each gene across all samples in the dataset. This metric provides an indication of the gene's overall expression level.

```
## Calculate the expression levels of the genes ----- using the counts --------
gene_means <- rowMeans(assay(rse_gene_SRP193734, "counts"))
summary(gene_means)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       0.0       0.0       1.3    1111.1     204.7 1699954.0
```

```
## Delete genes with low expression levels (under 0.1)
rse_gene_SRP193734 <- rse_gene_SRP193734[gene_means > 0.1, ]

## Defining the finals dimension
dim(rse_gene_SRP193734)
```

```
## [1] 36504   143
```

```
## Percentage of genes that we keep
round(nrow(rse_gene_SRP193734) / nrow(rse_gene_SRP193734_unfiltred) * 100, 2)
```

```
## [1] 65.87
```

Section 3: Normalization of the data

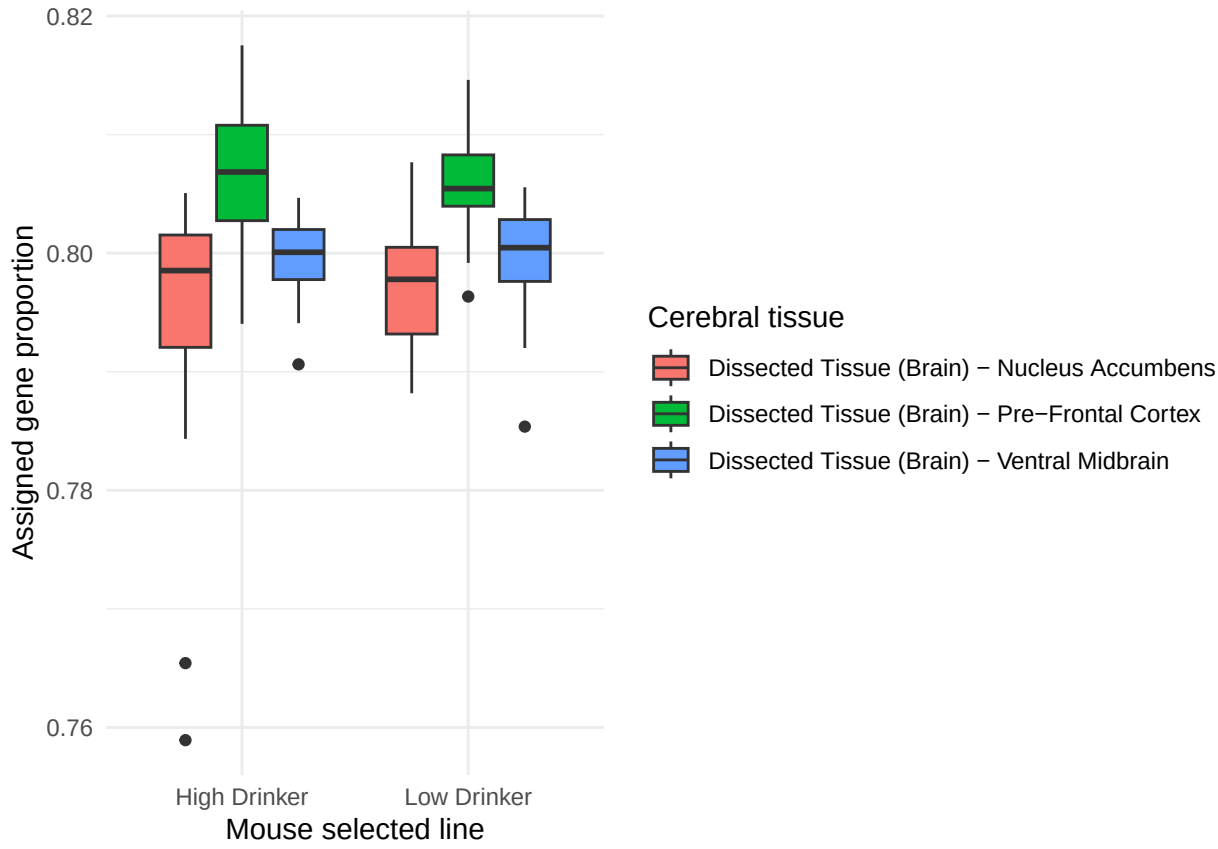
Normalize the data is a crucial step in RNA-Seq analysis, as it adjusts for differences in sequencing depth and other technical factors that can confound downstream analyses.

```
## Load the libraries
library(edgeR)
library(ggplot2)
```

```
## Normalize the data
dge <- DGEList(
  counts = assay(rse_gene_SRP193734, "counts"),
  genes = rowData(rse_gene_SRP193734)
)

dge <- calcNormFactors(dge)

## Visualize the data after normalization
ggplot(as.data.frame(colData(rse_gene_SRP193734)),
       aes(x = sra_attribute.selected_line, y = assigned_gene_prop, fill = sra_attribute.tissue)) +
  geom_boxplot() +
  labs(x = "Mouse selected line", y = "Assigned gene proportion", fill = "Cerebral tissue") +
  theme_minimal()
```



Section 4: Differential expression analysis

In this step, we're going to perform a differential gene expression analysis to identify genes that are differentially expressed between mice bred for high and low methamphetamine consumption.

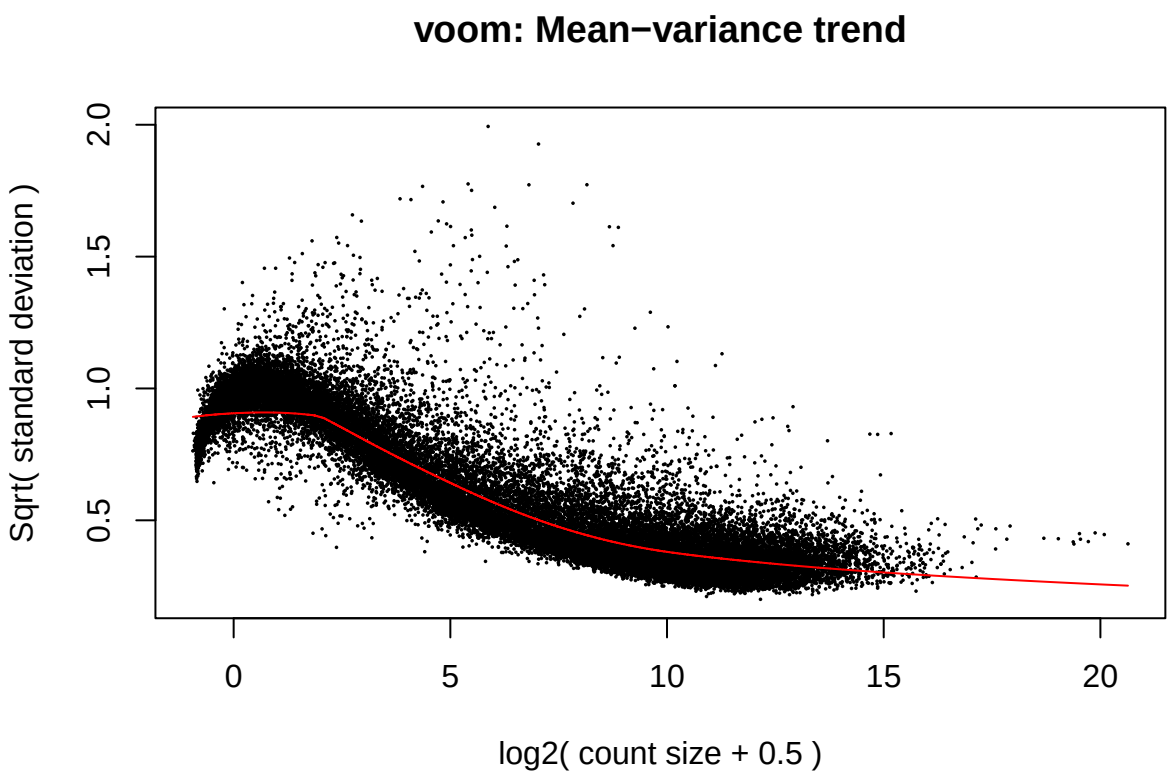
```
## Load the library
library("limma")

## Implement the statistical model
mod <- model.matrix(~ sra_attribute.selected_line + sra_attribute.tissue,
                    data = colData(rse_gene_SRP193734))
colnames(mod)
```

```
## [1] "(Intercept)"
## [2] "sra_attribute.selected_lineLow Drinker"
## [3] "sra_attribute.tissueDissected Tissue (Brain) - Pre-Frontal Cortex"
## [4] "sra_attribute.tissueDissected Tissue (Brain) - Ventral Midbrain"
```

```
## Visualize the matrix
vd <- ExploreModelMatrix::VisualizeDesign(
  sampleData = colData(rse_gene_SRP193734),
  designFormula = ~ sra_attribute.selected_line + sra_attribute.tissue,
  textSizeFitted = 4
)
```

```
## Aply the voom transformation (to stabilize the variance)
vGene <- voom(dge, mod, plot = TRUE)
```




```
## Fit the linear model for each gene and perform the empirical Bayes moderation
eb_result <- eBayes(lmFit(vGene))

## Extract the results, sorting the genes without any particular order
de_results <- topTable(
  eb_result,
  coef = 2,
  number = nrow(rse_gene_SRP193734),
  sort.by = "none"
)

## Inspect the dimensions and the first rows of the results
dim(de_results)
```

```
## [1] 36504    17
```

```
head(de_results, n = 3)
```

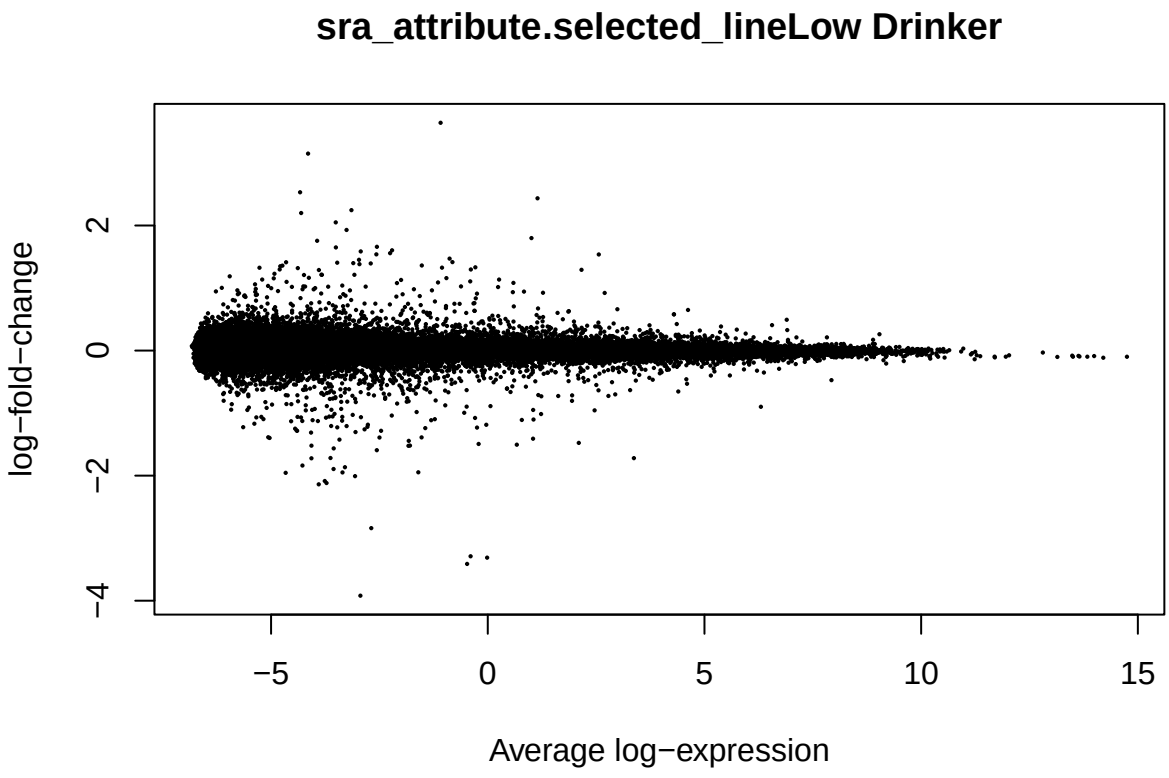
```
##                        source type bp_length phase              gene_id
## ENSMUSG00000079794.2 ENSEMBL gene       255    NA ENSMUSG00000079794.2
```

```
## ENSMUSG00000079190.3 ENSEMBL gene      1473    NA ENSMUSG00000079190.3
## ENSMUSG00000079808.3 ENSEMBL gene      1910    NA ENSMUSG00000079808.3
##                           gene_type  gene_name level mgi_id havana_gene  tag
## ENSMUSG00000079794.2 protein_coding AC125149.2     3   <NA>        <NA> <NA>
## ENSMUSG00000079190.3 protein_coding AC133103.1     3   <NA>        <NA> <NA>
## ENSMUSG00000079808.3 protein_coding AC168977.1     3   <NA>        <NA> <NA>
##                            logFC   AveExpr          t   P.Value adj.P.Val
## ENSMUSG00000079794.2 -0.01943436 -6.720592 -0.2001054 0.8416800 0.9409269
## ENSMUSG00000079190.3  0.02610546 -6.665950  0.2334203 0.8157661 0.9313419
## ENSMUSG00000079808.3  0.01793784 -6.245177  0.1196203 0.9049504 0.9658381
##                              B
## ENSMUSG00000079794.2 -5.805166
## ENSMUSG00000079190.3 -5.797287
## ENSMUSG00000079808.3 -5.805011
```

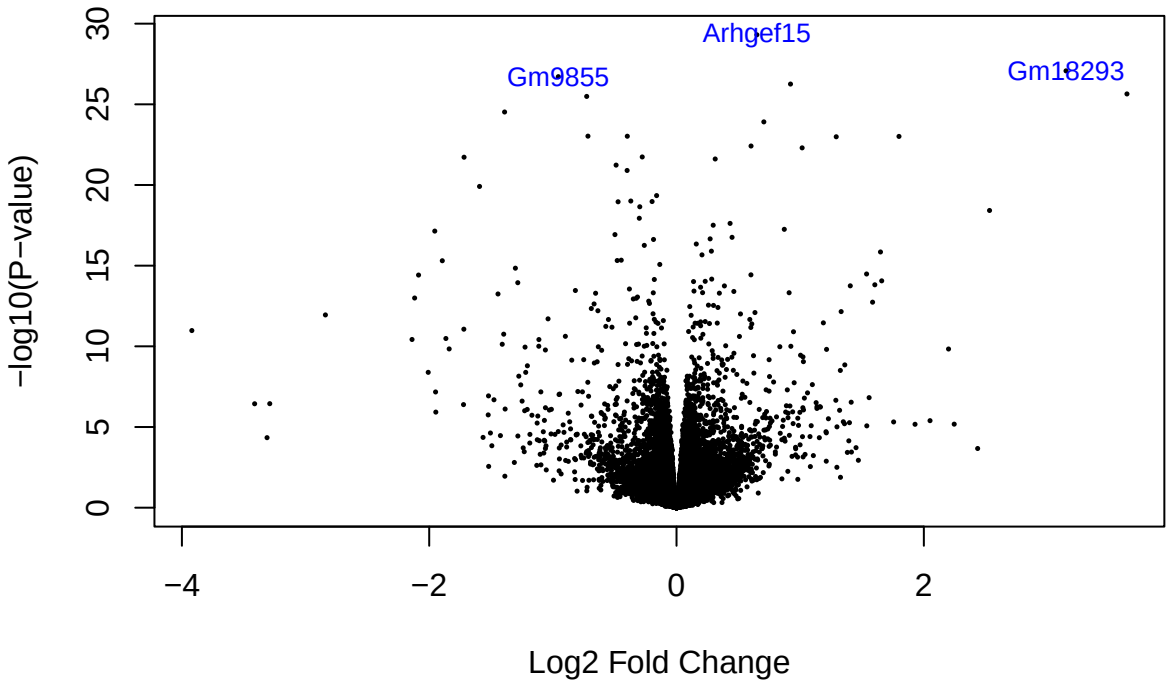
```r
## Count the genes with adjusted p-value < 0.05 (significant genes)
table(de_results$adj.P.Val < 0.05)
```

```
##
## FALSE  TRUE
## 32442  4062
```

```r
## Visualization of the changes in the differential expression
plotMA(eb_result, coef = 2)
```



```r
## Volcano plot highlighting the top 3 genes more significant
volcanoplot(eb_result, coef = 2, highlight = 3, names = de_results$gene_name)
```




```r
de_results[de_results$gene_name %in% c("Gm9855", "Arhgef15", "Gm1829"), ]
```

```
##                         source type bp_length phase               gene_id
## ENSMUSG00000085666.3  HAVANA gene      1194    NA  ENSMUSG00000085666.3
## ENSMUSG00000052921.13 HAVANA gene      6483    NA ENSMUSG00000052921.13
##                                  gene_type gene_name level      mgi_id
## ENSMUSG00000085666.3  processed_pseudogene    Gm9855     2 MGI:3704357
## ENSMUSG00000052921.13       protein_coding  Arhgef15     2 MGI:3045246
##                                  havana_gene  tag      logFC  AveExpr         t
## ENSMUSG00000085666.3  OTTMUSG00000026610.1 <NA> -0.9559158 2.465096 -13.52844
## ENSMUSG00000052921.13 OTTMUSG00000005952.2 <NA>  0.6495559 4.620562  14.53132
##                            P.Value    adj.P.Val        B
## ENSMUSG00000085666.3  1.934106e-27 2.353421e-23 51.79000
## ENSMUSG00000052921.13 4.900335e-30 1.788818e-25 57.64387
```

Section 5: Visualization of the results

Finally, here we're creating a heatmap with the expression values of the top 50 genes that are differentially expressed between mice bred for high and low methamphetamine consumption. This visualization provides a comprehensive overview of the gene expression patterns across the samples and highlights the genes that are most relevant to the study.

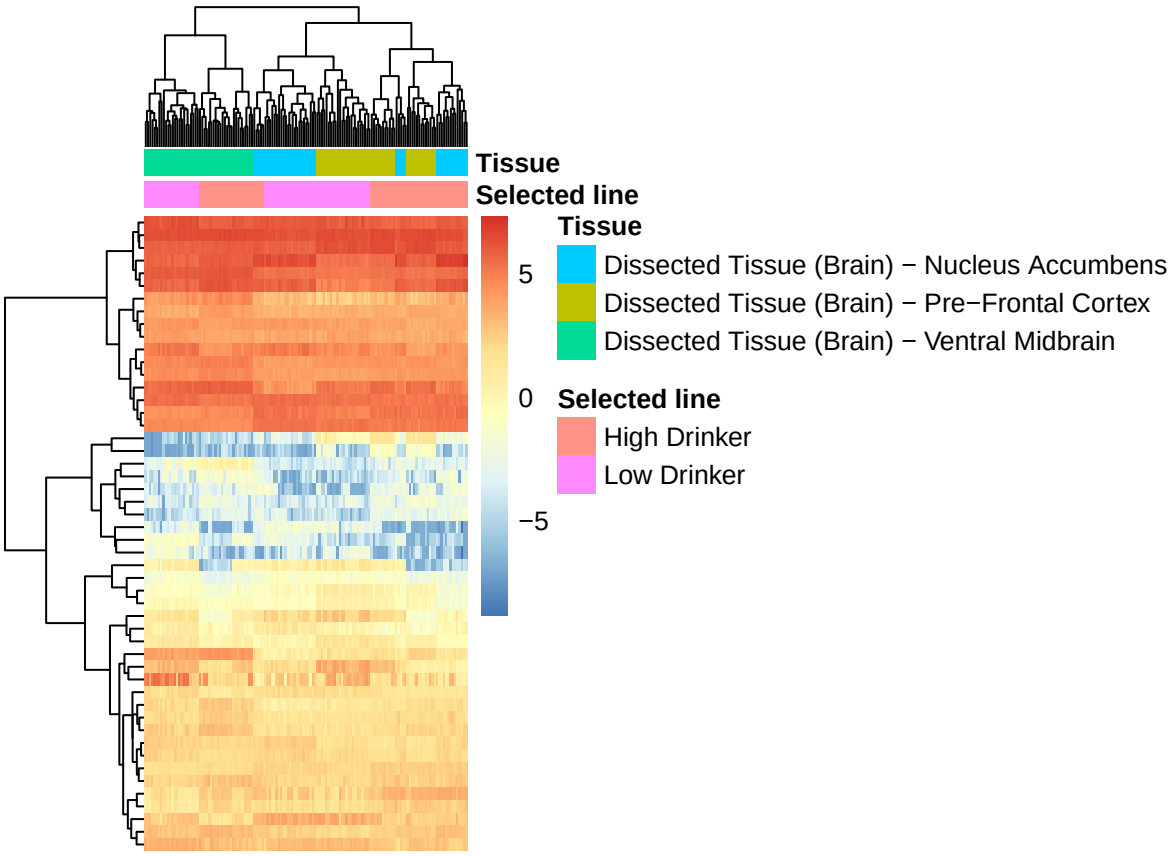
```
## Load the library
library("pheatmap")

## Extract the values of the genes of interest
exprs_heatmap <- vGene$E[rank(de_results$adj.P.Val) <= 50, ]

## Create a table with the information of the samples
df <- as.data.frame(colData(rse_gene_SRP193734))

# Subset the columns of interest
df <- df[, c("sra_attribute.selected_line", "sra_attribute.tissue")]
colnames(df) <- c("Selected line", "Tissue")

## Create a heatmap with the expression values of the top 50 genes
pheatmap(exprs_heatmap,
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         show_rownames = FALSE,
         show_colnames = FALSE,
         annotation_col = df
         )
```



Interpretation of the results

In the quality control and data cleaning section, the proportion of assigned genes in the samples was relatively high (between 0.76 and 0.82). This indicates that the majority of reads were successfully mapped to known genes, suggesting that the sequencing data is of good quality. No samples were of low quality, as none fell below the threshold of 0.3; therefore, no samples were removed from the dataset. Additionally, genes with very low expression levels (below 0.1) were filtered out, resulting in the retention of approximately 65.87% of the genes. This filtering step helps focus on the most informative genes and reduces noise in the analysis.

The voom transformation was applied to stabilize the variance of the data. The graph shows that the relationship between variance and mean follows the expected trend, and a linear model was fitted for each gene to perform empirical Bayes moderation. A total of 4,062 genes were found to be differentially expressed between mice bred for high and low methamphetamine consumption, with an adjusted p-value < 0.05. The volcano plot highlights the top three most significant genes in the analysis: Gm9855, Arhgef15, and Gm1829.

Finally, the heatmap visualization of the top 50 differentially expressed genes reveals clear differences between the high and low methamphetamine consumption groups across the three brain regions analyzed (NAc, PFC, and VMB).

Discussion

The genes identified in this analysis provide valuable insights into the molecular mechanisms underlying methamphetamine consumption in mice. The brain regions studied (NAc, PFC, and VMB) are known to play a critical role in reward circuitry and addiction-related behaviors. The differential expression of genes in these regions may contribute to the observed differences in methamphetamine consumption between the high and low consumption groups. Thus, the results support the idea that susceptibility to methamphetamine addiction involves molecular reconfiguration in brain regions associated with reward and addiction-related behaviors.

Arhgef15 is a gene that encodes a Rho guanine nucleotide exchange factor. Rho GTPases play a fundamental role in numerous cellular processes initiated by extracellular stimuli that act through G protein-coupled receptors. This gene encodes a protein that functions as a specific guanine nucleotide exchange factor for RhoA and interacts with ephrin A4 in vascular smooth muscle cells. Two alternatively spliced transcript variants encoding the same protein have been identified for this gene (National Center for Biotechnology Information 2023a).

Gm1829 (now *Gm6365*) (National Center for Biotechnology Information 2023b) and *Gm9855* (thymine DNA glycosylase) (National Center for Biotechnology Information 2023c) are both pseudogenes and remain poorly characterized. These findings could serve as a starting point for further research to elucidate the roles of these genes in methamphetamine consumption and addiction.

Conclusion

In conclusion, this analysis identified 4,062 genes differentially expressed between mice bred for high and low methamphetamine consumption. The top three most significant genes in the analysis—Gm9855, Arhgef15, and Gm1829—are established as potential candidates for further research into the molecular mechanisms underlying methamphetamine addiction.

References

Hitzemann, Robert, Ovidiu D. Iancu, Christopher Reed, Hiroki Baba, Denesa R. Lockwood, and Tamara J. Phillips. 2019. “Regional Analysis of the Brain Transcriptome in Mice Bred for High and Low

Methamphetamine Consumption." *Brain Sciences* 9 (7): 155. https://doi.org/10.3390/brainsci9070155.
Mayo Clinic. 2023. "Drug Addiction (Substance Use Disorder)." Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/drug-addiction/symptoms-causes/syc-20365112.
Moszczynska, Anna, and Sean P. Callan. 2017. "Molecular, Behavioral, and Physiological Consequences of Methamphetamine Neurotoxicity: Implications for Treatment." *Journal of Pharmacology and Experimental Therapeutics* 362 (3): 474–88. https://doi.org/10.1124/jpet.116.238501.
National Center for Biotechnology Information. 2023a. "Gene ID: 22899 - ARHGEF15 Rho guanine nucleotide exchange factor 15." https://www.ncbi.nlm.nih.gov/gene/22899.
———. 2023b. "Gene ID: Gm1829 - Gm6365 Predicted Pseudogene 6365." https://www.ncbi.nlm.nih.gov/gene/?term=Gm1829.
———. 2023c. "Gene ID: Gm9855 - Tdg-Ps2 Thymine DNA Glycosylase, Pseudogene." https://www.ncbi.nlm.nih.gov/gene/?term=Gm9855.
Wang, Jing, Fang Li, Xiaoyuan Xu, Xiaohong Liu, and Chengyu Jiang. 2019. "The Role of MicroRNAs in the Immune Response to Influenza Infection." *Journal of Immunology Research* 2019: 1–10. https://doi.org/10.1155/2019/8028725.