Differential Expression and the Mouse Gut-Brain Axis

This module introduces you to differential expression analyses using the R programming language.

This activity will walk through differential expression analysis and use bioinformatics tools (R) to understand how gut bacteria can influence the expression of genes in the brain (the gut-brain axis). You will work with real data from a mouse RNA-seq experiment to understand how researchers explore how human disorders might be impacted by gut-brain axis using model organisms.

## Part 1

|  |
| --- |
| Check Your Knowledge |
| 1. What is the gut-brain axis? 2. Why do you think the researchers looked at gene expression in the prefrontal cortex and the striatum? 3. Why are mice useful as model organisms for human disease studies? |

### Package Install and Load

We will need to install the tidyverse package for this activity.

Type the following into your console and press return to run the code. You will see a lot of red text, but usually that’s a good thing!

install.packages("readr")

Installing package into '/usr/local/lib/R/site-library'  
(as 'lib' is unspecified)

install.packages("dplyr")

Installing package into '/usr/local/lib/R/site-library'  
(as 'lib' is unspecified)

Next, we will load the package so it’s ready to use:

library(readr)  
library(dplyr)

We’ll read in some data from the GEMs website. Run the following code in your console:

prefrontal <- read\_csv("https://genomicseducation.org/data/mouse\_gutbrain\_de\_autismVcontrol\_in\_prefrontalcortex.csv")

Rows: 55421 Columns: 7  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): GeneID  
dbl (6): baseMean, log2FoldChange, lfcSE, stat, pvalue, padj  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

You have just loaded differential expression data comparing ASD-transplant mice and control-transplant mice for the prefrontal cortex region of the brain.

## Part 2

Let’s take a look at the data you loaded into R. Type the following code into the console:

head(prefrontal)

# A tibble: 6 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000090303 23.2 -25.1 1.92 -13.1 3.75e-39 1.36e-34  
2 ENSMUSG00000086112 17.5 -24.7 1.93 -12.8 2.63e-37 4.74e-33  
3 ENSMUSG00000101734 20.2 -24.9 2.06 -12.1 8.68e-34 1.04e-29  
4 ENSMUSG00000118036 17.6 -24.7 2.07 -11.9 6.87e-33 6.20e-29  
5 ENSMUSG00000111968 25.0 -25.2 2.18 -11.6 5.79e-31 3.64e-27  
6 ENSMUSG00000080716 25.4 -25.2 2.18 -11.6 6.05e-31 3.64e-27

The log2FoldChange value gives us an idea of which genes show the greatest differential expression between the ASD mice and the control mice. It is a ratio of how many gene transcripts were found in ASD-type mice compared to control-type mice.

We can arrange the table based on these values (going from smallest to largest) by copying the following code into your console:

prefrontal\_ascending <- arrange(prefrontal, log2FoldChange)  
head(prefrontal\_ascending)

# A tibble: 6 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000080716 25.4 -25.2 2.18 -11.6 6.05e-31 3.64e-27  
2 ENSMUSG00000111968 25.0 -25.2 2.18 -11.6 5.79e-31 3.64e-27  
3 ENSMUSG00000090303 23.2 -25.1 1.92 -13.1 3.75e-39 1.36e-34  
4 ENSMUSG00000062818 21.2 -25.0 2.36 -10.6 3.54e-26 1.07e-22  
5 ENSMUSG00000101734 20.2 -24.9 2.06 -12.1 8.68e-34 1.04e-29  
6 ENSMUSG00000110465 20.0 -24.9 2.56 -9.72 2.57e-22 3.31e-19

The genes at the top of this table have the most negative log2FoldChange values. Negative log2FoldChange values means a **decreased expression** (or, a downregulation in gene expression) in the experimental group versus the control group.

We can use a similar command to look at the genes with the largest positive log2FoldChange. Positive log2FoldChange values indicate genes that are upregulated, or have **increased expression** in the experimental group versus the control group.

prefrontal\_descending <- arrange(prefrontal, desc(log2FoldChange))  
head(prefrontal\_descending)

# A tibble: 6 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000113093 17.0 24.6 2.93 8.37 5.63e-17 2.86e-14  
2 ENSMUSG00000112524 15.3 24.4 2.70 9.05 1.44e-19 1.06e-16  
3 ENSMUSG00000069301 13.9 24.3 2.73 8.91 5.30e-19 3.36e-16  
4 ENSMUSG00000104341 12.5 24.1 2.93 8.23 1.88e-16 9.06e-14  
5 ENSMUSG00000073974 15.0 24.1 2.70 8.92 4.87e-19 3.14e-16  
6 ENSMUSG00000054136 11.9 24.1 2.50 9.61 7.22e-22 7.66e-19

|  |
| --- |
| Check Your Knowledge |
| 1. What gene shows the greatest upregulation (increase in gene expression) in ASD-type mice compared to control mice? Give both the gene ID and the gene name. 2. What molecular functions does this gene do? 3. What gene shows the greatest downregulation (decrease in gene expression) in ASD-type mice compared to control mice? Give both the gene ID and the gene name. 4. What biological processes is this gene involved in? |

## Part 3

In part 3, we’ll create a list of genes that are differentially expressed in the prefrontal cortex, then examine whether they show the same differential expression pattern in the striatum.

In the previous section, we identified genes with the highest and lowest log2FoldChange. However, this metric isn’t always the best indicator of differential expression. The adjusted p-value, or padj, is preferred over log2FoldChange for identifying statistically significant changes in gene expression because padj accounts for multiple testing. This is crucial when analyzing a large number of genes simultaneously, as each additional statistical test increases the risk of identifying a false positive. The adjusted p-value provides a statistical correction that minimizes this risk. log2FoldChange provides information about the magnitude of expression change, while padj indicates its statistical significance.

Let’s create a list of all the genes that are differentially expressed between ASD and control mice using the filter command. We’ll set the padj threshold to 0.05.

prefrontal\_sig <- filter(prefrontal, padj < 0.05)  
head(prefrontal\_sig)

# A tibble: 6 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000090303 23.2 -25.1 1.92 -13.1 3.75e-39 1.36e-34  
2 ENSMUSG00000086112 17.5 -24.7 1.93 -12.8 2.63e-37 4.74e-33  
3 ENSMUSG00000101734 20.2 -24.9 2.06 -12.1 8.68e-34 1.04e-29  
4 ENSMUSG00000118036 17.6 -24.7 2.07 -11.9 6.87e-33 6.20e-29  
5 ENSMUSG00000111968 25.0 -25.2 2.18 -11.6 5.79e-31 3.64e-27  
6 ENSMUSG00000080716 25.4 -25.2 2.18 -11.6 6.05e-31 3.64e-27

dim(prefrontal\_sig)

[1] 103 7

It might be interesting to look at the expression of a gene across all the possible regions. To do this, we will need to load the differential expression data for the striatum.

striatum <- read\_csv("https://genomicseducation.org/data/mouse\_gutbrain\_de\_autismVcontrol\_in\_striatum.csv")

Rows: 55421 Columns: 7  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): GeneID  
dbl (6): baseMean, log2FoldChange, lfcSE, stat, pvalue, padj  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Then we’ll filter out the gene in which we’re interested from each object. Let’s take a look at gene ENSMUSG00000079516, which is the reg3a gene we previously looked up on MGI.

reg3a\_prefrontal <- filter(prefrontal, GeneID == "ENSMUSG00000079516")  
  
reg3a\_striatum <- filter(striatum, GeneID == "ENSMUSG00000079516")

Finally, take a look at the differential expression of reg3a in each region.

In the prefrontal cortex:

reg3a\_prefrontal

# A tibble: 1 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000079516 10.6 -22.6 2.11 -10.7 7.30e-27 2.63e-23

In the striatum:

reg3a\_striatum

# A tibble: 1 × 7  
 GeneID baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000079516 0.371 2.27 2.96 0.765 0.444 NA

|  |
| --- |
| Check Your Knowledge |
| 1. How many genes did you keep after filtering out any genes with a padj > 0.05? 2. How many genes in #8 show increased expression in ASD-type mice? 3. Did the genes you identified in part 2 (the one with the greatest positive log2FoldChange, as well as the one with the greatest negative log2FoldChange) meet the padj threshold? That is, did you keep these genes in your dataset after filtering? 4. Is the differential expression pattern of Reg3a the same in the striatum as in the prefrontal cortex? |