**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

In Part 1, we’ll go over some background context for this activity. Then, we’ll get set up on Posit Cloud for the next steps.

The data for this activity includes gene expression data from two different brain regions (striatum and prefrontal cortex) in mice. Mice in this experiment received fecal transplants. Some mice received transplants from humans who have been diagnosed with Autism Spectrum Disorder (ASD), while other mice received transplants from humans who did not have any diagnosis (control). This allowed researchers to control the composition of the gut microbiome in each mouse. Researchers then bred the mice and looked for differences in gene expression in the brains of the offspring.

### Background

*Autism spectrum disorder (ASD)* is a neurological disorder that affects behavioral and social interactions, among other things. Although ASD can be diagnosed at any age, it’s considered a neurodevelopmental disorder because symptoms usually show up within the first two years of life. Individuals diagnosed with ASD can experience a wide range of symptoms, including differences in social behaviors and communication styles, as well as intellectual disabilities and physical issues like sensory sensitivities or gastrointestinal problems.

The *striatum* is part of the brain involved in motor control and cognitive tasks like reward processing, decision-making, and social interactions (often called executive functions). It lies deep within the center of the brain and is composed of both gray matter (which can be thought of as the “processing” part of brain tissue) and white matter (which is the brain structure involved in transporting messages); the combination of the gray and white matter give this region of the brain a striped appearance, resulting in the name “striatum”. The striatum is involved in both reflexive movement - that is, involuntary movement that happens as an immediate response to a stimulus - and slower, planned movement like walking. In Parkinson’s disease, some patients experience degeneration of parts of the striatum, resulting in spastic, uncontrollable movement.

The *prefrontal cortex* is the part of the brain that is primarily in charge of decision making, reasoning, personality, maintaining social appropriateness, and other complex behaviors that fall under the umbrella of executive functions. This can include planning, self-control, and working towards long-term goals. The prefrontal cortex is located in the very front of the brain, just behind your forehead. One of the most famous brain injury patients was Phineas Gage, a railroad worker who survived an iron rod through his forehead. His prefrontal cortex was destroyed in this accident, and doctors noted huge behavioral and personality changes. You can read more about his case [here](https://pmc.ncbi.nlm.nih.gov/articles/PMC1114479/).

#### The original study

The original study, *Human Gut Microbiota from Autism Spectrum Disorder Induces Behavioral Deficits in Mice*, was published in 2019. Gut microbiota are known to be different between individuals with ASD and individuals who are considered typically-developing. Additionally, some individuals with ASD also experience gastrointestinal symptoms, and their gut microbiota show the greatest difference when compared to the gut microbiota of typically-developing individuals. Some researchers have proposed that gut bacteria can influence some of the symptoms of ASD. The relationship between the intestinal microbiome and the development and function of the human brain is known as the *gut-brain axis.*

In this study, researchers explored whether they could induce ASD-like behaviors in mice by changing their gut microbiome. They transplanted stool from either humans with ASD or controls into germ-free mice and discovered that colonization with gut microbiota was enough to induce ASD-like behaviors in the mice. They also let the mice breed and collected gene expression data from the brains of their offspring to explore whether changing the gut microbiota could result in changed gene expression. In particular, they discovered that the offspring of mice who received stool from ASD donors showed different gene splicing and expression profiles of certain ASD-relevant genes.

|  |
| --- |
| Note |
| It is important to note that researchers are **not** suggesting that ASD is entirely induced by gut bacteria. There is a strong genetic component to ASD. Scientists have known for years that there are **both genetic and environmental components** to the development or severity of **some** ASD symptoms. This research explores one possible environmental component. |

You can read the original research paper [here](https://www.cell.com/cell/fulltext/S0092-8674(19)30502-1).

#### The mouse as a model organism

The mouse is the most commonly-used model organism in laboratory work. In fact, mice and rats make up 95% of the lab animal population, and more than 80% of the research that has been awarded the Nobel Prize for Medicine was done at least in part with mouse models (https://www.cshl.edu/of-mice-and-model-organisms/, https://fbresearch.org/medical-advances/nobel-prizes).

So what makes mice such good model organisms for biomedical research? Well, first, they’re economical and relatively easy to keep. Since mice are small, they don’t require a huge amount of space or food. They also have fast reproductive cycles, so researchers can study multiple generations within only a few years. Most importantly, though, mice and humans are both mammals and have about 85% of their protein-coding genome in common. As a result, mouse physiology is quite similar to human physiology. The mouse circulatory, reproductive, digestive, hormonal, and nervous systems are frequently used as models to study how humans grow, age, and develop chronic diseases. They are particularly important model organisms for cancer research and neuroscience.

You can find additional information about how the mouse is used in research [here!](https://www.google.com/url?q=https://www.informatics.jax.org/mgihome/other/homepage_IntroMouse.shtml)

|  |
| --- |
| 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? |

### Start Posit Cloud

In the next steps, we will explore the data from the publication described above.

We will use Posit Cloud for this activity. Go to <https://posit.cloud/plans/free> and follow the steps to sign up. You can also Log In using the button on the top right if you have used Posit Cloud in the past.

|  |
| --- |
| Tip |
| What is cloud computing? |

TODO: Add figure

Once logged in, select “New Project” and “New RStudio Project”. It will take a few seconds to deploy and load.

TODO: Add figure

Rename your project “Differential Expression Mouse GutBrain” by clicking on “Untitled Project”.

TODO: Add figure

### 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("tidyverse")

TODO: Add figure

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

library(tidyverse)

TODO: Add figure

|  |
| --- |
| What are R Packages? |
| Packages are collections of R code, data, and documentation that extend the base functionality of R. Think of them like “expansion packs” on top of your basic R software.  Packages are developed by the R community and made available through repositories like CRAN (Comprehensive R Archive Network), Bioconductor, and GitHub. They are especially useful if you want to do a specialized kind of analysis, such as genomic analysis!  We use the library command to load and attach packages to the R environment. This means links the package you downloaded to your current session of R.  The “tidyverse” package that you loaded is useful for loading, wrangling, and exploring data. |

### Load Differential Expression Data

**Differential expression** refers to the process where different cells within an organism, even those with the same DNA, express different genes, with some genes being expressed or “turned on”, while others are “turned off”. This expression difference is the basis of cell type diversity in multicellular organisms (like mice and humans). Because some diseases are thought to be the result of improper gene expression, researchers will use differential gene expression experiments to pinpoint the genes responsible for the disease symptoms.

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): gene  
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  
 gene 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

|  |
| --- |
| What do the column names mean? |
| When working with someone else’s data, it can be difficult to figure out what a particular column or variable name means. A **data dictionary** is a key to decode these!  **gene**: the geneID (see below for more detail). **baseMean**: the average of the normalized count values for this gene, taken over all samples. **log2FoldChange**: a measure of the change in gene expression for this gene between the ASD and control mice. A positive value indicates upregulation (increased expression), while a negative Log2FC indicates downregulation (decreased expression). **lfcSE**: the standard error calculated for log2FoldChange. **stat**: t-value calculated to determine if the log2FoldChange is significantly different from zero. **pvalue**: the p-value for the log2FoldChange. **padj**: the p-value, adjusted for multiple testing. |

In this dataset, genes are identified using their Ensembl Gene ID code. Every gene has an ID that looks something like this:

ENSMUSG00000000001

**ENS** stands for “Ensembl”. Ensembl is a [genome database project](https://ensemblgenomes.org/) managed by the European Bioinformatics Institute. It’s one of several databases like this. Others include NIH’s National Center for Biotechnology Information (NCBI; the organization that manages GenBank and PubMed) and the University of California, Santa Cruz (UCSC) Genome Browser. When a gene code starts with “ENS”, it means you should look up the gene code in the Ensembl database.

**MUS** stands for “Mus”, which is the genus for the mouse.

**G** stands for “Gene”. When you see “G” in the ID name, you know you are working with a gene. There are also codes for transcripts (“T”), exons (“E”), and proteins (“P”).

**00000000001** is the numerical code associated with the gene.

### Ranking the genes by log2FoldChange

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  
 gene 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  
 gene 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

### Using the Mouse Genome Database

We have identified the genes (by their geneID) that show the greatest differential expression between ASD and control mice. Now let’s find out what those genes do!

The [Mouse Genome Informatics](https://www.informatics.jax.org) database that tracks mouse genes and expression data. A full introduction to everything available through the MGI can be found [here](https://www.informatics.jax.org/mgihome/other/homepage_usingMGI.shtml). We’ll reproduce some of it below.

To look up information on a particular gene of interest, choose the “Genes” button.



Next, type the gene ID into the “Search” bar up top. We’ll look up “ENSMUSG00000079516”.



After you type the gene ID into the Search bar and hit enter, you should see a new page with basic information about the gene. Click on the gene symbol (in this example, Reg3a) to get more detailed information.



|  |
| --- |
| Tip |
| You might come across some unexpected terms when you search the MGI for your gene ID. In addition to genes, Ensembl gene IDs are also given to “pseudogenes”, “putative genes”, and “lncRNA”.  **pseudogene**: This is a stretch of DNA that looks like a gene but doesn’t actually code for any protein products. It’s essentially a copy of a gene that contains mutations that prevent translation into a protein product. The mutations can include partial deletions, missing promoters, missing start codons, premature stop codons, frameshift mutations, or missing introns. Any of these are enough to result in a pseudogene.  **putative gene**: This is a DNA segment that is believed to be a gene, but its function and protein product has not been confirmed. They are frequently identified based on the presence of an Open Reading Frame. (An Open Reading Frame, or ORF, is a stretch of DNA between a start codon and a stop codon that can be translated into a protein.) Putative genes are not given names until they become confirmed genes.  **lncRNA**: This stands for “long non-coding RNA”. lncRNA is a type of RNA molecule that is transcribed from DNA but does not code for proteins. These RNA molecules are at least 200-500 nucleotides long and play roles in various biological processes, like gene regulation. |

On the new page that you open, details about the gene are organized into categories. Down the left-hand side of the page, you will see sections about the chromosomal location, homology, gene ontology, expression data, and more.

If you scroll down the page, you can also examine the pathways and processes the gene product is involved in under the “gene ontology” section. Gene ontology refers to a common, controlled set of vocabulary and terms that researchers use when categorizing gene and protein function, biological processes, and location within the cell. Reg3a is categorized as a protein involved in carbohydrate derivative binding, signaling receptor activity, and signaling receptor binding. Biologically, the Reg3a protein helps with cell differentiation, cell population proliferation, the immune system, and responding to stimuli.



|  |
| --- |
| 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.

### Creating a gene list

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  
 gene 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

### Compare differential gene expression across regions

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): gene  
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, gene == "ENSMUSG00000079516")  
  
reg3a\_striatum <- filter(striatum, gene == "ENSMUSG00000079516")

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

In the prefrontal cortex:

reg3a\_prefrontal

# A tibble: 1 × 7  
 gene 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  
 gene 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? |

POSSIBLE ADDITION: add code/question where they look up the expression of the genes from part 2?