**Differential Expression and the Mouse Gut-Brain Axis**

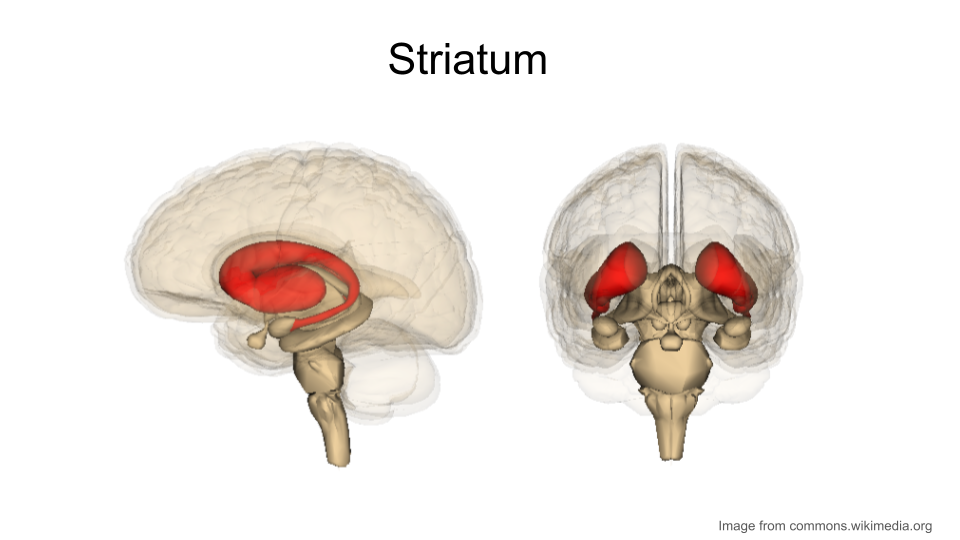
This module provides data for an independent project exploring differential expression in RNA-seq data.

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.

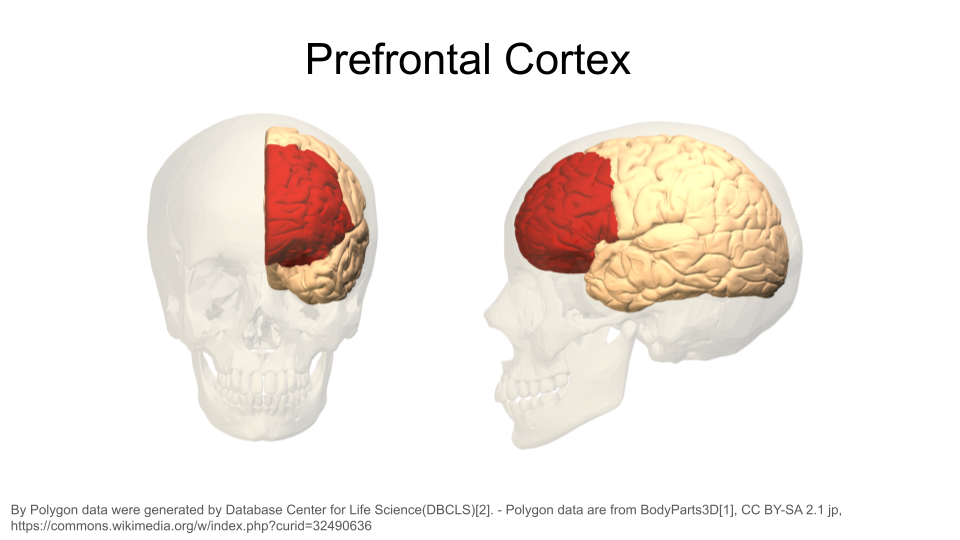
## Background

*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 data for this activity includes gene expression data from two different brain regions (striatum and prefrontal cortex) in mice.



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. Mice in this experiment received fecal transplants. Some mice received transplants from humans who have been diagnosed with 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. They 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)

### Exploring genes of interest with the Mouse Genome Informatics database

The [Mouse Genome Informatics](https://www.informatics.jax.org/genes.shtml) 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.



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. 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 familiar categories. Down the left-hand side of the page, you will see sections about the chromosomal location, homology, gene ontology, expression data, and more. Most sections are expanded by default, but you’ll need to expand the “homology” section yourself.



Once this section is expanded, you can find information about possible human homologs to the mouse gene, including alternate names and where the human homolog is located in the human genome.



If you continue scrolling down the page, you can also examine the pathways and processes the gene product is involved in under the “gene ontology” section. Clicking on the blue squares takes you to a page with more information about how that particular gene was assigned to a pathway or molecular process.



Directly underneath the “ontology” section is information about when the gene is expressed during development. You can learn more information by clicking on the blue squares, or by clicking on the links in the upper right-hand corner. These links will take you to other websites.



## Analysis

Before starting this activity, you should have completed the miniCURE activities and identified genes of interest for your individual project. Remember, the basic steps of your analysis are:

1. Examine expression of your gene of interest in this dataset.
2. Explore whether your gene of interest is differentially expressed between the experimental groups.
3. Characterize the genes that are differentially expressed between the experimental groups based on their molecular activity.

Below, we’ve included a cheatsheet of some of the analysis steps you might want to do and the R code that helps you do it.

### Package Install and Load

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

Type the following into the SciServer console and press return to run the code.

install.packages("tidyverse")

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

library(tidyverse)

|  |
| --- |
| 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. |

### Loading the gene expression count data

The first analyses you might want to do involve exploring dataset that shows how much each gene is expressed overall. This is the gene expression count dataset.

You can load this dataset into R using the following code:

gutbrain\_genes <- read\_csv("https://genomicseducation.org/data/mouse\_gutbrain\_de\_counts.csv")

New names:  
Rows: 55421 Columns: 140  
── Column specification  
──────────────────────────────────────────────────────── Delimiter: "," chr  
(1): ...1 dbl (139): SRR6652436, SRR6652443, SRR6652469, SRR6652471,  
SRR6652505, SRR66...  
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.  
• `` -> `...1`

This dataset has 140 columns. Each row represents a gene. Each column is an individual mouse (sample), except for the first column (“GeneID”, which is the Ensembl gene IDs) and the last column (“total\_counts”, which is the sum across all mice).

#### Examining count data

When you first open the gene expression data, you want to verify that the data has loaded correctly. You can do this with the head and tail commands. Remember, head prints the first 6 rows of a dataset, while tail prints the last 6 rows.

head(gutbrain\_genes)

# A tibble: 6 × 140  
 ...1 SRR6652436 SRR6652443 SRR6652469 SRR6652471 SRR6652505 SRR6652506  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG0000… 0 0 0 0 0 0  
2 ENSMUSG0000… 0 0 0 0 0 0  
3 ENSMUSG0000… 0 0 0 0 0 0  
4 ENSMUSG0000… 0 0 0 0 0 0  
5 ENSMUSG0000… 0 0 0 0 0 0  
6 ENSMUSG0000… 0 0 0 0 0 0  
# ℹ 133 more variables: SRR6652507 <dbl>, SRR6652508 <dbl>, SRR6652509 <dbl>,  
# SRR6652510 <dbl>, SRR6652511 <dbl>, SRR6652512 <dbl>, SRR6652513 <dbl>,  
# SRR6652514 <dbl>, SRR6652515 <dbl>, SRR6652521 <dbl>, SRR6652522 <dbl>,  
# SRR6652389 <dbl>, SRR6652390 <dbl>, SRR6652391 <dbl>, SRR6652392 <dbl>,  
# SRR6652393 <dbl>, SRR6652394 <dbl>, SRR6652395 <dbl>, SRR6652396 <dbl>,  
# SRR6652397 <dbl>, SRR6652398 <dbl>, SRR6652399 <dbl>, SRR6652400 <dbl>,  
# SRR6652401 <dbl>, SRR6652402 <dbl>, SRR6652403 <dbl>, SRR6652404 <dbl>, …

tail(gutbrain\_genes)

# A tibble: 6 × 140  
 ...1 SRR6652436 SRR6652443 SRR6652469 SRR6652471 SRR6652505 SRR6652506  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG0000… 0 0 0 0 0 0  
2 ENSMUSG0000… 18605 12561 14910 17809 6377 14750  
3 ENSMUSG0000… 4469 2944 3748 5384 2409 7057  
4 ENSMUSG0000… 117349 27498 68321 62645 65449 126271  
5 ENSMUSG0000… 175 36 87 100 148 18  
6 ENSMUSG0000… 3478 775 2437 1510 1665 5238  
# ℹ 133 more variables: SRR6652507 <dbl>, SRR6652508 <dbl>, SRR6652509 <dbl>,  
# SRR6652510 <dbl>, SRR6652511 <dbl>, SRR6652512 <dbl>, SRR6652513 <dbl>,  
# SRR6652514 <dbl>, SRR6652515 <dbl>, SRR6652521 <dbl>, SRR6652522 <dbl>,  
# SRR6652389 <dbl>, SRR6652390 <dbl>, SRR6652391 <dbl>, SRR6652392 <dbl>,  
# SRR6652393 <dbl>, SRR6652394 <dbl>, SRR6652395 <dbl>, SRR6652396 <dbl>,  
# SRR6652397 <dbl>, SRR6652398 <dbl>, SRR6652399 <dbl>, SRR6652400 <dbl>,  
# SRR6652401 <dbl>, SRR6652402 <dbl>, SRR6652403 <dbl>, SRR6652404 <dbl>, …

#### Summarizing data

You may want to explore the distribution of your count data using summary. This command lets you see the mean, median, minimum, and maximum for each numeric column. You can specify which column you’d like to get summary data for, or look at summary data for all the columns.

summary(gutbrain\_genes$total\_counts)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
0.000e+00 9.600e+01 6.666e+03 4.130e+06 7.888e+05 7.732e+09

You can also find the total number of reads in the dataset.

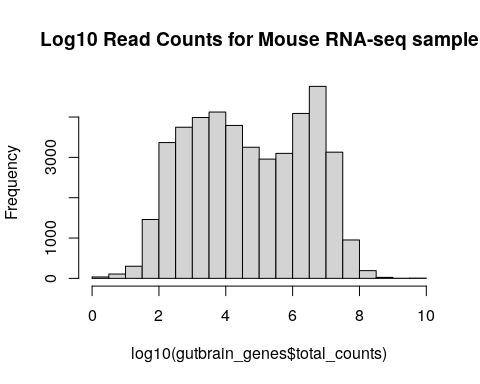
sum(gutbrain\_genes$total\_counts)

[1] 2.28913e+11

#### Plotting histograms and dealing with zero counts

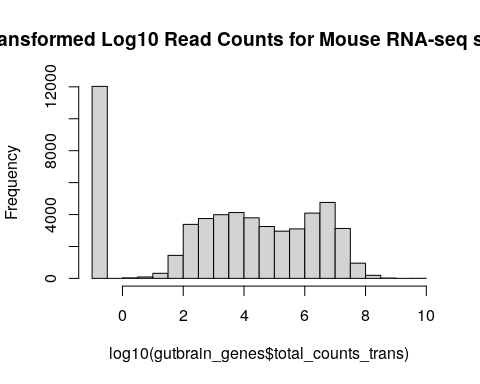
A histogram is a good way to visually explore the distribution of the read counts, as well as get an idea of how many genes have few or no reads. Use a log10 scale on the x-axis.

hist(log10(gutbrain\_genes$total\_counts), main="Log10 Read Counts for Mouse RNA-seq sample")



You may choose to deal with zero-count genes by adding 0.1 to each count before the log10 transformation.

gutbrain\_genes$total\_counts\_trans <- gutbrain\_genes$total\_counts + 0.1  
  
hist(log10(gutbrain\_genes$total\_counts\_trans), main="Transformed Log10 Read Counts for Mouse RNA-seq sample")



#### Looking at read count for a single gene

You may want to look at the read count for a specific gene in your dataset. This can be done with the filter command. Let’s look at the read counts for the reg3a gene, across all mice. Remember, this is the gene we looked up earlier in the MGI database. The Ensembl gene ID is “ENSMUSG00000079516”.

filter(gutbrain\_genes, GeneID == "ENSMUSG00000079516")

### Loading the differential expression data

You have the option to look at gene expression in control vs ASD mice, as well as gene expression in both striatum and prefrontal cortex. All the mice were male and sacrificed at the same age (45 days).

Additionally, you can look at gene expression in striatum vs prefrontal cortex in only ASD mice or only control mice. Likewise, you also have the option of looking at gene expression in ASD and control mice, focusing only on striatum or only on prefrontal cortex.

Let’s say you want to open the dataset that compares gene expression between ASD and control mice in both brain regions and call it asd\_vs\_c. We will do this using read\_csv. Copy and paste this command into your console:

asd\_vs\_c <- read\_csv("https://genomicseducation.org/data/mouse\_gutbrain\_de\_autismVcontrol.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.

Here are the URLs for all the possible comparisons you can examine with this dataset:

**Comparing gene expression between ASD and control mice**

Both brain regions: <https://genomicseducation.org/data/mouse_gutbrain_de_autismVcontrol.csv>

Prefrontal cortex only: <https://genomicseducation.org/data/mouse_gutbrain_de_autismVcontrol_in_prefrontalcortex.csv>

Striatum only: <https://genomicseducation.org/data/mouse_gutbrain_de_autismVcontrol_in_striatum.csv>

**Comparing gene expression between prefrontal cortex and striatum**

All mice: <https://genomicseducation.org/data/mouse_gutbrain_de_tissuetype.csv>

Only ASD mice: <https://genomicseducation.org/data/mouse_gutbrain_de_tissuetype_in_ASDmice.csv>

Only control mice: <https://genomicseducation.org/data/mouse_gutbrain_de_tissuetype_in_controlmice.csv>

#### Arranging dataset based on padj

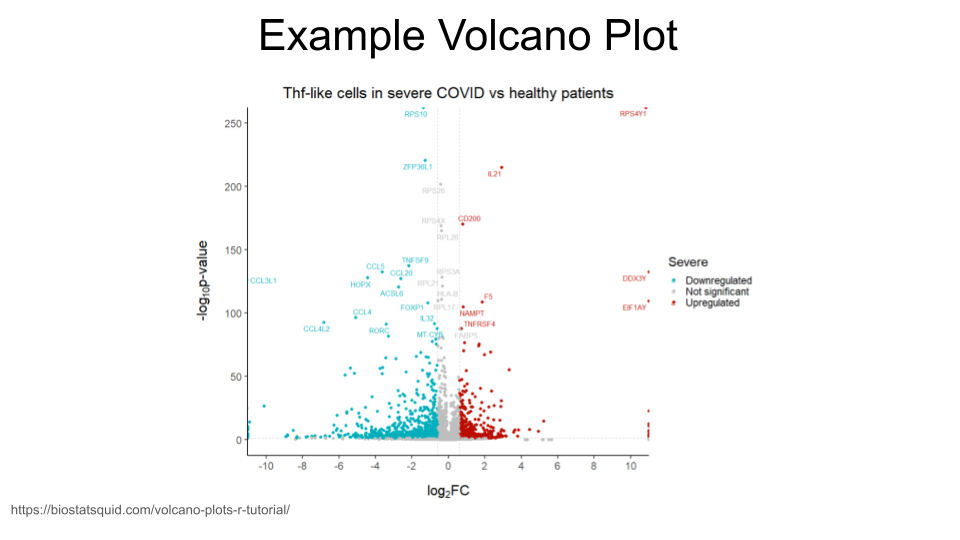
You may be interested in identifying the genes that have been differentially expressed between your two groups (in this case, ASD mouse brains and control mouse brains). You can use the arrange command to sort the results to put the smallest padj values first.

asd\_vs\_c\_sorted <- arrange(asd\_vs\_c, padj)  
head(asd\_vs\_c\_sorted)

# A tibble: 6 × 7  
 gene baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000089657 16.9 -25.5 1.81 -14.1 2.18e-45 7.29e-41  
2 ENSMUSG00000083812 14.6 -25.4 1.80 -14.1 3.82e-45 7.29e-41  
3 ENSMUSG00000094151 11.2 -25.0 1.79 -14.0 2.97e-44 3.77e-40  
4 ENSMUSG00000102414 14.9 -25.4 1.89 -13.4 4.00e-41 3.81e-37  
5 ENSMUSG00000074445 11.5 -25.0 1.89 -13.3 3.44e-40 2.62e-36  
6 ENSMUSG00000053773 10.6 -24.9 1.99 -12.5 6.66e-36 4.23e-32

#### Creating a volcano plot

A volcano plot is a great way to visualize differential gene expression between two datasets. These types of plots have log2FoldChange on the x-axis and a log transformed padj on the y-axis.



We will first need to do a log transformation of the padj variable. We do this in order to spread our values out along the y-axis - otherwise, all the dots that represent genes will be too scrunched together.

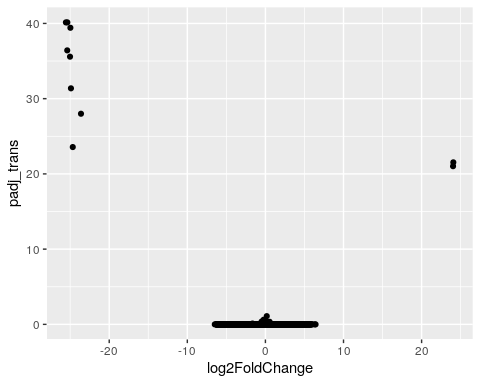
asd\_vs\_c$padj\_trans <- -log10(asd\_vs\_c$padj)

Now, we just create a scatterplot. The code for this is kind of complicated, so you will want to paste the code below into your console.

Here’s the code for your basic volcano plot. data tells R what data to use, x tells R what the x-axis values should be, and y tells R what the y-axis values should be. Notice that we can just use the column names for x and y!

ggplot(data = asd\_vs\_c, aes(x = log2FoldChange, y = padj\_trans)) +  
 geom\_point()

Warning: Removed 17318 rows containing missing values or values outside the scale range  
(`geom\_point()`).



We can also color code the genes (points) which are significantly upregulated or downregulated. In order to do this, we will need to create a new column in our dataset that identifies whether a gene is upregulated in ASD-type mice, downregulated in ASD-type mice, or has the same regulation in ASD-type mice compared to control-type mice.

In this new column, we can decide the thresholds that a log2FoldChange value must meet in order to be considered upregulated or downregulated. We will consider genes to be upregulated if the log2fold change is greater than 0.6 and downregulated if it is less than -0.6. **You can change these thresholds to whatever makes sense for you.**

log2foldchange is just a representation of a ratio, so the sign indicates the direction of regulation. It’s important to remember which group is your comparison group!

For example, if you are look at regulation in group A compared to group B, you might have a log2foldchange value of 1 (indicating the gene is expressed twice as much in A as it is in B).

However, if you are looking at regulation in group B compared to group A, the log2foldchange value will be -1 (because the gene is expressed half as much in B as it is in A).

The comparison group matters with ratios!

We are also setting a significant p-value as less than 0.05 (but you can also change this based on your best judgment).

We start by creating a new column called diffexpressed and setting everything to “NO” (as in “no expression differences between ASD-type and control-type mice”). We then specify the genes for which this column should be changed to “UP” or “DOWN” (based on log2FoldChange and padj).

To recap: A gene is labeled “UP” if its log2FoldChange value is greater than 0.6 *and* its adjusted p-value is less than 0.05.

A gene is labeled “DOWN” if its log2FoldChange value is less than -0.6 *and* its adjusted p-value is less than 0.05.

asd\_vs\_c$diffexpressed <- "NO"  
  
asd\_vs\_c$diffexpressed[asd\_vs\_c$log2FoldChange > 0.6 & asd\_vs\_c$padj < 0.05] <- "UP"  
  
asd\_vs\_c$diffexpressed[asd\_vs\_c$log2FoldChange < -0.6 & asd\_vs\_c$padj < 0.05] <- "DOWN"  
  
head(asd\_vs\_c)

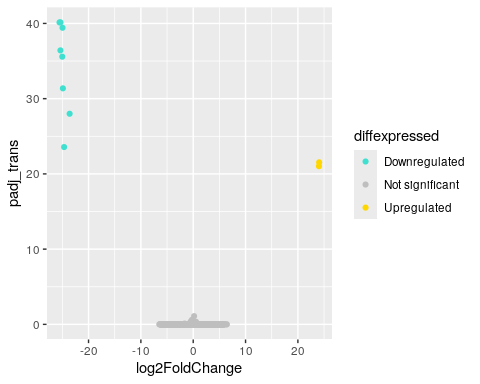
# A tibble: 6 × 9  
 gene baseMean log2FoldChange lfcSE stat pvalue padj padj\_trans  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000… 16.9 -25.5 1.81 -14.1 2.18e-45 7.29e-41 40.1  
2 ENSMUSG00000… 14.6 -25.4 1.80 -14.1 3.82e-45 7.29e-41 40.1  
3 ENSMUSG00000… 11.2 -25.0 1.79 -14.0 2.97e-44 3.77e-40 39.4  
4 ENSMUSG00000… 14.9 -25.4 1.89 -13.4 4.00e-41 3.81e-37 36.4  
5 ENSMUSG00000… 11.5 -25.0 1.89 -13.3 3.44e-40 2.62e-36 35.6  
6 ENSMUSG00000… 10.6 -24.9 1.99 -12.5 6.66e-36 4.23e-32 31.4  
# ℹ 1 more variable: diffexpressed <chr>

Okay! We can now add some color to the volcano plot we made earlier. We do this by adding a line to our code that tells R to color the data points based on the diffexpressed column (col = diffexpressed). We also add a line about the colors we want to use (scale\_color\_manual). In this case, we have chosen to make the points for downregulated genes turquoise and the points for upregulated genes gold. The additional command labels just creates the labels on the legend.

You can find an extensive list of all the colors available in R at <https://www.datanovia.com/en/blog/awesome-list-of-657-r-color-names/>.

ggplot(data = asd\_vs\_c, aes(x = log2FoldChange, y = padj\_trans, col = diffexpressed)) +  
 geom\_point() +  
 scale\_color\_manual(values = c("turquoise", "grey", "gold"),   
 labels = c("Downregulated", "Not significant", "Upregulated"))

Warning: Removed 17318 rows containing missing values or values outside the scale range  
(`geom\_point()`).



We can also label genes of interest in our plot. Let’s label the Reg3a gene (the same one we looked up in the MGI database), as well as Pcdh12 (or protocadherin 12). The two gene IDs we need to know are ENSMUSG00000079516 and ENSMUSG00000024440.

We will create another new column in our dataset. We will then tell R In this new column, the genes in our list (Reg3a and Pcdh12) are named, which everything else will have a missing name. When we remake the volcano plot, R will *only* label those points that do not have missing names in the gene\_label column.

This is going to require a little bit of R trickery and something called an “ifelse” statement.

In R, an “ifelse” statement tells R to do one of two things. If the condition stated right after ifelse( is true, then R should do the second thing in the statement. If the condition is not true, then R should do the third thing.

In our command, R is going through the asd\_vs\_c row by row. For each row, we are telling R to first check whether the geneID of that row is “ENSMUSG00000079516”. If it is, then R will fill in the gene\_label column with the gene symbol “Reg3a”. If it is not, R moves onto the next part of our command, which is a second (or nested) ifelse loop.

When R moves onto this second ifelse loop, it’s now checking to see if the geneID for the row is “ENSMUSG00000024440”. If it is, then R fills in the gene\_label column with the gene symbol “Pcdh12”. If it is not, R fills in the gene\_label column with “NA”, which basically means the information for this column is missing.

asd\_vs\_c$gene\_label <- ifelse(  
 asd\_vs\_c$gene == "ENSMUSG00000079516", "Reg3a",  
   
 ifelse(  
 asd\_vs\_c$gene == "ENSMUSG00000024440", "Pcdh12", NA   
 ))

Great! You can check to see what the new column gene\_label looks like by using the head command.

head(asd\_vs\_c)

# A tibble: 6 × 10  
 gene baseMean log2FoldChange lfcSE stat pvalue padj padj\_trans  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000… 16.9 -25.5 1.81 -14.1 2.18e-45 7.29e-41 40.1  
2 ENSMUSG00000… 14.6 -25.4 1.80 -14.1 3.82e-45 7.29e-41 40.1  
3 ENSMUSG00000… 11.2 -25.0 1.79 -14.0 2.97e-44 3.77e-40 39.4  
4 ENSMUSG00000… 14.9 -25.4 1.89 -13.4 4.00e-41 3.81e-37 36.4  
5 ENSMUSG00000… 11.5 -25.0 1.89 -13.3 3.44e-40 2.62e-36 35.6  
6 ENSMUSG00000… 10.6 -24.9 1.99 -12.5 6.66e-36 4.23e-32 31.4  
# ℹ 2 more variables: diffexpressed <chr>, gene\_label <chr>

Now we just need to remake our volcano plot and add some to code that tells R to label the points we’re interested in. The command label = gene\_label goes right after the code that tells R how to color the points. We also need to add something at the end of our code that tells R to put the labels on top of the plot that we’ve made. That’s the + geom\_text\_repel(max.overlaps = Inf) command.

In order for this command to work, we do need to install and load a new library called ggrepel.

install.packages('ggrepel')

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

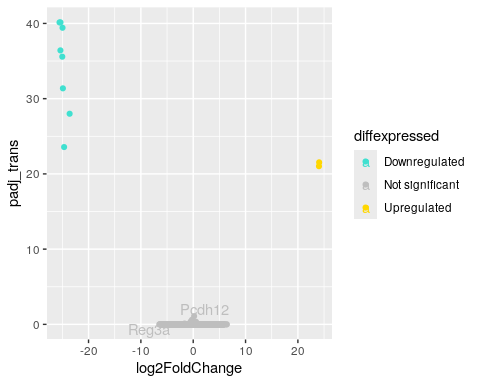
library(ggrepel)

Excellent. Now that we have installed and loaded this package (which you only need to do once for each R session!), we can run the code to create the volcano plot. Remember, we added label = gene\_label to tell R to add labels to the points, and geom\_text\_repel(max.overlaps = Inf) to tell R to put the labels on top of the plot so we can see them.

ggplot(data = asd\_vs\_c, aes(x = log2FoldChange, y = padj\_trans, col = diffexpressed, label = gene\_label)) +  
 geom\_point() +  
 scale\_color\_manual(values = c("turquoise", "grey", "gold"),   
 labels = c("Downregulated", "Not significant", "Upregulated")) +  
 geom\_text\_repel(max.overlaps = Inf)

Warning: Removed 17318 rows containing missing values or values outside the scale range  
(`geom\_point()`).

Warning: Removed 55419 rows containing missing values or values outside the scale range  
(`geom\_text\_repel()`).

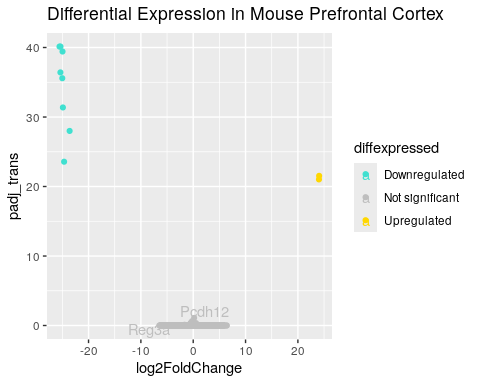


You may want to add a title to your plot. This can be done with the ggtitle command.

ggplot(data = asd\_vs\_c, aes(x = log2FoldChange, y = padj\_trans, col = diffexpressed, label = gene\_label)) +  
 geom\_point() +  
 scale\_color\_manual(values = c("turquoise", "grey", "gold"),   
 labels = c("Downregulated", "Not significant", "Upregulated")) +  
 geom\_text\_repel(max.overlaps = Inf) +  
 ggtitle("Differential Expression in Mouse Prefrontal Cortex")

Warning: Removed 17318 rows containing missing values or values outside the scale range  
(`geom\_point()`).

Warning: Removed 55419 rows containing missing values or values outside the scale range  
(`geom\_text\_repel()`).

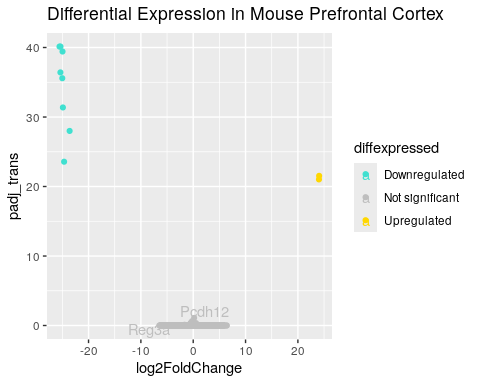


Here’s all the code we used to build a volcano plot in one place so that you can easily run it again (and make modifications).

#log transform the padj values so the y-axis is formatted for plotting  
asd\_vs\_c$padj\_trans <- -log10(asd\_vs\_c$padj)  
  
#Create the diffexpressed column so that we can color the upregulated and downregulated genes on the plot  
asd\_vs\_c$diffexpressed <- "NO"  
  
asd\_vs\_c$diffexpressed[asd\_vs\_c$log2FoldChange > 0.6 & asd\_vs\_c$padj < 0.05] <- "UP"  
  
asd\_vs\_c$diffexpressed[asd\_vs\_c$log2FoldChange < -0.6 & asd\_vs\_c$padj < 0.05] <- "DOWN"  
  
#Create the gene\_label column so that we can label individual points  
asd\_vs\_c$gene\_label <- ifelse(  
 asd\_vs\_c$gene == "ENSMUSG00000079516", "Reg3a",  
   
 ifelse(  
 asd\_vs\_c$gene == "ENSMUSG00000024440", "Pcdh12", NA   
 ))  
  
#Make the final plot  
ggplot(data = asd\_vs\_c, aes(x = log2FoldChange, y = padj\_trans, col = diffexpressed, label = gene\_label)) +  
 geom\_point() +  
 scale\_color\_manual(values = c("turquoise", "grey", "gold"),   
 labels = c("Downregulated", "Not significant", "Upregulated")) +  
 geom\_text\_repel(max.overlaps = Inf) +  
 ggtitle("Differential Expression in Mouse Prefrontal Cortex")

Warning: Removed 17318 rows containing missing values or values outside the scale range  
(`geom\_point()`).

Warning: Removed 55419 rows containing missing values or values outside the scale range  
(`geom\_text\_repel()`).



#### Creating a gene list

You might want to create a list of all the genes that are differentially expressed between your two groups using the filter command. Make sure to check that your gene list has enough genes on it using head and dim (and that you have a good ratio of “real” positive versus false positive results).

This step requires some thought as to what you want to set as your padj threshold! Remember, you want to make sure you have a minimal number of false positive results, while also still keeping enough genes for future analyses.

asd\_vs\_c\_sig <- filter(asd\_vs\_c, padj < 0.05)  
head(asd\_vs\_c\_sig)

# A tibble: 6 × 10  
 gene baseMean log2FoldChange lfcSE stat pvalue padj padj\_trans  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000… 16.9 -25.5 1.81 -14.1 2.18e-45 7.29e-41 40.1  
2 ENSMUSG00000… 14.6 -25.4 1.80 -14.1 3.82e-45 7.29e-41 40.1  
3 ENSMUSG00000… 11.2 -25.0 1.79 -14.0 2.97e-44 3.77e-40 39.4  
4 ENSMUSG00000… 14.9 -25.4 1.89 -13.4 4.00e-41 3.81e-37 36.4  
5 ENSMUSG00000… 11.5 -25.0 1.89 -13.3 3.44e-40 2.62e-36 35.6  
6 ENSMUSG00000… 10.6 -24.9 1.99 -12.5 6.66e-36 4.23e-32 31.4  
# ℹ 2 more variables: diffexpressed <chr>, gene\_label <chr>

dim(asd\_vs\_c\_sig)

[1] 10 10

#### Running a gene set analysis

You can use the special command runClusterProfiler to figure out the types of processes genes on your gene list are involved in. You can also create a dotplot to visualize your results. (NOTE: this is only possible if you are using the C-MOOR environment in SciServer.)

asd\_vs\_c\_clusters <- runClusterProfiler(asd\_vs\_c\_sig)  
  
dotplot(asd\_vs\_c\_clusters, showCategory=34, title="asd vs control in prefrontal cortex", font.size=10, label\_format = 50)

#### Compare differential gene expression of a SINGLE GENE across regions

It might be interesting to look at the expression of a gene across all the possible regions. To do this, we will first load the region-specific datasets.

asd\_vs\_c\_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.

asd\_vs\_c\_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(asd\_vs\_c\_prefrontal, gene == "ENSMUSG00000079516")  
  
reg3a\_striatum <- filter(asd\_vs\_c\_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

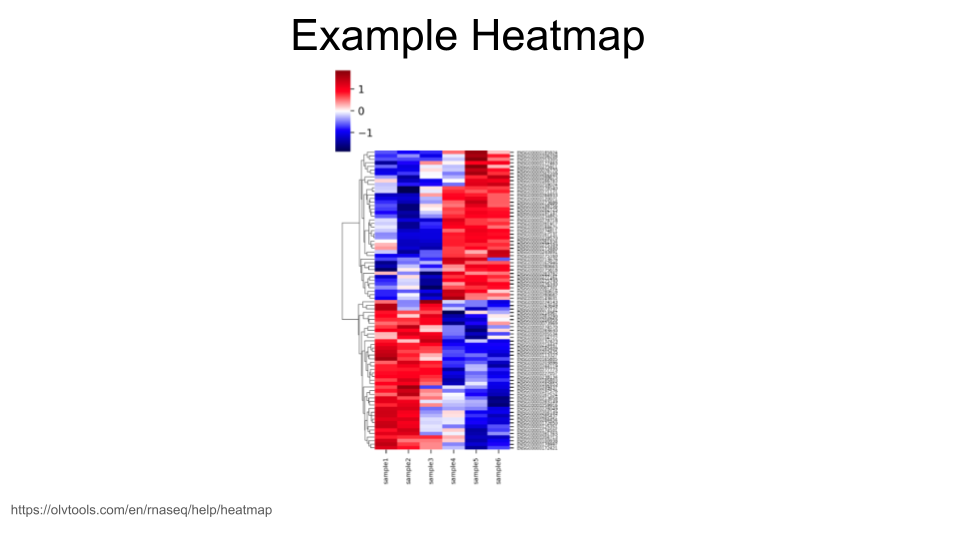
# A tibble: 1 × 7  
 gene baseMean log2FoldChange lfcSE stat pvalue padj  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG00000079516 0.371 2.27 2.96 0.765 0.444 NA

You can create a histogram of the gene expression counts for this gene using the special plotAcrossRegions command. (NOTE: this is only available in SciServer.)

plotAcrossRegions("ENSMUSG00000079516")

#### Visualize differential gene expression of MULTIPLE GENES across regions

We can also look at the differential gene expression of a set of genes at once! This is usually done by creating a **heatmap**. In RNA-seq analyses, we usually put individual samples on the x-axis and normalized gene counts on the y-axis.



When creating heatmaps, we want to work with a dataset that includes normalized gene expression data for each gene. We loaded a dataset like this above that we called gutbrain\_genes. If you don’t have an object in your environment called this, scroll back up to the “**Loading the gene expression count data**” section and run the code again.

The dataset you loaded has expression data for 55,421 different genes, which is too many to visualize. (You can try, but R will usually throw an error if you do.) In general, though, we are mostly interested in looking at the visualization of a particular group of genes, or a gene set.

Let’s say we have identified 6 mouse genes () that have been linked to the development of motor neurons. We’re interested in whether these genes show differential gene expression between the striatum and the prefrontal cortex. We can use the filter command to pull out normalized gene expression data for just these six genes.

These are the genes (with gene IDs) that we’re including in our gene set:

Sema3a: ENSMUSG00000028883 Mapk8: ENSMUSG00000021936 Nrcam: ENSMUSG00000020598 Dlg4: ENSMUSG00000020886 Slit1: ENSMUSG00000025020 Creb1: ENSMUSG00000025958

First, create a vector that contains these gene IDs. (You can replace the gene IDs with the ones for your genes of interest, but make sure you use Ensembl gene IDs! They should all start with “ENSMUSG”.)

gene\_set <- c("ENSMUSG00000028883", "ENSMUSG00000021936", "ENSMUSG00000020598", "ENSMUSG00000020886", "ENSMUSG00000025020", "ENSMUSG00000025958")

Now we simply filter the gutbrain\_genes dataset so that we only keep the rows for the gene IDs in our gene set. Remember, we want to use the filter command.

gutbrain\_motor <- filter(gutbrain\_genes, gutbrain\_genes$...1 %in% gene\_set)  
  
head(gutbrain\_motor)

# A tibble: 6 × 141  
 ...1 SRR6652436 SRR6652443 SRR6652469 SRR6652471 SRR6652505 SRR6652506  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 ENSMUSG0000… 86704 93818 62879 116582 50148 93087  
2 ENSMUSG0000… 2220069 1537477 1521896 1355721 945194 1790682  
3 ENSMUSG0000… 462615 362226 557979 714703 396624 750340  
4 ENSMUSG0000… 113380 104242 156926 223405 111427 211342  
5 ENSMUSG0000… 26284 19265 74037 56337 49613 93139  
6 ENSMUSG0000… 44034 30884 26225 45861 26901 46580  
# ℹ 134 more variables: SRR6652507 <dbl>, SRR6652508 <dbl>, SRR6652509 <dbl>,  
# SRR6652510 <dbl>, SRR6652511 <dbl>, SRR6652512 <dbl>, SRR6652513 <dbl>,  
# SRR6652514 <dbl>, SRR6652515 <dbl>, SRR6652521 <dbl>, SRR6652522 <dbl>,  
# SRR6652389 <dbl>, SRR6652390 <dbl>, SRR6652391 <dbl>, SRR6652392 <dbl>,  
# SRR6652393 <dbl>, SRR6652394 <dbl>, SRR6652395 <dbl>, SRR6652396 <dbl>,  
# SRR6652397 <dbl>, SRR6652398 <dbl>, SRR6652399 <dbl>, SRR6652400 <dbl>,  
# SRR6652401 <dbl>, SRR6652402 <dbl>, SRR6652403 <dbl>, SRR6652404 <dbl>, …

We also want to filter out some of the samples. Because we are interested in the gene expression of these six genes in the striatum compared to the prefrontal cortex, we want to control for any possible differences caused by the sample coming from an ASD-type mouse vs a control-type mouse.