Introduction to RNA-Seq & Transcriptome Analysis

Jessica Holmes

Introduction

Steps in gray have been performed already. For this lab, you'll focus on the last step in **black**

a) Filter and trim reads using fastp or trimmomatic

b) Create Salmon index & prep files

c) Use Salmon to pseudo-align RNA-Seq reads to mouse transcriptome

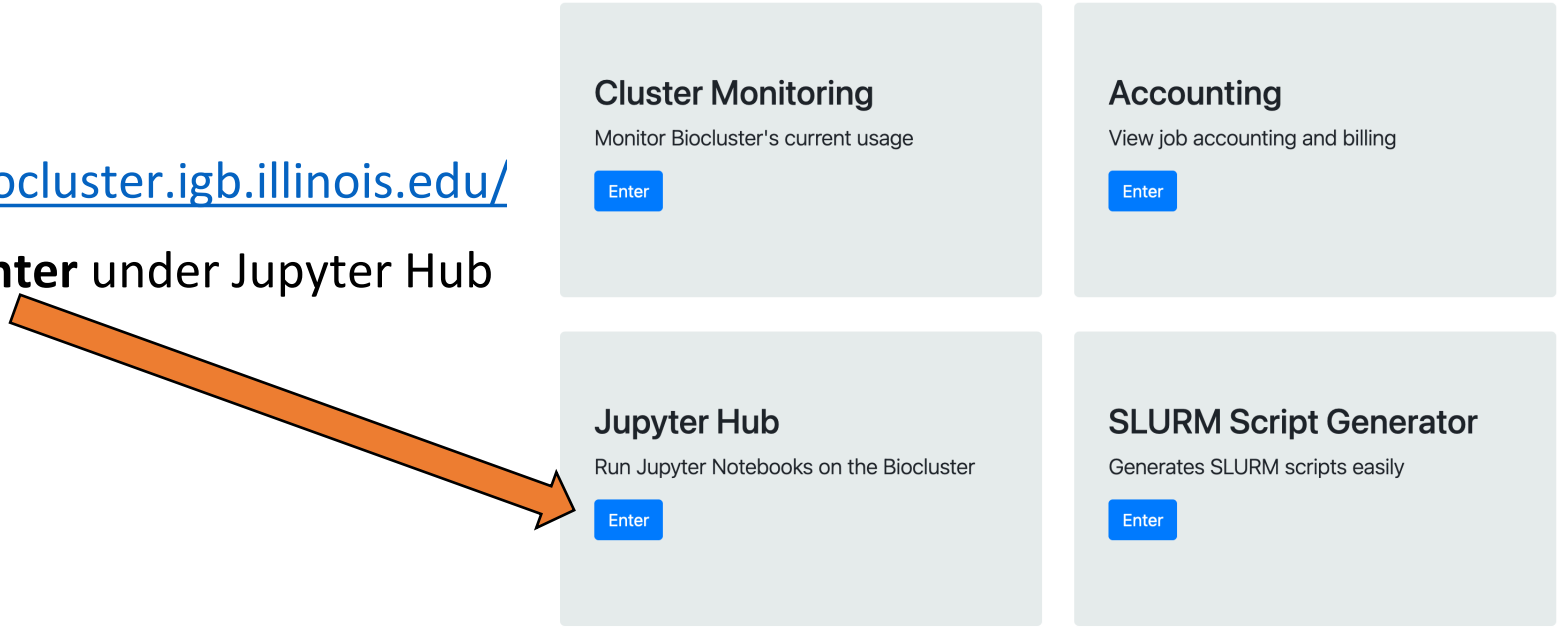
d) Use multiqc to summarize results of each step

e) Use R packages to clean, sort, and filter data

f) Use edgeR and limma to find differentially expressed genes

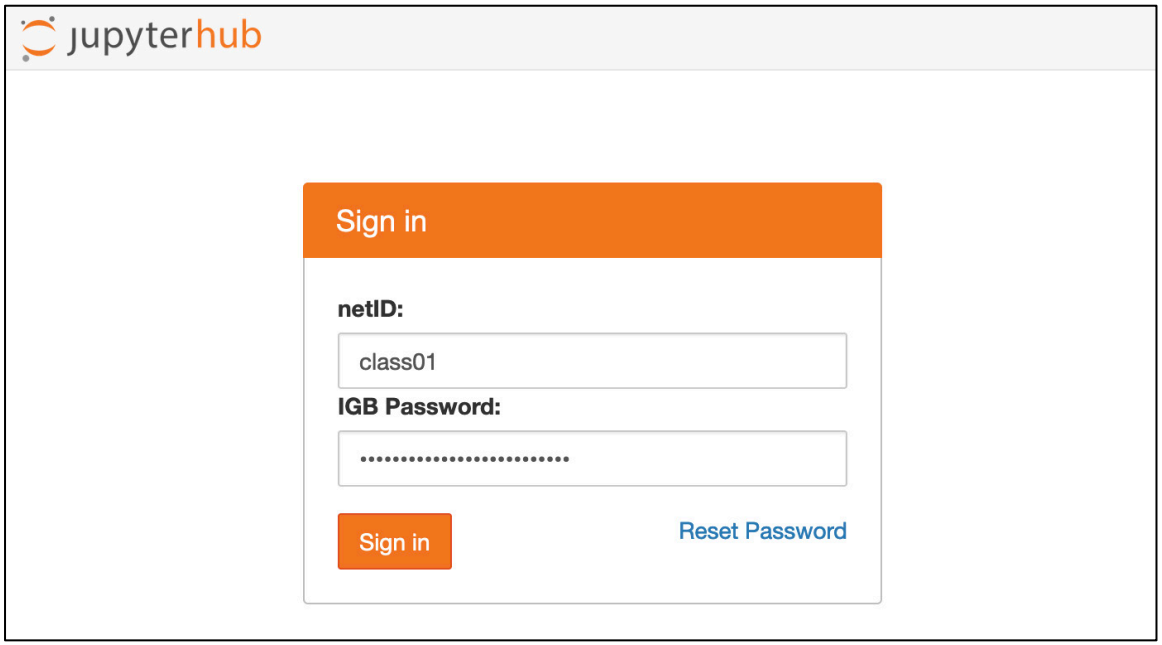
Step 1: Start Jupyter Hub

1. Go to https://biocluster.igb.illinois.edu/
2. Click on **Enter** under Jupyter Hub



Step 1: Start Jupyter Hub

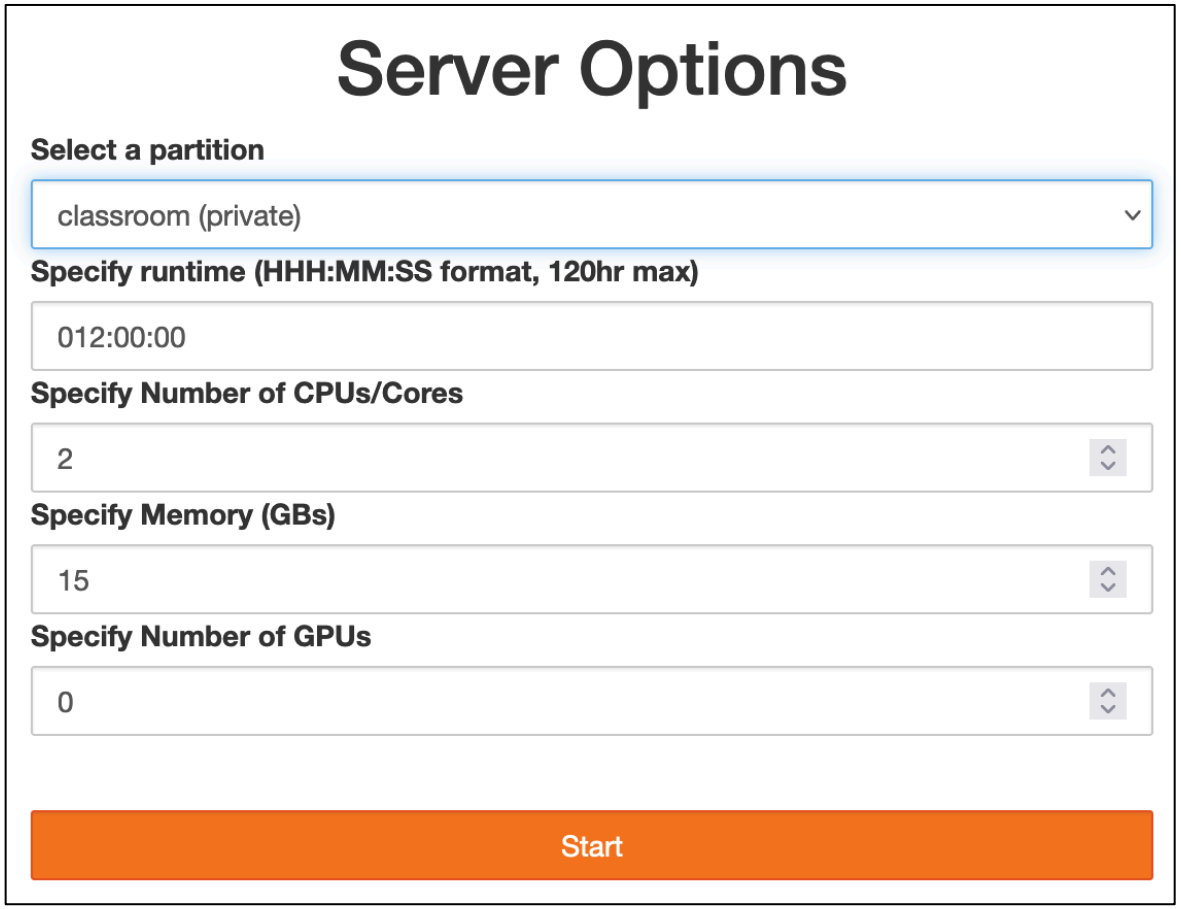
- Enter login credentials assigned to you.
- Example username: **class01**



- **If you've logged in during a previous lab, then you may not see this screen. Move on to next slide.**
- **I you have not logged in during a previous lab and you don't see this, then you may be logged in with a different account. Please log out first (upper right corner) and then sign in with your assigned credentials.**

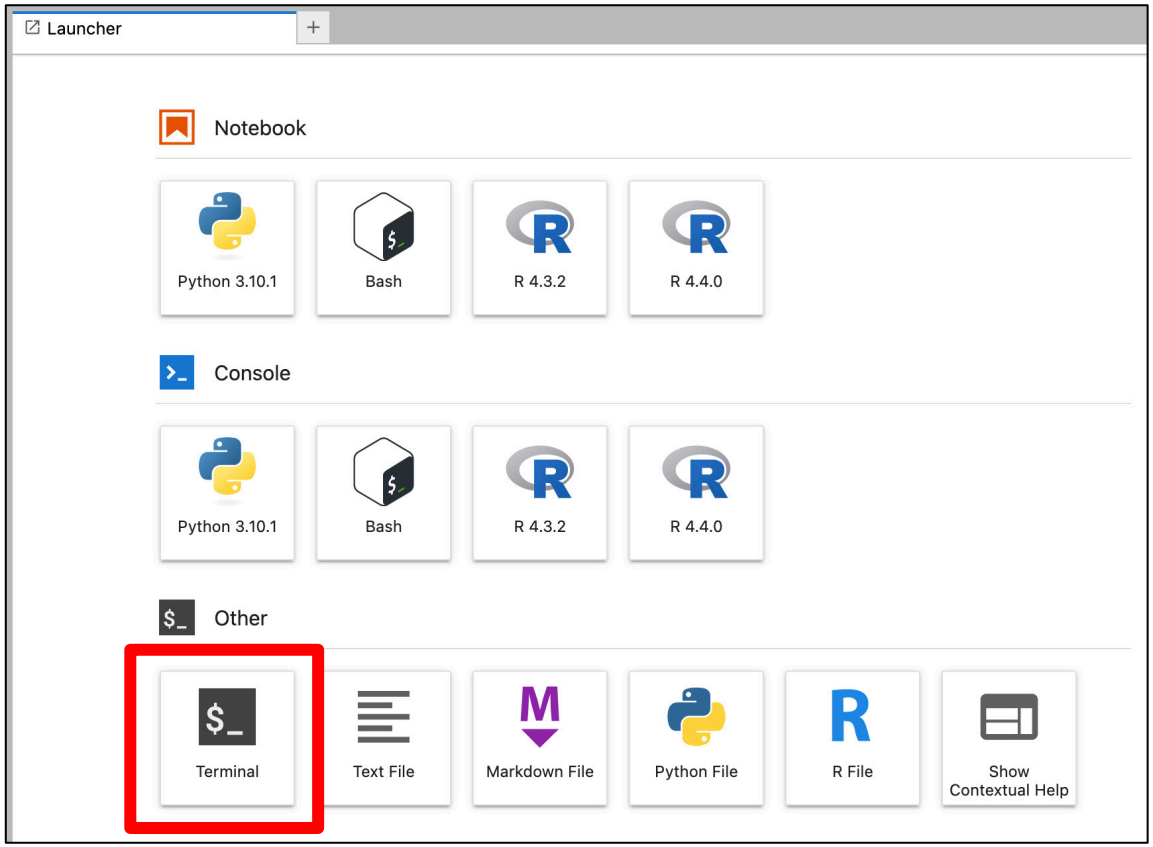
Step 1: Start Jupyter Hub

- Select partition: **classroom**
- Specify runtime: **012:00:00**
- Specify CPUs: **2**
- Click on **Start**



Step 2: Start Terminal & Copy Materials

- Click on **Terminal**



Step 2: Start Terminal & Copy Materials

The following commands will copy a folder filled with necessary files for today's lab.

```
# Note: ~ is a symbolic of your home directory

$ cd ~/

# Copies exercise materials to your home directory

$ cp -r /home/classroom/mayo/2025/04-Bulk-RNA-Sequencing/mouse-rnaseq-2025/ .
```

Step 2: Start Terminal & Copy Materials

Navigate to the mouse-rnaseq-2025/ directory and look at what the folder contains:

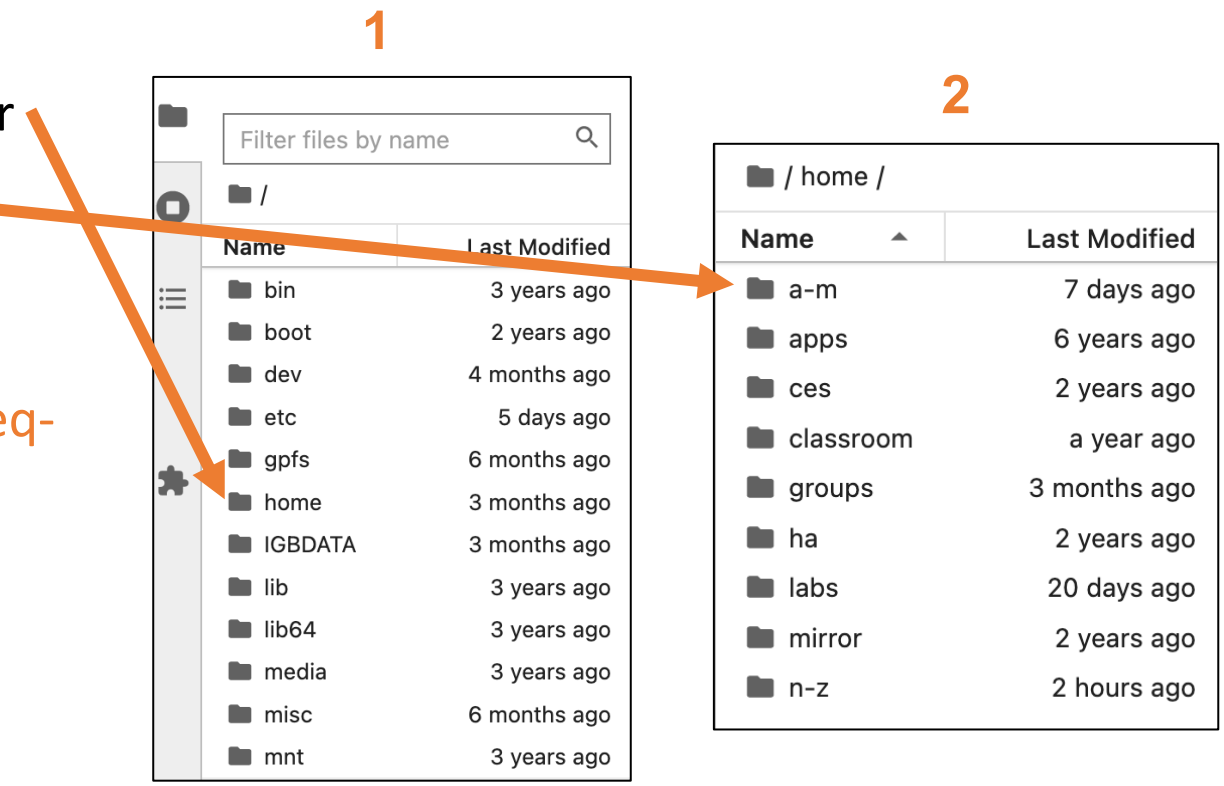
```
$ tree mouse-rnaseq-2025/
mouse-rnaseq-2025/
├── data
│   └── reference
│       └── tx_info_GRCm39_RS_2024_02.csv
├── results
│   ├── salmon
│   │   └── SalmonSummarizedOutput.RData
│   ├── stats
│   └── Targets0.txt
├── RNA-Seq_R_stats_2025.ipynb
└── src
    └── summarizeFit.R
```

Name	Description
tx_info_GRCm39_RS_2024_02.csv	Contains mouse transcript IDs, gene IDs, gene symbols, and gene products. This info was manually pulled from an annotation file (GFF/GTF) to aid in summing transcripts to the gene level.
SalmonSummarizedOutput.Rdata	Salmon transcript-level counts and Salmon run metadata.
Targets0.txt	Metadata about the mouse dataset that we've collected.
RNA-Seq_R_stats_2025.ipynb	R notebook that contains code we'll run to complete our statistical analysis. This needs to be in the directory you wish to work in.
summarizeFit.R	An R script that we'll utilize later.

Step 3: Statistical Analysis in R notebook

First navigate to the mouse-rnaseq-2025 folder double clicking on the icons available on the left-hand side of Jupyter. If you are already in your home directory, you can skip to step 4.

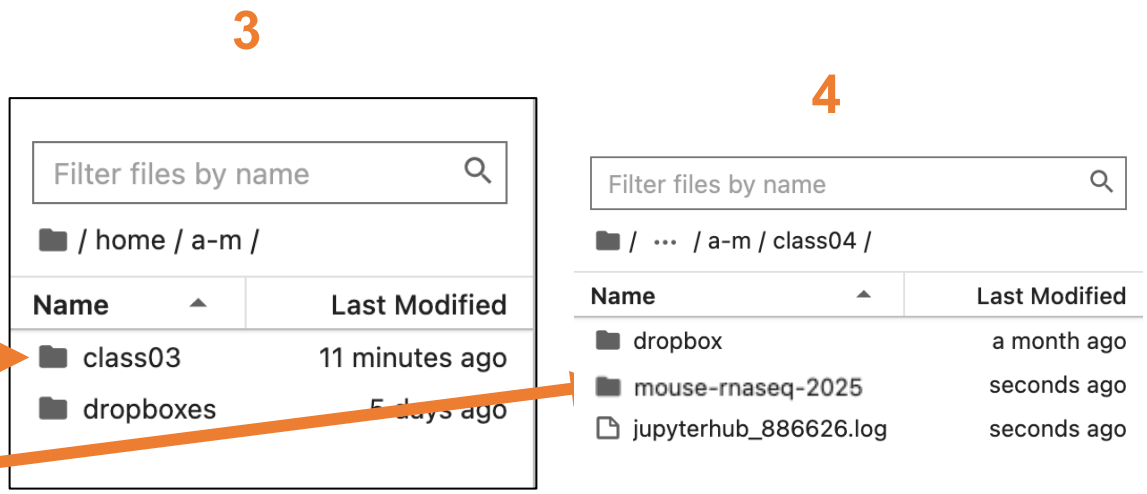
1. Go to home/ folder
2. Go to a-m/ folder
3. Go to your unique class##/ folder
4. Go to mouse-rnaseq-2025/ folder



Step 3: Statistical Analysis in R notebook

First navigate to the mouse-rnaseq-2025 folder double clicking on the icons available on the left-hand side of Jupyter. If you are already in your home directory, you can skip to step 4.

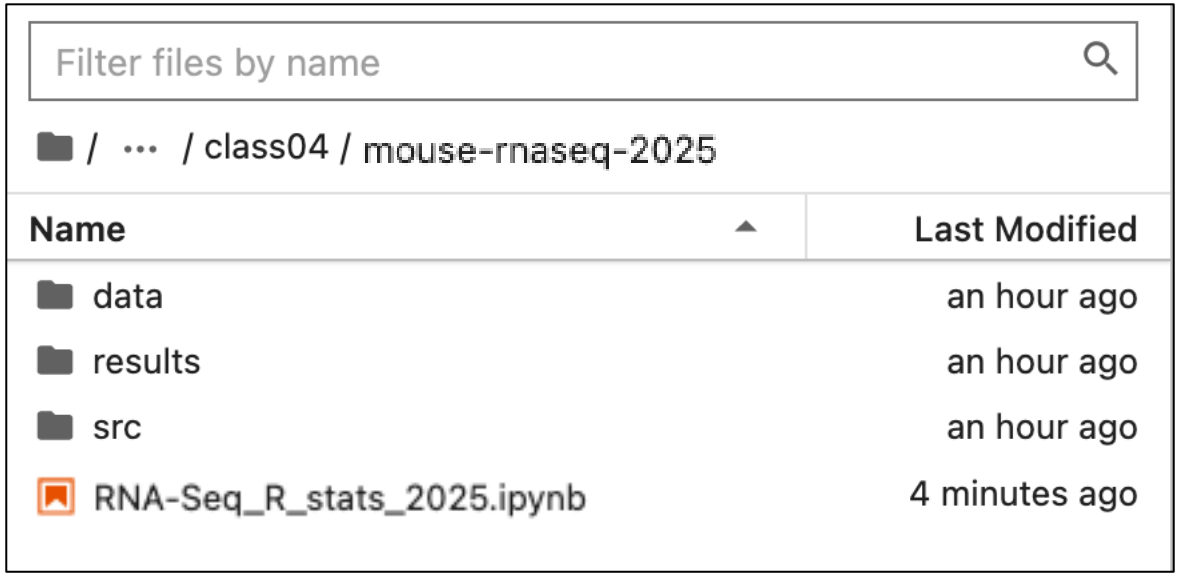
1. Go to home/ folder
2. Go to a-m/ folder
3. Go to your unique class##/ folder
4. Go to mouse-rnaseq-2025/ folder



Step 4: Statistical analysis in R notebook

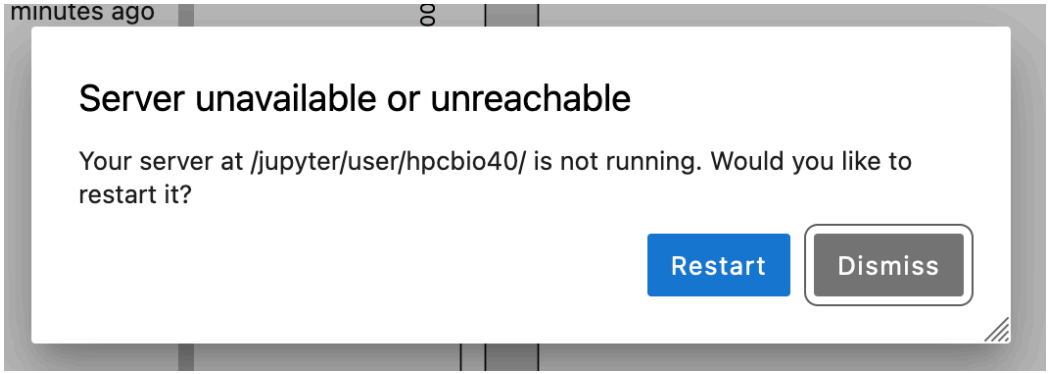
Now open the R notebook, RNA-Seq_R_stats_2025.ipynb, to begin your analysis. Everything will be performed and explained within this notebook. We'll come back to these slides when we need to end this lab session.

Double click



Step 5: Ending Jupyter Hub

- If you set the runtime of your Jupyter Hub instance to be for a shorter time than the workshop day, you may see a message like this:



- Before this happens make sure to save the R notebook if you made any changes to it.
- If you still have time left on your Jupyter hub session, you can end it early by going to File -> Hub Control Panel. Then click the Stop My Server button. This helps free up resources so that others can use the Biocluster.

Key Insights

- Arguably the most time-consuming and important step is getting all your data into R and in the right format and order. This ensures that everything going forward will be accurate and valid.
- Salmon transcript counts should be summed to the gene-level to improve accuracy.
- You can preview your raw data, but the counts need to be filtered and normalized before performing differential gene expression (DGE) analysis.
- An MDS plot can allow us to see if our samples are clustering by our treatment group or other metadata.
- We can create our own unique contrasts to test for our DGE analysis, and pull out detailed information on our up and down-regulated genes.
- This is a short snippet of what you can do. Other analyses include venn diagrams, heatmaps, WGCNA, annotation, and more!