**Eun-Young and JM RNAseq 2024**

**Library Prep**

These libraries were made with the NEBNext® Ultra II Directional RNA Library Prep Kit for Illumina with NEBNext® Globin & rRNA Depletion Kit (Human/Mouse/Rat). It uses the “dUTP method”.

**Pipeline**

Fastqc > STAR\_2.7.8a > Samtools(1.21)Index > htseq\_2.0.7 > Deseq2

**Mouse genome –**

**Genome** *- /data/NHGRIVenditti/Payal/Genomes/Mouse/GRCm39/STAR\_2.7.8a/*

GTF *- /data/NHGRIVenditti/Payal/Genomes/Mouse/GRCm39/Mus\_musculus.GRCm39.104.gtf*

**Long Term Storage**

1. Raw Files in Biowulf (Fastq, BAM, htseq\_counts) - /data/NHGRIVenditti/Payal/EunYoung\_2024\_RNAseq/Part1\_JM
2. Deseq2\_results (Teams/Sharepoint) - [Part1\_JM](https://nih.sharepoint.com/:f:/r/sites/NHGRI-VendittiGroup/Shared%20Documents/General/2024_EunYoung_RNAseq/Part1_JM?csf=1&web=1&e=LZN9Ol)

**Comparisons**

* **Untreated**
* Wild Type(Group A) vs Heterozygote(Group B) vs Mutant (Group C) : Adolescent (33-41 DOL)
  + Wild Type(Group A) vs Heterozygote(Group B)
  + Wild Type(Group A) vs Mutant (Group C)
  + Heterozygote(Group B) vs Mutant (Group C)
* Heterozygote(Group D) vs Mutant (Group E) : Adult (150-155 DOL)
* Adolescent (33-41 DOL) vs Adult (150-155 DOL)
  + - Heterozygote(Group B) vs Heterozygote(Group D)
    - Mutant (Group C) vs Mutant (Group E)
* **Untreated vs Treated**

* Untreated (Group C) vs 1E11 @ neonate (Group F)
* Untreated (Group C) vs 5E13 @ DOL10 (Group G)
* Untreated (Group C) vs 5E13 @ DOL30 (Group H)
* Untreated (Group C) vs 1E13 @ DOL30 +HPD (Group I)
* Mutant (Group E) vs 5E13 @ DOL30 (Group H)
* Mutant (Group E) vs 1E13 @ DOL30 +HPD (Group I)

Each Deseq2 folder in Teams contain the following files:

1. The .csv file is the main differential gene expression result file. These are all the genes that were detected with p-adjusted value of 0.1. The column named **“padj**” can be filtered down to more optimal values Eg: 0.5. Also, the column “**log2 FoldChange”** can be filtered and sorted according to preference. Ascending will sort the values with downregulated genes first. Descending will sort the values with the upregulated genes first.

The columns are :

* 1. **ensembl\_gene\_id** of the genes
  2. baseMean - Is the average of the normalized count values, dividing by size factors, taken over all samples.
  3. **log2 FoldChange (LFC)** – It tells us how much the gene’s expression seems to have changed due to treatment in comparison to control. This value is reported on a logarithmic scale to base 2: for example, a log2 fold change of 1.5 means that the gene’s expression is increased by a multiplicative factor of 2\*1.5 ≈ 2.82.

Here the base group is control. So log foldchange is mutants over control. That means if a gene is positive, it means that gene is upregulated in mutants over control/other groups. Again, if a gene is negative, that means the gene is downregulated in mutants with respect to controls/other groups.

* 1. log2 FoldChange Standard Error ( lfcSE) - the standard error estimate for the log2 fold change estimate
  2. **pvalue -** Indicates whether the observed difference between treatment and control is significantly different.
  3. **padj** – Multiple testing correction values using Benjamini Hochberg method
  4. **mgi\_symbol** – Gene symbols
  5. **external\_gene\_name** – Gene symbols (same as mgi\_symbol)
  6. description – Description of genes

1. **PCA plot** – It shows how the samples are grouped in general between the two conditions
2. **Volcano plot** – Visulaization of top differentially expressed genes above foldchange 1 and above -log10(pvalue= 10e-5).
3. **Heatmap** – Top 100 differentially expressed genes sorted by padj values