Soodabeh Ghaffari

There are two analysis paradigms:

1. Quantify against a genome
2. Classify against a transcriptome

Steps for both “control” state and “perturbed” state:

1. Produce sequencing data from a transcriptome
2. Match sequencing reads to the genome or the transcriptome.
3. Count how many reads align to a region

**Understand the reference data**

Diagram, timeline

Description automatically generated with medium confidence

From the command above, we realize that the reference genome has a single chromosome of size of 128,765 bp. There are 92 transcripts. The total transcript size is 92\*899=82708.

The scripts are attached to the homework.

A few lines of the results table (file b.txt attached)

Graphical user interface

Description automatically generated with low confidence

The results show:

1. The genes starting with A such as AAA, ABA and so on are all up-regulated 4 fold in the Excited state compared to the Bored state.
2. The genes starting with B such BAB, BBB, and so on stay the same in both the Bored and Excited states.
3. The genes starting C such as CAC, CBC, and so on are all down-regulated 0.67 foled in the Excited state compared to the Bored state.

**Issues that I have encountered:**

The file “counts.txt” does not have the value of p-values.

deseq2-results.csv and the corresponding deseq2-heatmap.pdf have zero size, but edger-heatmap.pdf worked and its file is attached to the homework. The heatmap shows how consistent the inter-replicate and intra-sample variations are.

For classification part, I ran the code (attached to homework), but the following error I faced:

Text

Description automatically generated