**Bioinformatic Skillset Evaluation**

**Background:**

We have RNA-seq expression profiles from 8 mouse samples (provided in **RNAseqExpressionProfile.txt**), comprising:

4 samples from control mice: Control1, Control2, Control3, and Control4.

4 samples from knockout mice: KO1, KO2, KO3, and KO4.

**Questions:**

1. Could you develop your scripts to identify differentially expressed genes? (Consider genes with a log2 Fold Change >= 1 and an adjusted P value <= 0.05 as differentially expressed).

Note: Compute the log2 FC and adjusted P value using your codes rather than relying on published software. I would suggest using Wilcoxon test to measure P value.

Answer:

A PCA was first performed on the dataset, indicating that Control2 is abnormal. Control2 was then removed for downstream analysis as directed by Dr. Wang. Extremely low expressed genes were filtered out with a threshold of sum 100, based on the histogram of log-transformed sum of counts.

A graph with numbers and dots

Description automatically generatedA graph of a number of numbers

Description automatically generated

The filtered expression data was then normalize the with size factors calculated from geometric mean for each gene, mean\_normalized\_counts > 10 and variance\_normalized\_counts > 5 were used for the final filtering. log2 FC and adjusted P value were calculated with my code using the Wilcoxon test as suggested, Adjust p-values using the Benjamini-Hochberg method. However, the lowest adjusted P value is 0.2714. No significantly DE genes can be identified with the original requirement. I switched to use P value to identify the significantly DE genes as directed by Dr. Wang. In total, 106 genes were identified as differentially expressed genes, a diff\_genes\_data.csv was generated, with their log2FC and P value listed.

1. Can you generate a volcano plot to illustrate the differentially expressed genes and label the top 10 genes (5 up-regulated, 5 down-regulated) in the plot? For the volcano plot, please use the red color for up-regulated genes, blue color for down-regulated genes, gray color for not significant genes. As for the top 10 genes, please ranked by the adjusted P values.

Answer:

Volcano plot was generated as requested. Please be noted that, it is the P value instead of adjusted P value was used in the analysis, the distribution of -log10 P value present a more discrete distribution. I attached here a plot by using adjusted P value from the result generated from another pipeline.

A screen shot of a graph

Description automatically generatedA graph of red and blue dots

Description automatically generated

1. Can you create an expression heatmap for the top 10 up-regulated and 10 down-regulated genes? Please select these top 20 genes ranked by log2 FC.

Answer:

The heatmap for the top 10 up-regulated and 10 down-regulated genes shown here as requested. The top genes were ranked by log2FC. The complete method was used for clustering.

A screenshot of a computer

Description automatically generatedA chart with different colored squares

Description automatically generated

1. Please extract all Uroplakin genes (GeneID starting with Upk) and examine their correlated expressions. Generate a correlation matrix of these Upk gene expression signatures from the recurrences.

Answer: The expression profile of 6 Upk genes was extracted for examining their correlated expressions. A correlation matrix for these Upk genes (correlation\_matrixs.csv) and plot have been generated as requested.

A red and blue squares

Description automatically generated