Assignment 2 Report – CGS4144

A graph of a number of lines

Description automatically generated

**Density of Per-Gene Median Expression (log2)**: The plot shows the density distribution of the median gene expression values. The data was calculated from gene expression values across multiple samples in the dataset.

**Key Observations:**

* Peak near Zero Expression:

There is a significant spike in the distribution near 0 on the log2 scale, indicating that a large portion of the genes have low to very low median expression across the samples. This could represent genes that are either not expressed or have very low expression in the studied conditions.

* Secondary Peak around 1-2 (log2 scale):

A smaller peak appears around log2 values of 1-2, indicating a subset of genes with moderate expression levels.

* Long Tail of High Expression:

suggests that a smaller fraction of genes exhibit high levels of expression, with a gradual decrease as the expression levels increase.

A screen shot of a grid

Description automatically generated

**Heatmap (significant differentially expressed genes)**: The plot demonstrates the correlation between male and female differentially expressed genes in the set

**Key Observations:**

* The red areas in the top left and bottom right quadrants of the heatmap represent the genes with the highest cross correlation between male and female samples.
  + You can visibly see bigger clusters of where the genes are highly expressed for one sex compared to the other (e.g. red for males at bottom right)
* We have for each row represent a gene (Ensembl IDs) and each column a sample.
* Heatmap helps show us normalization from low blue to high red, and between sexes
* We see multiple clusters of one or multiple color on different regions, showing the differences of expression values and their profiles

A graph with red and blue dots

Description automatically generated

**PCA Plot (female vs male expression of low grade glioma):** This plot shows the principle components from a PCA decomposition of the set, and color codes them based on sex.

**Key Observations:**

* Inverse correlation: seeing as many of the datapoints are directionally opposite through the origin, it is likely that there is some sort of inverse correlation between the two groups. Additionally, it looks like each subgroup has some correlation with it’s other members based on the grouping relative to origin.

A graph of a volcano plot

Description automatically generated

**Volcano Plot (gene differential expression):** Shows gene expression differences between different expression sets. We used

* Data has their outliers within the volcano plot (with genes ending in: ENSG…9807 & 9824)
* We also see a common value of our p-value and NS plots to be in the same threshold: between -1, and 1
* X as the log\_2 to see the change within our genes and -log\_10(p values) to determine larger values (more significant)

A white background with orange dots

Description automatically generated

**Enriched Pathways Volcano Plot:** This volcano plot shows the significance of multiple enriched pathways depending on their p-values. We have the X with an enrichment score and our Y as -log10(*p-value*) where we can see larger values (significant)

* We have extremely smaller p-values shown meaning more enrichment in our data
* You can see the path of interest with lower p-values (and plot a line through)

A graph with a green line and a line

Description automatically generated  
**GSEA Enrichment Score:** This plot shows the enrichment score of the pathway: HALLMARK\_FATTY\_ACID\_METABOLISM. The black line through the pathway marks the positions of the genes and the green line through the plot is the enrichment score through the pathway.

* The results suggest that the fatty acids is downregulated with the common pattern of a decrease
* Negative enrichment score
* So fatty acids aren’t in the condition of interest

A graph with red and blue dots

Description automatically generated

**GO Enrichment Dot Plot** This plot shows the number of genes involved in each process from the set. The larger the dot, the more genes are involved with the corresponding y axis process set. The warmth of each dot represents statistical significance from a p value adjustment, with warmer p values representing more significant gene clusters.

**Key Observations:**

* Antigen Receptor Gene Clusters Prominent

From this plot, we can see that many of the genes involved in immune response and antigen receptor pathways show significance relative to the presence of low-grade gliomas in our population. Furthermore, they have the highest gene ratio, furthering this hypothesis by indicating that their gene clusters are more dense.

A graph of a number of dots

Description automatically generated with medium confidence

**KEGG Enrichment Dot Plot:** This plot shows the frequency of genes involved in each of the processes outlined on the y axis. The larger the dots, the more genes from the set are involved in the expression. Red dots are more likely to be significant. The x axis represents the ratio of genes in an expression cluster to genes in the entire set.

**Key Observations:**

* Th1 and Th2 Cell Differentiation: It could potentially be significant that these Th1 and Th2 regulatory genes have correlation with sample low grade gliomas. Th1 is important in modulating anti-tumor activity in humans.

A graph with dots and numbers

Description automatically generated

**t-SNE Plot – Gene Expression Data:** The plot shows the density and correlation of similarities between genes. The female genes (coded with purple) seem to trend slightly more clustered, but both groups do have defined boundaries.

**Key Observations:**

* Potential Negative Correlation

Based on the angles of the female and male data points and clustering, it seems like there are defined differences between the two, and that they may have a negative correlation across the two subsets.

A graph with dots and numbers

Description automatically generated

**UMAP Plot – Gene Expression Data:** The plot also shows the density and correlation of similarities between genes using the UMAP paradigm. The female genes (coded with purple) seem to trend slightly more clustered, but both groups do have defined boundaries. While we hoped the UMAP would give us a more defined clustering, it seems relatively similar with many of the same features.

**Key Observations:**

* Potential Negative Correlation

Similarly to the t-SNE plot, the UMAP plot suggests that there could be a negative correlation across the two subsets.

A graph with numbers and a line

Description automatically generated

**Key Features:**

* **X-axis**: Likely represents log2 fold change (LFC) or another test statistic. Values range from around -4 to +6, with most genes centered around 0.
* **Y-axis**: Represents the frequency or count of genes. The values on the Y-axis suggest that over 14,000 genes fall very close to 0 in terms of their fold change or statistic.
* **Legend**: The two colors differentiate "Upregulated" and "Downregulated" genes:
  + **Upregulated genes**: Likely have positive LFC values.
  + **Downregulated genes**: Likely have negative LFC values.

**Observations:**

* **Concentration at 0**: Both upregulated and downregulated genes are tightly concentrated around a log2 fold change of 0, implying that for most genes, there may be little to no change in expression, or the statistical values hover around the null hypothesis.
* **Limited Variation**: The bars show no clear distribution across the X-axis, which could indicate that only a small fraction of genes is significantly differentially expressed, or that the test failed to capture wide variations between groups.