# Summary Report: Differential Gene Expression Analysis

## Project Goal

To identify differentially expressed genes between metastasis and non-metastasis samples using RNA-Seq data from the GEO dataset GSE64850, and generate a filtered list of exploratory genes.

## Summary of Activities

Installed required Python libraries such as GEOparse, pandas, scipy, and statsmodels.

* Defined and loaded sample files (metastasis and non-metastasis groups).
* Merged gene expression data from multiple samples.
* Performed independent t-tests to find statistically significant genes.
* Calculated log2 fold changes for expression levels between the two groups.
* Filtered results based on p-value (< 0.05) and |log2FC| (> 0.5).
* Saved the final list of exploratory genes to 'exploratory\_gene\_list.csv'.

## Step-by-Step Task List

Set up environment and import necessary libraries.

* Define paths to expression files for metastasis and non-metastasis groups.
* Extract and merge gene expression data on 'Gene' column.
* Conduct statistical analysis using t-tests and compute log2 fold change.
* Filter the gene list based on statistical significance and effect size.
* Sort and export the top exploratory genes to a CSV file.

## Top Exploratory Genes

|  |  |  |
| --- | --- | --- |
| Gene | log2FC | p-value |
| RERG | -2.437042 | 7e-05 |
| SLC25A25 | 0.8655 | 0.000178 |
| RNF13 | -1.002352 | 0.000404 |
| SORBS2 | -3.522439 | 0.000413 |
| RABL5 | -1.899339 | 0.000417 |
| DUSP14 | -1.704473 | 0.000427 |
| ZBTB8A | -1.528991 | 0.000493 |
| MAML3 | -1.678754 | 0.000513 |
| NDST1 | -1.300696 | 0.000519 |
| XPO6 | 3.445299 | 0.00055 |

**Key Learnings**

* Developed hands-on experience in RNA-Seq data handling and preprocessing.
* Understood the statistical considerations (p-values, fold change thresholds) in gene expression analysis.
* Learned how to extract biologically meaningful signals from high-dimensional data.

**Ideas for Future Work**

* Perform multiple testing correction for more robust significance.
* Visualize results with volcano plots and heatmaps.
* Annotate top genes using databases like Ensembl or Gene Ontology.
* Explore machine learning models for classification based on expression profiles.