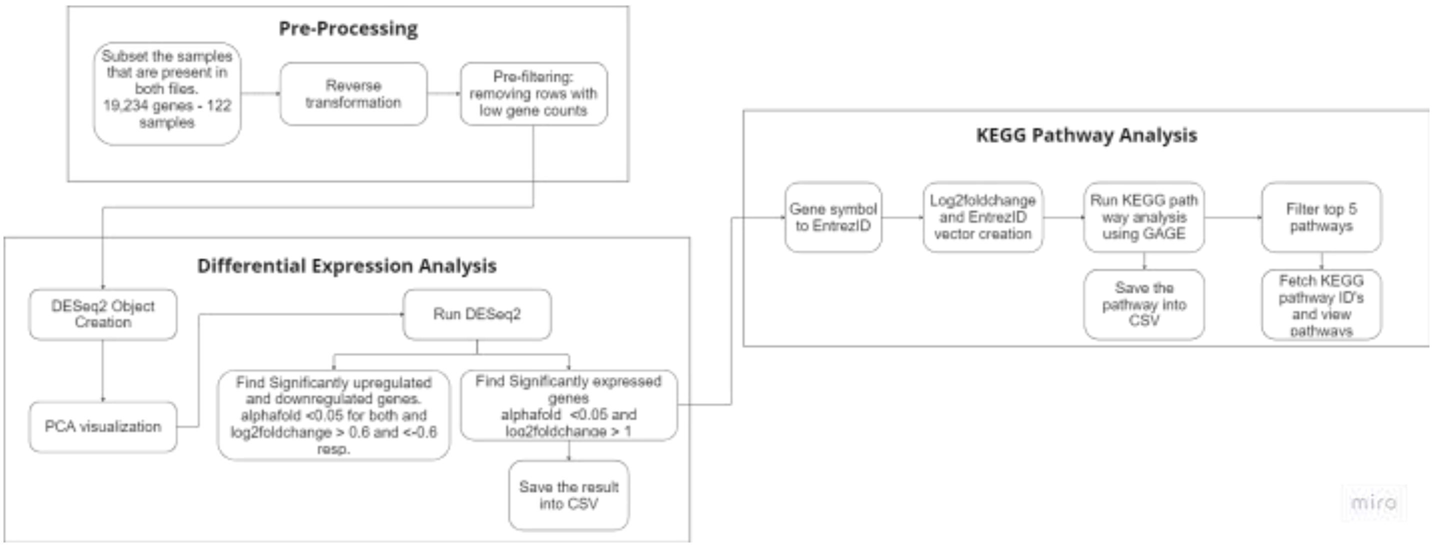
Differential Gene Expression and KEGG Pathway Enrichment Analysis

Introduction:

The objective of this analysis is to perform a comprehensive investigation of differential gene expression and pathway enrichment between "Tumor" and "Normal" samples. Leveraging the DESeq2 package [1], differential gene expression analysis is conducted. Subsequently, the GAGE package [3] is employed to perform pathway enrichment analysis using the KEGG pathway database. The gene expression data sourced from "TCGA-BRCA.htseq_counts_gene_name.tsv" is in log2-transformed RNAseq count format, with corresponding sample labels available in the "TCGA-BRCA.pheno.tsv" file.

Materials and Methods:


Pre-Processing
Subset the samples that are present in both files. 19,234 genes - 122 samples
Reverse transformation
Pre-filtering: removing rows with low gene counts
Differential Expression Analysis
DESeq2 Object Creation
PCA visualization
Run DESeq2
Find Significantly upregulated and downregulated genes. alphafold <0.05 for both and log2foldchange > 0.6 and <-0.6 resp.
Find Significantly expressed genes alphafold <0.05 and log2foldchange > 1
Save the result into CSV
KEGG Pathway Analysis
Gene symbol to EntrezID
Log2foldchange and EntrezID vector creation
Run KEGG path way analysis using GAGE
Save the pathway into CSV
Filter top 5 pathways
Fetch KEGG pathway ID's and view pathways
miro

Data Preprocessing:

Initial data preprocessing involves the harmonization of gene expression data from "TCGA-BRCA.htseq_counts_gene_name.tsv" and sample labels from "TCGA-BRCA.pheno.tsv". A stringent criterion is applied, selecting samples that are shared between both datasets to ensure data integrity. Notably, count matrix values undergo reverse transformation, a pivotal step to enable accurate differential expression analysis using the DESeq2 package. Furthermore, pre-filtering is executed to ensure robustness in subsequent analyses.

Differential Gene Expression Analysis:

The DESeq2 package is harnessed to facilitate the identification of genes exhibiting differential expression between "Tumor" and "Normal" samples. The gene count data undergoes a sequence of key steps, encompassing normalization, dispersion estimation, and rigorous statistical analysis. Genes demonstrating a fold change greater than 1 and an adjusted p-value below 0.05 are considered differentially expressed, serving as indicators of potential biological significance.

Pathway Enrichment Analysis:

In parallel, pathway enrichment analysis is conducted employing the esteemed KEGG pathway database. The differential expression profiles are mapped onto KEGG pathways, enabling the identification of pathways exhibiting enrichment. This step provides a holistic understanding of the biological processes potentially perturbed in the context of the analyzed samples.

Results and Discussion:

During the preprocessing stage, a refined dataset comprising 19,234 genes and 122 samples was generated. Remarkably, all genes exhibited a minimum count of 10, thereby obviating the need for further filtration. Subsequently, by employing the criterion "which(sig_genes$log2FoldChange > 0.6)," a subset of genes exhibiting a log2 fold change greater than 0.6 was extracted, signifying upregulated genes. Correspondingly, the criterion "which(sig_genes$log2FoldChange < -0.6)" facilitated the selection of genes with a log2 fold change below -0.6, indicative of downregulated genes. However, it is essential to underscore that the chosen log2 fold change threshold is contingent upon the biological context of interest. In this investigation, a threshold of 0.6 was adopted, translating to an approximate 1.5-fold change in gene expression. Consequently, the analysis pinpointed the upregulation of 3,785 genes and the downregulation of 2,924 genes within the tumor samples. Furthermore, for the broader cohort of differentially expressed genes, a threshold of 1 was established, approximating a 2-fold change in gene expression, unveiled a total of 3,935 differentially expressed genes discernible between "Tumor" and

"Normal" samples. Detailed information concerning these genes, including their statistical characteristics, is encapsulated within the "diff_result_Kiran_Franklin.csv" file.

Pathway Enrichment Results:

Pathway enrichment analysis highlighted several significantly enriched KEGG pathways. 3935 genes which were differentially expressed were used for pathway enrichment analysis. The top enriched pathways, along with their adjusted p-values, are detailed in Table 3 of the attached report. For full details report of all pathways please refer pathway_result_Kiran_Franklin.csv. The pathways hsa04110 Cell cycle and hsa04114 Oocyte meiosis show significant enrichment in the pathways with geometric p-value of 0.000107 and 0.00686 respectively, for additional evidence stats mean i.e log2foldchanges is also high with 3.9689 and 2.551006 respectively. Full pathway image for this can be found in hsa04110.pathview.png and hsa04114.pathview.png

Table 1: Stats for top two enriched pathways

	greater.p.geomean	**greater.stat.mean**	**greater.p.val**	**greater.q.val**	**greater.set.size**	**greater.exp1**
hsa04110 Cell cycle	0.000107537	3.96897	0.000108	0.00828	36	0.000108
hsa04114 Oocyte meiosis	0.006862916	2.551006	0.006863	0.264222	27	0.006863

Conclusion and Future Perspective:

In conclusion, this analysis successfully identified differentially expressed genes between "Tumor" and "Normal" samples using Deseq2. The pathway enrichment analysis provided insights into the biological pathways that are potentially dysregulated in tumor samples. These findings contribute to a deeper understanding of the molecular mechanisms underlying the disease.

Future work could involve validating the identified differentially expressed genes and pathways through experimental techniques, improve the strategy to find the biologically significant differentially expressed genes. Additionally, exploring the functional significance of the differentially expressed genes in the context of cancer biology would be valuable.

References:

1. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, 550. doi:10.1186/s13059-014-0550-8.
2. Luo, Weijun, Friedman, Michael, Shedden, Kerby, Hankenson, Kurt, Woolf, Peter (2009). “GAGE: generally applicable gene set enrichment for pathway analysis.” *BMC Bioinformatics*, **10**, 161.

Code Availability:

1. Diff_analysis.ipynb
   Note: All graphs and matrices can be found in the notebooks.