Transcriptomic Analysis of Basal vs Luminal Subtypes in TCGA-BRCA Dataset

-Renuka Reddy Namala

Objective

The aim of this study is to identify differentially expressed genes (DEGs) between Basal and Luminal A subtypes of breast cancer using RNA-seq data from The Cancer Genome Atlas (TCGA-BRCA). This analysis helps uncover molecular signatures associated with these clinically distinct subtypes and highlights subtype-specific immune markers.

Dataset Description

- **Source**: TCGA-BRCA project from GDC
- **Sample Type**: Primary Tumor
- **Subtypes Analyzed**: Basal and Luminal A (LumA)
- **Gene Expression Workflow**: STAR - Counts
- **Number of Samples**: 768 (197 Basal, 571 Luminal A)

Tools & Libraries Used

- **R version**: 4.3.2
- **Packages**: TCGAbiolinks, DESeq2, biomaRt, EnhancedVolcano, SummarizedExperiment

Methodology

1. Data Acquisition

- Queried and downloaded STAR-counts from TCGA-BRCA using TCGAbiolinks::GDCquery()
- Prepared SummarizedExperiment object using GDCprepare()

2. Sample Filtering

- Filtered metadata to retain only samples annotated as Basal or Luminal A using the paper_BRCA_Subtype_PAM50 column
- Matched expression count matrix with filtered metadata

3. Differential Expression Analysis

- Constructed DESeq2 dataset using filtered counts and subtype metadata
- DESeq2 was run with design formula: ~ subtype_group
- Extracted DEGs using: results(dds, contrast = c("subtype_group", "Basal", "Luminal"))
- Filtered DEGs using adjusted p-value < 0.05

4. Gene Annotation

- Removed Ensembl version numbers from gene IDs
- Annotated Ensembl IDs with HGNC symbols using biomaRt
- Merged annotations with DEG results

5. Immune Gene Subset

- Extracted log2FoldChange and padj values for immune markers: CD8A, GZMB, CXCL10, PDCD1, CD274, IFNG, CTLA4, LAG3

Results Summary

- Total genes tested: 57,938
- Significant DEGs (padj < 0.05): 32,060
- Upregulated in Basal: 22,152
- Downregulated in Basal: 11,779

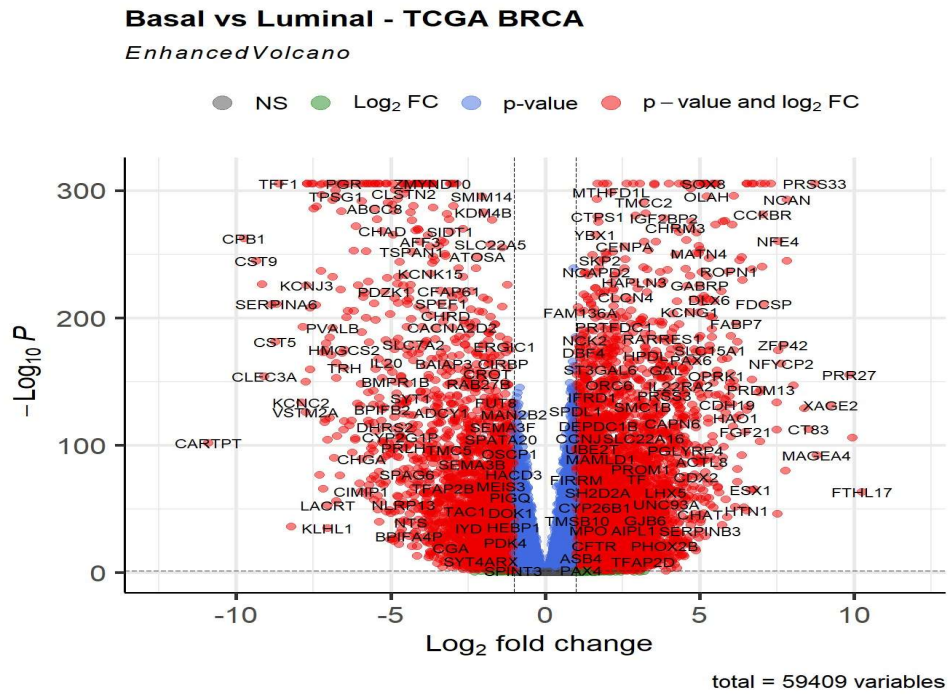
6. Key Differentially Expressed Immune Genes

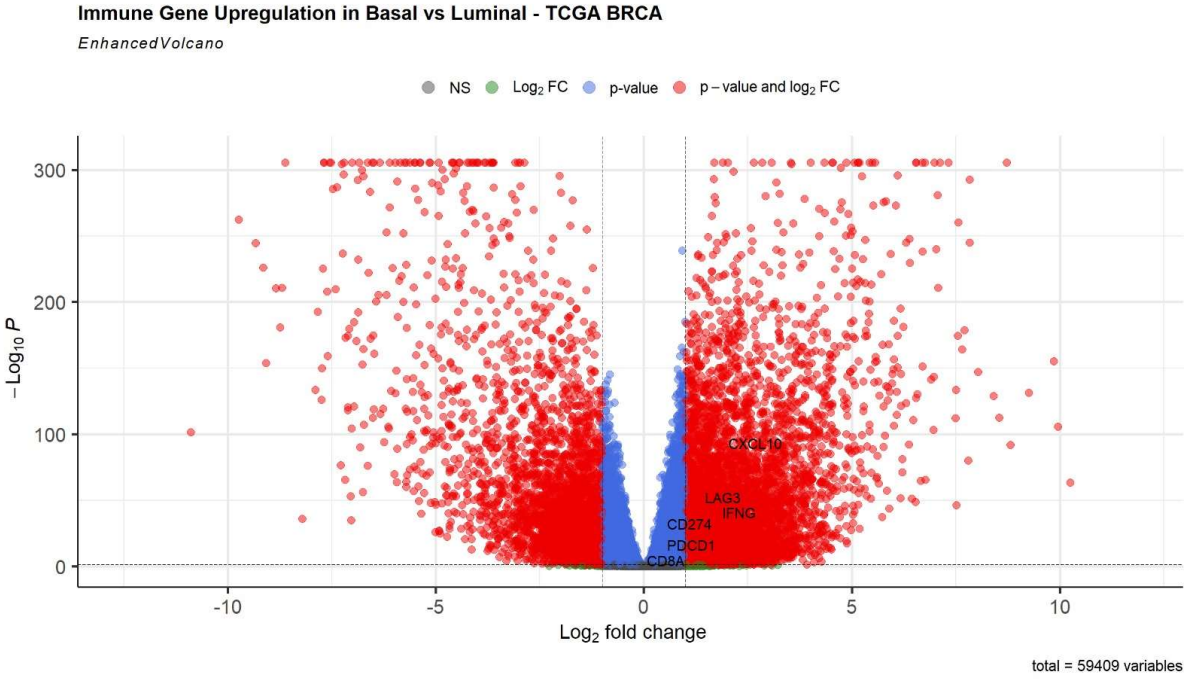
Ensembl ID	**HGNC Symbol**	**log2 Fold Change (log2FC)**	**Adjusted p-value (padj)**
ENSG00000004838	LAG3	1.89	5.04e-53
ENSG00000100453	GZMB	2.18	1.05e-50
ENSG00000111537	IFNG	2.29	2.24e-41
ENSG00000120217	CD274	1.10	6.79e-33
ENSG00000153563	CD8A	0.52	2.52e-05

Visualizations to Include

1. Volcano Plot of All DEGs

- Generated using EnhancedVolcano()
- X-axis: log2FoldChange
- Y-axis: -log10(adjusted p-value)
- Thresholds: padj < 0.05, FC > 1





2. Immune Marker Volcano Plot

- Highlighted genes: **CD8A, CXCL10, PDCD1, CD274, GZMB, CTLA4, IFNG, LAG3**
- Custom labels and points sized for visibility

Conclusion

This analysis revealed a large number of differentially expressed genes between Basal and Luminal A breast cancer subtypes. Many immune-related genes (e.g., CXCL10, PDCD1, CD8A) are significantly upregulated in Basal subtype, suggesting a distinct immune microenvironment. These results may aid in subtype-specific therapeutic targeting and biomarker discovery.

Files Included in GitHub Repository

- scripts/TCGA_DEA_pipeline.R: Full analysis code
- data/Basal_vs_Luminal_Annotated_DEGs.csv: Final annotated DEG table
- plots/volcano_all_genes.png: Full volcano plot
- plots/volcano_immune_genes.png: Immune-specific plot
- README.md:project overview and instructions
- report/TCGA_Basal_vs_Luminal_Report.pdf: Full report