

COMPUTATIONAL BIOLOGY GROUP
INSTITUTE FOR BIOMEDICAL RESEARCH

Differential Gene Expression Analysis in Breast Cancer Subtypes

Technical Report

Project ID: TCGA-BRCA-2024-001

Lead Analyst: Dr. Jane Doe

Collaborators: Dr. John Smith, Dr. Alice Johnson

Analysis Date: January 15, 2024

Data Source: TCGA (The Cancer Genome Atlas)

Pipeline Version: v2.3.1

Software: R 4.3.2, DESeq2 1.40.0, clusterProfiler 4.8.0

Repository: <https://github.com/example/brca-analysis>

Executive Summary

This report presents a comprehensive differential gene expression analysis comparing four molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, and Basal-like) using RNA-seq data from the TCGA-BRCA cohort.

Key Finding

- **2,847 differentially expressed genes** identified (adj. $p < 0.01$, $|\log_2 \text{FC}| > 1$)
- **ESR1** and **ERBB2** expression patterns strongly associate with clinical subtypes ($R^2 = 0.78$)
- **Basal-like tumors** show enrichment for cell cycle and DNA repair pathways (FDR $< 10^{-15}$)
- **Machine learning classifier** achieves 94.2% accuracy (5-fold CV) distinguishing subtypes

Contents

Executive Summary	1
1 Introduction	2
1.1 Background	2
1.2 Objectives	2
1.3 Data Source	2
2 Methods	2
2.1 Data Preprocessing	2
2.2 Differential Expression Analysis	3
2.3 Pathway Enrichment Analysis	3
2.4 Machine Learning Classification	4
3 Results	4
3.1 Sample Distribution	4
3.2 Differential Expression Results	4
3.3 Top Differentially Expressed Genes	4
3.4 Pathway Enrichment	5
3.4.1 Basal-like Enriched Pathways	5
3.4.2 KEGG Pathway Analysis	5
3.5 Classification Performance	5
4 Discussion	5
4.1 Biological Interpretation	5
4.2 Clinical Implications	6
4.3 Limitations	6
5 Conclusions	7
5.1 Future Directions	7
A Supplementary Tables	7

B Session Information

7

1. Introduction

1.1 Background

Breast cancer is a heterogeneous disease with distinct molecular subtypes defined by gene expression profiles [1, 2]. The PAM50 classifier identifies four intrinsic subtypes based on 50 genes:

1. **Luminal A:** ER+/HER2-, low proliferation, best prognosis
2. **Luminal B:** ER+/HER2- or ER+/HER2+, high proliferation
3. **HER2-enriched:** HER2+, ER-/PR-
4. **Basal-like:** ER-/PR-/HER2-, worst prognosis

Understanding the transcriptomic differences between subtypes is essential for developing targeted therapies and prognostic biomarkers.

1.2 Objectives

1. Identify differentially expressed genes (DEGs) between breast cancer subtypes
2. Perform pathway enrichment analysis to characterize biological processes
3. Develop a predictive model for subtype classification
4. Validate findings against independent datasets

1.3 Data Source

Dataset Summary

Project:	TCGA-BRCA (Breast Invasive Carcinoma)
Samples:	$n = 1,097$ primary tumors
Platform:	Illumina HiSeq RNA-seq
Normalization:	FPKM → TPM
Genes:	20,531 protein-coding genes
Clinical data:	Age, stage, subtype, survival

2. Methods

2.1 Data Preprocessing

Raw RNA-seq counts were obtained from the GDC Data Portal. Preprocessing steps included:

Listing 1: Data preprocessing pipeline in R

```

1 library(DESeq2)
2 library(edgeR)
3
4 # Load count matrix and metadata
5 counts <- read.csv("TCGA_BRCA_counts.csv", row.names=1)
6 metadata <- read.csv("TCGA_BRCA_clinical.csv")
7
8 # Filter low-expression genes (CPM > 1 in at least 10% of samples)
9 keep <- rowSums(cpm(counts) > 1) >= 0.1 * ncol(counts)

```

```

10 counts_filtered <- counts[keep, ]
11 # Retained: 15,847 genes
12
13 # Create DESeq2 object
14 dds <- DESeqDataSetFromMatrix(
15   countData = counts_filtered,
16   colData = metadata,
17   design = ~ subtype
18 )
19
20 # Variance stabilizing transformation for visualization
21 vsd <- vst(dds, blind = FALSE)

```

2.2 Differential Expression Analysis

Pairwise comparisons between subtypes were performed using DESeq2 with the following thresholds:

- Adjusted p -value (Benjamini-Hochberg FDR) < 0.01
- $|\log_2 \text{fold change}| > 1.0$

The statistical model accounts for the negative binomial distribution of count data:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i), \quad (1)$$

where K_{ij} is the count for gene i in sample j , μ_{ij} is the expected value, and α_i is the dispersion parameter.

2.3 Pathway Enrichment Analysis

Gene Ontology (GO) and KEGG pathway enrichment were performed using clusterProfiler [4]:

Listing 2: Pathway enrichment analysis

```

1 library(clusterProfiler)
2 library(org.Hs.eg.db)
3
4 # Convert gene symbols to Entrez IDs
5 gene_list <- bitr(DEGs$gene_symbol,
6                     fromType = "SYMBOL",
7                     toType = "ENTREZID",
8                     OrgDb = org.Hs.eg.db)
9
10 # GO Biological Process enrichment
11 go_bp <- enrichGO(gene = gene_list$ENTREZID,
12                      OrgDb = org.Hs.eg.db,
13                      ont = "BP",
14                      pAdjustMethod = "BH",
15                      pvalueCutoff = 0.05,
16                      qvalueCutoff = 0.05)
17
18 # KEGG pathway enrichment
19 kegg <- enrichKEGG(gene = gene_list$ENTREZID,
20                      organism = 'hsa',
21                      pvalueCutoff = 0.05)

```

2.4 Machine Learning Classification

A random forest classifier was trained to predict subtype from gene expression:

- **Features:** Top 500 DEGs (by variance)
- **Model:** Random Forest (500 trees, mtry = 22)
- **Validation:** Stratified 5-fold cross-validation
- **Metrics:** Accuracy, precision, recall, F1-score

3. Results

3.1 Sample Distribution

Table 1 shows the distribution of samples across subtypes and clinical stages.

Table 1: Sample distribution by molecular subtype and tumor stage.

Subtype	Stage				Total
	I	II	III	IV	
Luminal A	98	187	72	8	365 (33.3%)
Luminal B	61	124	58	5	248 (22.6%)
HER2-enriched	29	68	41	3	141 (12.9%)
Basal-like	37	89	55	7	188 (17.1%)
Normal-like	23	41	27	4	95 (8.7%)
Unclassified	18	29	11	2	60 (5.5%)
Total	266	538	264	29	1,097

3.2 Differential Expression Results

A total of **2,847 differentially expressed genes** were identified in at least one pairwise comparison (adj. $p < 0.01$, $|\log_2 \text{FC}| > 1$).

Table 2: Number of differentially expressed genes in pairwise subtype comparisons.

Comparison	Up	Down	Total
Basal vs. Luminal A	1,284	967	2,251
Basal vs. Luminal B	1,031	742	1,773
Basal vs. HER2	687	412	1,099
HER2 vs. Luminal A	534	398	932
HER2 vs. Luminal B	321	287	608
Luminal B vs. Luminal A	187	154	341

3.3 Top Differentially Expressed Genes

Table 3 lists the top 15 genes ranked by adjusted p -value across all comparisons.

Table 3: Top 15 differentially expressed genes (Basal vs. Luminal A comparison).

Gene	Description	$\log_2 FC$	SE	adj. p	Sig.
ESR1	Estrogen receptor 1	-4.82	0.18	$< 10^{-100}$	***
GATA3	GATA binding protein 3	-3.94	0.15	$< 10^{-95}$	***
FOXA1	Forkhead box A1	-3.67	0.14	$< 10^{-87}$	***
KRT5	Keratin 5	+4.21	0.19	$< 10^{-78}$	***
KRT14	Keratin 14	+3.89	0.17	$< 10^{-72}$	***
FOXC1	Forkhead box C1	+3.54	0.16	$< 10^{-65}$	***
MKI67	Marker of proliferation Ki-67	+2.87	0.12	$< 10^{-58}$	***
ERBB2	HER2/neu receptor	+0.42	0.08	3.2×10^{-7}	***
PGR	Progesterone receptor	-3.21	0.14	$< 10^{-55}$	***
CDH1	E-cadherin	-1.54	0.09	$< 10^{-42}$	***
CCNB1	Cyclin B1	+2.34	0.11	$< 10^{-40}$	***
TOP2A	Topoisomerase II alpha	+2.67	0.12	$< 10^{-38}$	***
BRCA1	BRCA1 DNA repair	+1.12	0.07	$< 10^{-28}$	***
TP53	Tumor protein p53	+0.87	0.06	$< 10^{-22}$	***
MYC	MYC proto-oncogene	+1.34	0.08	$< 10^{-19}$	***

*** adj. $p < 0.001$; ** adj. $p < 0.01$; * adj. $p < 0.05$

3.4 Pathway Enrichment

3.4.1 Basal-like Enriched Pathways

Genes upregulated in Basal-like tumors were significantly enriched in:

Table 4: Top GO Biological Process terms enriched in Basal-like tumors.

GO Term	Genes	Fold Enrich.	FDR
Cell cycle (GO:0007049)	187	3.4	$< 10^{-45}$
DNA replication (GO:0006260)	89	4.2	$< 10^{-32}$
Mitotic nuclear division (GO:0140014)	124	3.8	$< 10^{-28}$
DNA repair (GO:0006281)	112	2.9	$< 10^{-21}$
Chromosome segregation (GO:0007059)	67	4.1	$< 10^{-18}$

3.4.2 KEGG Pathway Analysis

3.5 Classification Performance

The random forest classifier achieved excellent performance in distinguishing subtypes:

The top 10 most important features (by Gini importance) were: ESR1, GATA3, FOXA1, KRT5, KRT14, MKI67, FOXC1, PGR, CDH1, and ERBB2.

4. Discussion

4.1 Biological Interpretation

The differential expression results are consistent with known biology of breast cancer subtypes:

- **Luminal tumors:** High expression of ESR1, GATA3, FOXA1 reflects estrogen receptor signaling dependency.

Table 5: Significantly enriched KEGG pathways.

Pathway ID	Pathway Name	Category	Genes	FDR
hsa04110	Cell cycle	Cell growth	78	$< 10^{-25}$
hsa03030	DNA replication	Replication	34	$< 10^{-18}$
hsa04115	p53 signaling pathway	Signal trans.	45	$< 10^{-12}$
hsa03440	Homologous recombination	Repair	28	$< 10^{-10}$
hsa04512	ECM-receptor interaction	Adhesion	52	$< 10^{-8}$

Table 6: Classification performance metrics (5-fold cross-validation).

Subtype	Precision	Recall	F1-Score	Support
Luminal A	0.96	0.94	0.95	365
Luminal B	0.89	0.91	0.90	248
HER2-enriched	0.92	0.88	0.90	141
Basal-like	0.98	0.99	0.98	188
Macro avg	0.94	0.93	0.93	942
Weighted avg	0.94	0.94	0.94	942
Overall Accuracy:				94.2%

- Basal-like tumors:** Upregulation of basal keratins (KRT5, KRT14) and proliferation markers (MKI67) aligns with their aggressive phenotype.
- HER2-enriched:** ERBB2 amplification drives the distinct expression pattern.

The enrichment of cell cycle and DNA repair pathways in Basal-like tumors explains their sensitivity to platinum chemotherapy and PARP inhibitors.

4.2 Clinical Implications

Key Finding

The gene signature identified here could serve as:

1. A diagnostic tool for subtype classification
2. Prognostic markers for survival prediction
3. Targets for subtype-specific therapies

4.3 Limitations

Limitation

- Batch effects:** TCGA data was collected across multiple centers; while we applied batch correction, residual effects may persist.
- Tumor heterogeneity:** Bulk RNA-seq averages expression across cell types; single-cell analysis would provide higher resolution.
- External validation:** Results should be validated in independent cohorts (e.g., METABRIC, GSE96058).

5. Conclusions

1. We identified 2,847 differentially expressed genes distinguishing breast cancer subtypes.
2. Basal-like tumors show strong enrichment for cell cycle and DNA repair pathways.
3. A machine learning classifier achieves 94.2% accuracy in subtype prediction.
4. ESR1, GATA3, and FOXA1 are the strongest discriminators of luminal vs. non-luminal subtypes.

5.1 Future Directions

- Integration with proteomics and metabolomics data
- Single-cell RNA-seq analysis for tumor microenvironment characterization
- Survival analysis incorporating the identified gene signature

References

- [1] C. M. Perou et al., “Molecular portraits of human breast tumours,” *Nature* **406**, 747–752 (2000).
- [2] T. Sørlie et al., “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,” *PNAS* **98**, 10869–10874 (2001).
- [3] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology* **15**, 550 (2014).
- [4] G. Yu et al., “clusterProfiler: an R package for comparing biological themes among gene clusters,” *OMICS* **16**, 284–287 (2012).
- [5] D. C. Koboldt et al., “Comprehensive molecular portraits of human breast tumours,” *Nature* **490**, 61–70 (2012) [TCGA].

A. Supplementary Tables

Full gene lists and pathway enrichment results are available at:

<https://github.com/example/brca-analysis/supplementary>

B. Session Information

Listing 3: R session info for reproducibility

```

1 R version 4.3.2 (2023-10-31)
2 Platform: x86_64-pc-linux-gnu (64-bit)
3
4 attached packages:
5 - DESeq2_1.40.0
6 - clusterProfiler_4.8.0
7 - org.Hs.eg.db_3.17.0
8 - ggplot2_3.4.4
9 - dplyr_1.1.4
10 - randomForest_4.7-1.1

```