BT3041 – Project Report

Identification of Biomarkers for Breast Cancer Through Differential Expression Analysis

Nikshep Grampurohit, Anurag N, Abdul Rahman, Saathvik Sista, Sarath

Abstract

Breast cancer remains a leading cause of mortality among women worldwide, necessitating the identification of reliable biomarkers for early detection, prognosis, and therapeutic targets. This study aims to identify **potential biomarkers** in breast cancer tissue through **differential expression analysis**. We employed high-throughput **RNA sequencing data** of 6 normal blood samples and 8 tumour blood samples from Sequence Read Archive, (Juile Wang et al). Eligibility was limited to female patients with biopsy-proven stage II–III breast cancer without prior therapy. All libraries were sequenced using an Illumina HiSeq 2500 using standard protocol.

To identify differentially expressed genes we used a negative binomial model to identify those that are significantly impacted. Our study revealed 2286 up regulated and 67 downregulated genes in cancer blood as compared to normal blood. Among those that were differentially expressed, novel candidates like COX3, CYTB, ND1, and ND6 emerged as potential biomarkers that were the most **significantly affected** (adjusted p value < e-10). **Pathway enrichment analysis** indicated that these differentially expressed genes are involved in critical cancer-related pathways. The most affected biological processes in the upregulated genes were negative regulation of epithelial cell proliferation and collagen fibril organization which is in line with symptoms of cancer. In those that were downregulated the immune system and antimicrobial response were most affected. This information can help design further studies and therapeutics for breast cancer.

Introduction

Breast cancer is the most frequently diagnosed cancer and a leading cause of cancer-related death among women globally, representing about 25% of all cancer cases and 15% of cancer deaths. Despite advances in detection and treatment, the heterogeneity of breast cancer complicates prognosis and therapy, underscoring the need for reliable biomarkers for diagnosis, prognosis, and therapeutic guidance.

Biomarkers, biological indicators of disease states, are crucial for cancer management, aiding in risk assessment, early detection, prognosis, and therapy monitoring. While traditional biomarkers like estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (et al Michael J Duffy, 2015) have improved breast cancer management, additional biomarkers are essential to address the complexity of the disease.

High-throughput technologies such as RNA sequencing (RNA-seq) have revolutionized biomarker discovery by enabling comprehensive gene expression profiling. Differential expression analysis

compares gene expression levels between different conditions, making it a powerful tool for identifying potential biomarkers in cancers, including breast cancer.

In this study, we identified potential biomarkers in breast cancer tissue through differential expression analysis and functional annotation. Using RNA-seq, we profiled the transcriptomes of breast cancer tissues and adjacent normal tissues. Bioinformatic analyses, including normalization, differential expression analysis, and functional analysis, were conducted to identify and understand the roles of differentially expressed genes (DEGs) in breast cancer. Our study identified both known and novel biomarkers, enhancing our understanding of breast cancer biology.

Methodology

Sample Collection, RNA Sequencing, and Data Processing

We used data from the paper - *Lang JE, Ring A, Porras T, Kaur P, Forte VA, Mineyev N, Tripathy D, Press MF, Campo D. RNA-Seq of Circulating Tumor Cells in Stage II-III Breast Cancer. Ann Surg Oncol. 2018 Aug;25(8):2261-2270. doi: 10.1245/s10434-018-6540-4. Epub 2018 Jun 4. PMID: 29868978; PMCID: PMC7065419.*

This study included female patients with biopsy-proven stage II–III breast cancer who had not received prior therapy. Approved by the Institutional Review Board of the University of Southern California (USC) and compliant with REMARK criteria, the study collected 20 mL of peripheral blood (PB) from each patient into EDTA tubes at baseline, with an aliquot stored in RNAlater at -80°C. Clinical pathology reports followed the 2010 ASCO and CAP breast cancer biomarker guidelines. Formalin-fixed, paraffin-embedded primary tumour sections were obtained for RNA extraction. RNA-seq library preparation for PB and tumour samples was performed using the NuGEN Ovation RNA Systems, with sequencing on an Illumina HiSeq 2500. Data were stored on USC's High Performance Computer Cluster. NanoString nCounter assays using the PAM50 CodeSet were conducted for breast cancer subtype classification, analyzed with the genefu package in R.

Data Pre-Processing

To ensure accurate and comparable gene expression data across samples, we first filtered out genes not expressed in all samples. This step eliminated noise and bias, focusing on consistently expressed genes. We then normalized the gene expression counts using log Counts Per Million (log CPM), which adjusts for differences in sequencing depth and stabilizes variance across expression levels. To visualize the data, we performed Multi-Dimensional Scaling (MDS) to understand the similarity between samples.

Differential Expression Analysis

After preprocessing the sequence counts, we analyzed the data using edgeR, a software package designed for differential expression analysis of RNA-seq count data. EdgeR employs a negative binomial (NB) model to handle the overdispersion typical in RNA-seq data, where the variance exceeds the mean. This model accurately accounts for biological variability by estimating a dispersion parameter for each gene, providing robust estimates of differential expression. By

incorporating this parameter, edgeR effectively models the distribution of count data, ensuring reliable results.

To identify differentially expressed genes (DEGs), we compared gene expression profiles between normal and cancer blood samples using a contrast matrix. EdgeR applied a likelihood ratio test or a quasi-likelihood F-test to determine the statistical significance of observed differences. We used an adjusted p-value cutoff of 0.05, applying the Benjamini-Hochberg method to control the false discovery rate (FDR). This adjustment reduced the likelihood of false positives, ensuring that the identified DEGs were truly associated with the condition studied, thereby enhancing the reliability of our biomarker discovery process.

Functional Analysis

To understand the biological significance of differentially expressed genes (DEGs), we performed functional annotation using the enrichR. Pathway enrichment analysis was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases to identify pathways and biological processes significantly associated with the DEGs.

Results

Differentially Expressed Genes

Our study revealed a significant alteration in gene expression profiles between cancer blood and normal blood samples. The initial MDS plot (figure 1) showed clear clustering of normal samples and tumour samples. Specifically, we identified 2,286 genes that were upregulated and 67 genes (figure 2) that were downregulated in cancer blood compared to normal blood. This substantial number of differentially expressed genes highlights the extensive molecular changes associated with breast cancer.

Among the differentially expressed genes, several novel candidates emerged as potential biomarkers. Notably, genes such as COX3, CYTB, ND1, and ND6 were among the most significantly affected, with adjusted p-values less than 10^{-10} (figure3). These genes, which are involved in mitochondrial function and energy metabolism, exhibited marked changes in expression levels, suggesting their critical role in the pathophysiology of breast cancer. The significant upregulation of these genes underscores their potential utility as biomarkers for early detection and targeted therapies in breast cancer. Their consistent and significant differential expression across samples indicates that they could serve as reliable indicators of disease presence and progression, warranting further investigation into their functional roles and clinical applicability.

Functional Analysis

The genes that were up-regulated exhibited associations with key biological processes and pathways related to Extracellular Matrix (ECM) Organization, Collagen Fibril Assembly, ECM receptor interaction, and Focal Adhesion, as documented in both the KEGG and Reactome Databases. Moreover, within the Gene Ontology Biological Processes, these genes demonstrated involvement in

the negative regulation of epithelial cell proliferation and migration, indicating their potential roles in modulating cellular behaviour and tissue homeostasis. (figure 4, 5 and 6)

Conversely, the downregulated genes were linked to pathways crucial for immune system function, including the Innate Immune System, general Immune System processes, and leukocyte transendothelial migration. This downregulation hints at potential impacts on immune response dynamics. Additionally, the biological processes associated with these genes encompassed antimicrobial humoral response and superoxide anion generation, further underscoring their involvement in immune function and defence mechanisms. (figure 7, 8, and 9)

Discussion

The findings of our study hold significant implications for advancing the field of breast cancer research and guiding future investigations towards improved diagnostic, prognostic, and therapeutic strategies.

First and foremost, the identification of novel biomarker candidates such as COX3, CYTB, ND1, and ND6 underscores the importance of comprehensive validation studies in independent patient cohorts. Validating these biomarkers across diverse populations, cancer subtypes, and disease stages is essential to assess their robustness, reliability, and clinical utility.

Furthermore, the functional characterization of identified biomarkers represents a critical area for future investigation. Understanding the molecular mechanisms underlying the dysregulation of COX3, CYTB, ND1, and ND6 in breast cancer can elucidate their roles in tumorigenesis, tumour progression, and therapeutic resistance. Experimental studies, including in vitro assays and animal models, can shed light on the biological functions of these biomarkers and their interactions with key signalling pathways, offering opportunities for targeted therapeutic interventions. Overall, our study sets the stage for a multifaceted approach to breast cancer biomarker research, encompassing validation studies, functional characterization, mechanistic investigations, and clinical translation.

Conclusion

In conclusion, our study identifies novel biomarker candidates and sheds light on critical pathways dysregulated in breast cancer. The findings underscore the importance of further validation, functional characterization, and mechanistic studies to elucidate the roles of these biomarkers in tumorigenesis and progression. Translation of these discoveries into clinical practice holds promise for improving diagnostic accuracy, prognostic assessment, and personalized therapeutic interventions in breast cancer. Through collaborative efforts and interdisciplinary approaches, we can advance the understanding of breast cancer biology and pave the way for more effective management strategies, ultimately enhancing patient outcomes and quality of life.

Annexe

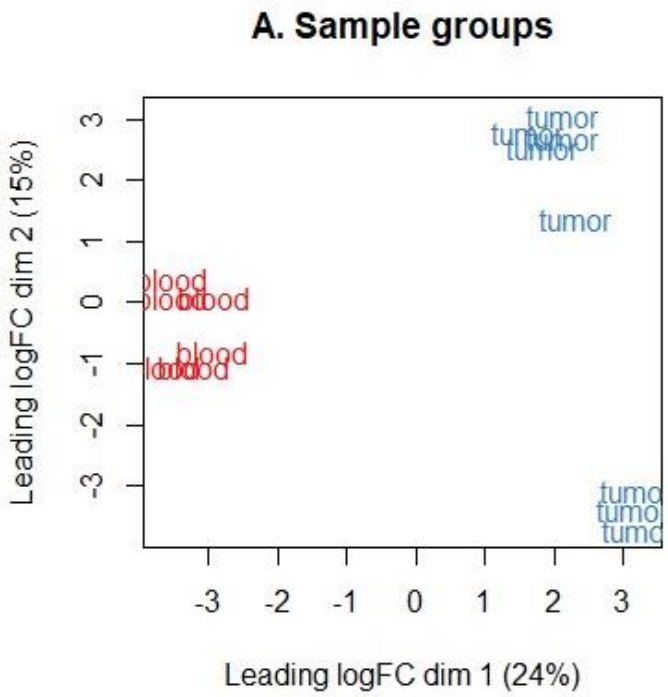

Figure 1: MDS Plot of all samples

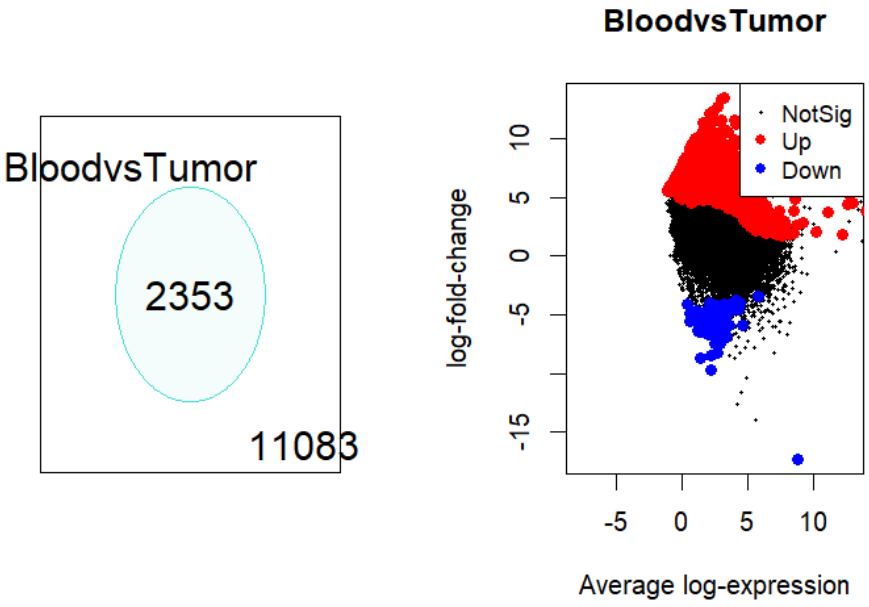

Figure 2: Differentially expressed genes in tumor blood vs normal

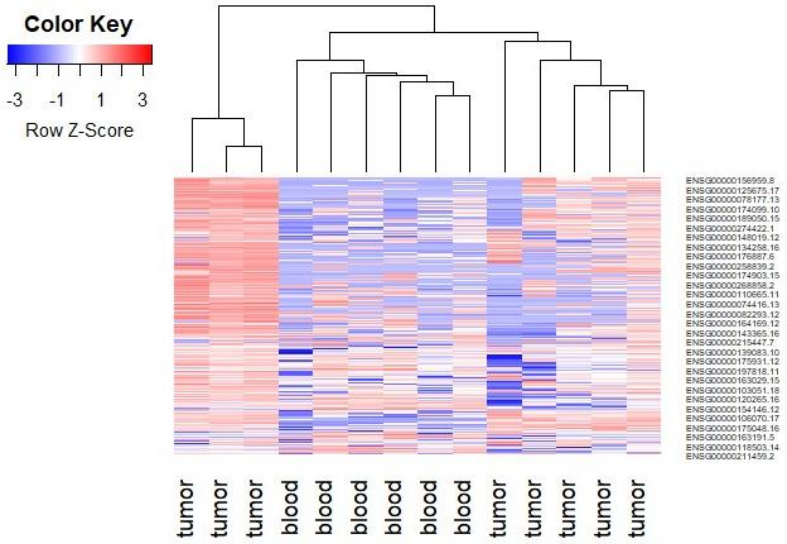

Figure 3: Heat map of top 100 differentially expressed genes

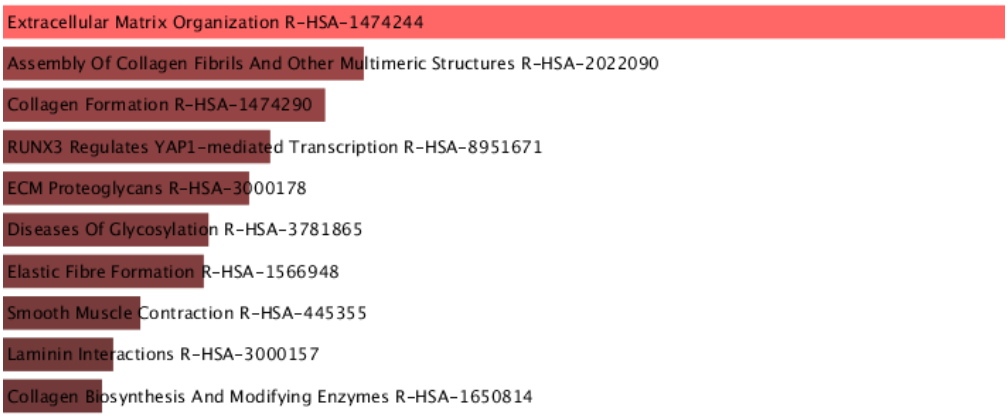

Figure 4: Reactome 2022 pathways enriched in upregulated genes

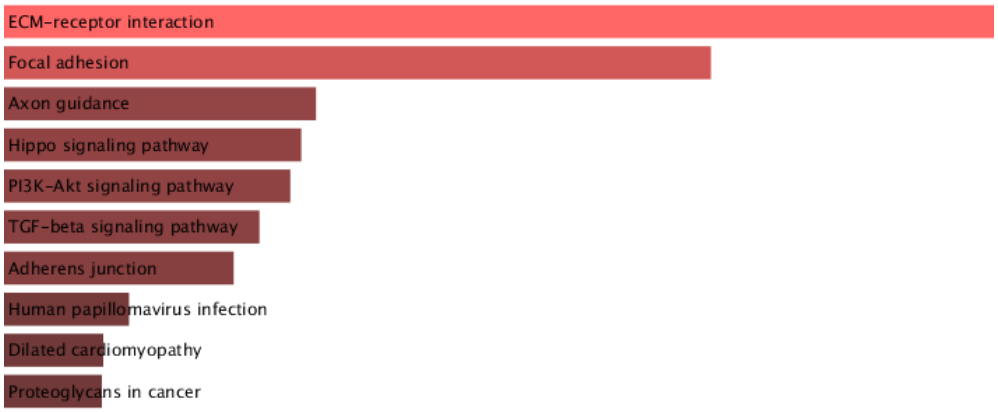

Figure 5: KEGG 2021 pathways enriched in upregulated genes

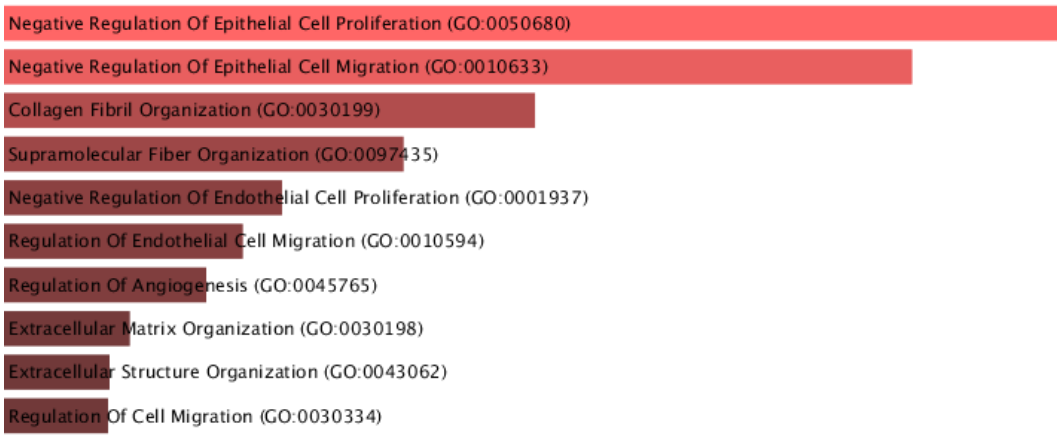

Figure 6: GO Biological Processes enriched in upregulated genes

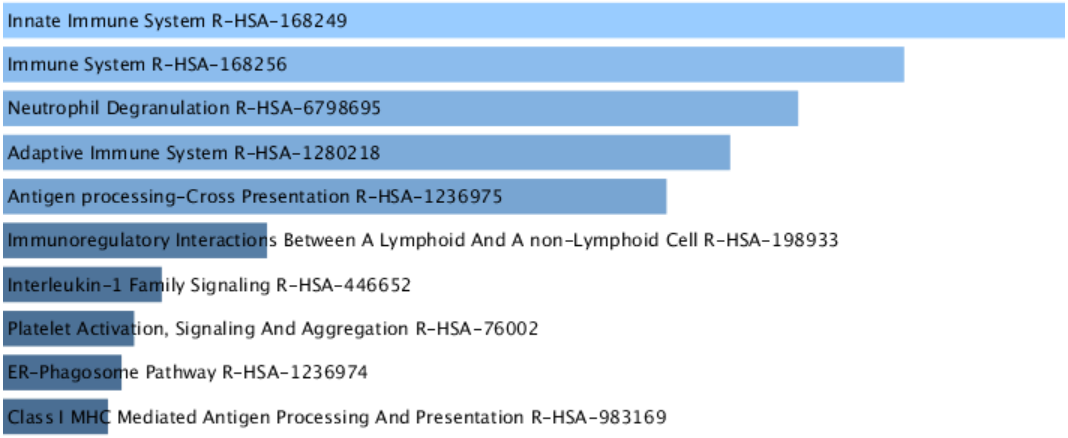

Figure 7: Reactome 2022 pathways enriched in downregulated genes

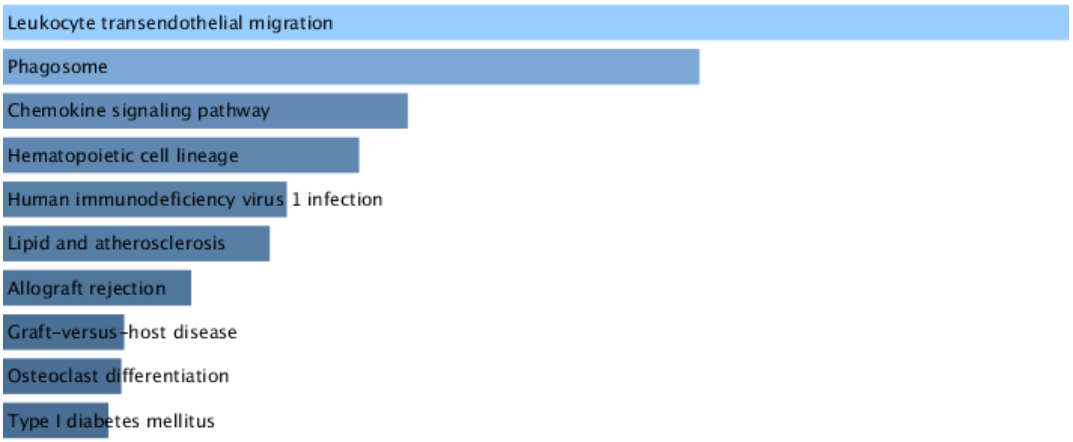

Figure 8: KEGG 2021 pathways enriched in downregulated genes

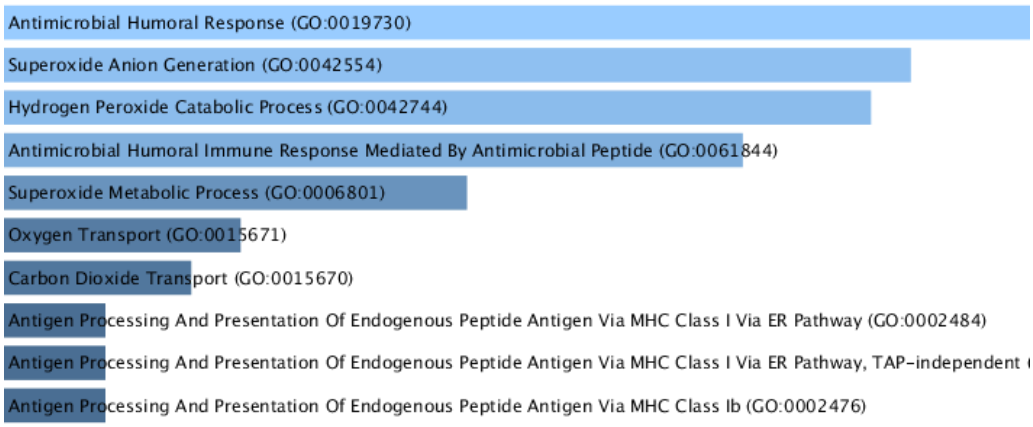

Figure 9: GO Biological Processes enriched in downregulated genes