Identification of Biomarkers for Pregnancy Associated Breast Cancer (PABC) Using Gene Expression Data from GSE31192

Mahima Mahabaleshwar Siddheshwar
Department of Bioinformatics
Luddy School of Informatics, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, IN, USA

Abstract

Pregnancy Associated Breast Cancer (PABC) represents a subtype of breast cancer diagnosed during or shortly after pregnancy. It is characterized by distinct biological features and a typically aggressive course. This study aimed to identify molecular biomarkers for PABC by analyzing gene expression profiles from the GSE31192 dataset. Through robust normalization procedures, differential expression analysis, and gene ontology enrichment analysis, I have identified significant genes biomarkers that differentiate PABC from normal breast tissue. These findings pave the way for potential therapeutic targets and diagnostic markers.

Background

Pregnancy Associated Breast Cancer (PABC), diagnosed during pregnancy or within a year postpartum, represents a uniquely challenging subset of breast cancer due to its rarity and aggressive nature. The increasing incidence of PABC, influenced by the rising maternal age across populations, underscores the urgent need for focused research to unveil its complex interplay between pregnancy-induced hormonal changes and cancer progression (Amant et al., 2012). These physiological changes during pregnancy, such as increased breast density, complicate early detection, and diagnosis, making PABC particularly perilous (Loibl et al., 2012).

The complexity of managing PABC lies in balancing effective cancer treatment while ensuring the safety of both mother and fetus, highlighting the critical need for identifying precise molecular biomarkers for better therapeutic targeting. Despite significant advances in breast cancer research, PABC remains under-studied, with limited tailored therapeutic strategies available (Johansson et al., 2011). This study leverages advanced bioinformatics to analyze the GSE31192 dataset, aiming to identify distinct molecular biomarkers that could revolutionize the diagnostic and therapeutic approaches for this challenging cancer subtype, paving the way for personalized treatment innovations (Cardonick et al., 2014).

Materials and Methods

In my exploration of the GSE31192 dataset from the Gene Expression Omnibus (GEO), I employed the R package GEOquery for systematic data retrieval, ensuring the reproducibility and integrity of my workflow. The dataset underwent log2 transformation to achieve normalization of gene expression values, crucial for mitigating technical variation and enhancing comparability across the 20 normal and 13 tumor breast tissue samples. Phenotypic annotations were meticulously refined, establishing a robust framework for subsequent comparative analyses. Utilizing Principal Component Analysis (PCA), we reduced the dataset's dimensionality, which facilitated visual discrimination between normal and tumor samples, revealing distinct clustering patterns indicative of intrinsic biological differences.

Hierarchical clustering was then applied to the normalized data, with the generation of a heatmap providing an intuitive visual representation of gene expression correlations. This step was pivotal in illustrating the distinct molecular profiles characteristic of PABC. To pinpoint differentially expressed genes, I implemented the limma

package, a statistical approach tailored for such high-throughput data, enabling us to discern significant expression differences with precision. Completing my analysis, Gene Ontology (GO) enrichment was conducted to contextualize the biological relevance of our findings, drawing attention to overrepresented molecular functions and pathways potentially implicated in PABC's pathogenesis.

This suite of bioinformatic techniques underscored the potential for identifying novel biomarkers and therapeutic targets, propelling forward the molecular understanding of PABC.

Results and Interpretation

Box Plot Analysis:

The box plot (Fig.1) provided an initial view of the log-normalized gene expression data across several samples from the GSE31192 dataset. The uniformity of median expression levels and interquartile ranges across samples indicated successful normalization. This uniformity is crucial for the accuracy of downstream differential expression analysis, ensuring that variations are due to biological differences rather than technical disparities.

Heatmap Clustering:

The heatmap (Fig.2) displayed shows a visual representation of the gene expression data, with clustering evident on two levels: samples and genes. On the sample side (vertical dendrogram), we see a bifurcation, segregating samples into what are labeled 'Normal' and 'Tumor' groups. This division is a critical finding because it visualizes the extent to which the gene expression profiles of normal breast tissue diverge from those associated with tumor tissue. The dendrogram at the top (horizontal) shows the clustering of gene expression patterns across the samples, indicating groups of genes that exhibit similar expression levels. The color gradient, with red representing higher expression and blue representing lower expression, illustrates the variance in gene activity across the samples. These patterns are fundamental for understanding PABC at the molecular level because they suggest a signature of gene expression that is characteristic of the cancerous state during or after pregnancy.

Principal Component Analysis (PCA):

The PCA scatter plot (Fig.3) revealed a clear segregation between normal and tumor samples, primarily along the first principal component (PC1). This separation suggests that PC1 captures a significant variance associated with the disease state, reinforcing that gene expression profiling can distinguish between normal and cancerous cells in PABC.

Identification of Up-Regulated and Down-Regulated Gene:

Out of 36,589 total number of genes I found 1801 genes to be up-regulated and 2322 genes to be down-regulated genes.

Volcano Plot - Differential Expression:

The volcano plot (Fig.4) provides a striking visualization of significant differences in gene expression between normal and tumor cells, identifying potential biomarkers for PABC. Key genes such as DLK1, MYH11, and CXCL11 exhibit significant fold changes and are highlighted as promising candidates for further investigation. The plot shows a clear distinction in the expression profiles of the two groups, with a predominance of differentially expressed genes on both sides of the log-fold change axis.

Gene Ontology (GO) Enrichment Analysis:

The Gene Ontology (GO) enrichment analysis shown in the bar (Fig.5) and dot plots (Fig.6) provides an in-depth information about the biological processes overrepresented among the differentially expressed genes in PABC. This analysis sheds light on key functional categories that might be disrupted or altered in PABC.

The most significant GO term is "neuropeptide receptor binding," highlighted in red to indicate its prominence. This term's enrichment suggests a potential role of neuropeptide signaling in PABC, perhaps influencing tumor behavior or patient response to hormonal changes during pregnancy.

"G protein-coupled receptor binding" also shows significant enrichment, reinforcing the idea that signaling pathways associated with these receptors might be integral to PABC pathogenesis. G protein-coupled receptors are involved in various physiological processes and are known to play roles in various cancers (Rachel Bar-Shavit et al., 2016). Other notable terms include "actin receptor binding," "CXCR chemokine receptor binding," and "transmembrane receptor protein serine/threonine kinase binding." These terms point towards the involvement of intricate cellular signaling networks in PABC, with potential implications for how the disease develops and how it might be targeted therapeutically.

Significance of Findings

The box plot confirmed the quality and comparability of the data, a foundational step in any gene expression analysis.

The heatmap provided visual evidence of distinct gene expression patterns in PABC, which could guide the development of diagnostic tests or treatments.

The PCA plot served as a powerful tool for reducing the complexity of gene expression data, enabling the visualization of the most significant patterns that may be diagnostic of PABC.

The volcano plots were key in identifying specific genes that could serve as biomarkers for PABC, offering a starting point for targeted therapies or detailed molecular studies.

The GO enrichment analysis underscored the involvement of specific signaling pathways and receptor interactions in PABC, offering insights into potential mechanisms driving the disease and possible intervention points.

Collectively, these findings provide a robust foundation for further research into PABC, pointing to specific genes and biological processes that may be integral to the disease. The visualizations not only present the data effectively but also highlight the potential for these molecular markers and processes to be leveraged in the pursuit of better diagnostic and therapeutic strategies for PABC.

Discussion:

The gene expression analysis provided a molecular distinction between PABC and normal breast tissue. The data-driven approach highlighted several candidate biomarkers and implicated signaling pathways in the disease's etiology, which could lead to novel therapeutic interventions.

Figure 1: Box Plot

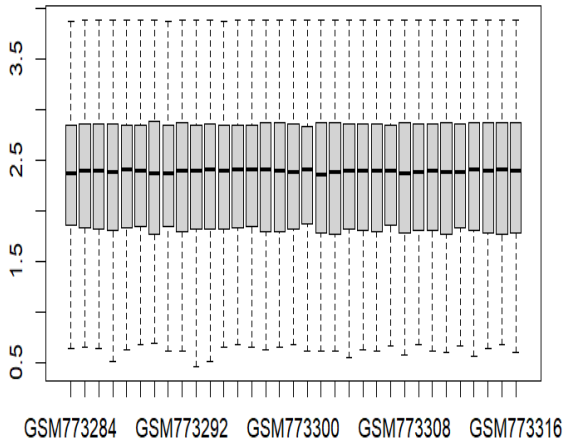


Figure 2: Heatmap Clustering

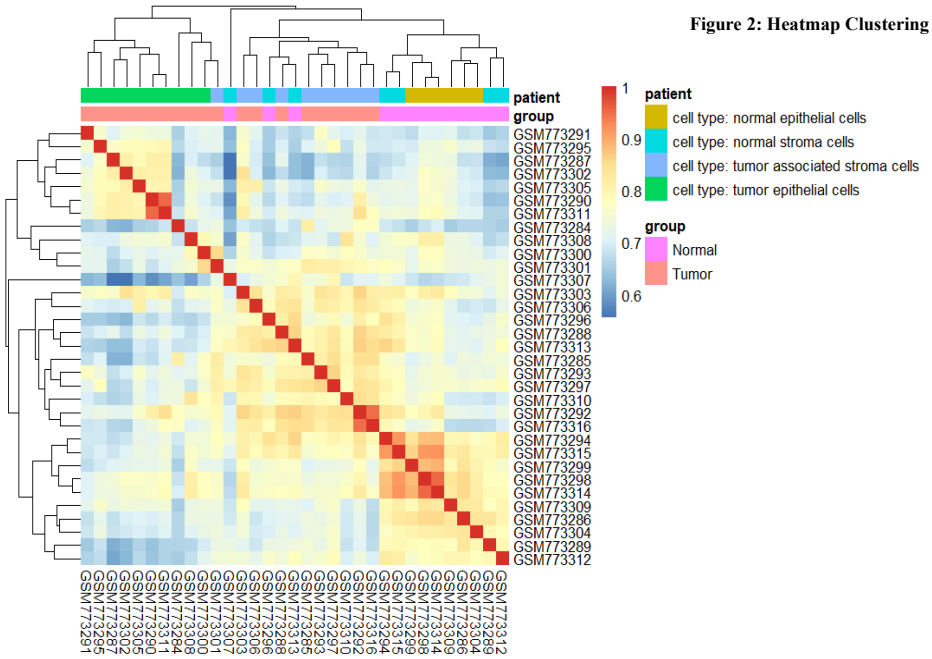


Figure 3: PCA scatter plot

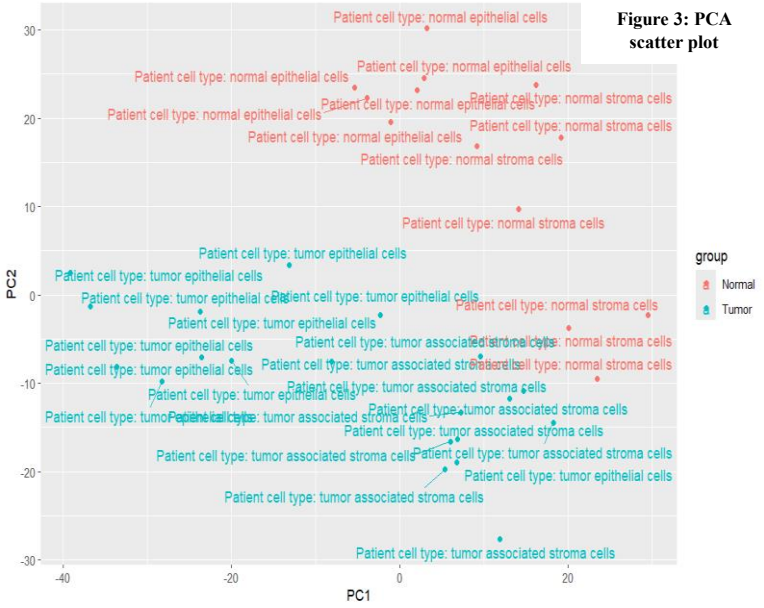


Figure 4: Volcano Plot

Figure 5: GO Bar Plot

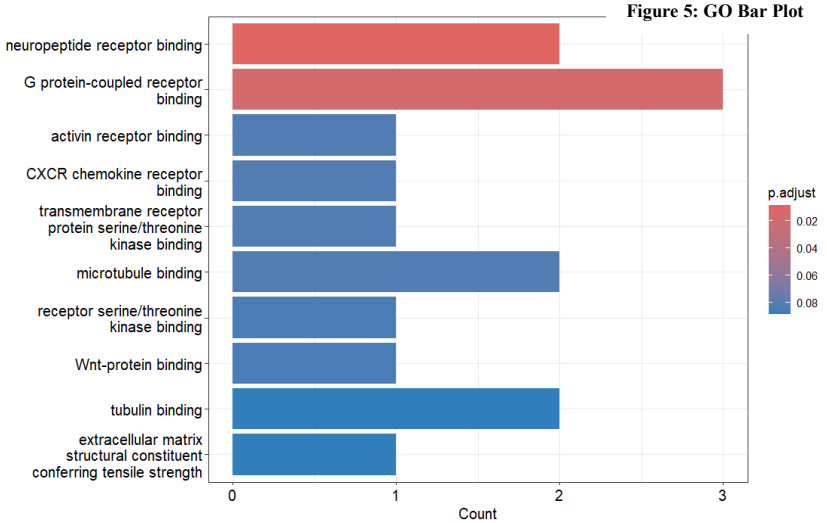
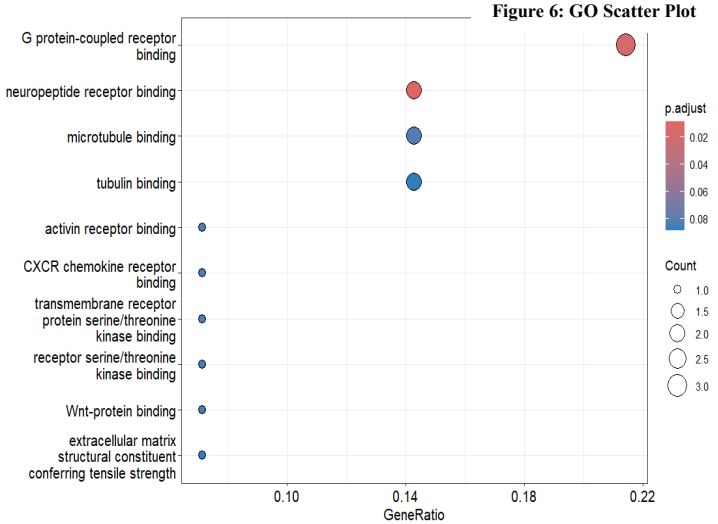


Figure 6: GO Scatter Plot



Conclusion

This study's comprehensive analysis of the GSE31192 dataset has illuminated distinct biomarkers that hold the potential to profoundly alter our understanding and management of Pregnancy Associated Breast Cancer (PABC). By employing a robust bioinformatics framework, the research identified key molecular signatures—namely DLK1, MYH11, and CXCL11—that are differentially expressed between normal and tumor tissues associated with PABC. DLK1, a protein linked with fetal development and various malignancies, acts as a Notch family protein that may inhibit differentiation, suggesting its disruptions could contribute to cancer development, including PABC (Pittaway et al., 2021). Meanwhile, MYH11's involvement in cell motility may hint at mechanisms that drive cancer metastasis. CXCL11's role in modulating the immune response underscores its potential impact on tumor-immune system interactions in the unique context of pregnancy. The significant differentiation of these genes between normal and PABC tissues not only underscores their potential as diagnostic and prognostic biomarkers but also highlights them as viable targets for future therapeutic strategies. Further research is required to fully understand the specific roles these biomarkers play in PABC, and to explore their potential for clinical application, thus paving the way for targeted, more effective interventions for this challenging subtype of breast cancer.

Future Directions

Building on the findings of this study, future research should pivot towards experimental validation of the identified biomarkers DLK1, MYH11, and CXCL11, using in vitro and in vivo models to explore their roles in PABC progression and response to treatment. Clinical trials are essential to assess the efficacy of these biomarkers in clinical settings, particularly for early diagnosis and monitoring of PABC. Additionally, longitudinal studies monitoring these gene expressions throughout pregnancy and the postpartum period could provide critical insights into their temporal dynamics and influence on PABC development. An integrated omics approach, combining genomic, transcriptomic, and proteomic data, would further elucidate the complex molecular interplay in PABC, paving the way for personalized medicine strategies tailored to the unique biology of PABC patients. This comprehensive approach promises to transform the understanding and treatment of PABC, leading to significantly improved outcomes for affected women.

*****Note: Readme file, images, and code to generate the plots are provided in the GitHub link. [7]**

References

1. Amant, F., Loibl, S., Neven, P., & Van Calsteren, K. (2012). Breast cancer in pregnancy. The Lancet, 379(9815), 570-579. https://doi.org/10.1016/S0140-6736(11)61092-1
2. Loibl, S., Han, S. N., von Minckwitz, G., Bontenbal, M., Ring, A., Giermek, J., Fehm, T., Van Calsteren, K., Linn, S. C., Schlehe, B., Mhallem Gziri, M., Westenend, P. J., Müller, V., Heyns, L., Rack, B., Van Calster, B., Harbeck, N., Lenhard, M., Halaska, M. J., Kaufmann, M., … Amant, F. (2012). Treatment of breast cancer during pregnancy: An observational study. The Lancet Oncology, 13(9), 887-896. https://doi.org/10.1016/S1470-2045(12)70261-9
3. Johansson, A. L. V., Andersson, T. M. L., Hsieh, C. C., Cnattingius, S., & Lambe, M. (2011). Increased mortality in women with breast cancer detected during pregnancy and different periods postpartum. Cancer Epidemiology, Biomarkers & Prevention, 20(9), 1865-1872. https://doi.org/10.1158/1055-9965.EPI-11-0515
4. Cardonick, E. (2014). Pregnancy-associated breast cancer: Optimal treatment options. International Journal of Women's Health, 6, 935-943. https://doi.org/10.2147/IJWH.S52381
5. Bar-Shavit, R., Maoz, M., Kancharla, A., Nag, J. K., Agranovich, D., Grisaru-Granovsky, S., & Uziely, B. (2016). G Protein-Coupled Receptors in Cancer. International Journal of Molecular Sciences, 17(8), 1320. https://doi.org/10.3390/ijms17081320
6. Pittaway, J. F. H., Lipsos, C., Mariniello, K., & Guasti, L. (2021). The role of delta-like non-canonical Notch ligand 1 (DLK1) in cancer. Endocrine-Related Cancer, 28, R271-R287.
7. *** Link to GitHub: https://github.iu.edu/msiddhe/MahimaMS_HTP_Project.git