Thomas Rucinski | 12/10/25 | BIOL-616

Manuscript Title: Differential Gene Expression of Luminal A Invasive Ductal Carcinoma in Comparison to Unaffected Breast Tissue

1. Introduction

Breast cancer remains the most diagnosed cancer among women worldwide, affecting approximately one in eight (13.1%) women in their lifetime [1]. Within this broad classification, molecular subtyping has become essential for prognosis and effective treatment planning. The Luminal A subtype, characterized by estrogen receptor positivity (ER+), Human Epidermal growth factor Receptor 2 negativity (HER2-), and low expression of the proliferation marker Ki-67, is the most prevalent, accounting for 47.7% of diagnoses in the TCGA-BRCA cohort and between 50 to 60% of all breast cancer diagnoses [2]. While considered the most favorable subtype due to its responsiveness to hormone therapy and its relatively low proliferation rate, patient outcomes vary, and the mechanisms driving this variation remain incompletely understood [6].

Current research heavily emphasizes the hormonal drivers of Luminal A, such as estrogen signaling, often overshadowing the tumor microenvironment's metabolic and immune landscape. Specifically, there is a gap in understanding how Luminal A tumors actively suppress homeostatic pathways to ensure survival. While many hormonal drivers are well defined, the downregulation of specific metabolic and immune pathways in Luminal A Invasive Ductal Carcinoma (IDC) compared to unaffected breast tissue has not been sufficiently characterized. Understanding how these suppression mechanisms work is critical, as they may reveal how these tumors minimize oxidative stress and evade immune surveillance, contributing to their survival and growth.

This study addresses this gap by analyzing transcriptomic data from The Cancer Genome Atlas (TCGA) to characterize the dysregulated landscape of Luminal A IDC. I hypothesize that differentially expressed genes in Luminal A IDC will reveal not only the expected enrichment in hormone-regulated and cell-cycle pathways but also a distinct and significant downregulation of metabolic and immune defense pathways compared to unaffected breast tissue. By identifying these "cold" immune features and key downregulated pathways, this study aims to provide a more holistic view of Luminal A biology beyond simple hormonal proliferation.

2. Data and Methods

2.1 Dataset Acquisition and Preprocessing

Publicly available RNAseq data (IlluminaHiSeq pancan normalized) and patient metadata (Phenotype) were acquired from the University of California, Santa Cruz (UCSC) Xena database, specifically the TCGA-BRCA cohort [3,4]. To ensure a robust analysis, the sample size was expanded significantly from the initial proposal, increasing from 50 samples to over 400 samples, consisting of ~300 Luminal A IDC primary tumor samples and ~100 unaffected solid tissue samples. Data preprocessing was performed using Python. The pipeline involved harmonizing sample IDs and intersecting datasets to ensure overlap between gene expression matrix and clinical phenotype metadata matrix. Samples were rigorously filtered to include only "Primary Tumor" samples with a "Luminal A" subtype and "Invasive Ductal Carcinoma" histological type, and the control "Solid Tissue Normal" samples. This curation isolated the test and control samples for use in the DEG Analysis.

2.2 Differential Expression of Genes (DEG)

DEG analysis was conducted using the R programming language, primarily utilizing the limma package. Limma was selected for its

renowned linear modeling and empirical Bayes moderation functionalities, which improve variance estimation and statistical power. A design matrix was constructed to contrast Tumor versus Normal groups. To ensure statistical significance, genes were considered differentially expressed only if they met a False Discovery Rate (FDR) q-value threshold of < 0.05 and a magnitude threshold of $|\log2 \text{ Fold Change}| > 1$.

2.3 Gene Set Enrichment Analysis (GSEA)

To biologically interpret gene expression signatures, Gene Set Enrichment Analysis (GSEA) was performed using the University of California, San Diego (UCSD) GSEA 4.4.0 software. A ranked list of genes was generated based on differential expression scores, then input into the software. Three distinct gene set databases, each with a unique purpose, were utilized to provide a multi-layered perspective:

- **Hallmark Gene Set** (h.all.v2025.1.Hs.symbols.gmt)**:** Used for a broad, well-defined overview of specific biological states.
- **KEGG (Kyoto Encyclopedia of Genes and Genomes) Legacy** (c2.cp.kegg_legacy.v2025.1.Hs.symbols.gmt)**:** Used for its standard metabolic and signaling pathway definitions.
- **KEGG Medicus** (c2.cp.kegg_medicus.v2025.1.Hs.symbols.gmt)**:** Selected specifically for its focus on disease pathway dysregulation and immunological drivers.

Pathways were ranked by Normalized Enrichment Score (NES) to visualize the magnitude and direction of dysregulation; the NES represents the degree to which a pathway's genes are clustered at the extremes of the ranked gene list. However, to ensure statistical validity and control the expected rate of false positive findings across all tested pathways, only those pathways satisfying a False Discovery Rate (FDR) q-value < 0.25 were considered significant for discussion [5]. Figure 2 illustrates the three major steps of the pipeline, broken down into a flowchart.

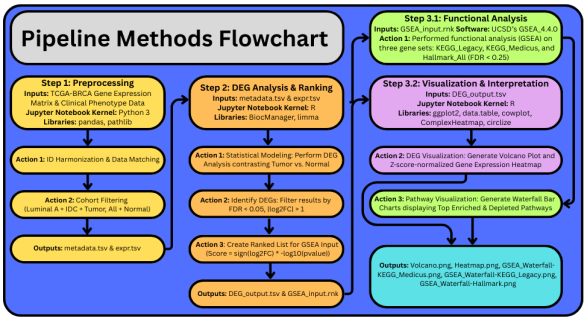

Figure 2: Data Analysis Methods Flowchart

3. Results

3.1 DEG Analysis Results

DEG analysis identified 4191 statistically significant genes that distinguished Luminal A IDC tumor samples from unaffected breast tissue. 1,421 of those significant DEGs were found to be upregulated in Luminal A IDC primary tumor samples compared to the unaffected breast tissue samples, and 2,770 DEGs were found to be downregulated.

The distribution of these DEGs is summarized in the volcano plot (Figure 1), which illustrates the strong statistical power of the analysis, and clearly shows the correlation of the log2 Fold Change and the FDR q-value for all 20,530 genes within the gene expression dataset.

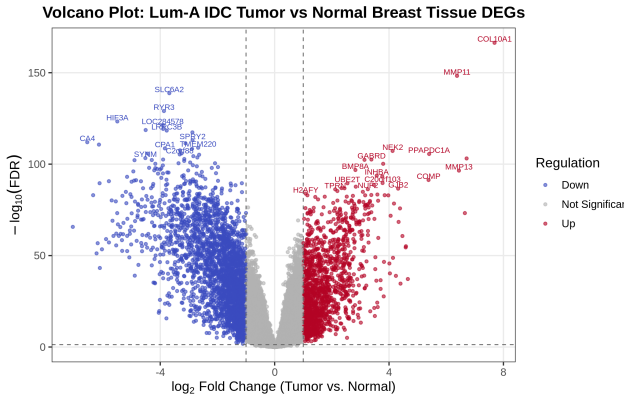

Figure 1: Volcano Plot from DEG Analysis

The most significant DEGs were labeled on the volcano plot using their respective gene symbols. The heatmap in Figure 3, generated in the visualization step of the pipeline, displays the top 25 upregulated and downregulated genes compared to the unaffected breast tissue samples, and confirms the success of the DEG analysis, displaying the clear separation of the Luminal A IDC and normal gene groups. The yellow columns at the top represent the Luminal A IDC (test) group, and the green columns display the normal (control) group. The blue rows represent the top 25 downregulated genes in the tumor samples, and the red rows represent the top 25 upregulated genes.

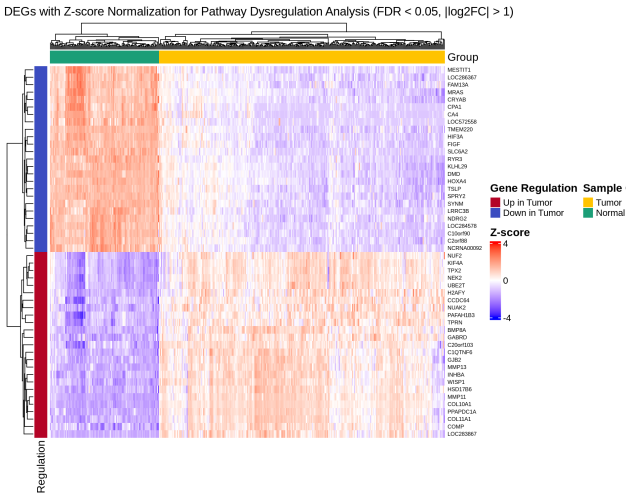

Figure 3: Heatmap of Gene Expression Values

3.2 Gene Set Enrichment Analysis Results

GSEA was conducted using three distinct gene set databases to provide a multi-layered functional interpretation of the computed ranked DEG list. A total of 60 pathways were selected across the three databases, the top 10 enriched (NES>0) and top 10 depleted (NES<0) for each database (Figure 4, 5, 6). The core findings demonstrated an extremely consistent and clear transcriptional signature defined by significant enrichment in cell cycle and proliferation pathways, coupled with suppression of metabolic and immune-related functions.

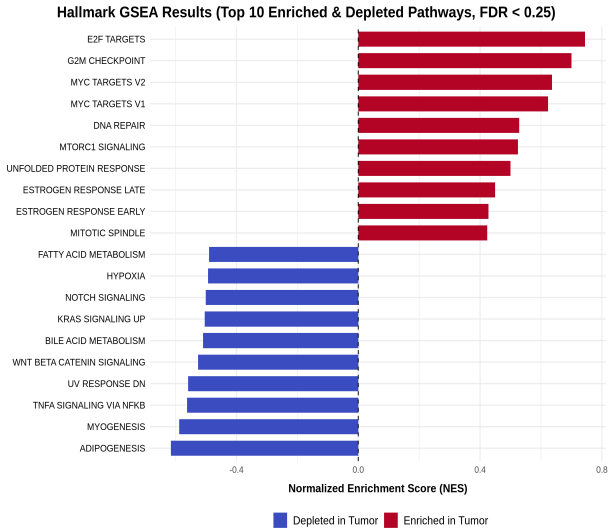

Figure 4: Hallmark GSEA Pathway Enrichment and Depletion Waterfall Chart

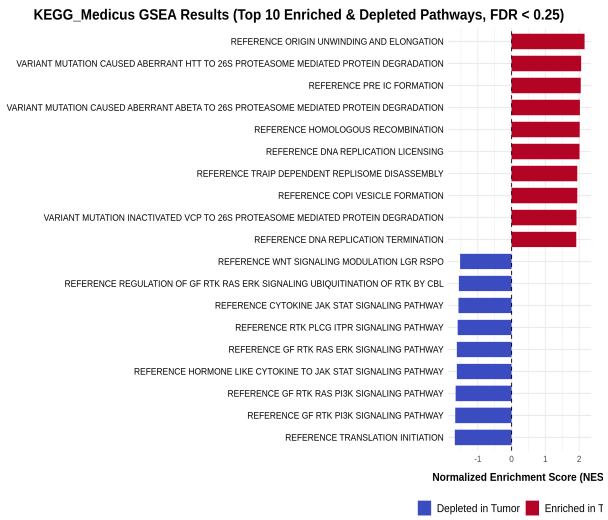

Figure 5: KEGG Medicus GSEA Pathway Enrichment and Depletion Waterfall Chart

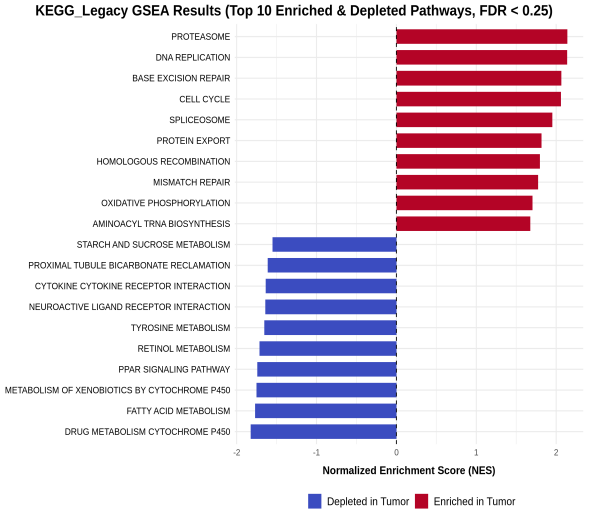

Figure 6: KEGG Legacy GSEA Pathway Enrichment and Depletion Waterfall Chart

3.2.1 Core Pathway Enrichment:

The most significant enrichment scores across all three databases were the molecular drivers for uncontrolled cell division, which is characteristic of cancers.

Cell Cycle Enrichment: The Hallmark set provides the clearest view of proliferation drivers, with the highest enrichment scores observed for E2F Targets, G2M Checkpoint, and MYC Targets V1/V2 (Figure 4).

Genomic Integrity and Replication: The KEGG databases confirmed a state of high genomic activity. KEGG Legacy showed high scores for DNA Replication and Base Excision Repair, while the KEGG Medicus set showed the highest magnitude of enrichment for pathways like Origin Unwinding and Elongation, and DNA Replication Licensing (Figure 5, 6). Evaluation of multiple gene lists in GSEA highlights the commonality of active DNA repair and synthesis mechanisms across independent reference databases. This strong, multi-database pathway functional analysis supports the claim that the tumors are experiencing severe genomic stress, which is necessary to maintain their rapid cellular output.

Protein Management and ER Stress: The enrichment of the Proteasome and MTORC1 confirms an elevated rate of protein synthesis and degradation necessary for the rapid cell growth within a tumor. The presence of Unfolded Protein Response (UPR) suggests that the high translational demand in the tumors is inducing estrogen receptor (ER) stress, causing protein folding machinery such as UPR to upregulate its function to prevent apoptosis (Figure 4).

3.2.2 Core Pathway Depletion:

The depleted pathways reveal a consistent pattern of metabolic simplification and a reduced reliance on complex external signaling pathways, supporting the study's central hypothesis.

Metabolic Reprogramming: The most significant suppression based on Normalized Enrichment Score (NES) was observed in lipid homeostasis, consistent with the expected metabolic shift in ER+ tumors away from normal tissue function. “Adipogenesis” was the single most depleted pathway in the Hallmark set, and

"Fatty Acid Metabolism" was significantly depleted in both Hallmark and KEGG Legacy (Figure 4, 5). Furthermore, detoxification pathways like "Drug Metabolism Cytochrome P450" were suppressed (Figure 5).

Suppression of Signaling Cascades: The KEGG Medicus analysis provided key insight into suppressed regulation, with the strongest depletion signals coming from complex receptor tyrosine kinase RTK signaling networks, including "Reference Regulation of GF RTK RAS ERK Signaling" and "Cytokine JAK STAT Signaling Pathway" (Figure 6). This pattern suggests that the tumor is either relying on constitutive internal signals or has decoupled from external growth factor regulation entirely.

Translation and Differentiation: The depletion of "Reference Translation Initiation" in the Medicus set, paired with the enrichment of "Unfolded Protein Response" and the "Proteasome" suggests the tumor may rely on some alternative, high-efficiency protein production or remodeling processes rather than generic, high-volume ribosomal biogenesis (Figure 6). Pathways linked to differentiation, such as "WNT Beta Catenin Signaling" and "Myogenesis" (Figure 4), were also suppressed, further indicating a loss of normal tissue identity.

4. Conclusion and Discussions

4.1 Overall Conclusion:

This study successfully utilized a bioinformatics pipeline to characterize the transcriptomic landscape of Luminal A Invasive Ductal Carcinoma in comparison to unaffected breast tissue, revealing a signature defined by two opposing transcriptional occurrences: accelerated cell proliferation and metabolic/immune suppression.

Hypothesis Support: The findings supported the hypothesis that differentially expressed genes would show not only cell-cycle enrichment but also clear downregulation of key homeostatic, immune, and metabolic pathways.

Interpretation of DEGs: Differential expression analysis identified 4191 DEGs that cluster Luminal A IDC tumors away from normal tissue, confirming the profound transcriptional shift in the cancerous state.

Interpretation of GSEA Results: Gene Set Enrichment Analysis consistently highlighted the dominance of cell cycle drivers "E2F Targets", "MYC Targets", "DNA Replication" and the suppression of specialized functions "Adipogenesis", "Fatty Acid Metabolism", and "RTK Signaling" across three independent gene set databases.

4.2 Significant and Novel Biological Insight:

Biological Novelty: The most significant finding is the direct confirmation of the metabolic shift and loss of external control in Luminal A IDC. The profound depletion of pathways like "Adipogenesis" and "Fatty Acid Metabolism", combined with the suppression of multiple RTK and JAK STAT signaling cascades, suggests the tumor aggressively discards non-essential cell activity and relies on internal drivers.

Statistical Strength: The multi-database GSEA approach (Hallmark, KEGG Legacy, and KEGG Medicus) provided evidence across independent gene and pathway databases, lending high confidence to the identified biological themes.

4.3 Assumptions and Limitations:

Assumption: The analysis assumes that the ~100 unaffected breast tissue samples represent a true baseline for comparison against the Luminal A Invasive Ductal Carcinoma tumors.

Limitation 1 (Data Source): TCGA IluminaHiSeq pancan normalized RNA-seq data provides bulk tissue analysis. This cannot differentiate between gene expression changes in the cancer cells themselves versus those in the tumor microenvironment, such as immune cells or fibroblasts, potentially diluting or entirely missing specific tumor derived signals.

Limitation 2 (Sample Heterogeneity): Although the cohort was filtered for Luminal A IDC, biological heterogeneity exists within this large group. The analysis treats all Luminal A IDC tumors as a single homogeneous entity, which may dilute sub-group specific transcriptional signatures that could correlate with variations in patient outcome. For example, the heatmap in Figure 3 illustrates some Luminal A IDC samples that do not follow the same gene expression patterns as other Luminal A samples for certain genes.

Limitation 3 (Protein Activity): The study is based purely on mRNA levels (transcriptomics). The level of mRNA does not always directly correlate with the final level of functional protein (proteomics) or its post-translational changes due to phosphorylation (phosphoproteomics). The observed dysregulation (such as in MTORC1 Signaling) may not perfectly reflect the actual protein activity in the cell, or even the proteins' function.

4.3. Future Direction:

Subtype Heterogeneity: Conduct a deeper analysis of the Luminal A cohort to identify any underlying transcriptional heterogeneity (sub-clusters), which could explain the patient outcome variation mentioned in the Introduction. This would help to address limitation 2, and this additional work could start by looking at the sub-clusters within the gene expression heatmap that don't appear to fit in with the obvious current clusters.

Metabolic Target Validation: Focus on the highly suppressed “Fatty Acid Metabolism” and “Adipogenesis” pathways. Future wet-lab work could involve inhibiting key regulatory enzymes within these depleted pathways to confirm if their suppression is essential for Luminal A tumor cell survival and proliferation.

Therapeutic Targeting: Investigate the relationship between the enriched Unfolded Protein Response (Estrogen Receptor Stress) pathway and the depleted RTK Signaling. Targeting the UPR machinery could be a promising therapeutic strategy for overcoming the inherent resistance mechanisms of Luminal A Invasive Ductal Carcinoma tumors.

Combining Multi-omics Analysis with Convolutional Neural Networks for Histological Image Processing for More Accurate Diagnostics in Cancer: Finally, combining multi-omics approaches, especially transcriptomics and proteomics, with machine learning models for histological image processing will improve cancer detection and prognosis by leveraging two of the most promising diagnostic tools available today. This integration represents the next step in addressing a critical gap in precision oncology, and its success will allow clinical decisions to be informed by both molecular function and tissue morphology, improving patient outcomes across the board.

Works Cited:

[1]"Breast Cancer Facts & Figures," *American Cancer Society*, 2024. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf (accessed Oct. 01, 2025).

[2]D. Schlamadinger, "Luminal A Breast Cancer Explained | BCRF," *Breast Cancer Research Foundation*, Aug. 04, 2025. https://www.bcrf.org/about-breast-cancer/luminal-a-breast-cancer/ (accessed Oct. 1, 2025).

[3]"UCSC Xena," *Xenabrowser.net*, 2025. https://xenabrowser.net/datapages/?dataset=TCGA.BRCA.sampleMap%2FHiSeqV2_PANCAN&host=https%3A%2F%2Ftcga.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443 (accessed Sep. 26, 2025).

[4]"UCSC Xena," *Xenabrowser.net*, 2024. https://xenabrowser.net/datapages/?dataset=TCGA.BRCA.sampleMap%2FBRCA_clinicalMatrix&host=https%3A%2F%2Ftcga.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443 (accessed Sep. 26, 2025).

[5]J. Li *et al.*, "Identification of differentially expressed genes-related prognostic risk model for survival prediction in breast carcinoma patients," *Aging*, vol. 13, no. 12, pp. 16577–16599, Jun. 2021, doi: https://doi.org/10.18632/aging.203178. (accessed Sep. 26, 2025).

[6]R. R. Bastien *et al.*, "PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers," *BMC Medical Genomics*, vol. 5, no. 1, Oct. 2012, doi: https://doi.org/10.1186/1755-8794-5-44. (accessed Dec. 1, 2025).

[7]Aleksandra Mordzińska-Rak, Grégory Verdeil, Y. Hamon, E. Błaszczak, and Tomasz Trombik, "Dysregulation of cholesterol homeostasis in cancer pathogenesis," *Cellular and Molecular Life Sciences*, vol. 82, no. 1, Apr. 2025, doi: https://doi.org/10.1007/s00018-025-05617-9. (accessed Dec. 1, 2025).

[8]T. Du *et al.*, "Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism," *Scientific Reports*, vol. 8, no. 1, p. 7205, May 2018, doi: https://doi.org/10.1038/s41598-018-25357-0. (accessed Dec. 10, 2025).