**Supplementary Material (SM) for**

Identifying Key Genes Involved in Axillary Lymph Node Metastasis in Breast Cancer Using Advanced RNA-Seq Analysis: A Methodological Approach with GLMQL and MAS

Mostafa Rezapour, PhD1\*; Robert Wesolowski, MD2; Metin Nafi Gurcan, PhD1

1Center for Artificial Intelligence Research, Wake Forest University School of Medicine, Winston-Salem, NC, United States of America

2The Ohio State University Comprehensive Cancer Center, Columbus, OH, United States of America

**\* Correspondence:** Mostafa Rezapour, PhD

Email: [mrezapou@wakehealth.edu](mailto:mrezapou@wakehealth.edu)

A blue grid with many colored lines

Description automatically generated

***Figure S1:*** *Hierarchical Clustering Heatmap of Gene Expression in Breast Cancer Samples. This figure illustrates the differential expression patterns of the top 100 genes selected using the GLMQL-MAS approach, filtered with a LogFC threshold of >1. Samples are grouped into ALNM- (negative for axillary lymph node metastasis, green bar) and ALNM+ (positive for axillary lymph node metastasis, purple bar), showcasing distinct molecular signatures that differentiate metastatic from non-metastatic status. Each row represents a gene, and each column a sample, with color intensity reflecting the log2 expression level as indicated by the color scale (blue: low expression, red: high expression).*

A graph of different colored squares

Description automatically generated with medium confidence

***Figure S2:*** *Stability of GLMQL-RMAS When Applying RMAS to Raw p-Values. This figure illustrates the consistent performance of the GLMQL-RMAS methodology in identifying significant genes, demonstrating its robustness by applying the RMAS to raw p-values, a testament to the method's reliability across different statistical treatments.*

A graph of different colored squares

Description automatically generated with medium confidence

***Figure S3:*** *Effectiveness of GLMQL-MAS with Bonferroni Correction Method. This figure illustrates the performance of the GLMQL-MAS when implementing a Bonferroni correction rather than the Benjamini-Hochberg approach.*

A diagram of a test

Description automatically generated with medium confidence

***Figure S4.*** *LDA1 projection illustrating the separation of ALNM+ and ALNM- samples using the top-n GLMQL-MAS selected genes, with n ranging from 20 to 100. Perfect separation (100%) is achieved at n=100, highlighting the discriminative power of the selected genes in distinguishing between the two groups.*

|  |  |  |
| --- | --- | --- |
| A graph with red and blue dots  Description automatically generated | A graph of a number of dots  Description automatically generated with medium confidence | A graph with red and blue dots  Description automatically generated |

***Figure S5.*** *Logistic regression decision boundaries depicted using TSNE1 and TSNE2 coordinates for samples projected with top-n GLMQL-MAS selected genes. From left to right, the panels represent n=100, n=200, and n=300, showcasing the progressive improvement in differentiating ALNM+ from ALNM- samples through the GLMQL-MAS gene selection process.*

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

***Figure S6.*** *Distribution of the Top 10 Genes by BH-Significance Occurrence Using GLMQL-MAS. This figure illustrates the top-10 genes that achieved the highest frequency of BH-significance in an analysis of 500 iterations, with |LogFC| upper thresholds set at 1, 2, 3, and 4. It highlights the genes that consistently demonstrate significant differential expression in lymph node metastasis of breast cancer.*

A graph of data with different colored lines

Description automatically generated with medium confidence

***Figure S7.*** *Top 10 GO Processes Related to Lymph Node Metastasis. This figure details the GO processes most intimately connected with the pathology of lymph node metastasis, providing insights into the molecular functions and cellular components affected.*

A graph of a graph

Description automatically generated with medium confidence

***Figure S8.*** *This figure illustrates the significant pathways identified from the GSEA using Hallmark gene sets, highlighting the predominant biological mechanisms influenced by the GLMQL-MAS selected genes between 42 ALNM positive (ALNM+) and 65 ALNM negative (ALNM-) breast cancer samples.*

|  |
| --- |
| A screen shot of a chart  Description automatically generated |

***Figure S9.*** *Significant GO Processes for Upregulated BH-Significant Genes. Showcasing the top q-value-based GO processes for genes identified as BH-significant and upregulated in at least 250 of 500 iterations. This visualization emphasizes the biological processes most consistently associated with lymph node metastasis.*

A graph of a bar graph

Description automatically generated with medium confidence

***Figure S10:*** *Gene Set Enrichment Analysis (GSEA) of Hallmark Gene Sets. This figure illustrates the enrichment results for Hallmark gene sets in the context of lymph node metastasis. Each bar represents a specific pathway, highlighting the pathways most significantly associated with the genetic changes observed in the ALNM+ group compared to the ALNM- group, based on q-values. This analysis underscores the biological relevance of the identified genes and their roles in critical cancer-related pathways.*

A graph of a bar graph

Description automatically generated with medium confidence

***Figure S11****.* ***Gene Set Enrichment Analysis (GSEA) Using Hallmark Gene Sets for Common Genes Across Two Cases:*** *This figure presents the results of the Gene Set Enrichment Analysis (GSEA) using Hallmark gene sets for genes that are common between two comparative cases: Case 1 and Case 2. It highlights significant biological pathways that are consistently implicated across different sample sets, elucidating potential mechanisms of lymph node metastasis in breast cancer.*