Transcriptomic Profiling of Melanoma Progression: Unsupervised and Supervised Analysis of the GSE8401 Dataset.

Caterina Sanetti

Jun 7/2025

Abstract

Cutaneous melanoma is among the most lethal skin cancers, characterized by high metastatic potential and rising global incidence. This study utilizes the GSE8401 dataset to dissect transcriptomic differences between primary and metastatic melanoma using an integrative bioinformatics framework. Unsupervised techniques (PCA, k-means, hierarchical clustering) and supervised models (Random Forest, LDA, Lasso regression) were employed to identify gene expression signatures predictive of metastasis. Distinct separation between tumor stages emerged across clustering approaches, highlighting transitional expression patterns in select metastatic samples. Feature selection revealed genes involved in cytoskeletal organization, spliceosomal function, and extracellular matrix remodeling. Functional enrichment underscored critical pathways such as keratinization and estrogen signaling, implicating their roles in metastatic adaptation and plasticity. These findings offer insight into the molecular mechanisms of melanoma progression and identify candidate biomarkers for early metastatic transition.

---

1. Introduction

Melanoma is a malignant neoplasm arising from melanocytes, the pigment-producing cells primarily located in the basal layer of the epidermis but also found in hair follicles, mucosal surfaces, and the uveal tract, and it is recognized as the most aggressive form of skin cancer due to its high metastatic potential and associated mortality [2]. Indeed, cutaneous melanoma remains one of the most aggressive malignancies, with global incidence rates increasing by 44% between 1990 and 2020. While early-stage disease is associated with a favorable prognosis, metastatic dissemination leads to devastating outcomes, reducing median survival to just 6–9 months and limiting 5-year survival rates to below 5% [7]. To investigate the molecular mechanisms driving this aggressive progression, the GSE8401 dataset was developed through comparative transcriptomic analysis of human melanoma cell lines with differing metastatic potential, selected via rigorous in vivo serial passaging in immunodeficient mouse models.

This study analyzes the GSE8401 dataset to characterize transcriptomic alterations associated with melanoma progression. An integrative framework combining exploratory analysis, unsupervised clustering, supervised machine learning and network analysis , is applied to uncover molecular subtypes and identify gene expression patterns predictive of metastasis. Functional enrichment of selected features further contextualizes the results within biological pathways relevant to tumor evolution and metastatic potential.

2. Methods

The gene expression dataset GSE8401 was retrieved from the Gene Expression Omnibus (GEO) using the GEOquery package in R. The dataset consisted of 83 samples, including 31 primary tumors and 52 melanoma metastases, profiled on the Affymetrix HG-U133A platform. Raw expression values were log2-transformed using log1p to stabilize variance, followed by scaling to standardize the data for downstream

analyses. Principal Component Analysis (PCA) was performed using prcomp to assess sample separation and identify potential outliers. Samples were visualized in PCA space and colored by tumor type (primary vs. metastasis). To evaluate sample grouping, both K-means clustering (k=2) and hierarchical clustering (maximum distance with average linkage) were applied. For hierarchical clustering, all possible combinations of dissimilarity measures and agglomeration methods were systematically tested. The optimal configuration was selected based on alignment with ground truth labels.

Using t-tests with a significance threshold of $p < 0.05$, 8680 differentially expressed genes (DEGs) from 22283 initial genes were identified.

Predictive models were trained on the subset of differentially expressed genes using cross-validation (implemented via the caret and glmnet packages). Three classifiers—Linear Discriminant Analysis (LDA), Random Forest (RF), and Lasso regression—were evaluated for their classification accuracy.

Model performance was assessed through repeated cross-validation (5 repeats of 10-fold), with accuracy as performance metrics.

For additional validation, the LDA model was assessed also with a hold-out procedure by retaining 15 samples from each group for testing; while the RF model was also trained on the DEGs, using 1000 decision trees.

For downstream analysis, subsets of differentially expressed genes were selected from each method using specific criteria. Random Forest identified the top 25 most important genes based on variable importance scores from the 1000-tree model; for LDA, the 200 genes with the strongest discriminatory power for the metastatic class (based on absolute coefficient magnitude) were selected from the cross-validated model. While for Lasso regression, 25 genes with non-zero coefficients at the optimal lambda value were retained.

SCUDO (Signature-based Clustering for Diagnostic and Prognostic Understanding) was implemented using the rScudo package to identify sample-specific gene signatures. For each sample, signatures were derived from the top and bottom 25 most differentially expressed genes ($\alpha = 0.05$), with network analysis (threshold $N = 0.2$) and classification performance evaluated through hold-out validation and spin-glass clustering (spins = 2).

Functional enrichment of overlapping top-ranked genes from LDA and Lasso, was performed using g:Profiler. For network -based analysis, separate analyses were conducted for upregulated and downregulated top-ranked genes to identify metastasis-specific biological processes. Pathway networks were constructed with pathfindR. Generating clustering enriched terms graphs and visualizing gene-pathway interactions.

To further characterize these networks, a protein-protein interaction (PPI) network was generated using downregulated genes identified from LDA-selected metastatic cell data through STRING database. The resulting PPI network was then analyzed using Cytoscape with cytoHubba plug-in to extract key network topological measures.

3. Results

3.1 Exploratory analysis

The gene expression dataset was initially examined for quality control and preprocessing. A histogram of the raw data revealed a non-normal distribution, with a significant number of low-expression values and a long right tail, suggesting the need for transformation. (Figure 1)

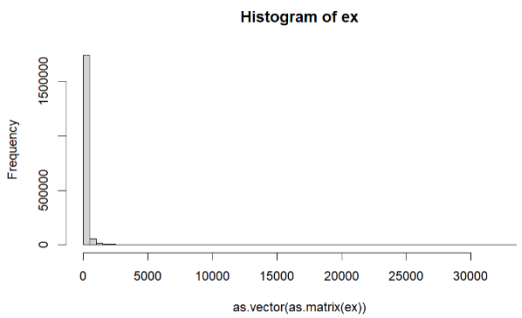


Figure 1 : Distribution of raw gene expression values across all samples, showing a right-skewed pattern with abundant low-expression values. The long tail indicates the need for log-transformation to normalize the data.

To normalize the data, a log1p transformation was applied to stabilize variance and reduce skewness. Following this, the data were scaled to facilitate comparative analysis across samples. The transformed and scaled data displayed a more symmetric distribution, as confirmed by the boxplot. Indeed, the boxplot (Figure 2) of the processed data demonstrated improved comparability between samples, with reduced technical variability.

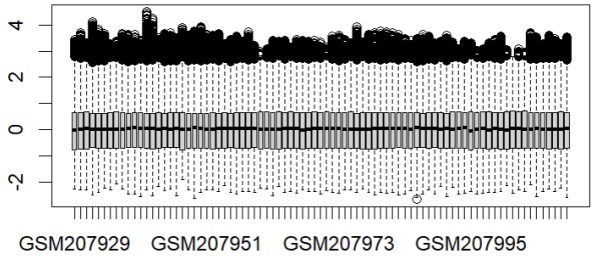


Figure 2 : Boxplot of pre-processed data. Post-transformation (log1p + scaling) expression values demonstrating reduced skewness and improved comparability across samples. Boxes represent interquartile ranges (IQR), with whiskers extending to 1.5× IQR.

3.2 Unsupervised Learning

3.2.1 PCA

Principal component analysis of the normalized expression data revealed a gradual decrease in variance contribution across components, as evidenced by the scree plot (Figure 3). The first principal component (PC1) accounted for 12.17% of the total variance (standard deviation = 20.85), while PC2 explained an additional 8.33% (SD = 17.25). Together, these first two components captured 20.50% of the cumulative variance in the dataset, with subsequent components contributing progressively smaller proportions. Visualization of the first two principal components demonstrated distinct clustering of primary (blue) and metastatic (red) samples, with minimal overlap between groups (Figure 4). This separation indicates significant differences in global gene expression patterns between primary and metastatic samples. Examination of PC1 versus PC3 revealed partial overlap between groups. (Supplementary 1)

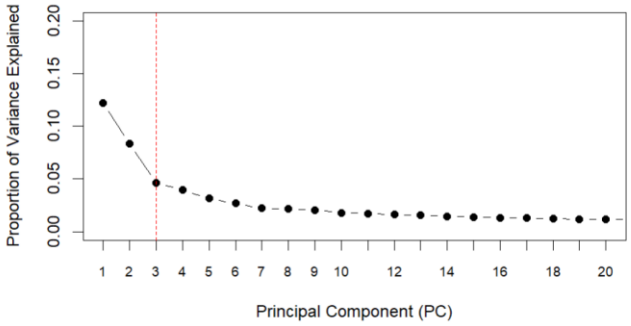


Figure 3 : *Scree plot of PCA variance. Proportion of variance explained by first 20 principal component (PC). PC1 (12.17%) and PC2 (8.33%) collectively capture 20.5% of total variance, with subsequent PCs contributing diminishing proportions. Red line at the elbow.*

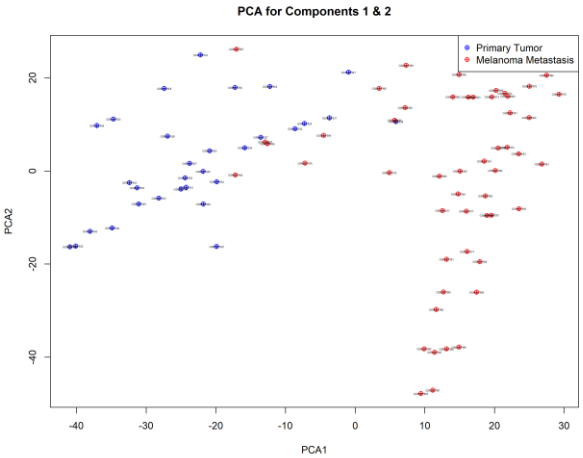


Figure 4 : *PCA plot (PC1 vs. PC2). Separation of primary (blue) and metastatic (red) samples along the first two principal components. Minimal overlap suggests distinct transcriptomic profiles between disease states.*

3.2.2 K -means

The k-means clustering (k=2) of log-transformed expression data effectively separated samples into two distinct groups that largely corresponded to their tumor progression status. Cluster 1 contained 48 samples, with 46 correctly classified as metastases and only 2 misclassified primaries. Cluster 2 comprised 35 samples, including 29 true primaries and 6 misclassified metastases. Visualization in PCA space (PC1 vs PC2) showed clear spatial separation between clusters, with most misclassified samples located near the decision boundary, suggesting that their expression profiles exhibit transitional features, sharing characteristics of both primary and metastatic tumors (Figure 5) .

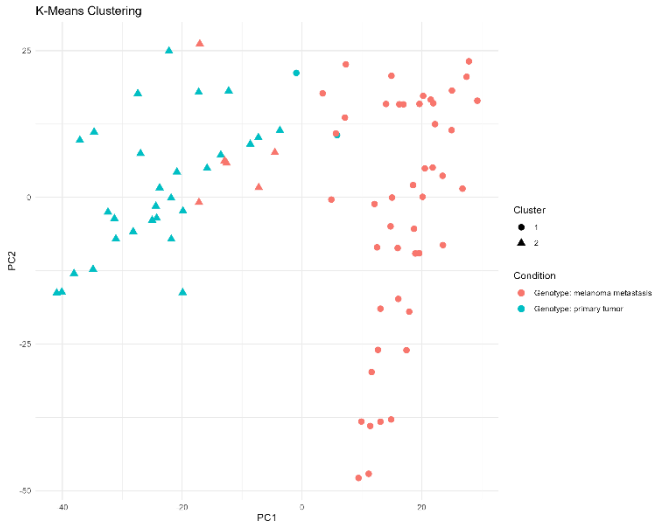


Figure 5 : *K-means clustering (k=2) of log-transformed expression data separates samples into primary (skyblue) and metastatic (salmon) groups. Cluster 1 contains 46/52 metastases, while Cluster 2 captures 29/31 primaries.*

3.2.3 Hierarchical clustering

To assess the robustness of sample clustering, different combinations of distance measures (Euclidean, Manhattan) and linkage methods (Ward.D2, average, complete, single) were evaluated (Supplementary 2). The optimal combination— maximum distance with average linkage —was selected as it produced the most biologically meaningful separation between primary tumor and metastatic samples. Hierarchical clustering based on the selected parameters revealed two main clusters: one composed of all the primary tumor samples (n=31) and another dominated by metastatic samples (n=45). However, seven metastatic samples were misclassified into the primary tumor cluster. (Figure 6). Notably, five of these—GSM207967, GSM207994, GSM207999, GSM208006, and GSM208007—were consistently misclassified also by k-means clustering, suggesting intrinsic similarities to primary tumors.

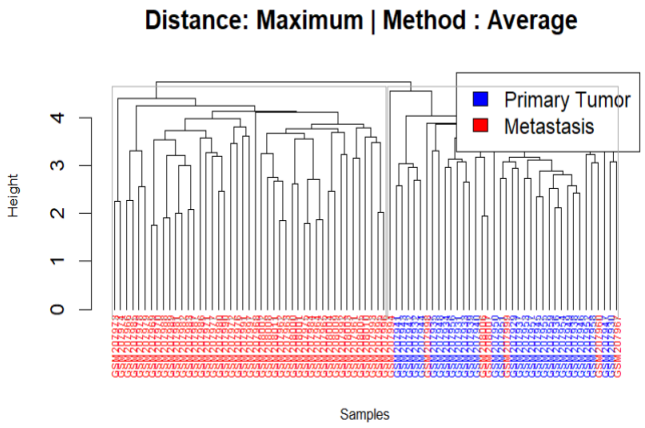


Figure 6 *: Dendrogram from hierarchical clustering. Hierarchical clustering (maximum distance + average linkage) reveals two dominant branches: primary tumors (blue) and metastases (red).*

The spatial distribution of misclassified samples in both clustering methods reveals biologically meaningful patterns. Most misclassifications (4/5) occur near the decision boundary (Figure 7) , strongly suggesting these represent transitional states where metastatic tumors retain some molecular signatures of their primary origins. In contrast, GSM207967 appears far from the decision boundary, suggesting either biological ambiguity or technical factors

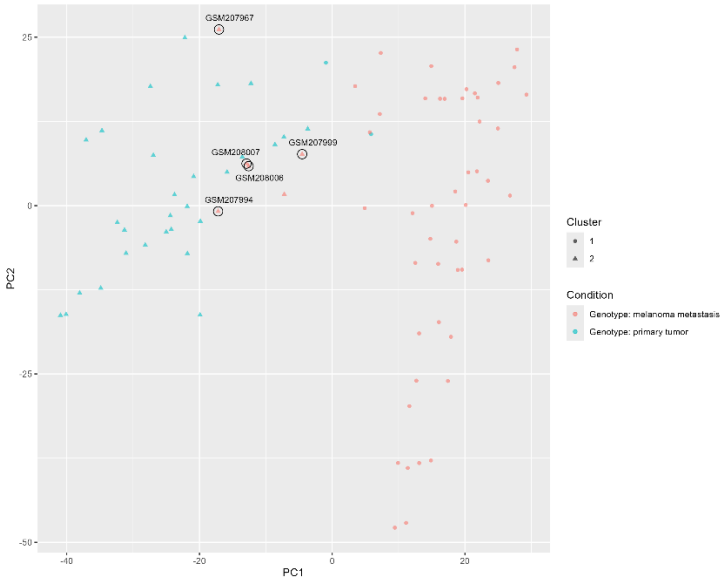


Figure 7: *PCA projection highlighting misclassified samples near the decision boundary (PC1 vs. PC2). Their intermediate positions suggest hybrid expression profiles with features of both primary and metastatic tumors***.**

3.3 Supervised Learning

The random forest model, trained with 1000-trees , effectively distinguished between primary and metastatic tumor samples based on gene expression patterns. The error rate plot demonstrated the model's stability, showing that classification accuracy improved rapidly with increasing trees and plateaued after approximately 400 trees, maintaining a consistently low error rate between 0.05 and 0.10 (Figure 8) . The feature importance analysis identified the first 200 genes to have the strongest predictive power (importance scores: 0.2–0.4). While genes ranked 201–500 exhibited moderate contributions (scores: 0.1–0.2), the remaining genes (beyond 500) had negligible importance (Figure 9).

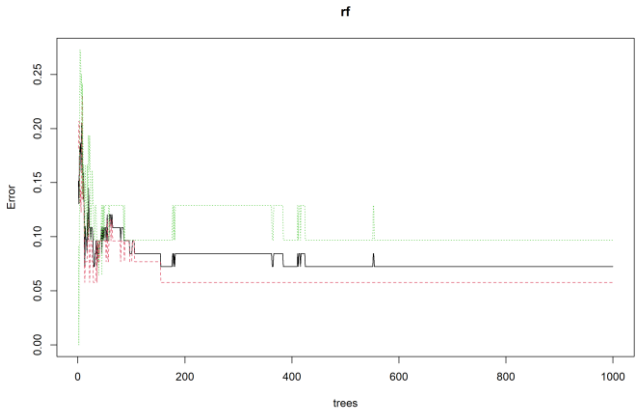


Figure 8 : *Random Forest error rate***.** *Stabilization of classification error (~0.05–0.10) beyond 400 trees, indicating model convergence.*

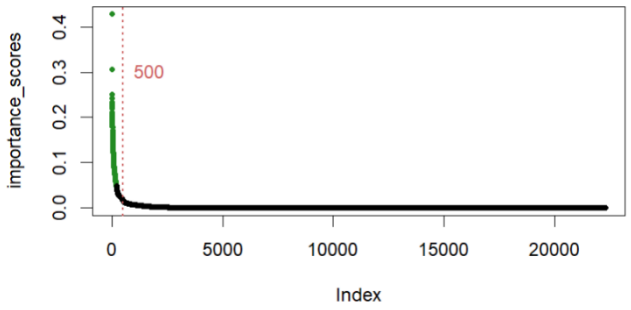


Figure 9 : *Random Forest Importance Plot. Top 200 genes (scores: 0.2–0.4) dominate classification.*

The heatmap analysis of the first 25 most important genes, reveals a distinct expression pattern where metastatic samples tend to show downregulation, while primary tumors exhibit upregulation of these key genes. Notably, three primary tumor samples (GSM207947, GSM207930, and GSM207952) clustered with metastatic lesions. Of these, two samples (GSM207947 and GSM207930) were also misclassified by k-means clustering, while GSM207952 was correctly assigned to the primary tumor group in the k-means analysis. Conversely, three metastatic samples (GSM208007, GSM207994, and GSM208006) clustered with primary tumors, all of which had been consistently misclassified by other analytical methods. The heatmap visualization demonstrated these samples exhibit intermediate expression levels of the discriminatory genes - showing neither the characteristic upregulation of typical primary tumors nor the complete downregulation seen in metastases (Figure 10). This intermediate molecular signature likely explains their recurrent misclassification, as their hybrid expression profile falls between the discrete categories defined by clustering algorithms.

This observation was reinforced by PCA, where these samples—along with their nearest phylogenetic neighbours mapped near the decision boundary between the two groups (Figure 11).

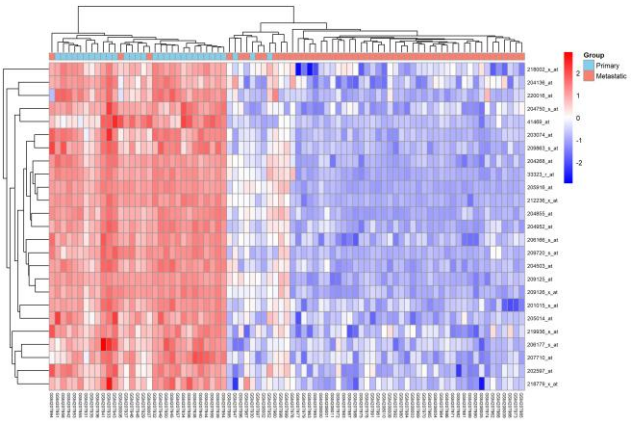


Figure 10 : *Heatmap of top 25 discriminatory genes , Metastatic samples (salmon) show downregulation (blue) while primaries (skyblue) exhibit upregulation (red) of key genes. Misclassified samples display intermediate expression*

(white).

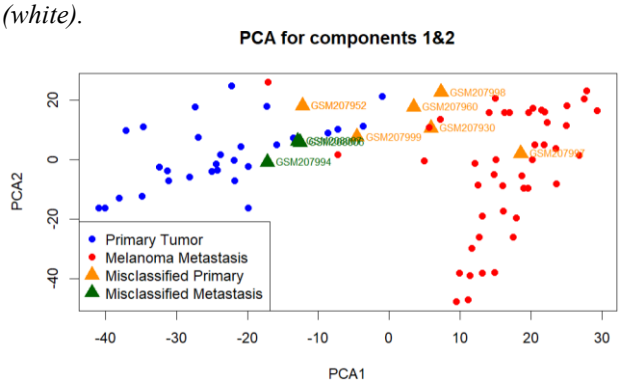


Figure 11 : *PCA of misclassified primary tumors and their nearest neighbors.*

The Linear Discriminant Analysis (LDA) model was trained on a subset of significantly differentially expressed genes. The training set consisted of a balanced subset, with 15 samples from each group held out for testing. On the test set, the model achieved an accuracy of 93.3%, correctly classifying 15 Primary and 13 Metastatic samples (Figure 12) . Receiver operating characteristic (ROC) analysis yielded an area under the curve (AUC) of 0.93733 (Supplementary 3) , reflecting strong discriminatory power.

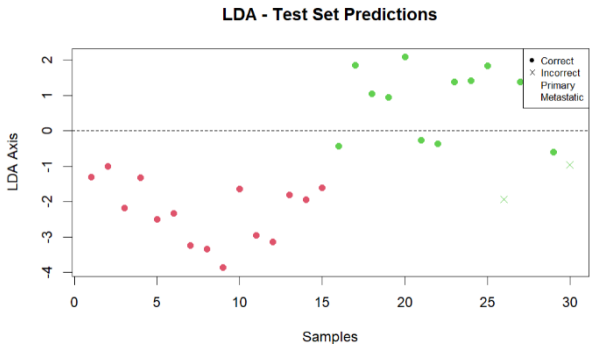


Figure 12: *Linear Discriminant Analysis (LDA) predictions on the held-out test set (n=30, 15 primary/15 metastatic).*

For the comparative evaluation of classification approaches, all three models (LDA, Lasso regression, and Random Forest) were trained and tested under identical conditions using repeated 10-fold cross-validation (5 repeats) on the full set of differentially expressed genes

The cross-validation results revealed LDA as the top-performing classifier with a mean accuracy of 0.959, and a median accuracy of 1.0, indicating perfect classification in the majority of cases. The model maintained this perfect performance in at least 75% of folds (3rd quartile = 1.0). Lasso regression followed closely with a mean accuracy of 0.950, while Random Forest trailed with 0.919 mean accuracy. (Figure 13)

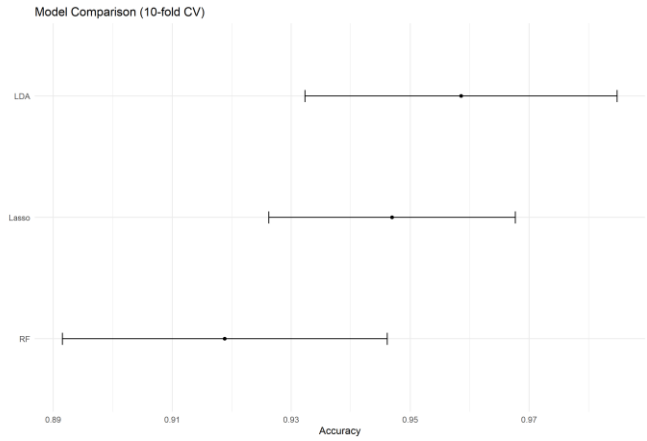


Figure 13 : *Model accuracy comparison. Mean classification accuracies across 10-fold cross-validation: LDA (95.9%), Lasso (95.0%), and Random Forest (91.9%). Error bars represent variability over 5 repeats*

From the best-performing LDA model, where extracted the top 200 discriminatory genes based on their metastatic classification importance scores (Figure 15). Among these, 33322_i_at emerged as the highest-ranked gene (importance score = 100), demonstrating the strongest contribution to distinguishing primary from metastatic tumors. The top-ranked genes (33322_i_at, 33323_r_at, and 212236_x_at) all exhibited importance scores >99, indicating their exceptional discriminatory power.

From the Lasso model at optimal lambda, were identified 25 genes with non-zero coefficients , with 11 showing positive associations (strongest in 220016_at) and 14 showing negative associations (strongest in 221041_s_at , followed by 200751_s_at) (Figure 14).

Among these 25 genes, nine ("212236_x_at" ,"220016_at" , "209125_at" , "209800_at" , "222223_s_at" , "209885_at" , "200751_s_at" "201042_at" , "207549_x_at ") overlapped with those identified by LDA.

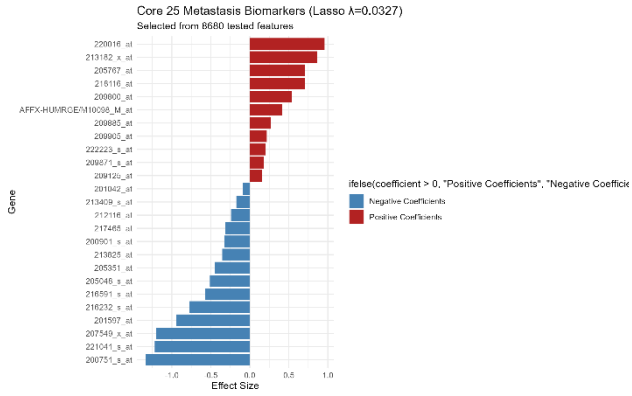


Figure 14 : *The plot shows 25 non-zero coefficient genes at the optimal lambda value, with 11 genes positively associated with metastasis (red) and 14 negatively associated (blue).*

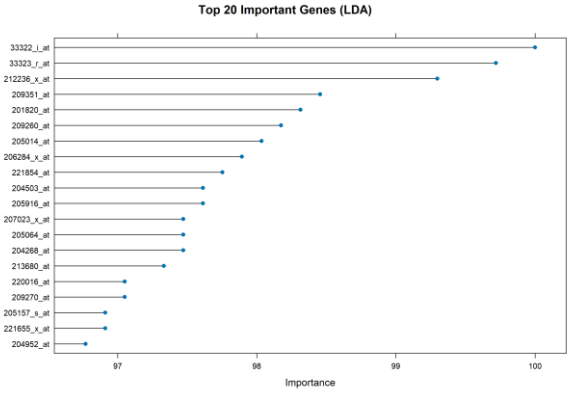
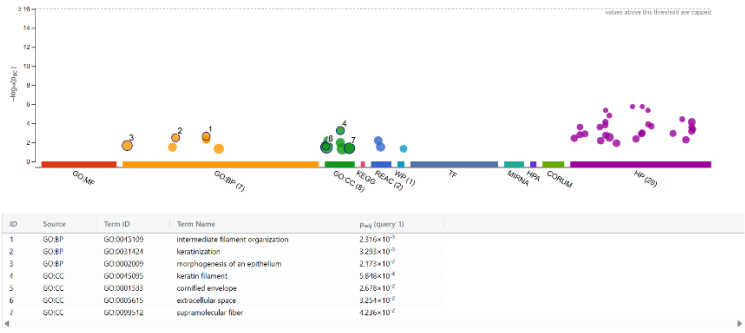


Figure 15 : *Ranked importance scores of genes identified by Linear Discriminant Analysis. Top candidates (scores >0.3) are potential biomarkers for metastasis.*

The SCUDO analysis also, effectively classified tumor samples into Metastatic and Primary groups with high accuracy (95.12%). Network visualization and community detection demonstrated robust separation between metastatic and primary tumors, with minimal misclassification (Supplementary 4).

3.5 Enrichment Analysis

Functional enrichment analysis of the nine shared genes between the LDA and Lasso models showed significant enrichment for biological processes and cellular involved in epithelial integrity and cytoskeletal organization. The most significantly enriched biological process was intermediate filament organization (GO:BP, p-value = 2.316 x 10^{-3}), Followed by keratinization with p-value of 3.293 x 10^{-3} and morphogenesis of an epithelial. While at cellular component level, keratin filament (GO:CC, p-value = 5.848 x 10^{-4}) emerged as the most strongly enriched term (Figure 16).



ID	Source	Term ID	Term Name	P_{adj} (query 1)
1	GO:BP	GO:0045109	intermediate filament organization	2.316×10^{-3}
2	GO:BP	GO:0031424	keratinization	3.293×10^{-3}
3	GO:BP	GO:0002009	morphogenesis of an epithelium	2.173×10^{-2}
4	GO:CC	GO:0045095	keratin filament	5.848×10^{-4}
5	GO:CC	GO:0001533	cornified envelope	2.676×10^{-2}
6	GO:CC	GO:0005615	extracellular space	3.254×10^{-2}
7	GO:CC	GO:0099512	supramolecular fiber	4.236×10^{-2}

Figure 16 *: Top enriched terms for shared LDA/Lasso genes, including intermediate filament organization and keratinization, implicating cytoskeletal remodeling in metastasis.*

Pathway analysis was performed using pathfindR, a network-based approach that identifies functionally enriched pathways by integrating protein-protein interaction data with gene expression patterns. For LDA-upregulated genes, the most significantly enriched pathways included the spliceosome (hsa03040), protein processing in the endoplasmic reticulum (hsa04141), and SNARE interactions in vesicular transport (hsa04130), with key genes such as HNRNPC, SRSF1, and HSP90B1 (Figure 17-a) . In contrast, LDA-downregulated genes were strongly associated with ECM-receptor interactions (hsa04512), central carbon metabolism in cancer (hsa05230), and Staphylococcus aureus infection (hsa05150) (Figure 17-b).

Comparative analysis (Figure 18) revealed significant overlap between the key genes identified by LDA and Lasso, with four particularly enriched pathways: the spliceosome pathway (hsa03040), Salmonella infection (hsa05132), cell cycle regulation (hsa04110) and estrogen signaling (hsa04915).

The protein-protein interaction (PPI) network constructed from LDA-identified downregulated genes demonstrated significant biological relevance, as evidenced by its strong clustering (average clustering coefficient = 0.53) and highly significant PPI enrichment (p < 1.0e-16) (Figure 19) .

Hub genes were identified using three centrality measures. Degree centrality analysis revealed PKP1, KRT5, and KRT14 as the most highly connected nodes (45, 44, and 43 interactions respectively), followed by LOR and IVL (41 interactions each). Closeness centrality identified KRT5, KRT14, PKP1, IVL and LOR as the most centrally positioned genes. Betweenness centrality showed EGFR as the top-ranked hub, followed by KRT5, KRT14, KLF4 and CTSG. (Figure 20).

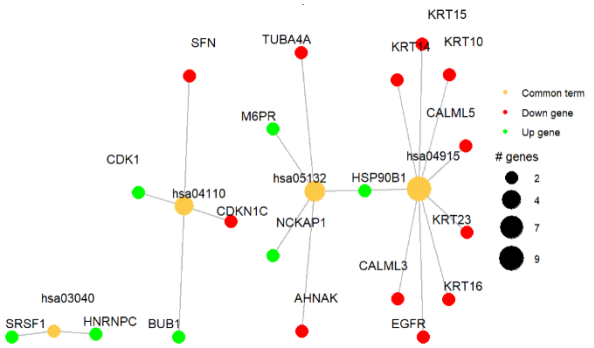
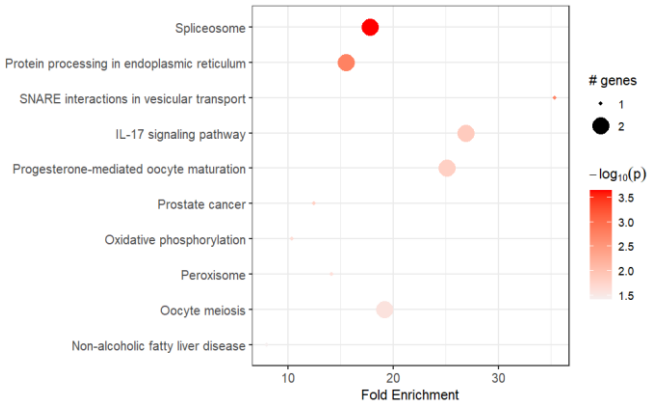


Figure 18 : *Shared enriched pathways between LDA and LASSO top genes. The four overlapping pathways include the spliceosome (hsa03040), Salmonella infection (hsa05132), cell cycle regulation (hsa04110), and estrogen signaling (hsa04915).*

a



b

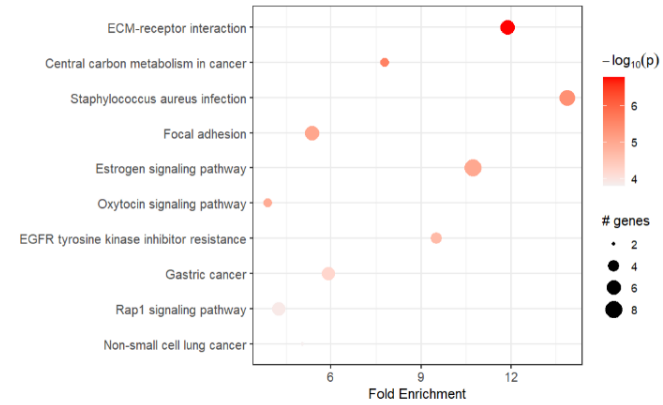


Figure 17 : *Pathway enrichment of LDA selected - upregulated (a) and LDA selected -downregulated (b) genes*

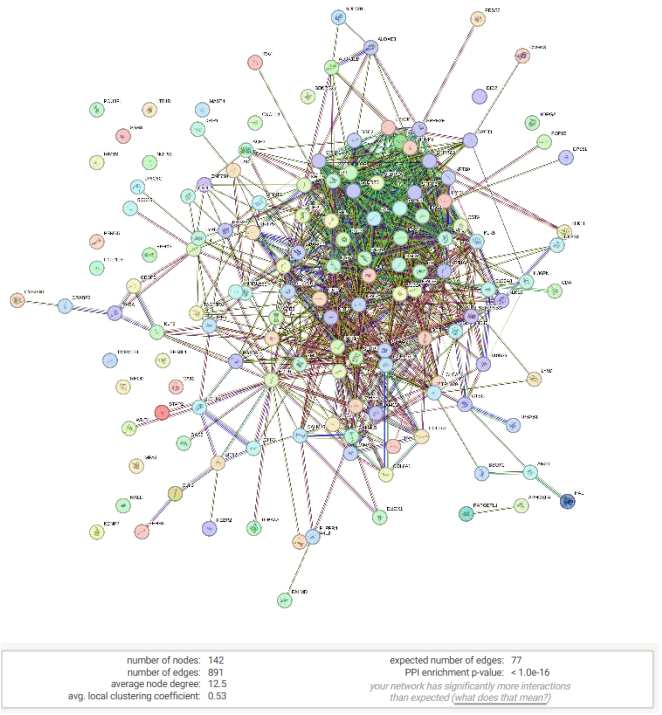


Figure 19 : *PPI interaction network constructed with the LDA – downregulated genes, generated using STRING.*

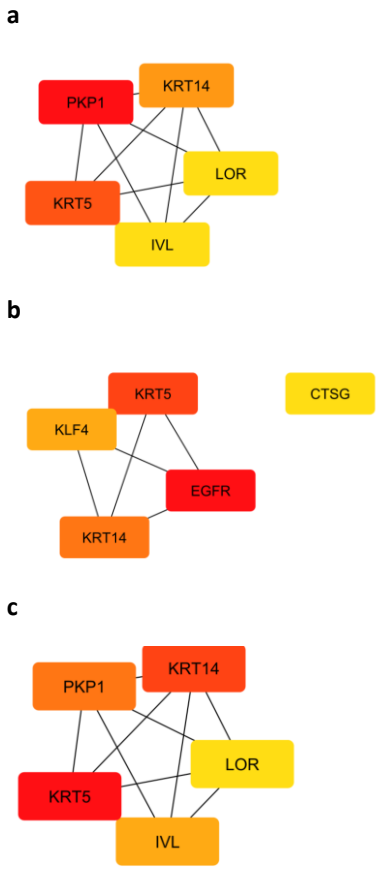


***Figure 20**: Network centrality analysis of key hub genes using three centrality measures: a) degree b) betweenness c) closeness. Color intensity reflects centrality scores (red = highest, yellow = lower).*

4. Discussion and conclusion

This study systematically analyzed the GSE8401 melanoma dataset, findings reveal coordinated transcriptional reprogramming during melanoma progression. Exploratory analyses demonstrated robust separation between primary and metastatic samples, with unsupervised methods revealing a subset of metastases exhibiting transitional expression profiles. These samples, consistently misclassified across clustering algorithms, may represent biologically distinct early metastatic variants, warranting further investigation.

The strong discriminatory power of supervised models—particularly LDA (95.9% accuracy) and Lasso (95.0%)—confirms that metastatic progression is associated with reproducible transcriptomic alterations.

Functional enrichment of the nine shared genes between LDA and Lasso revealed intermediate filament organization as the top enriched term, underlying the metastatic cellular need to restructure the cytoskeletal architecture to enable the morphological plasticity needed for invasion and motility. Moreover, intermediate filaments not only maintain cell integrity and mechanical stability but also participate in signal transduction pathways that regulate cell survival and migration. In melanoma, alterations in the organization of these filaments are associated with the aggressive behavior of cancer cells, facilitating their ability to detach, invade, and survive in distant microenvironments [6, 5].

The second most enriched term, keratinization, suggests the remodeling of keratinocyte characteristics and reactivation of keratin gene programs in metastatic samples. This process likely represents a re-adaptation of differentiation pathways whereby melanoma cells modify their epithelial features to acquire increased motility and invasive capacity [4]

LDA - selected upregulated genes were enriched for spliceosome and protein processing pathways (e.g., HNRNPC, SRSF1). This finding aligns with extensive evidence that aberrant RNA splicing is a hallmark of cancer, with tumors exhibiting up to 30% more alternative splicing events than normal tissues, often generating cancer-specific isoforms that drive malignant progression[8] .In fact , the HNRNPC's overexpression has been experimentally shown to drive metastatic phenotypes through multiple mechanisms: by stabilizing EMT factors (ZEB1/2), altering polyadenylation (APA), and rewiring mRNA processing. [10].

The coordinated upregulation of ER protein processing and SNARE-mediated vesicular transport reveals a dual adaptation strategy in metastatic cells. To counteract the stress and meet the functional demands of invasion and colonization, metastatic cells upregulate genes involved in protein processing in the endoplasmic reticulum (ER) to enhance their protein folding, quality control, and secretory functions, allowing these cells to sustain aggressive growth and invasion [10]. Simultaneously, elevated SNARE mediated vesicular

transport facilitates efficient exosome release, ensuring that metastatic melanoma cells can deliver oncogenic signals that remodel the extracellular matrix and promote cell invasion [1]

Concurrently those cells downregulate the ECM-receptor interaction pathway, reducing their affinity for the surrounding matrix, thereby facilitating their detachment and migration. [12]

Network analysis reveals that this process may be further amplified by the downregulation of three key hub genes-KRT5, KRT14, and PKP1-which are shared across multiple centrality measures. Normally, these genes support the mechanical stability of the epithelium: KRT5 and KRT14 stabilize intermediate filaments, while PKP1 assembles desmosomal complexes to reinforce cell-cell adhesion. However, in metastatic cells, the downregulation of these genes is thought to promote the loss of cell–cell adhesion, increased cellular motility, and ultimately, tumor dissemination. [12]

Of particular interest, the convergence of LDA and Lasso on genes associated with the estrogen signaling pathway, as it suggests a modulatory role of hormone signaling in melanoma metastasis. Although melanoma is traditionally viewed as hormone-insensitive, growing evidence highlights the protective role of ERβ and GPER, which can inhibit metastasis via immune modulation and tumor cell differentiation. Loss of ERβ or imbalanced ERα/ERβ expression may drive aggressiveness [9].

Perhaps most unexpectedly, the enrichment of the Salmonella infection pathway was observed. While this does not likely indicate an active infection it rather reflects the mimicry of cellular processes that are typically activated during such infections by metastatic melanoma cells. Indeed, metastatic cells frequently undergo alterations in signaling, immune modulation, cytoskeletal reorganization (via TUBA4A/AHNAK downregulation), vesicular trafficking (via M6PR/NCKAP1 upregulation), mirroring pathogen-induced host adaptations.

Notably, the spliceosome pathway emerged also as a shared pathway, confirming the importance of aberrant RNA splicing as a hallmark of cancer progression.

Lastly, the cell-cycle pathway shared enrichment reveals how metastatic melanoma subverts normal proliferation controls through coordinated alterations in key regulators: CDKN1C downregulation removes critical proliferation brakes, BUB1 upregulation promotes chromosomal instability, CDK1 overexpression drives unchecked G2/M progression, and downregulation of SFN compromises DNA damage responses. This collective dysregulation creates a permissive environment for uncontrolled proliferation by simultaneously disabling cell cycle checkpoints while hyperactivating progression signals, enabling metastatic spread despite genomic stress. [3]

In conclusion, this study reveals how metastatic melanoma cells coordinately rewire cytoskeletal organization, RNA splicing, secretory pathways and cell cycle to drive aggressive dissemination. The identification of transitional metastatic states and conserved pathway alterations provides new insights into melanoma plasticity.

References

1. Boussadia, Z., Lamberti, J., Mattei, F. et al. Acidic microenvironment plays a key role in human melanoma progression through a sustained exosome mediated transfer of clinically relevant metastatic molecules. J Exp Clin Cancer Res 37, 245 (2018). https://doi.org/10.1186/s13046-018-0915-z

2. Davey, M. G., Miller, N., & McInerney, N. M. (2021). A Review of Epidemiology and Cancer Biology of Malignant Melanoma. Cureus, 13(5), e15087. https://doi.org/10.7759/cureus.15087

3. Denaro, M. (2018). Molecular characterization of noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP): Analysis of miRNA expression and mutational status

4. Han, W., Hu, C., Fan, Z. J. et al. Transcript levels of keratin 1/5/6/14/15/16/17 as potential prognostic indicators in melanoma patients. Sci Rep 11, 1023 (2021). https://doi.org/10.1038/s41598-020-80336-8

5. Han, W., Xu, W.-H., Wang, J.-X., Hou, J.-M., Zhang, H.-L., Zhao, X.-Y., & Shen, G.-L. (2020). Identification, validation, and functional annotations of genome-wide profile variation between melanocytic nevus and malignant melanoma. BioMed Research International, 2020, Article 1840415. https://doi.org/10.1155/2020/1840415

6. Holle, A. W., Kalafat, M., Sales Ramos, A., Seufferlein, T., Kemkemer, R., & Spatz, J. P. (2017). Intermediate filament reorganization dynamically influences cancer cell alignment and migration. Sci Rep, 7, 45152. https://doi.org/10.1038/srep45152

7. Li, F. Z.; Dhillon, A. S.; Anderson, R. L.; McArthur, G.; Ferrao, P. T. Phenotype switching in melanoma: Implications for progression and therapy. Front. Oncol. 2015, *5*, 31. https://doi.org/10.3389/fonc.2015.00031

8. Lv, X., Sun, X., Gao, Y. et al. Targeting RNA splicing modulation: new perspectives for anticancer strategy?. J Exp Clin Cancer Res 44, 32 (2025). https://doi.org/10.1186/s13046-025-03279-w

9. Marrapodi, R., & Bellei, B. (n.d.). The keratinocyte in the picture cutaneous melanoma microenvironment. Laboratory of Cutaneous Physiopathology and Integrated Center of Metabolomics Research, San Gallicano Dermatological Institute, IRCCS.

10. Mo, L., Meng, L., Huang, Z. et al. An analysis of the role of HnRNP C dysregulation in cancers. Biomark Res 10, 19 (2022). https://doi.org/10.1186/s40364-022-00366-4

11. Szadai, L., Velasquez, E., Szeitz, B., de Almeida, N. P., Domont, G., Betancourt, L. H., Gil, J., Marko-Varga, M., Oskolas, H., Jánosi, Á. J., Boyano-Adánez, M. d. C., Kemény, L., Baldetorp, B., Malm, J., Horvatovich, P., Szász, A. M., Németh, I. B., & Marko-Varga, G. (2021). Deep proteomic analysis on biobanked paraffin-archived melanoma with prognostic/predictive biomarker read-out. Cancers, 13(23), 6105. https://doi.org/10.3390/cancers13236105

12. Xie, R., Li, B., Jia, L., & Li, Y. (2022). Identification of Core Genes and Pathways in Melanoma Metastasis via Bioinformatics Analysis. International Journal of Molecular Sciences, 23(2), 794. https://doi.org/10.3390/ijms23020794

Supplementary

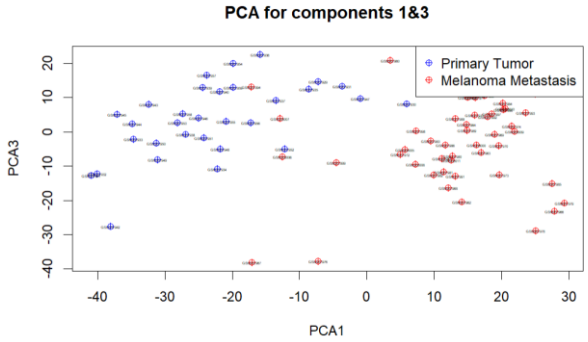


Figure 1: *Principal Component Analysis (PC1 vs. PC3)*. PCA plot showing sample distribution across PC1 and PC3. Partial overlap between primary and metastatic samples along PC3

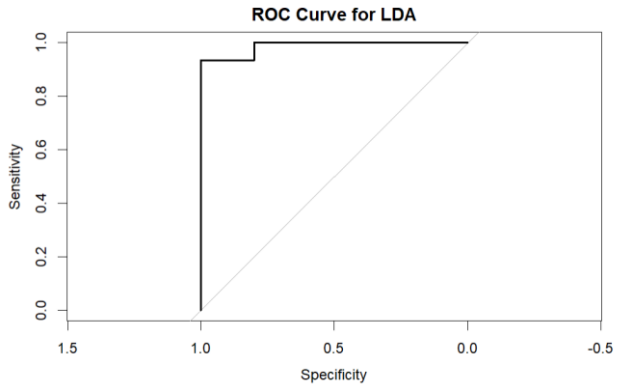
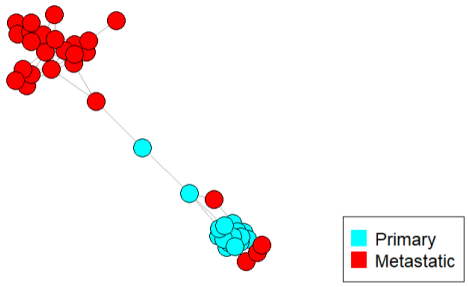


Figure 3 : *ROC Curve for LDA Model* .



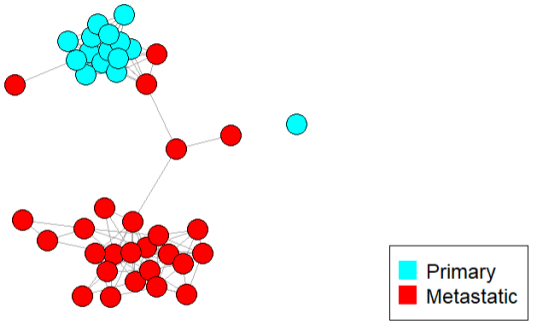


Figure 4: Train (a) and test (b) network generated by SCUDO