

Network-based Data Analysis: Identification of Autoantibody Biomarkers for Lung Cancer Using Protein Microarrays

Seyed Pooria Sajadi Parsa
`seyed.sajadiparsa@studenti.unitn.it`

July 23, 2024

Abstract

Lung cancer is a leading cause of cancer-related mortality, underscoring the need for improved diagnostic and prognostic tools. Autoantibodies, antibodies targeting the body's own proteins, have shown promise as biomarkers for various diseases, including cancer. In this study, We identified autoantibody signatures in lung cancer patients using protein microarrays and machine learning. These signatures accurately distinguished cancer patients from healthy controls, suggesting their potential as diagnostic biomarkers. Functional and network analyses revealed enriched pathways related to exosomes, immune response, and cellular stress, providing insights into lung cancer pathogenesis. These findings pave the way for the development of novel non-invasive diagnostic tests and targeted therapies for lung cancer.

1 Introduction

Lung cancer remains a formidable global health challenge due to its high incidence and mortality rates. Early detection is paramount for improving patient outcomes, yet current diagnostic methods often involve invasive procedures and are not always reliable.

Autoantibodies, produced by the immune system in response to tumor antigens, hold promise as non-invasive biomarkers for lung cancer detection. Their stability in serum and potential for detection in easily accessible bodily fluids make them particularly attractive for diagnostic applications. Protein microarrays, enabling simultaneous measurement of thousands of proteins, provide a powerful platform for discovering and validating autoantibody biomarkers.

This study aimed to harness protein microarray technology, coupled with advanced computational analysis, to identify autoantibody signatures that differentiate between lung cancer patients and healthy controls.

2 Methods

2.1 Dataset

The dataset we used (GSE168198) details a protein profiling experiment using protein microarrays to investigate autoantibody biomarkers for lung cancer. The study employed a two-stage design:

- Phase I (Discovery): This phase included 8 control serum samples from healthy individuals and 8 lung cancer serum samples. The primary goal of this phase was to identify potential autoantibody biomarkers that are differentially expressed between the two groups.
- Phase II (Validation): This phase expanded the study to include 60 negative control samples, 60 disease control samples, and 60 cancer samples. The purpose of this phase was to validate the findings from Phase I and assess the specificity of the identified biomarkers.

In this analysis, we focused solely on Phase I data due to the significantly reduced number of genes (features) in Phase II, which could limit the scope of our biomarker discovery efforts.

Each serum sample in Phase I was analyzed using the Human Protein Atlas protein microarray, which features 46,066 proteins. The microarrays were incubated with serum samples, allowing autoantibodies to bind to their corresponding target proteins. The types of autoantibodies specifically targeted were IgM antibodies: Labeled with the fluorescent dye Cy5.

2.2 Principal Component Analysis

PCA helps reduce the dimensionality of the data while capturing most of the variance. This is useful for visualization purposes in high-dimensional datasets such as gene data. In this method, we transform the data into a new set of variables (principal components) that capture the most significant sources of variation. By focusing on the first few principal components, which explain the majority of the variance, we can visualize the relationships between samples in a lower-dimensional space.

2.3 Data Clustering

Two unsupervised machine learning algorithms, K-means clustering and hierarchical clustering, were applied to the data. These algorithms aim to group samples based on the similarity of their gene expression profiles. In K-means clustering, the number of clusters (K) is predefined, and the algorithm iteratively assigns samples to clusters to minimize the within-cluster variance. Hierarchical clustering, on the other hand, builds a tree-like structure (dendrogram) that reveals the hierarchical relationships between samples.

2.4 Classification Models

To develop predictive models based on gene expression profiles for classifying samples as either control or cancer, we employed three supervised machine learning algorithms: Random Forest (RF), Linear Discriminant Analysis (LDA), and Lasso regression.

- Random Forest (RF): RF is an ensemble learning method that harnesses the power of multiple decision trees. Each tree is trained on a random subset of the data and features, and their predictions are aggregated to form a final prediction. This approach mitigates the risk of overfitting and enhances the model's robustness.
- Linear Discriminant Analysis (LDA): LDA is a statistical technique that identifies linear combinations of features that optimally discriminate between classes. It assumes that the data are normally distributed and that the classes share a common covariance matrix. Particularly useful when dealing with two classes (cancer vs. control in our case).

- **Lasso Regression:** Lasso, or Least Absolute Shrinkage and Selection Operator, is a regression analysis method that performs both variable selection and regularization. It adds a penalty term to the loss function, encouraging the model to select only the most important features and shrink the coefficients of less important ones. This can improve model interpretability and reduce overfitting.

The choice of these three models was motivated by their distinct strengths. RF is known for its ability to handle high-dimensional data and capture complex interactions between features. LDA is a simpler model that is often effective when the classes are linearly separable. Lasso regression can improve model interpretability by selecting only the most important features, which is particularly valuable in the context of genes.

The performance of these algorithms was compared against each other using repeated cross-validation. By comparing the performance of these models, we aimed to identify the most suitable algorithm for predicting lung cancer status based on autoantibody expression profiles.

2.5 Functional Enrichment Analysis

- **DAVID:** To gain deeper insights into the biological functions and pathways associated with the differentially expressed genes in our lung cancer dataset, we performed Functional Annotation Clustering using DAVID (Database for Annotation, Visualization, and Integrated Discovery). DAVID is a widely used bioinformatics resource that provides a comprehensive set of tools for interpreting large gene lists.

In contrast to standard enrichment analysis, which treats each gene list annotation as independent, functional annotation clustering leverages the relationships between annotations to group them into functionally related clusters. This approach can reveal broader biological themes and provide a more holistic view of the underlying biology.

- **g:Profiler:** In order to gain a more comprehensive understanding of the biological pathways and processes involved in lung cancer, we conducted functional enrichment analysis using g:Profiler. This tool enables the identification of pathways and functional terms that are statistically over-represented in our list of differentially expressed genes (DEGs) compared to a background set.

2.6 Network-Based Analysis

Network-based analysis is a powerful approach that leverages prior knowledge of biological networks to gain deeper insights into gene expression data. Unlike traditional enrichment analysis, which focuses on individual genes or gene sets, network-based analysis considers the interactions between genes, providing a more comprehensive view of the underlying biology. In this study, we employed three network-based analysis tools:

- **STRING:** STRING is a database of known and predicted protein-protein interactions (PPIs). It integrates information from various sources, including experimental data, text mining, and computational predictions. By mapping our differentially expressed genes onto the STRING network, we can visualize their interactions and identify functional modules or pathways that are enriched in our dataset.
- **PathfindR:** PathfindR is an R package that identifies active subnetworks within a PPI network. An active subnetwork is a group of interconnected genes that are significantly altered in a

given condition. PathfindR then performs enrichment analysis on these subnetworks to identify biological processes and pathways that are associated with the observed changes in gene expression.

- **EnrichNet:** EnrichNet is a web-based tool that performs network-based gene set enrichment analysis. It calculates the distance between gene sets and pathways in a PPI network and ranks them based on their association with the input gene list. This allows us to identify pathways that are not only enriched in our dataset but also closely connected to the differentially expressed genes.

3 Results

3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed to create a scatter plot of the first two principal components to visualize potential clusters based on sample classification (control vs. cancer).

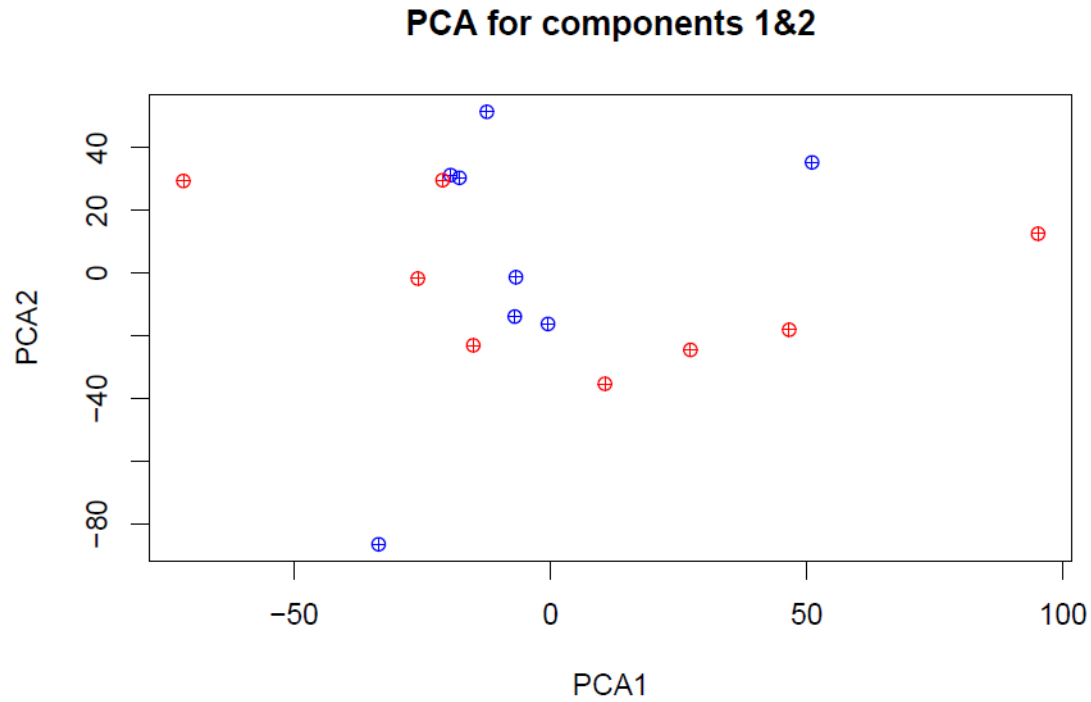


Figure 1: PCA Plot of Control and Lung Cancer Samples

The resulting PCA plot for components 1 and 2 (Figure 1) reveals a partial separation between control (red) and cancer (blue) samples. This suggests that there are underlying differences in the autoantibody profiles between the two groups. However, the separation is not complete, indicating some degree of overlap in the expression patterns of certain autoantibodies.

3.2 K-means Clustering

K-means clustering (K=2) was applied to the dataset to explore natural groupings. The results, visualized on a scatter plot based on the first two principal components (Figure 2), show a general trend for control and cancer samples to cluster together, but with significant overlap. This suggests heterogeneity in autoantibody profiles within both groups and the need for supervised learning approaches for accurate classification.

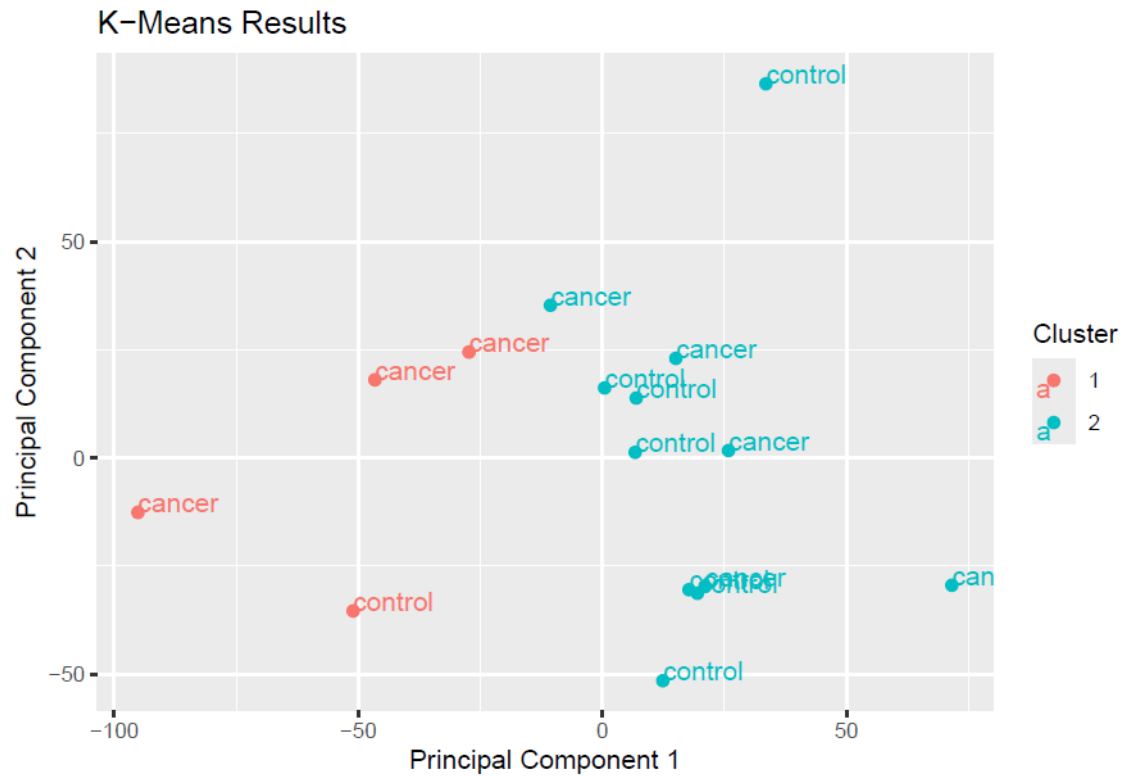


Figure 2: K-means Clustering of Control and Lung Cancer Samples

3.3 Hierarchical Clustering

Hierarchical clustering was performed to further investigate the relationships between samples based on their autoantibody expression profiles. The resulting dendrogram (Figure 3), reveals a primary division of the samples into two main clusters, but with heterogeneity within each cluster, indicating individual variation and the limitations of unsupervised clustering in perfectly separating the groups.

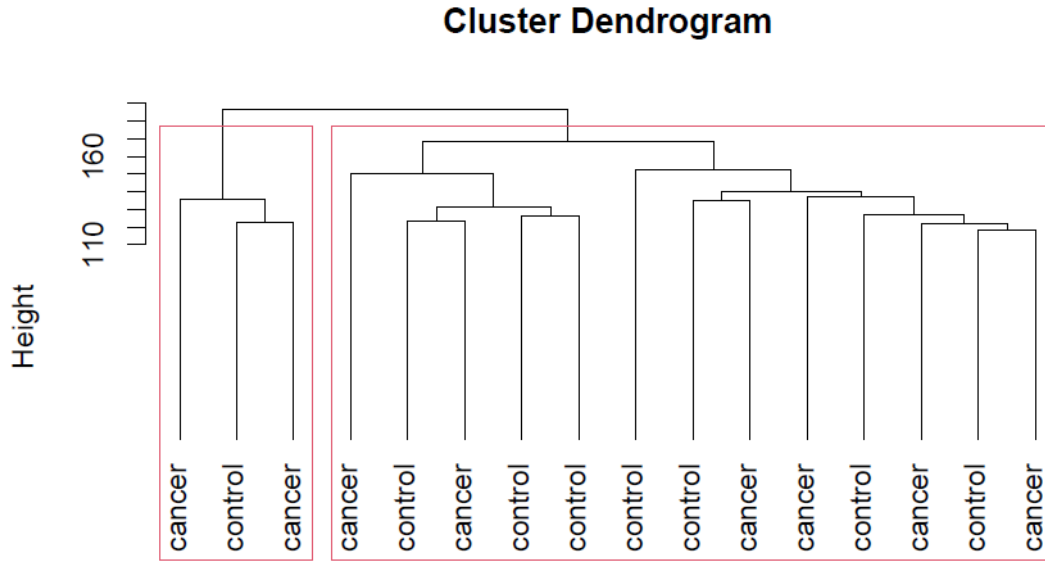


Figure 3: Dendrogram of Hierarchical Clustering for Control and Lung Cancer Samples

3.4 Classification Models

To identify the most effective model for predicting lung cancer status based on autoantibody expression profiles, we compared the performance of Random Forest (RF), Linear Discriminant Analysis (LDA), and Lasso regression using repeated cross-validation. The results, presented in Figure 4, showcase the accuracy of each model across multiple iterations of cross-validation.

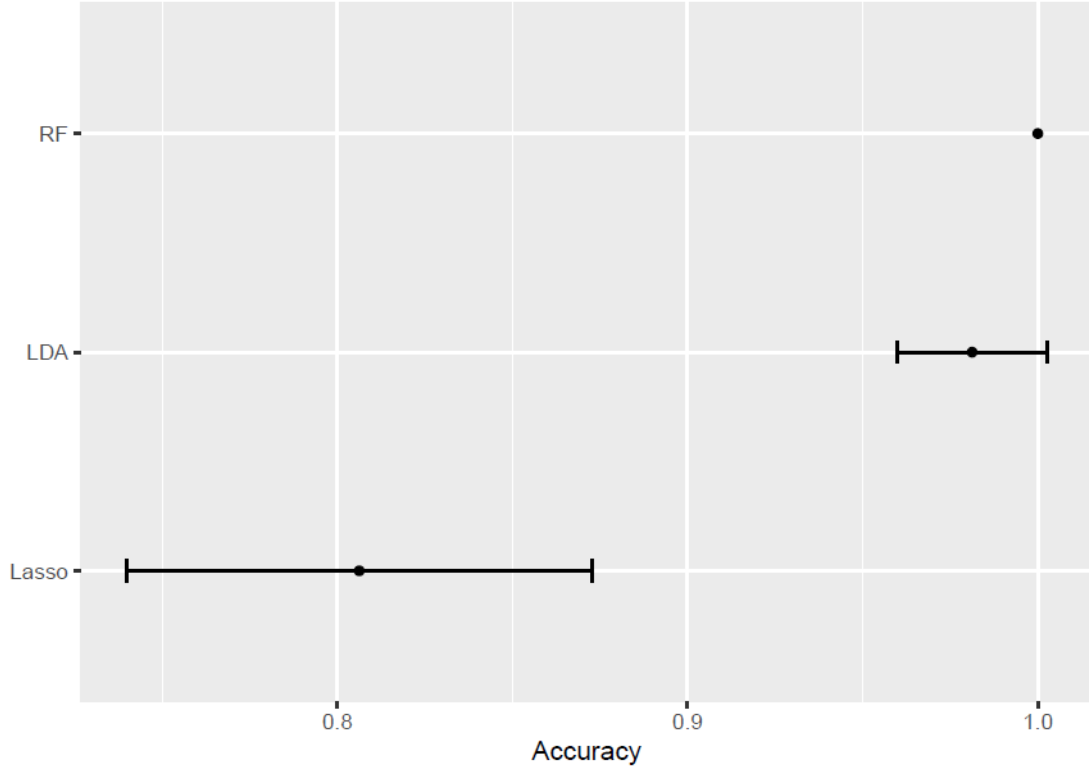


Figure 4: Model Comparison: Accuracy of RF, LDA, and Lasso in Classifying Lung Cancer Samples

The RF model demonstrated the highest accuracy, consistently achieving near-perfect performance (accuracy close to 1.0) with minimal variability across repetitions. LDA also performed well, with slightly lower accuracy but still exceeding 95%. Lasso regression exhibited the lowest accuracy among the three models, though its performance was still respectable.

This comparison underscores the suitability of both RF and LDA for classifying lung cancer samples based on autoantibody expression. While both models exhibited high accuracy, the RF model's superior performance and lower variability suggest it may be the more reliable option for future applications. The slightly lower accuracy of LDA could be attributed to its assumption of linear relationships between features and class labels, which may not fully capture the complex interactions in the autoantibody data.

Lasso regression, while less accurate than RF and LDA, still holds value due to its ability to identify a subset of informative features. This can be particularly useful in biomarker discovery, as it helps to prioritize autoantibodies that are most relevant for distinguishing between lung cancer and control samples.

3.5 DAVID

To gain deeper insights into the biological processes associated with differentially expressed proteins, we performed functional annotation clustering using DAVID. This analysis revealed significant

enrichment of several gene clusters, suggesting the involvement of specific biological processes in lung cancer.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_DIRECT	extracellular exosome	RT		42	25.8	3.4E-8	8.9E-6
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytosol	RT		63	38.7	1.9E-4	2.1E-2
<input type="checkbox"/>	GOTERM_CC_DIRECT	secretory granule lumen	RT		7	4.3	2.4E-4	2.1E-2
<input type="checkbox"/>	GOTERM_MF_DIRECT	protein binding	RT		127	77.9	1.2E-4	3.9E-2
<input type="checkbox"/>	KEGG_PATHWAY	Epstein-Barr virus infection	RT		9	5.5	4.4E-4	1.1E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytoplasm	RT		60	36.8	1.8E-3	1.2E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	ficolin-1-rich granule lumen	RT		6	3.7	2.5E-3	1.3E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	phagocytic vesicle membrane	RT		5	3.1	3.6E-3	1.6E-1
<input type="checkbox"/>	REACTOME_PATHWAY	Neutrophil degranulation	RT		15	9.2	2.6E-4	1.7E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	extracellular region	RT		28	17.2	5.4E-3	1.8E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	mitochondrial envelope	RT		3	1.8	5.4E-3	1.8E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	secretory granule membrane	RT		5	3.1	7.2E-3	2.1E-1
<input type="checkbox"/>	REACTOME_PATHWAY	Innate Immune System	RT		23	14.1	6.4E-4	2.2E-1
<input type="checkbox"/>	GOTERM_MF_DIRECT	identical protein binding	RT		27	16.6	2.1E-3	3.3E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	extracellular space	RT		24	14.7	1.6E-2	4.3E-1
<input type="checkbox"/>	KEGG_PATHWAY	Metabolic pathways	RT		25	15.3	3.7E-3	4.5E-1
<input type="checkbox"/>	GOTERM_BP_DIRECT	antigen processing and presentation	RT		5	3.1	4.6E-4	4.8E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	MLL1 complex	RT		3	1.8	2.4E-2	5.3E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleoplasm	RT		41	25.2	2.4E-2	5.3E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	myofibril	RT		3	1.8	2.7E-2	5.4E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	tRNA methyltransferase complex	RT		2	1.2	2.9E-2	5.4E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	azurophil granule lumen	RT		4	2.5	3.1E-2	5.4E-1
<input type="checkbox"/>	KEGG_PATHWAY	Type 1 diabetes mellitus	RT		4	2.5	6.7E-3	5.4E-1
<input type="checkbox"/>	KEGG_PATHWAY	Hematopoietic cell lineage	RT		5	3.1	1.2E-2	7.3E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	vesicle	RT		5	3.1	4.5E-2	7.4E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	dendrite	RT		8	4.9	4.8E-2	7.4E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	cell junction	RT		5	3.1	5.2E-2	7.6E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	lamellipodium	RT		5	3.1	5.5E-2	7.6E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	proteasome regulatory particle_lid subcomplex	RT		2	1.2	5.8E-2	7.6E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	endosome membrane	RT		6	3.7	6.3E-2	7.7E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleus	RT		55	33.7	6.4E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Tuberculosis	RT		6	3.7	2.3E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		10	6.1	2.7E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Pertussis	RT		4	2.5	3.1E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Asthma	RT		3	1.8	3.1E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Leishmaniasis	RT		4	2.5	3.2E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Prion disease	RT		7	4.3	3.5E-2	7.7E-1
<input type="checkbox"/>	KEGG_PATHWAY	Fluid shear stress and atherosclerosis	RT		5	3.1	3.6E-2	7.7E-1

Figure 5: Functional Annotation Clustering with DAVID for Lung Cancer

Based on Figure 5, the most enriched cluster was related to extracellular exosome terms (p-value = 3.4E-8), indicating a potential role for exosomal proteins in lung cancer development or progression. Exosomes are small vesicles secreted by cells that can carry proteins, nucleic acids, and lipids, and have been implicated in various biological processes, including cell-to-cell communication, immune response, and tumor progression.

The second most enriched cluster was associated with the cytosol (p-value = 1.9E-4), highlighting the involvement of intracellular proteins in lung cancer. This cluster was closely followed by terms related to the secretory granule lumen (p-value = 2.4E-4) and protein binding (p-value = 1.2E-4), suggesting that proteins involved in protein secretion and interaction play a significant role in lung cancer.

Other significantly enriched terms included Epstein-Barr virus infection (p-value = 4.4E-4), phagocytic vesicle membrane (p-value = 3.6E-3), and mitochondrial envelope (p-value = 5.4E-3). These terms suggest potential links between viral infection, immune response, and mitochondrial dysfunction in lung cancer.

The Epstein-Barr virus (EBV) is a common human virus that has been associated with several types of cancer, including some forms of lymphoma and nasopharyngeal carcinoma. While its role in lung cancer is less clear, some studies have suggested a potential link. The enrichment of EBV-related terms in our analysis could indicate that EBV infection may play a role in the development or progression of lung cancer in some patients, or it could reflect a general immune response to viral infection in the context of cancer.

3.6 g:Profiler

Functional enrichment analysis was performed using g:Profiler to identify biological pathways and processes significantly associated with the differentially expressed proteins in lung cancer samples. This analysis shown in Figure 6 revealed the enrichment of several Gene Ontology (GO) terms, including:

GO:0005515 (protein binding): This term, with a highly significant p-value ($7.09\text{e-}104$), indicates that proteins involved in protein-protein interactions are over-represented among the differentially expressed proteins in lung cancer. This highlights the crucial role of protein interactions in the molecular mechanisms underlying lung cancer development and progression.

GO:0006950 (response to stress): The enrichment of this term (p-value = $1.05\text{e-}10$) suggests that cellular stress response pathways are activated in lung cancer cells. This could be due to various stressors, such as oxidative stress, DNA damage, or hypoxia, which are often associated with tumorigenesis. The activation of stress response pathways may contribute to tumor survival and resistance to therapy.

GO:0030488 (tRNA methylation): This term, with a p-value of $2.23\text{e-}10$, indicates the involvement of tRNA methylation pathways in lung cancer. tRNA modifications play a crucial role in translation regulation, and their dysregulation can have profound effects on gene expression and protein synthesis, potentially contributing to tumorigenesis.

GO:0032071 (regulation of endodeoxyribonuclease activity): The enrichment of this term (p-value = $1.28\text{e-}10$) suggests that the regulation of DNA repair mechanisms is altered in lung cancer. This could lead to genomic instability, a hallmark of cancer cells, and contribute to tumor development and progression.

GO:0005737 (cytoplasm) and GO:0140535 (intracellular protein-containing complex): These terms, with p-values of $1.51\text{e-}10$ and $4.63\text{e-}10$, respectively, indicate that the majority of differentially expressed proteins are located within the cytoplasm and are likely involved in the formation of protein complexes. This highlights the importance of intracellular processes and protein interactions in lung cancer biology.

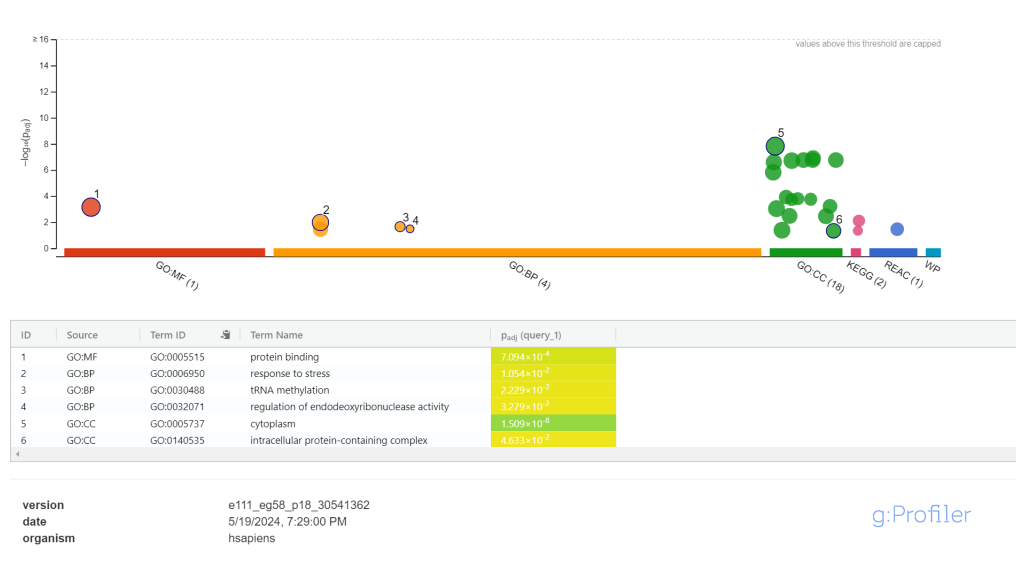


Figure 6: Functional Enrichment Analysis with g:Profiler for Lung Cancer

These findings from the g:Profiler analysis underscore the complex molecular landscape of lung cancer, involving dysregulation of various biological processes, including protein interactions, stress responses, RNA modifications, DNA repair, and intracellular signaling. The identification of these enriched pathways provides valuable insights into the potential mechanisms driving lung cancer development and progression, and may pave the way for the development of novel therapeutic targets.

It is important to note that the g:Profiler analysis did not reveal significant enrichment of KEGG pathways or WikiPathways, which may be due to the limited number of genes in our dataset or the specific focus of these databases on metabolic and signaling pathways. However, the enriched GO terms provide a comprehensive overview of the biological processes involved in lung cancer.

3.7 STRING

A protein-protein interaction (PPI) network analysis was conducted using the STRING database to explore the relationships between the differentially expressed proteins in lung cancer. The network (Figure 7) reveals several interconnected clusters of proteins, highlighting potential functional modules and pathways relevant to lung cancer pathogenesis.

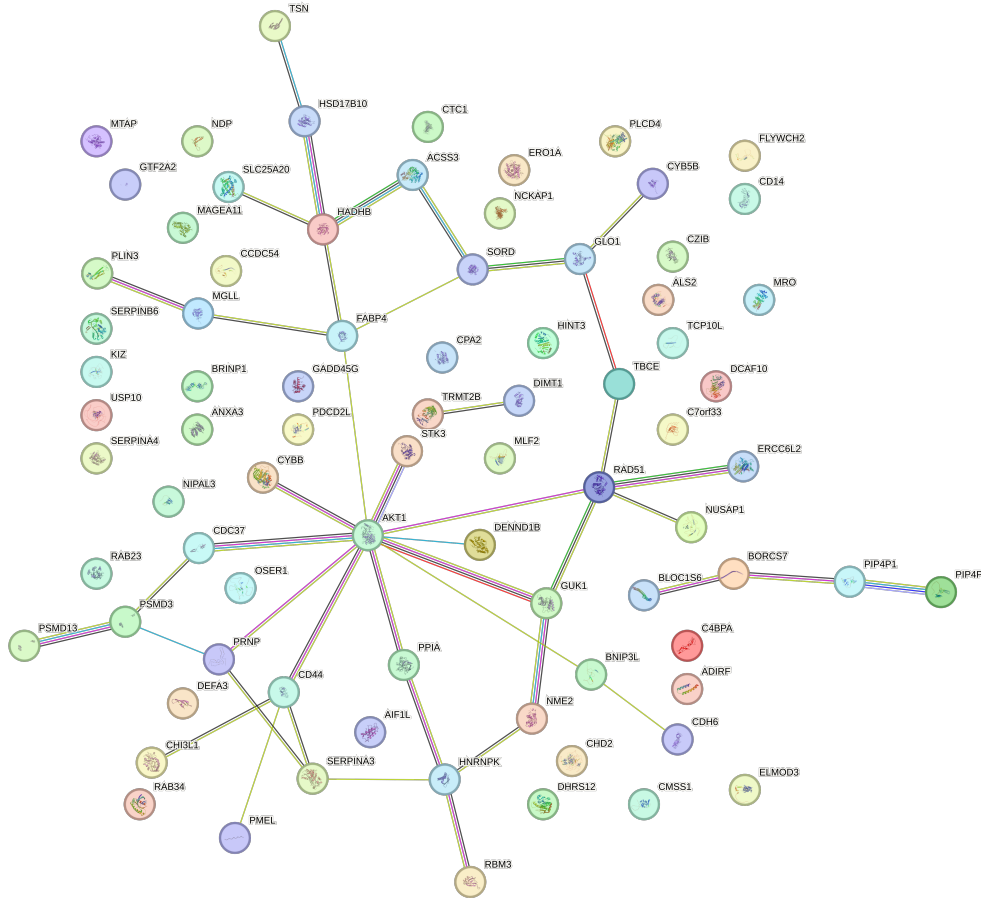


Figure 7: STRING PPI Network of Differentially Expressed Proteins in Lung Cancer

The network analysis revealed a central cluster of highly interconnected proteins, including several key players in cell cycle regulation, DNA repair, and apoptosis. These proteins, such as ANXA6, SERPINA1, and SERPINA3, are known to be involved in cancer-related processes and their interactions suggest a complex interplay between these processes in lung cancer.

The network also revealed several smaller clusters of proteins, which may represent distinct functional modules involved in lung cancer. For example, a cluster of proteins involved in immune response was identified, suggesting a potential role for autoantibodies in modulating the immune response to lung cancer.

3.8 PathfindR

PathfindR was employed to identify active subnetworks within the protein-protein interaction (PPI) network constructed using the differentially expressed proteins in lung cancer samples. This analysis shown in Figure 8 revealed significant enrichment of several pathways, providing further insights into the molecular mechanisms underlying lung cancer. The most enriched pathways included:

Glioma: This pathway, with a high fold enrichment and significant p-value, suggests a potential

link between glioma-related mechanisms and lung cancer. While glioma is a type of brain tumor, the shared pathways may point to common dysregulated processes in both types of cancer.

Endocrine resistance: The enrichment of this pathway suggests that mechanisms related to endocrine resistance, which is often observed in hormone-dependent cancers, may also play a role in lung cancer. This could have implications for the development of therapies targeting endocrine pathways in lung cancer.

Melanoma: Similar to glioma, the enrichment of melanoma-related pathways suggests shared mechanisms between these two distinct cancer types. This finding could open up new avenues for exploring potential therapeutic targets for lung cancer based on the knowledge gained from melanoma research.

Other enriched pathways included the AMPK signaling pathway, biosynthesis of amino acids, shigellosis, FoxO signaling pathway, human cytomegalovirus infection, microRNAs in cancer, and endocytosis. These pathways encompass a wide range of biological processes, including cell signaling, metabolism, immune response, viral infection, and gene regulation, highlighting the complexity and interconnectedness of the molecular mechanisms underlying lung cancer.

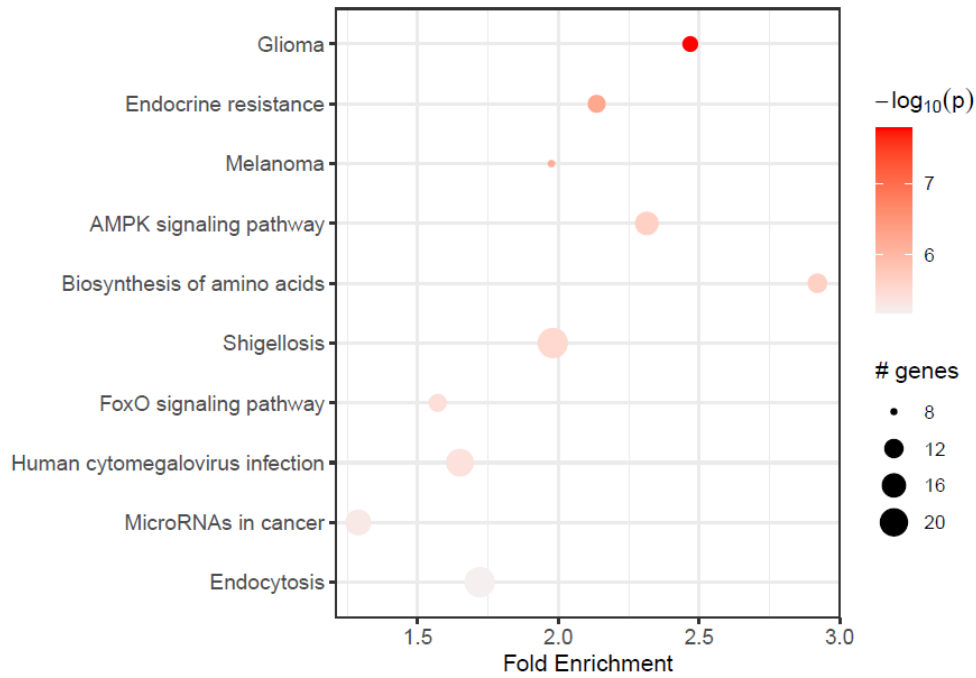


Figure 8: PathfindR Pathway Enrichment Analysis for Lung Cancer

3.9 EnrichNet

EnrichNet, a tool designed for network-based gene set enrichment analysis, was employed to identify pathways significantly associated with differentially expressed proteins in lung cancer samples while considering their interactions within the protein-protein interaction network. The results (Figure 9) indicate the enrichment of several pathways, with the top four being:

Response to amine stimulus, Cellular response to dexamethasone stimulus, and Urea cycle: These pathways share a network distance XD-score of 1.70038 and an overlap-based q-value of 0.74. Each involves 2 genes from the dataset overlapping with 10 genes in the reference pathway gene set. While these q-values are not statistically significant (typically requiring <0.05), the consistent XD-score suggests a potential biological relevance to lung cancer that warrants further investigation.

Dexamethasone is a glucocorticoid with anti-inflammatory properties, and its cellular response pathway may be implicated in lung cancer development or progression. The urea cycle is primarily involved in ammonia detoxification, but its potential connection to lung cancer may lie in its role in amino acid metabolism or other cellular processes.

Inflammatory response to antigenic stimulus: This pathway shows a slightly lower XD-score of 1.40038 but a higher q-value of 0.89, indicating a weaker association with the differentially expressed genes compared to the previous pathways. However, the involvement of inflammatory response in cancer development is well-documented, suggesting this pathway may still be relevant to lung cancer.

Overall, the EnrichNet analysis suggests a potential link between lung cancer and pathways related to amine and dexamethasone stimulus response, the urea cycle, and inflammatory response. While the statistical significance of these findings is limited by the small sample size and high q-values, the consistent network distance scores across these pathways hint at a possible biological relevance to lung cancer. Further investigation with larger datasets and experimental validation is needed to confirm these associations and elucidate the underlying mechanisms.






















Annotation (pathway/process) 	Significance of network distance distribution (XD-Score) 	Significance of overlap (Fisher-test, q-value) 	Dataset size (uploaded gene set) 	Dataset size (pathway gene set) 	Dataset size (overlap) 	Tissue-specific XD-scores 
<u>response to amine stimulus</u>						
 compute graph visualization  see mapped genes	1.70038	0.74	137	10	2 (show)	 show tissue specificity
<u>cellular response to dexamethasone stimulus</u>						
 compute graph visualization  see mapped genes	1.70038	0.74	137	10	2 (show)	 show tissue specificity
<u>urea cycle</u>						
 compute graph visualization  see mapped genes	1.70038	0.74	137	10	2 (show)	 show tissue specificity
<u>inflammatory response to antigenic stimulus</u>						
 compute graph visualization  see mapped genes	1.40038	0.89	137	12	2 (show)	 show tissue specificity
<u>positive regulation of protein serine/threonine kinase activity</u>						
 compute graph visualization	1.18610	1.00	137	14	2 (show)	 show tissue specificity

Figure 9: EnrichNet Pathway Enrichment Analysis for Lung Cancer

In addition to pathway enrichment, EnrichNet was also used to assess the tissue-specific expression patterns of the differentially expressed proteins. This analysis revealed a significant enrichment of proteins expressed in adipose tissue, fetal lung, kidney, fetal liver, and bone marrow (Figure 10). The

high XD-scores for these tissues suggest that the identified autoantibodies may target proteins that are preferentially expressed in these tissues. This finding could have implications for understanding the tissue origins of lung cancer, as well as the potential role of these tissues in tumor development and progression.

Overlap genes	
Tissue type	Tissue XD-Score:
adipose tissue	8.84
fetal lung	5.77
kidney	4.40
fetal liver	4.30
bone marrow	4.22
liver	4.17
dorsal root ganglion	2.16
atrioventricular node	1.44
trigeminal ganglion	1.20

Figure 10: EnrichNet Tissue-Specific Enrichment Analysis for Lung CancerEnrichNet Gene-Pathway Network Visualization for Lung Cancer

The EnrichNet analysis also generated a network representation of gene-pathway relationships (Figure 11). This network visualization provides further insights into the connectivity and interplay between the differentially expressed genes and enriched pathways. The network reveals that the response to amine stimulus pathway is interconnected with several other pathways, including the cellular response to dexamethasone stimulus, the urea cycle, and the inflammatory response to antigenic stimulus. This suggests that these pathways may share common regulatory mechanisms or participate in coordinated biological processes in the context of lung cancer.

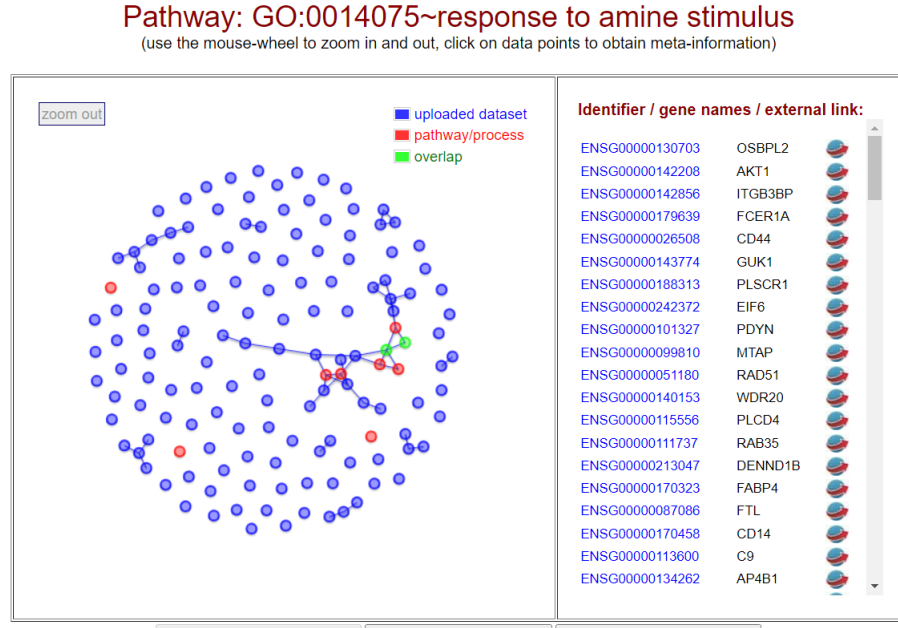


Figure 11: EnrichNet Gene-Pathway Network Visualization for Lung Cancer

4 Discussion

Our study demonstrates the feasibility and potential of using protein microarrays and machine learning to identify autoantibody biomarkers for lung cancer. The high classification accuracy achieved by Random Forest and Linear Discriminant Analysis models highlights the potential of autoantibody signatures for diagnostic purposes. The functional annotation clustering and pathway enrichment analyses provide valuable insights into the biological processes associated with lung cancer, suggesting potential targets for therapeutic intervention. The network-based analysis further elucidates the complex interactions between genes and pathways involved in lung cancer pathogenesis.

While our study focused on a limited sample size, the findings provide a strong foundation for future research. Larger and more diverse cohorts are needed to validate our results and assess the generalizability of the identified biomarkers. Additionally, integrating shrinkage methods like ridge regression and lasso in future analyses could enhance model robustness and interpretability.

The identified autoantibody biomarkers have the potential to be translated into non-invasive diagnostic tests for lung cancer, enabling earlier detection and improved patient outcomes. Furthermore, the enriched pathways and gene networks could inform the development of targeted therapies that specifically address the molecular mechanisms underlying lung cancer.

In conclusion, this study demonstrates the potential of protein microarrays and machine learning in lung cancer biomarker discovery. The identified autoantibody biomarkers offer a promising avenue for developing non-invasive diagnostic tests and personalized treatment strategies for lung cancer. Further research is warranted to validate these findings in larger, more diverse cohorts and translate them into clinical applications.