IDENTIFICATION AND VALIDATION OF ARSENIC-ASSOCIATED GENES AND RISK MODEL FOR PREDICTING LUNG CANCER

An Honors Thesis

Presented by

Jack Lathrop Robert Pendleton

Completion Date:
April 2025

Approved By:

Richard Schur, Honors Director

Joshua Kennedy, Department of Chemistry and Physics

Drury University, 2025

© 2025, Jack Pendleton

ABSTRACT

Title: **Identification and Validation of Arsenic-Associated Genes and Risk Model for Predicting Lung Cancer**
Author: **Jack Pendleton**
Thesis/Project Type: **Thesis**

Lung cancer is the leading cause of cancer-related mortality worldwide. Tobacco smoking is the primary risk factor, contributing to 90% of cases. Arsenic trioxide, a carcinogen in tobacco smoke, is linked to DNA damage and tumorigenesis, yet its role in lung cancer remains underexplored. This study investigates arsenic-associated genes in lung cancer to identify predictive biomarkers and therapeutic targets. A dataset of 959 lung tissue samples (842 carcinomas, 117 normal) and associated clinical data from the Gene Expression Omnibus (GEO) was analyzed and used to train a predictive model. Gene expression data from the Affymetrix U133 Plus 2.0 Array platform were normalized using frozen robust multiarray analysis (fRMA). Principal component analysis (PCA) confirmed distinct separation between tumor and normal tissues. Differential expression analysis identified 88 significantly altered arsenic-associated genes. Functional analysis highlighted six key genes (ARHGEF10, ADARB1, SEC14L1, CBX7, GYPC, and CRIM1) involved in tumor progression. A predictive model using four of these genes achieved high accuracy (AUC = 0.886, $p < 0.05$). The predictive model was validated using a second test dataset comprised of 1182 samples of RNA sequencing data (1105 tumor, 77 normal). On the test dataset, the model displayed a predictive accuracy of 82.9%. These findings suggest arsenic-associated genes play a crucial role in lung cancer risk and could inform future diagnostic and therapeutic strategies. Further research is needed to better understand these findings before they can be implemented into clinical practice.

Identification and Validation of Arsenic-Associated Genes and Risk Model for Predicting Lung Cancer

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, and was responsible for 12.4% of all cancer diagnoses globally in 2022 (Bray et al., 2024). During that year, more than 1.8 million patients died due to lung cancer. Despite the high occurrence of this cancer, the 5-year relative survival rate was only 27% (Luo et al., 2019), much lower than other common types of cancer such as breast, prostate, and colon which all have 5-year survival rates between 50% and 85% (Mattiuzzi & Lippi, 2019). Given its high prevalence and poor prognosis, research aimed at improving early detection and identifying new therapeutic targets for lung cancer could result in significant improvements in clinical outcomes.

While a variety of different factors such as genetic predisposition can lead to formation of malignant tumors, lung tumors are especially linked with environmental exposure. As many as 90% of all cases of lung cancer are caused by exposure to toxic fumes, whether from cigarette smoking, poor air quality, or occupational hazards (Zhou, 2019). Arsenic is a ubiquitous element, the 20th most abundant on earth. It is recognized by the Environmental Protection Agency as a known carcinogen in humans (Speer et al., 2023), and is found in soil, water, industrial chemicals, pesticides, and even some medications. Of significant interest to lung cancer is the presence of arsenic-containing compounds in cigarette smoke. Arsenic trioxide has been associated with various types of cancer including skin, liver, and bladder cancers, and has been associated with DNA damage in a wide variety of tissues including lung epithelial cells (Cooper et al., 2022). Previous work has identified a set of 147 genes associated with arsenic exposure in relation to molecular mechanisms of bladder cancer (Singhal et al., 2022). Since smoking is a major risk factor for lung cancer, understanding the impact of arsenic exposure on lung tissue and its role in tumorigenesis could be especially impactful in identifying predictive biomarkers and new therapeutic targets. Despite this, arsenic-associated genes have not been studied sufficiently in relation to cancers of the lungs and bronchus.

The primary objective of this research is to investigate the potential role of arsenic-related gene expression changes in lung cancer development. By identifying predictive biomarkers, this study aims to improve early detection of lung cancer, which is crucial for increasing survival rates. The stage at which lung cancer is diagnosed is strongly correlated with patient outcomes, with early-stage detection significantly improving prognosis and treatment options. In one study, patients diagnosed at stage I or II had five-year average survival rates of 76.9% and 56.1%, while the rates for patients diagnosed at stages III or IV were much lower at 32.6% and 21.4% (He et al., 2022). Unfortunately, it is uncommon for patients with many types of lung cancer to be diagnosed before stages III or IV (Casal-Mouriño et al., 2020). Thus, discovering gene expression markers for lung cancer that allow for earlier detection has the capability to significantly improve patient outcomes and increase survival rates.

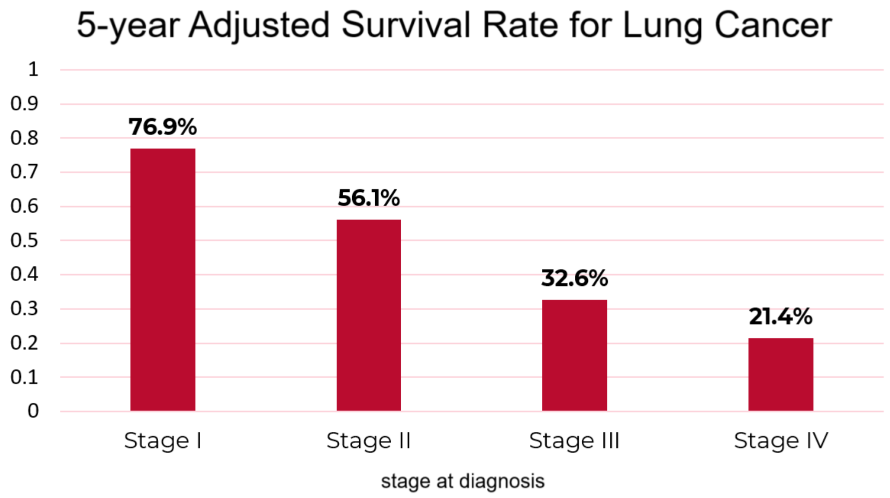


Figure 1: Adjusted 5-year average survival rates of lung cancer based on stage at diagnosis. (He et al., 2022)

Current lung cancer screening methods, including low dose computed tomography (CT), have shown promise in detecting lung cancer at early stages of tumor growth. However, these methods have many limitations, including high rates of false positives and the risks associated with radiation exposure. In one study, these methods were only able to identify lung cancer correctly 60% of the time (Oken et al., 2011). There are also many concerns related to the levels of radiation exposure required to image with CT. One model projected that implementing annual surveillance CT scans in high-risk patients could cause an approximately 5% increase in cancer incidence from the additional radiation exposure alone (Smith-Bindman et al., 2025). Additionally, current screening regimens primarily target individuals with significant smoking histories, which may lead to the underdiagnosis of lung cancer in never-smokers and individuals with alternative risk factors, such as environmental toxin exposure (Stephens et al., 2023). Biomarkers such as gene expression levels have been proposed as potential supplements or replacements for existing screening methods, offering a less invasive and more specific approach. However, many of the identified biomarkers require further validation before they can be implemented in clinical practice (Marmor et al., 2021). This further shows the need for continued research into novel biomarkers, including those linked to environmental carcinogens like arsenic, to enhance early detection strategies and improve lung cancer prognosis.

Methods

The first step in conducting this research was the curation of two large datasets of gene expression data. This was necessary so that any conclusions reached would have sufficient statistical power and significance. Collecting this data independently would be both cost-prohibitive and time-consuming, likely taking years of sample collection and requiring equipment, resources, and funding not available at the time this research was conducted. Therefore, existing datasets collected by other research groups around the world were aggregated and then analyzed. The primary sources of this data included the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), both of which provide high-quality datasets collected for use in previous studies done at other institutions. By combining samples from multiple studies, the aim was to improve the accuracy of the findings and correct for various types of statistical bias that can arise when using data from a single source. Metadata played a crucial role in this process, allowing adjustment for important confounding biological factors such as sex, smoking history, and age. To ensure the integrity of the analysis, only samples with complete metadata for these variables were included. Without proper metadata correction, the predictive model might inadvertently identify a genetic signature for certain demographic factors rather than distinguishing cancerous from normal expression.

The data for this research was organized into two groups: a training dataset comprising 959 microarray samples from six studies, and a testing dataset containing 1182 RNA sequencing samples. The first dataset contained 842 samples of lung tumor tissue from cancer patients, and 117 paired samples of non-cancerous lung tissue taken from a subset of the cancer patients. This dataset was used to identify important genes and create the predictive model. A second dataset was curated, containing 1182 total samples of RNA sequencing data. This was comprised of 604 samples from patients with adenocarcinoma tumors, 501 samples of squamous cell carcinoma, and 77 matched samples of non-cancerous lung tissue from some of the adenocarcinoma patients. This data was assembled from two TCGA datasets of tumor data: TCGA Lung Squamous Cell Carcinoma and TCGA Lung Adenocarcinoma, along with another GEO dataset containing both tumor and normal samples. In total, data from nine different studies was combined, normalized, and analyzed to train the predictive model and evaluate its performance.

Gene expression levels in the training dataset were measured using Affymetrix GeneChip™ U133 Plus 2.0 human genome microarrays. This method relies on silicon-based chips containing oligonucleotide probes, short DNA sequences designed to bind specifically to RNA transcripts. When RNA from a sample binds to its complementary DNA probe, the amount of bound RNA provides a quantitative measure of gene expression levels in the tissue. The intensity of hybridization signals directly correlates with the abundance of RNA

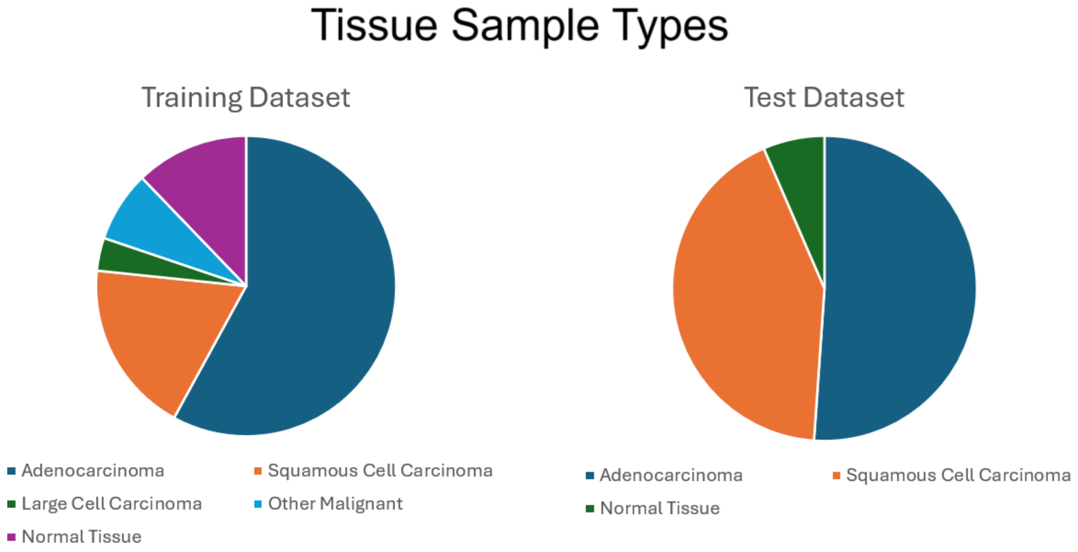


Figure 2: Graphs displaying the types of samples including in the training and test datasets.

transcripts, allowing for precise quantification of gene expression. This approach was chosen because microarrays are a widely-used and well established method for measuring gene expression, leading to a large amount of available data. This allowed the training dataset to be constructed from samples collected from six different studies made available on GEO.

The data for the testing set was collected using RNA sequencing technology (RNA-seq). Unlike microarrays, which rely on pre-designed probes, RNA-seq sequences the entire set of RNA transcripts produced by a genome, providing a more comprehensive view of gene expression. This method allows for the detection of novel and rare transcripts, offering a higher degree of flexibility and sensitivity compared to microarray-based analysis. The ability to capture the full transcriptome, including splice variants and low-abundance RNAs, makes RNA-seq a powerful tool for gene expression studies. However, because it is a newer and more expensive technology than microarrays, there have been significantly fewer studies conducted using this type of data. The testing dataset was created with tumor samples from two different TCGA studies, and normal tissue samples from one study available on GEO. The test dataset contained 1105 samples of tumor data and 77 matched samples of normal tissue. Collecting a different type of data for this set was a deliberate decision, as utilizing data collected using two completely different techniques provides additional evidence for the usefulness of the model. This is due to the fact that differential expression occurring due to artifacts of the data collection method can be found and corrected for.

Since the data used to perform the analyses for this project was sourced from several different experiments, correction was necessary to ensure that variations between experiments were corrected for. These confounding variables can be introduced by factors such as variations in microarray chip batches, differences in reagent quality, or variations in laboratory protocols. To eliminate these unwanted sources of variation, two normalization algorithms were considered: Robust Multi-array Average (RMA) and frozen RMA (fRMA). After comparison, fRMA was selected as the preferred method for normalization in the training dataset, as it demonstrated superior performance in reducing variability across different sample batches. Both algorithms work by performing background correction, quantile normalization, and summarization using a robust multi-array model. Using the RMA algorithm, this is accomplished by comparing samples relative to each other, which introduces a few limitations: smaller datasets can sway normalization, and all samples must be normalized at the same time (McCall et al., 2010). The fRMA algorithm addresses these limitations by comparing experimental data to precomputed probe-specific reference datasets, which are stored as the "frozen" parameters. This allows it to be used on datasets of any size, and for analysis to take place as samples are collected.

To assess the effectiveness of batch effect correction, density plots were used to visualize the variation between samples within the same dataset. Density plots provide a graphical representation of gene expression distribution across samples, allowing the evaluation of whether batch effects were successfully mitigated. If batch effect correction is successful, the density plots should show a more uniform distribution across all samples, indicating that technical variation has been minimized. This step was critical to ensuring that downstream analyses were not influenced by experimentally introduced differences between samples.

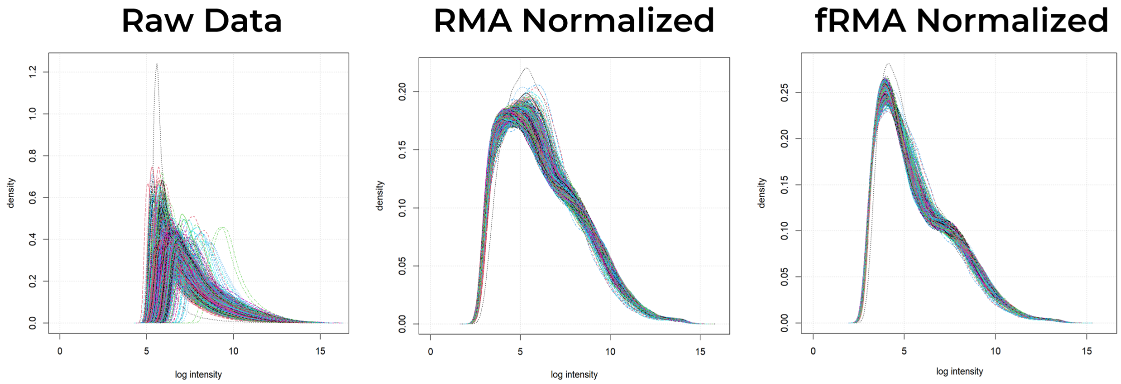


Figure 3: Logarithmically-scaled density plots of gene expression intensity showing the effects of the RMA and fRMA normalization algorithms on the training dataset.

Principle component analysis (PCA) was performed to reduce the dimensionality of the datasets while preserving the key sources of variation in gene expression data. This technique simplifies complex datasets by transforming them into principal components that capture the most significant patterns in the data (Groth et al., 2013). PCA was used to visualize clustering patterns and to identify potential sources of variation, such as study-related batch effects. By overlaying metadata variables onto the PCA plots, clusters of datapoints could be investigated to see if they were due to biological differences or confounding factors like batch effects. If the primary clustering pattern was associated with study origin rather than tumor status, it indicated that batch effects were still present. As this type of batch effect was observed in both the training and test datasets, the Limma algorithm was utilized on both of them. Following batch effect correction, additional PCA plots were created to verify the effectiveness of the correction process.

Differential gene expression analysis was then performed to identify changes in gene expression that occur due to lung cancer. This approach aimed to determine which genes exhibited significant differences in expression between tumor and normal tissue samples. This differential expression was then compared to a set of 147 arsenic-associated genes collected by Dr. Singhal's lab from bladder cancer samples (Singhal et al., 2022). These genes were used to

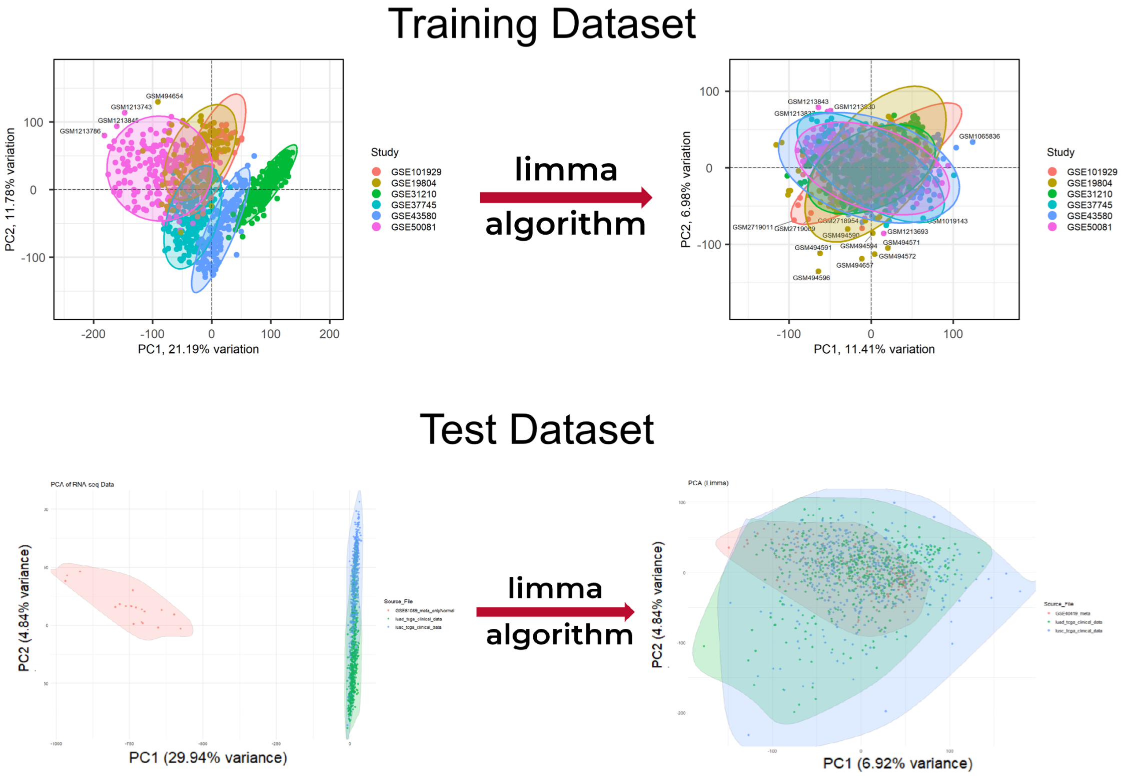


Figure 4: PCA plots of the training and testing datasets before and after batch effect correction with the Limma algorithm. Before correction, the datasets displayed significant clustering by study and a high degree of variance, while after correction the variance was significantly reduced and clustering by study was eliminated.

determine if any relevance was exhibited in predicting lung cancer. The analysis identified a subset of these genes that showed strong associations with lung cancer, marking them as good choices to be included in the predictive model. The selection of these genes was essential to the rest of the modeling process, as they formed the foundation for building a reliable model capable of distinguishing cancerous from normal tissue based on gene expression patterns.

To develop a predictive model for lung cancer classification, logistic regression was employed. This statistical approach models the relationship between a binary dependent variable and multiple independent variables (Stoltzfus, 2011). In this case, the dependent variable was whether or not the sample was cancerous, and the independent variables were the expression levels of the various genes making up the model. The logistic regression model was trained using the first dataset, which contained samples labelled whether they were data from tumor or normal tissue. The model's ability to classify new samples accurately depended on the expression levels of four selected genes, which were identified through differential gene expression analysis. The trained model was then used to predict cancer status in independent test samples.

The predictive performance of the logistic regression model was evaluated using receiver operating characteristic (ROC) analysis. This method of analysis measures the ability of a classifier to distinguish between two classes, such as tumor and normal tissue in this research (Obuchowski & Bullen, 2018). The primary metric used for evaluation was the Area Under the Curve (AUC), which quantifies the model's accuracy. An AUC value of 1 represents perfect classification, while an AUC of 0.5 indicates random guessing. The closer the AUC is to 1, the better the model's performance. ROC analysis was crucial for determining whether the logistic regression model was effective in predicting lung cancer based on gene expression profiles. A low AUC around 0.5 would indicate the model had no predictive power, while an AUC very close to 1 would imply unnatural accuracy and therefore problems with the model or the data used to construct it.

Discussion

The training dataset was first analyzed with principal component analysis to evaluate if batch effects were present and confirm that the normalization procedures effectively mitigated any potential unwanted technical variations. The first principal component (PC1) accounted for 9.47% of the total variance, while the second principal component (PC2) was responsible for 7.67%. The results demonstrated that batch effects were properly corrected, as there were no visible clustering patterns attributable to unexpected demographic or technical variables. It did reveal a distinct separation between tumor and normal samples, showing that the biological differences between these two groups were well-preserved in the dataset. This clear separation supports the robustness of the dataset for downstream differential gene expression analysis and predictive modeling.

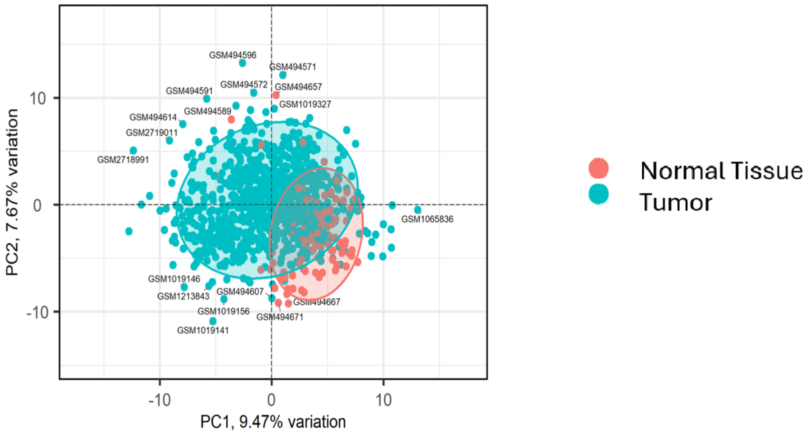


Figure 5: PCA plot of training dataset with datapoints grouped by study, showing separation by tissue type.

As various clinical factors were also collected along with the gene expression data, it was also possible to analyze the PCA plots utilizing these characteristics. Smoking history and biological sex did not appear to be associated with clustering patterns in the data, but distinct groupings did emerge when cancer stage was utilized to group the datapoints.

Differential gene expression analysis was then performed on the dataset, and 88 out of

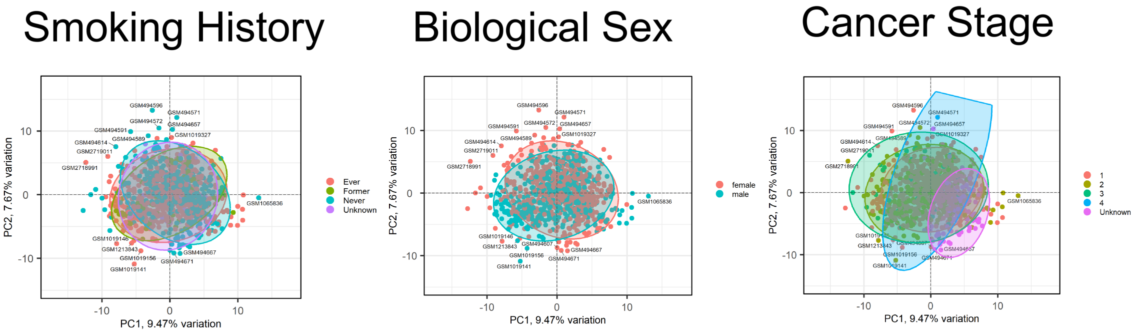


Figure 6: PCA plots of training dataset with datapoints grouped by smoking history, biological sex, and cancer stage, showing various degrees of seperation.

the 147 arsenic-associated genes were found to be differentially expressed between tumor and normal tissue. Among these, six genes emerged as key players in biological processes strongly associated with cancer progression and tumor biology. These genes included ARHGEF10, ADARB1, SEC14L1, CBX7, GYPC, and CRIM1, each of which has a well-established role in critical cellular processes such as tumor progression, signaling pathways, RNA editing, lipid transport, chromatin remodeling, cellular adhesion, and angiogenesis. The gene ARHGEF10 has been shown to induce tumor suppression in other types of cancers including pancreatic (Joseph et al., 2020) and gastric (Wang et al., 2020), so the down-regulation seen in this dataset aligns with what was found in previous works. ADARB1 is known to play important roles in RNA editing and efficacy of short interfering RNAs (W. Yang et al., 2005), an important regulatory mechanism of gene and protein expression. Another key gene, SEC14L1, has been implicated in vascular invasion and angiogenesis in breast cancer (Sonbul et al., 2018) and associated with a high risk of tumor recurrence in prostate cancer (Agell et al., 2012). Another tumor suppressor gene, CBX7, has previously been identified as a potential prognostic biomarker for lung cancer when it is down-regulated (Huang et al., 2022), and this has also been linked with poor clinical outcomes (Y. Yang et al., 2021). While not previously linked to lung cancer, changes in GYPC expression have been found to be associated with tumor progression in ovarian cancer by changing the infiltration abilities of immune cells into the tumor microenvironment (Guo et al., 2020). The final gene of the set, CRIM1, was shown to regulate cell adhesion and migration of lung cancer cells from immortal cell lines (Zeng et al., 2015).

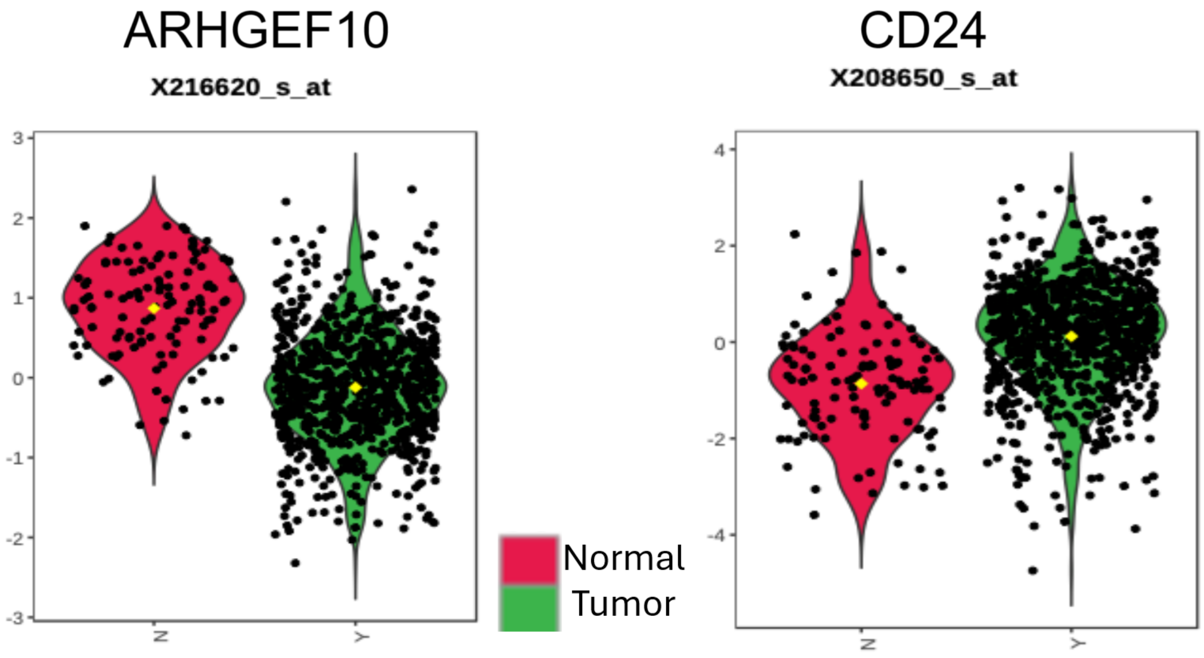


Figure 7: Expression levels of two selected genes, ARHGEF10 and CD24, showing differential expression between tumor and normal tissue. ARHGEF10 expression is shown to be down-regulated in tumor tissue, while CD24 is up-regulated.

The importance of these genes in lung cancer was analyzed by constructing a predictive classification model using logistic regression. The resulting predictive model used four of the genes and demonstrated strong classification performance. The genes used were ARHGEF10, ADARB1, SEC14L1, and CRIM1. Following receiver operating characteristic analysis, the area under the curve (AUC) of the model was of 0.886 (95% confidence interval: 0.838–0.932), meaning the preliminary predictive accuracy was 88.6%. All four of the selected genes were found to contribute equally to the predictive model ($p < 0.05$), indicating their strong association with the distinction between cancerous and non-cancerous tissue. These results suggest that these genes serve as valuable biomarkers for lung cancer classification and could potentially guide future diagnostic and therapeutic strategies.

The test dataset was used to verify the predictive model created on the training dataset.

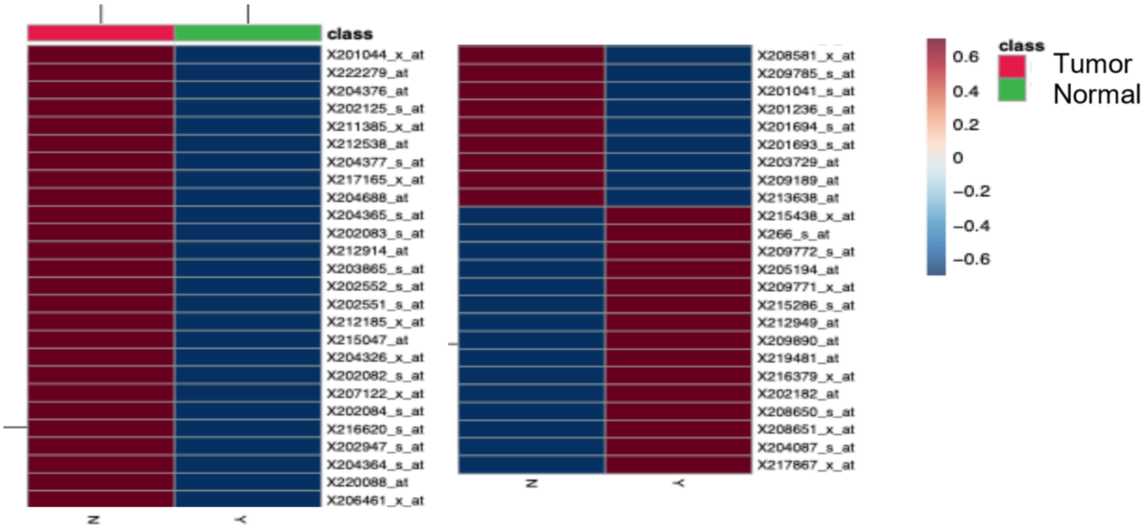
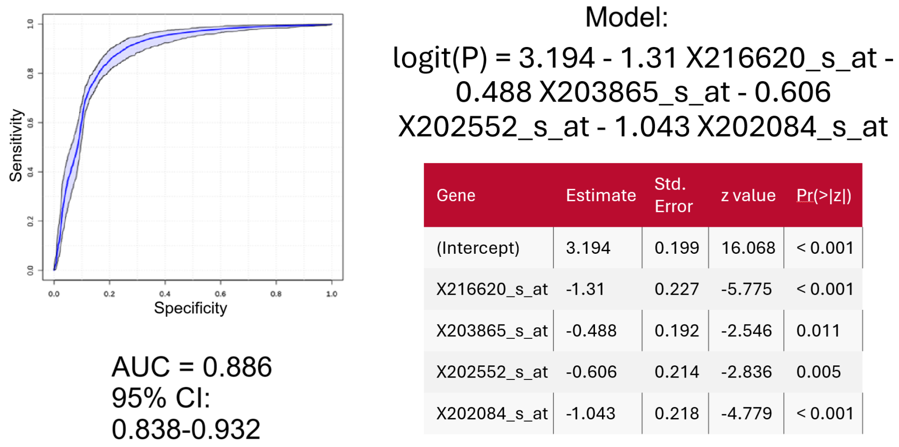


Figure 8: Plot showing full list of genes most differentially expressed between tumor and normal tissue. Purple in the first column with blue in the second means a gene was down-regulated in tumor tissue, while blue in the first column and purple in the second indicates up-regulation in the cancerous tissue.



AUC = 0.886
95% CI:
0.838-0.932

Model:
logit(P) = 3.194 - 1.31 X216620_s_at - 0.488 X203865_s_at - 0.606 X202552_s_at - 1.043 X202084_s_at

Gene	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.194	0.199	16.068	< 0.001
X216620_s_at	-1.31	0.227	-5.775	< 0.001
X203865_s_at	-0.488	0.192	-2.546	0.011
X202552_s_at	-0.606	0.214	-2.836	0.005
X202084_s_at	-1.043	0.218	-4.779	< 0.001

Figure 9: Predictive model developed from training dataset. The graph displays sensitivity and specificity of the model, while the table lists the four genes which make up the model and their statistical parameters.

As before, PCA was utilized to verify that the normalization and correction process was successful. After batch effect correction using fRMA and Limma, the first principal component accounted for 6.92% of the variance while the second principal component accounted for 4.84%. Similarly to the testing dataset, normal tissue samples were tightly clustered while tumor samples were distributed more broadly.

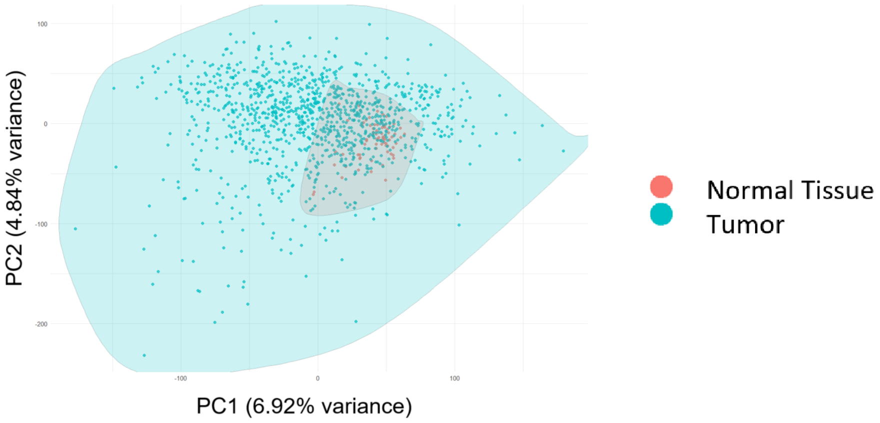


Figure 10: PCA plot of test dataset with datapoints grouped by study, showing separation by tissue type.

Utilizing the same four-gene model, samples from the test dataset were classified as either cancerous or noncancerous. The area under the curve for the model on this dataset was 0.829 (95% confidence interval, 0.775-0.871) This imputes a predictive accuracy of 82.9%. Similarly to the training dataset, all four genes contributed equally. This confirms that the four-gene model has true predictive accuracy on a wide array of samples from both microarray and RNA sequencing data.

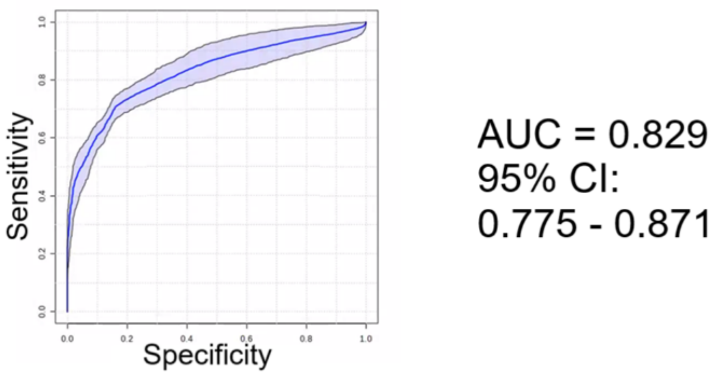


Figure 11: Performance of the predictive model on the test dataset comprised of RNA sequencing samples, displaying a predictive accuracy of 82.9%

Conclusion

This study demonstrates that arsenic-associated genes, previously studied in the context of bladder cancer, are also significantly altered in lung cancer and may serve as valuable biomarkers for diagnosis. Through examination of a large, heterogeneous collection of data from lung tissue samples using both microarray and RNA sequencing technologies, 88 differentially expressed genes related to arsenic were discovered, six of which were found to have strong functional association with tumor formation, progression, and maintenance. A predictive model was built using the expression levels of four of these genes: ARHGEF10, ADARB1, SEC14L1, and CRIM1; it showed strong classification power, achieving predictive accuracies of 88.6% and 82.9% on the training and test datasets, respectively. Validation with a different type of data further shows the robustness of the model, and that the predictive ability lies in actual biological changes due to arsenic exposure, not technical artifacts of a particular gene expression measurement technique. These results are compelling evidence that arsenic exposure leaves a detectable genetic signature in lung tissue and that this signature can be employed to distinguish cancer from normal samples with high accuracy.

This research extends current knowledge on lung cancer biomarkers by connecting environmental carcinogen exposure with gene expression signature patterns by creating a novel predictive model for lung cancer. Through the application of several bioinformatics and statistical techniques, this study provides a reproducible and interpretable framework for identifying cancer-related gene expression patterns among different data collection modalities.

In future work, incorporating demographic and clinical metadata such as sex, age, race, smoking history, and exposure to environmental toxins could refine the predictive model and uncover how gene expression changes differ across populations. This approach may reveal whether certain demographic groups are more susceptible to arsenic-induced gene expression changes or if particular genes are more predictive of cancer in specific subgroups. Such insights could help identify at-risk populations, guide targeted prevention strategies, and contribute to a more equitable understanding of how lung cancer disproportionately affects different communities.

While these findings are valuable on their own, additional research will need to be conducted in order to integrate them into clinical practice. Validation with more samples, functional studies to determine the mechanistic role of these genes in tumorigenesis, and examination of their detectability in less invasive sample types (e.g., blood or sputum) could all be next steps to utilize the findings discovered in this work. Yet, this study offers useful groundwork for the development of novel diagnostic agents that incorporate environmental risk factors such as arsenic exposure, and lays the groundwork for future studies to further improve early detection as well as patient outcomes for lung cancer.

BIBLIOGRAPHY

Agell, L., Hernandez, S., Nonell, L., Lorenzo, M., Puigdecanet, E., de Muga, S., Juanpere, N., Bermudo, R., Fernandez, P. L., Lorente, J. A., Serrano, S., & Lloreta, J. (2012). A 12-gene expression signature is associated with aggressive histological in prostate cancer: ¡em¿sec14l1¡/em¿ and ¡em¿tceb1¡/em¿ genes are potential markers of progression. *The American Journal of Pathology*, *181*(5), 1585–1594. https://doi.org/10.1016/j.ajpath.2012.08.005

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/https://doi.org/10.3322/caac.21834

Casal-Mouriño, A., Ruano-Ravina, A., Lorenzo-González, M., Rodríguez-Martínez, Á., Giraldo-Osorio, A., Varela-Lema, L., Pereiro-Brea, T., Barros-Dios, J. M., Valdés-Cuadrado, L., & Pérez-Ríos, M. (2020). Epidemiology of stage iii lung cancer: Frequency, diagnostic characteristics, and survival. *Translational Lung Cancer Research*, *10*(1). https://tlcr.amegroups.org/article/view/38800

Cooper, K. L., Liu, R., & Zhou, X. (2022). Particulate arsenic trioxide induces higher dna damage and reactive oxygen species than soluble arsenite in lung epithelial cells. *Toxicology and Applied Pharmacology*, *457*, 116320. https://doi.org/https://doi.org/10.1016/j.taap.2022.116320

Groth, D., Hartmann, S., Klie, S., & Selbig, J. (2013). Principal components analysis. *Methods in molecular biology (Clifton, N.J.)*, *930*, 527–547. https://doi.org/10.1007/978-1-62703-059-5_22

Guo, Y., Wang, Y. L., Su, W. H., Yang, P. T., Chen, J., & Luo, H. (2020). Three genes predict prognosis in microenvironment of ovarian cancer. *Frontiers in Genetics*, *Volume 11 - 2020*. https://doi.org/10.3389/fgene.2020.00990

He, S., Li, H., Cao, M., Sun, D., Yang, F., Yan, X., Zhang, S., He, Y., Du, L., Sun, X., Wang, N., Zhang, M., Wei, K., Lei, L., Xia, C., Peng, J., & Chen, W. (2022). Survival of 7,311 lung cancer patients by pathological stage and histological classification: A multicenter hospital-based study in china. *Translational Lung Cancer Research*, *11*(8). https://tlcr.amegroups.org/article/view/66962

Huang, J., Zhang, W., Lin, D., Lian, L., Hong, W., & Xu, Z. (2022). Chromobox homologue 7 acts as a tumor suppressor in both lung adenocarcinoma and lung squamous cell carcinoma via inhibiting erk/mapk signaling pathway. *Evidence-Based Complementary and Alternative Medicine*, *2022*(1), 4952185. https://doi.org/https://doi.org/10.1155/2022/4952185

Joseph, J., Radulovich, N., Wang, T., Raghavan, V., Zhu, C.-Q., & Tsao, M.-S. (2020). Rho guanine nucleotide exchange factor arhgef10 is a putative tumor suppressor in pancreatic ductal adenocarcinoma. *Oncogene*, *39*(2), 308–321. https://doi.org/10.1038/s41388-019-0985-1

Luo, Y.-H., Luo, L., Wampfler, J. A., Wang, Y., Liu, D., Chen, Y.-M., Adjei, A. A., Midthun, D. E., & Yang, P. (2019). 5-year overall survival in patients with lung cancer eligible or ineligible for screening according to us preventive services task force criteria: A prospective, observational cohort study. *The Lancet Oncology*, *20*(8), 1098–1108. https://doi.org/10.1016/S1470-2045(19)30329-8

Marmor, H. N., Zorn, J. T., Deppen, S. A., Massion, P. P., & Grogan, E. L. (2021). Biomarkers in lung cancer screening: A narrative review. *Current Challenges in Thoracic Surgery*, *5*(0). https://ccts.amegroups.org/article/view/49612

Mattiuzzi, C., & Lippi, G. (2019). Current cancer epidemiology. *Journal of Epidemiology and Global Health*, *9*, 217–222. https://doi.org/10.2991/jegh.k.191008.001

McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010). Frozen robust multiarray analysis (frma). *Biostatistics*, *11*(2), 242–253. https://doi.org/10.1093/biostatistics/kxp059

Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: Review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, *63*(7), 07TR01. https://doi.org/10.1088/1361-6560/aab4b1

Oken, M. M., Hocking, W. G., Kvale, P. A., Andriole, G. L., Buys, S. S., Church, T. R., Crawford, E. D., Fouad, M. N., Isaacs, C., Reding, D. J., Weissfeld, J. L., Yokochi, L. A., O'Brien, B., Ragard, L. R., Rathmell, J. M., Riley, T. L., Wright, P., Caparaso, N., Hu, P., . . . PLCO Project Team, f. t. (2011). Screening by chest radiograph and lung cancer mortality: The prostate, lung, colorectal, and ovarian (plco) randomized trial. *JAMA*, *306*(17), 1865–1873. https://doi.org/10.1001/jama.2011.1591

Singhal, S., Ruprecht, N. A., Sens, D., Tavakolian, K., Gardner, K. L., & Singhal, S. K. (2022). Association between arsenic level, gene expression in asian population, and in vitro carcinogenic bladder tumor. *Oxidative Medicine and Cellular Longevity*, *2022*(1), 3459855. https://doi.org/https://doi.org/10.1155/2022/3459855

Smith-Bindman, R., Chu, P. W., Azman Firdaus, H., Stewart, C., Malekhedayat, M., Alber, S., Bolch, W. E., Mahendra, M., Berrington de González, A., & Miglioretti, D. L. (2025). Projected lifetime cancer risks from current computed tomography imaging. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2025.0505

Sonbul, S. N., Aleskandarany, M. A., Kurozumi, S., Joseph, C., Toss, M. S., Diez-Rodriguez, M., Nolan, C. C., Mukherjee, A., Martin, S., Caldas, C., Ellis, I. O., Green, A. R., & Rakha, E. A. (2018). *Saccharomyces cerevisiae*-like 1 (sec14l1) is a prognostic factor in breast cancer associated with lymphovascular invasion. *Modern Pathology*, *31*(11), 1675–1682. https://doi.org/10.1038/s41379-018-0092-9

Speer, R. M., Zhou, X., Volk, L. B., Liu, K. J., & Hudson, L. G. (2023). Chapter six - arsenic and cancer: Evidence and mechanisms. In M. Costa (Ed.), *Environmental carcinogenesis* (pp. 151–202, Vol. 96). Academic Press. https://doi.org/https://doi.org/10.1016/bs.apha.2022.08.001

Stephens, E. K. H., Sigcha, J. G., Lopez-Loo, K., Yang, I. A., Marshall, H. M., & Fong, K. M. (2023). Biomarkers of lung cancer for screening and in never-smokers—a narrative review. *Translational Lung Cancer Research*, *12*(10). https://tlcr.amegroups.org/article/view/79853

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/https://doi.org/10.1111/j.1553-2712.2011.01185.x

Wang, D.-w., Tang, J.-y., Zhang, G.-q., & Chang, X.-t. (2020). Arhgef10l expression regulates cell proliferation and migration in gastric tumorigenesis. *Bioscience, Biotechnology, and Biochemistry*, *84*(7), 1362–1372. https://doi.org/10.1080/09168451.2020.1737503

Yang, W., Wang, Q., Howell, K. L., Lee, J. T., Cho, D.-S. C., Murray, J. M., & Nishikura, K. (2005). Adar1 rna deaminase limits short interfering rna efficacy in mammalian cells *. *Journal of Biological Chemistry*, *280*(5), 3946–3953. https://doi.org/10.1074/jbc.M407876200

Yang, Y., Hu, Z., Sun, H., Yu, Q., Yang, L., Yin, F., Sun, Y., Pu, L., Zhu, X., Li, S., Chen, X., & Zhao, Y. (2021). CBX7, a potential prognostic biomarker in lung adenocarcinoma. *Onco. Targets. Ther.*, *14*, 5477–5492.

Zeng, H., Zhang, Y., Yi, Q., Wu, Y., Wan, R., & and, L. T. (2015). Crim1, a newfound cancer-related player, regulates the adhesion and migration of lung cancer cells [PMID: 26653968]. *Growth Factors*, *33*(5-6), 384–392. https://doi.org/10.3109/08977194.2015.1119132

Zhou, G. (2019). Tobacco, air pollution, environmental carcinogenesis, and thoughts on conquering strategies of lung cancer. *Cancer Biology & Medicine*, *16*(4), 700–713. https://doi.org/10.20892/j.issn.2095-3941.2019.0180