Original Research Article

Differential Gene Expression Associated with Tumor Status and Smoking in Lung Tissue

Leyla Bac

Barrington High School: 220 Lincoln Ave, Barrington, RI 02806, United States

ABSTRACT

Smoking is a major contributor to lung adenocarcinoma, yet its specific gene expression signature remains unclear. This study, with a sample size of n=107, examined gene expression differences across current, former, and never smokers, as well as tumor and non-tumor lung samples, to better define smoking's molecular impact. Ordinary least squares regression and ANOVA were applied to the microarray GSE10072 data, followed by false discovery rate correction using the Benjamini–Hochberg method with a statistical threshold of q<0.05. The analysis identified many differentially expressed genes for each comparison: 10562 genes for tumor vs non-tumor, 627 genes for current vs never smokers, and 18 genes for former vs never smokers. Specific genes of interest include *CYP1B1*, upregulated by 110.8-148.8% across 3 probes in current smokers, and *ADARB1,* downregulated by *36.4%, 23.3%, and 73.2%* in current smokers, former smokers, and tumor samples, respectively. These genes, implicated in pathways such as AHR-CYP and PI3K/AKT signaling, highlight potential mechanistic links between smoking exposure and lung cancer development. The results suggest that smoking leaves distinct and lasting molecular signatures that may contribute to tumor progression, offering potential targets for improved therapeutic strategies.

Keywords: Differential expression; lung adenocarcinoma; OLS; ANOVA; molecular signatures; upregulation; downregulation; PI3K/AKT signaling pathways; downstream analysis; tumor progression

INTRODUCTION

Lung cancer contributes to the largest portion of cancer-related deaths each year (1), and its limitations in early diagnosis and treatment give it a 5-year survival rate of less than 15% (2). The abnormal division of cells in lung cancer is caused by the presence of mutations, either inherited or acquired during an individual's lifetime. To improve the efficiency and accuracy of cancer therapies, it is crucial to study the mutations that cause them and consider the factors that contribute to their development. Mutations have many potential causes, but this study primarily focuses on a specific mutagen - cigarette smoke. Cigarette smoke is a known mutagen that significantly increases an individual's risk of lung cancer. However, smoking's specific gene expression signature remains a topic of interest. Considering the impact of cigarette smoking on gene expression and resulting mutations is key to improving interventional therapies and reducing fatality rates in lung cancer.

Gene expression is the process that converts genetic information stored in DNA into a functional product, and studying its levels in different samples is crucial for determining the molecular impacts of cancer. Gene

Corresponding author: Leyla Bac, E-mail: leylabac@gmail.com.

Accepted December 22, 2025
https://doi.org/10.70251/HYJR2348.411222

expression levels refer to the amount of gene activity, specifically the amount of RNA and proteins produced by a gene. They can become dysregulated in the presence of cancer, making them a useful biomarker for measuring cancer development and progression. Measuring gene expression levels can also reveal signatures, or the molecular alterations associated with certain diseases. This is a powerful tool for understanding the factors that drive cancer development.

Cigarette smoking has already been found to affect gene expression levels, especially through DNA methylation (3). DNA methylation refers to the process by which methyl groups are added to DNA molecules, repressing them by either inhibiting transcription or recruiting proteins that silence gene activity. Several studies have identified multiple genes that exhibit changes in methylation and expression levels due to smoking (4, 5). Smoking has been found to alter DNA methylation levels by creating double-stranded breaks in DNA that become methylated during the repair process. Smoking also triggers an increase in the production of S-adenosylmethionine, a universal methyl donor involved in methylation (3). Finally, smoking can cause the downregulation of mRNA and protein expression of *DNMT1*, an enzyme crucial for maintaining stable methylation levels (3).

Despite the extensive research concerning smoking's impact on methylation and gene expression, the precise molecular pathway linking these changes to cancer development remains unclear. Identifying the specific molecular alterations caused by tobacco smoke and the genes affected can potentially reveal this pathway. Specifically, observing differing gene expression levels between current and never smokers, as well as between former and never smokers, can help conclude the lasting impacts of smoking at the molecular level. Comparing these various smoking statuses can effectively uncover underlying patterns in mutations that can be attributed to smoking through statistical analysis.

Gene expression levels for specific genes were measured in 105 current, former, and never-smokers (6). Measuring differentially expressed genes in cancerous and non-cancerous cells in these groups can establish a relationship between cigarette smoking and differences in gene expression. The objective of this study is to identify a statistically significant relationship between differences in gene expression levels among current, former, and never-smokers with differing cancer statuses, as well as identify key genes that can provide a mechanistic link between smoking and cancer.

METHODS AND MATERIALS

Dataset

The dataset includes subjects from the EAGLE (Environment and Genetics in Lung Cancer Etiology) study, a lung cancer study conducted in Italy. EAGLE provides 107 samples: 58 tumor and 49 non-tumor tissues (7). The report was based on 122 original samples; however, 15 duplicate and triplicate samples were averaged, resulting in a final sample size of 107 (6). Final expression values for each sample were found after running a standard multi-chip normalization pipeline on the processed cell intensity data created for each usable tissue. All arrays were generated on the Affymetrix Human Genome U133 Plus 2.0 platform (GPL570). The expression data for the final 107 samples, as well as the phenotypic profiles of each subject through GSE10072, were accessed to explore connections between the cancerous and non-cancerous samples and their resulting gene expression levels.

Preprocessing

The mean expression level for each gene was calculated and plotted in a histogram. To visualize their distribution, the gene expression variances were computed and displayed in a histogram as well. Principal component analysis (PCA; scikit-learn v1.4) was applied to mean-centered, variance-scaled expression (StandardScaler) to explore structure and potential confounding. PCA analysis, which decomposes each component's contribution to overall variance, was carried out to identify stratification between individuals based on sex, condition, and age differences. This helps to determine the effect of confounding factors on downstream differential expression analysis.

All analyses were performed in Python 3.12 using pandas (v2.x), statsmodels (v0.14), NumPy (v1.26+), scikit-learn (v1.4), and matplotlib (v3.8). Raw microarray intensities for each GEO sample ID were imported into Python and transposed to create columns corresponding to genes and rows to biological samples. When available, GEO series matrices were used as pre-normalized inputs; otherwise, processed intensity files supplied with the series were used. The accompanying phenotype data, which includes smoking status, gender, tumor/normal status, age at diagnosis, and textual sample titles, was aligned with the expression data by creating matching row indices. Missing values were handled by listwise deletion: rows with missing tumor status, gender, smoking, or per-gene expression were dropped prior to model fitting. To clean the phenotype data, categorical variables were created. Tumor status mapped each entry to either ‘tumor’, ‘normal’, or ‘other’. Smoking status was mapped to either N (never), F (former), or C (current), and gender became a pandas categorical. Categorical encodings used explicit pandas Categorical dtypes; for model formulas, statsmodels’ C() was used with treatment coding and clearly defined reference levels (e.g., “Never” as reference where contrasts were needed). GEO accession codes were used to merge the expression and phenotype tables and match samples with their metadata. To keep genes that carried the most information, the most highly variable genes, which were chosen to be those above the 5th percentile for both mean and variance, were retained, and each gene was z-scored using StandardScalar to normalize the data.

Statistical Analysis

For each gene, an ordinary least-squares (OLS) model was fit, a technique to model the relationship between one or more predictors by minimizing the squared differences between observed and predicted values. The model utilized expression, tumor status, gender, and smoking status. Specifically, type II ANOVA with dummy coding was applied to the fitted OLS to test the main effects of tumor status and smoking while accounting for other covariates. An adequate main-effects model, approximate normal and homoscedastic residuals, independence, and acceptable imbalance without severe collinearity were assumed. Assumptions were supported by log-scale modeling, balanced-enough group sizes with low collinearity, independence ensured via random effects for repeated patients, and sensitivity checks with interaction terms; batch was not modeled and is acknowledged as a limitation to be addressed in follow-up. Expression and phenotype tables were inner-joined on sample IDs, and confounding factors were addressed by including sex (and age where specified) as covariates, using explicit treatment coding with fixed reference levels (e.g., “Never Smoked”), and dropping rows with missing covariates or expression for the gene being tested. Genes causing fit errors were skipped, and for each retained gene, F-statistics and p-values for tumor and smoking effects were stored. Where effect sizes were summarized from dummy-coded coefficients (β) on the log2 expression scale, fold-change (FC) was computed as FC = 2**β, and percent change as (FC − 1) × 100. The β coefficients are an adjusted difference in mean gene expression on the log2 scale, representing the size and direction of the association.

Multiple Testing Correction

To control the false discovery rate, the Benjamini-Hochberg procedure was applied to each p-value, and the resulting q-values were recorded, enabling the creation of a list of tumor-responsive genes ranked by q-value.

RESULTS

To reduce the dataset’s dimensions, principal component analysis (PCA) was conducted, and the new principal components were plotted to visualize their relationships and identify potential confounding variables. The initial quality control analysis is provided (Figure 1). The analysis reveals clear clustering based on tumor phenotypes but not on age, gender, or smoking status. This means a correlation in the gene expression data with tumor status can be identified, but confounding variables such as age, gender, or smoking status are expected to have minimal effect, though PCA alone may not completely rule out confounding factors.

Next, differentially expressed genes between tumor and non-tumor samples were identified. The tumor effect for each gene was isolated and plotted against corrected significance (Figure 2). The top 5 statistically significant probe IDs, defined by those with the smallest Benjamini-Hochberg corrected p-values (q values), were identified (Table 1).

These genes have all previously been associated with lung adenocarcinoma development. The *FAM107A* gene, localized in the 3p14.3 chromosomal region, is significantly downregulated in lung tumor samples (8). In this analysis, comparable downregulation, with a 90% decrease in expression, was identified in tumor samples. CD36 has been repeatedly implicated in lung adenocarcinoma by supporting fatty-acid uptake and lipid-metabolism reprogramming that drives proliferation and metastasis (9). This analysis found an 81% decrease in expression in tumor samples. EDNRB is frequently hypermethylated and downregulated in non-small-cell lung cancer (10), consistent with the 78% decrease in expression found in this study**.** NPR1 signaling is increasingly connected to cancer biology, with some experimental models indicating NPR1 can modulate tumor growth and metastatic behavior (11). The decrease in expression found in this analysis corresponds with this potential lack of tumor modulation by NPR1.

Differentially expressed genes between current and never smokers, as well as between former and never smokers, were also found. A greater number of differentially expressed genes under the FDR threshold in

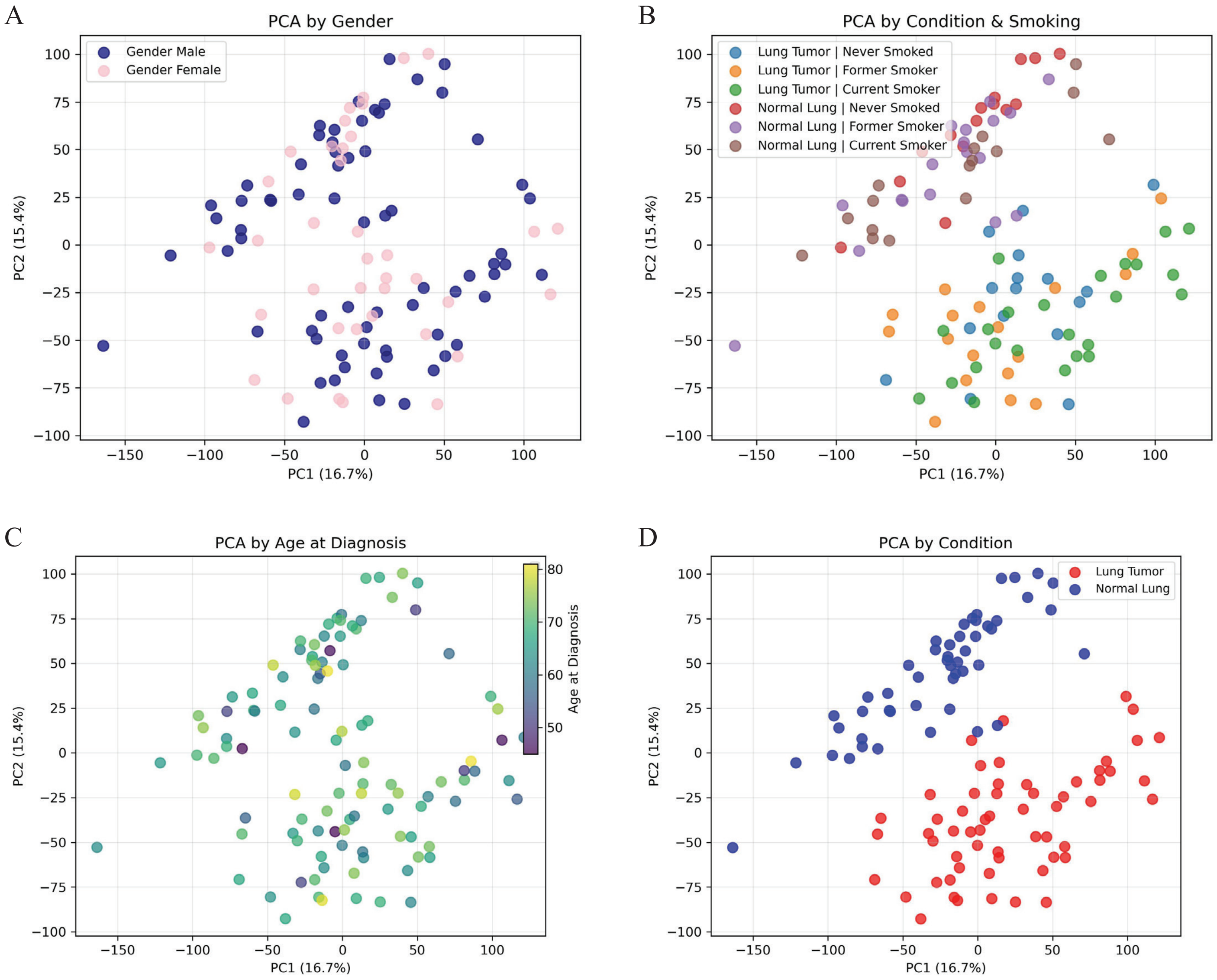

Figure 1. PCA plot of samples colored by A) gender, B) age at diagnosis, C) smoking status, and D) tumor vs normal tissue. Graphic created by the student researcher and Kristina Kordova using Anaconda, 2025.

Table 1. The top 5 most statistically significant probe IDs are matched to their respective genes, and their β tumor shift values and HB corrected p-values, fold change, and percent changes are displayed.

Probe ID	Gene	β-tumor	q-tumor	FC-tumor	% change tumor
209074_s_at	FAM107A	-3.39	1.118e-38	0.10x	-90.3%
209555_s_at	CD36	-2.41	4.489e-37	0.19x	-81.3%
204271_s_at	EDNRB	-2.22	2.734e-36	0.21	-78.6%
219719_at	HIGD1B	-1.36	3.417e-36	0.39x	-61.1%
32625_at	NPR1	-1.16	3.417e-36	0.44x	-55.9%

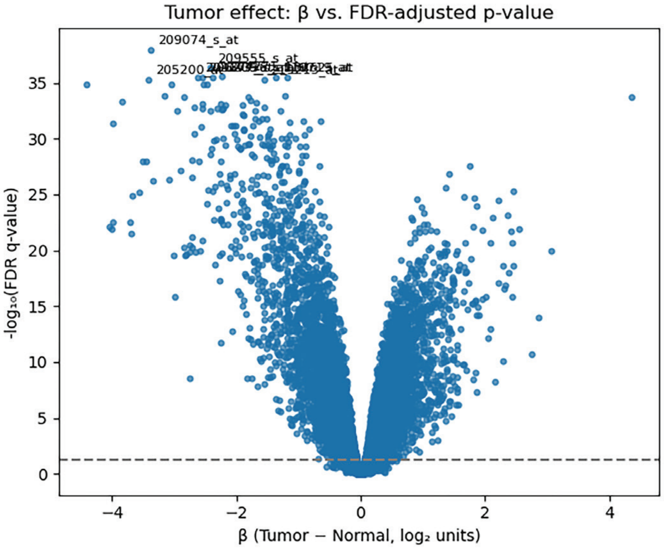

Figure 2. Volcano plot revealing genes above the 0.05 FDR significance threshold and their respective β shift, revealing size and direction for the tumor factor. The 5 most statistically significant probe IDs are labeled. Graphic created by the student researcher and Kristina Kordova using Anaconda, 2025.

the current vs never comparison were identified than in the former vs never comparison (Figure 3A). This aligns with the findings of the study where the data originated from, which found a greater number of significant genes between current and never smokers than between former and never smokers (6). The statistically significant genes ($p<0.05$) were plotted in volcano plots (Figures 3B and 3C). The 7 probe IDs with the lowest q-values for current vs never were identified, as well as 6 probe IDs for former vs never due to overlap in the genes they mapped to (Tables 2 and 3).

Many of these genes have been previously described to have associations with smoking. For example, the *CYP1B1* gene has been found to interact with NKK, a powerful lung carcinogen associated with tobacco smoke (12). 3 probes marked this gene, which, in this analysis, reveal a 110.8%, 148.8%, and 85.5% increase in expression in current smokers. The *XIST* gene is differentially expressed between former smokers and never smokers, and has been established to be associated with smoking (13). This gene was identified as downregulated in this analysis in both current vs never (-50.8%), and former vs never smokers (-81.5% and -85.4%). *RPS4Y1* has previously been found to be upregulated in smokers vs nonsmokers (14). In this analysis, a 577.7% increase in expression was present in former smokers, as well as a 90.7% increase in current smokers. The gene *GGH* was differentially expressed in the study this data originated from (6), and is associated with oxidative stress caused by smoking (15). *RPS4X* is downregulated in the effector club cells of smokers, which are cells in the lung (16). This analysis reveals a 52.8% decrease in *RPS4X* expression in former smokers, as well as a 45.3% decrease in current smokers. *CTNNBIP1* is being explored as a powerful prognostic marker in lung cancer, which, when overexpressed, may be useful in treating cancer (17). However, more research is necessary in this area. The gene *DDX3Y* is known to have a role in lung cancer that is dependent on smoking status, though the exact mechanisms of this remain unclear (18). The *DEXI* gene has been found overexpressed in patients with emphysema, a chronic lung condition causing shortness of breath (19), as well as differentially expressed when exposed to cigarette smoke (20). The gene *SMCY* is being researched for its

Table 2. The 7 most significant probe IDs (top 5 genes) for current vs never are matched to their respective genes, and their β tumor shift values and HB corrected p-values, fold change, and percent changes are displayed.

Probe ID	Gene	β-Current-Never	q-Current-Never	FC-Current-Never	% change
202436_s_at	CYP1B1	+1.185	1.283e-03	2.11x	+110.8%
202437_s_at	CYP1B1	+1.181	1.283e-03	2.49x	+148.8%
203733_at	DEXI	-1.217	1.283e-03	0.75x	-25.0%
203081_at	CTNNBIP1	-1.172	1.436e-03	0.76x	-23.7%
202435_s_at	CYP1B1	+1.111	1.918e-03	1.85x	+85.5%
203560_at	GGH	+1.108	1.92e-03	2.15x	+115.5%
221728_x_at	XIST	-1.022	1.92e-03	0.49x	-50.8%

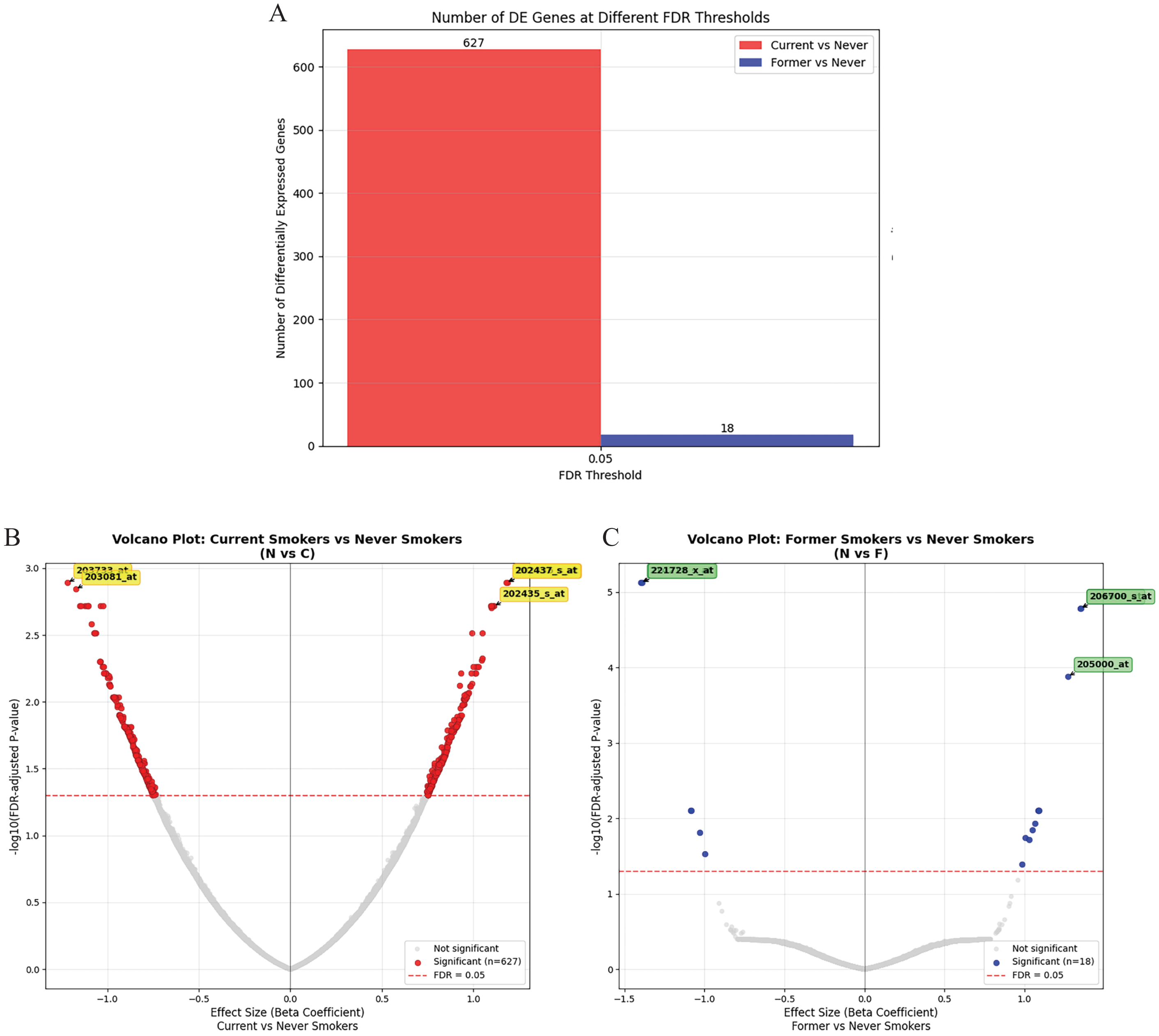

Figure 3. A) Bar graphs showing the number of differentially expressed genes for each smoking group comparison. B) Volcano plot visualizing the beta smoking effect, revealing size and direction of the association, and adjusted p-values of gene expression in current vs never smokers. C) Volcano plot visualizing the beta smoking effect, revealing size and direction of the association, and adjusted p-values of gene expression in former vs never smokers. Graphic created by the student researcher and Kristina Kordova using Anaconda, 2025.

Table 3. The 6 most significant probe IDs for former vs never are matched to their respective genes, and their β tumor shift values and HB corrected p-values, fold change, and percent changes are displayed.

Probe ID	Gene	β-Former-Never	q-Former-Never	FC-Former-Never	% change
214218_s_at	XIST	-1.397	8.00e-06	0.18x	-81.5%
221728_x_at	XIST	-1.391	8.00e-06	0.15x	-85.4%
201909_at	RPS4Y1	+1.351	1.60e-05	6.78x	+577.7%
206700_s_at	SMCY	+1.352	1.60e-05	2.62x	+162.1%
205000_at	DDX3Y	+1.272	1.31e-04	5.23x	+423.2%
213347_x_at	RPS4X	-1.082	7.88e-03	0.47x	-52.8%

potential role in prostate cancer, and its link to smoking or lung cancer is unclear (21).

To combine the two analyses, the genes that are simultaneously differentially expressed between smoking groups and cancer groups were identified. The FDR-adjusted p-values for smoking were plotted against the FDR-adjusted p-values for tumors, and the top 5 genes with the lowest p-values for both were identified (Figure 4).

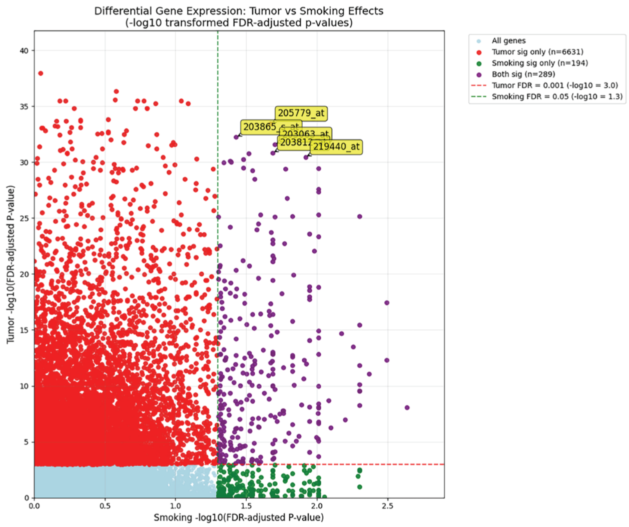

Figure 4. Smoking p-values are plotted against tumor p-values, with each colored section representing a different category of differentially expressed genes. Genes marked with red are differentially expressed between tumor groups, and those marked green are between smoking groups. Genes marked purple are significant for both, and those marked in blue were not found to be significantly associated with cancer or smoking in this dataset. The 5 most significant genes that are differentially expressed in both the tumor and smoking groups are labeled. Graphic created by the student researcher and Kristina Kordova using Anaconda, 2025.

RAMP2 has previously been found to be downregulated in smoking-associated lung cancer, and its overexpression can inhibit the proliferation and progression of cancer cells (22). This analysis reveals a 32% decrease in expression in current smokers, a 0.7% decrease in former smokers, and a 70.7% decrease in cancerous samples. *RAMP2* acts as a receptor activity-modifying protein that influences vascular signaling and controls cell growth with the help of *CALCRL*. *RAMP2* downregulation enables tumor growth by reducing cell growth control, which allows tumors to expand easily (23).

In lung squamous cell carcinoma, *ADARB1* is downregulated as well and is being investigated as a potential biomarker for cancer progression (24). *ADARB1* is an RNA editing enzyme that replaces adenosine with inosine. Low *ADARB1* expression leads to a decrease in these edits that allow pro-growth pathways to become overactive (24). Smoking has been proven to decrease *ADARB1* expression by causing an oxidant/antioxidant imbalance that alters the A-to-I RNA editing that *ADARB1* normally carries out (25). The results in Table 4 correspond with this downregulation, displaying a 36.4% decrease in expression in current smokers, a 23.3% decrease in former smokers, and a 73.2% decrease in tumor samples.

The *PPM1F* gene is upregulated in breast cancer and is associated with smoking behavior. *PPM1F* expression is correlated with the receptor α9-nAChR's expression, and nicotine has been proven to induce both α9-nAChR expression and *PPM1F* expression. (26). The same study concluded that *PPM1F* could work downstream of α9-nAChR to promote nicotine-induced carcinogenic signals for breast cancer (26). This divergence from lung cancer is plausible given tissue specificity, model differences (bulk primary lung tissue vs breast cell lines), and cell-composition effects in bulk arrays. Biologically, PPM1F's

Table 4. The top 5 most significant probe IDs for both tumors and smoking groups are matched to their respective genes, and their q values and % changes are displayed.

Probe ID	Gene	q-Tumor	q-Smoking	%-Current-Never	%-Former-Never	%-tumor
205779_at	RAMP2	3.651e-34	0.0213	-32.0%	-0.7%	-70.7%
203865_s_at	ADARB1	5.981e-33	0.037	-36.4%	-23.3%	-73.2%
203063_at	PPM1F	2.682e-32	0.020	-15.7%	-3.3%	-36.4%
203812_at	SLIT3	1.504e-31	0.021	-15.8%	+2.9%	-43.7%
219440_at	RAI2	3.560e-31	0.012	-31.9%	-5.2%	-63.9%

role as a stress-/signal-modulating phosphatase could be context-dependent, with tumor microenvironment, hypoxia, or cell-type shifts in the lung favoring reduced transcript abundance despite smoking exposure. The results in Table 4 reveal downregulation by 15.7% in current smokers, 3.3% in former smokers, and 36.4% in cancer patients.

SLIT3 was found to be downregulated in vitro lung cancer cells, and continues to be studied for its potential role in lung cancer development (27). *SLIT3* has also been linked to a disruption of the *SLIT3* function that may affect tobacco dependence (28). This analysis found downregulation of 15.8% in current smokers and 43.7% in tumor samples, as well as a 2.9% upregulation in former smokers. *SLIT3* is part of the Slit-Robo signaling pathway, which is involved in cellular migration, as well as the WNT/β-catenin pathway, which regulates cell growth (27). *SLIT3* underexpression makes these pathways less regulated, causing excessive growth and migration that speeds up lung cancer onset and progression (27).

High expression of the gene *RAI*2 has been shown to act as a tumor suppressor in non-small cell lung cancer by allowing microRNA miR-101-3p to suppress EZH2, a protein that drives cancer growth and survival (29). Conversely, *RAI2* expression is generally decreased in lung cancer patients, and this analysis calculated a 63.9% decrease in expression in tumor samples. Though *RAI2* expression's link to smoking remains unclear, this analysis found a 31.9% downregulation in current smokers and a 5.2% downregulation in former smokers.

DISCUSSION

In this analysis, 107 tissue samples, consisting of current smokers, former smokers, never smokers, tumor patients, and non-tumor patients, were used to identify differentially expressed genes that are statistically significant between each smoking and tumor group. Statistical analysis was conducted by running per-gene ANOVA tests and compiling these results with F-statistics, p-values, and FDR-adjusted q-values for each gene. Several differentially expressed genes were found in each group comparison. The primary genes identified were *FAM107A* and *HIGD1B* between tumor and non-tumor samples; *CYP1B1*, *XIST*, and *GGH* between current and never smokers; *RPS4Y1*, *DDX3Y*, and *XIST* between former and never smokers; and *RAMP2*, *ADARB1*, and *SLIT3* between smoking and tumor samples. Although this cohort has been analyzed previously, this paper aims to organize these markers into pathway-centered, drug-relevant programs and outline a translational route from gene-level signals to clinically testable future steps.

These genes are a part of several pathways that relate to both tumor progression and tobacco smoke exposure. The first is the AHR-CYP xenobiotic-response axis for genes like *CYP1B1*, suggesting it is a key gene in the connection between smoking and lung cancer. *CYP1B1* is part of cytochrome P450, a family of enzymes that metabolize foreign compounds, including carcinogens. Cytochrome P450 is regulated by the aryl hydrocarbon receptor (AHR), a transcription factor that plays a crucial role in regulating cellular processes (30). The PAHs in cigarette smoke activate AHR, inducing changes in cytochrome P450 function that have been linked to lung tumor development and may reveal a mechanistic bridge between smoking and tumor onset and progression (31). The second pathway is Wnt/β-catenin-linked migration/invasion, which is consistent with reported consequences of *SLIT3* loss of function in non-small cell lung cancer (27). The third pathway involves adrenomedullin/CALCRL-RAMP2 signaling, which regulates endothelial integrity and angiogenic tone. It is implicated in the process of adenocarcinoma development, with its disruption favoring vascular leak and metastatic potential (32). Finally, a possible mechanistic link between smoking and lung cancer is outlined in RNA editing via *ADARB1*, a regulatory layer that can modulate oncogenic cascades, including PI3K/AKT/mTOR (33). The PI3K/AKT pathway, in combination with the WNT signaling pathway, is already known to play a role in the development of almost every human cancer (34).

Targeting these pathways is already a focal point of cancer treatment and therapy, and framing these findings at the pathway level directly suggests therapeutic levers (e.g., AHR antagonists/CYP1 inhibitors, Wnt pathway modulators, and agents targeting the adrenomedullin axis). These pathways are associated with the carcinogens presented by smoking, which can cause dysfunction in the regulation of normal processes such as epithelial-mesenchymal transition (EMT). EMT is a process in development that involves tissue regeneration, organ fibrosis, and wound healing, which is normally tightly regulated. However, when the described pathways are altered, EMT becomes less controlled and promotes cancer metastasis, revealing a potential key link between smoking and lung cancer progression (35).

These discoveries demonstrate the high potential of microarray data in identifying genes known to be associated with smoking and cancer development. Further research could benefit from the application of some more advanced methodologies to provide further detail into the mechanistic links in smoking-associated carcinogenesis. For example, to strengthen translational relevance, pathway-level enrichment (e.g., Reactome/GSVA) should be conducted to verify module-level signals beyond single genes (36), and a compact "biological dial" panel with AHR-CYP, Wnt/β-catenin, AM-RAMP1, and RNA editing can be created after validation across external lung adenocarcinoma cohorts. Converting these dials into a clinically portable assay (RT-qPCR or targeted RNA-seq) tied to explicit treatment hypotheses would provide a potential route to clinical testing and impact.

Limitations

The study is limited by reliance on microarray rather than RNAseq technology, which has demonstrated utility in determining the effect of tobacco smoking on inducing a subset of activated tumor-resident T-regs in non-small cell lung cancer. These T-regs, which smoking can activate, will then inactivate anticancer immunity, an avenue worth integrating with the pathway signals identified here (37).

Multi-omics integration is also a powerful tool for determining the effect of smoking on patient prognosis and chemotherapeutic responses, as well as differentially expressed genes. Using multi-omics integration to identify the impact of smoking on patients of different smoking histories can reveal differences in genomic stability, methylation, and immune content that couldn't have been achieved with the microarray method used here (38).

Furthermore, this study is limited by its sample size of only 107 individuals. All samples were also obtained from the Lombardy region of Italy, representing a relatively small European population; hence, these results may not be replicated in other ethnicities. For example, Caucasian Americans and African Americans with lung cancer have differing biological profiles, including differences in the tumor microenvironment and immune responses (39). Associated comorbidities, such as diabetes, other cancers, and heart conditions, could also bias the results. Conditions such as diabetes, congestive heart failure, and peripheral vascular disease have been found to decrease lung cancer survival rates, likely by contributing to other immune responses at large, as well as altering the tumor microenvironment and, hence, tumor development (40).

Finally, sex-related confounding remains a concern, as several smoking-associated gene hits, such as *XIST*, *DEXI*, and *RPS4Y1*, map to X/Y chromosomes, despite no apparent separation in the sex-stratified PCA (Figure 1). Biologically, XIST is high in females (silences one X), while RPS4Y1 is male-specific. Although sex was adjusted for, residual confounding or true sex-smoking interactions and a synergy between the genes may exist. Apparent smoking effects on these genes can therefore arise from sex imbalance (e.g., more current smokers are male) or from residual sex effects that weren't effectively removed by adjustment. Further analysis should test sex-smoking relationships or exclude sex-chromosome probes to distinguish confounding variables from genuine sex-linked biology. Further limitations include the absence of batch corrections in the statistical modeling used, which could have inflated false positives and negatives and aliases with tumor or smoking status, potentially producing these misleading results.

CONCLUSION

This study identified and analyzed differentially expressed genes between 107 samples of varying smoking and tumor statuses. Differentially expressed genes between current, never, and former smokers were found, revealing the lasting impact of smoking on the human genome and providing insight into the mechanics of smoking associated with lung cancer. Specifically, *CYP1B1*, which acts on the AHR-CYP axis, and *ADARB1*, part of the PI3K/AKT pathway, provide insight into lung cancer's relation to smoking. Genes such as these that are revealed in this study are very prevalent in smoking-associated lung cancer and should continue to be researched for their potential role as powerful targets for future cancer therapies.

ACKNOWLEDGEMENTS

I would like to thank my mentor Kristina Kordova, PhD candidate at Christ's College, Cambridge, UK for help.

REFERENCES

1. World Health Organization. Cancer. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed 2025-08-29).
2. Liang C, Pan W, Zhou Z & Liu X. Identification of prognostic biomarkers of smoking-related lung cancer. *Journal for Thoracic Disease.* 2024; 16 (2): 1438–1449. https://doi.org/10.21037/jtd-23-1890
3. Lee KW & Pausova Z. Cigarette smoking and DNA methylation. *Frontiers in Genetics.* 2013; 4: 132. https://doi.org/10.3389/fgene.2013.00132
4. Tsai P-C, Glastonbury CA, Eliot MN, Bollepalli S, *et al*. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clinical Epigenetics.* 2018; 10: 126. https://doi.org/10.1186/s13148-018-0558-0
5. Arimilli S, Madahian B, Chen P, *et al.* Gene expression profiles associated with cigarette smoking and moist snuff consumption. *BMC Genomics.* 2017; 18: 156. https://doi.org/10.1186/s12864-017-3565-1
6. Landi MT, *et al.* Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLOS ONE.* 2008; 3 (2): e1651. https://doi.org/10.1371/journal.pone.0001651
7. National Cancer Institute, Division of Cancer Epidemiology and Genetics. EAGLE – Environment And Genetics in Lung cancer Etiology. Available from: https://dceg.cancer.gov/research/who-we-study/cancer-cases-controls/eagle-study (accessed 2025-11-06).
8. Ou D, *et al.* Identification of the putative tumor suppressor characteristics of FAM107A via pan-cancer analysis. *Frontiers in Oncology.* 2022; 12: 861281. https://doi.org/10.3389/fonc.2022.861281
9. Liu H, Guo W, Wang T, Cao P, *et al*. CD36 inhibition reduces non-small-cell lung cancer development through AKT-mTOR pathway. *Cell Biology and Toxicology.* 2024; 40 (1): 10. https://doi.org/10.1007/s10565-024-09848-7
10. Wei L, Ge Y, Li M, Wang R & Chen C. Role of endothelin receptor type B (EDNRB) in lung adenocarcinoma based on multi-omics analysis. *Thoracic Cancer.* 2020; 11 (7): 1885–1890. https://doi.org/10.1111/1759-7714.13474
11. Zhang J, Zhao Z & Wang J. Natriuretic peptide receptor A as a novel target for cancer. *World Journal of Surgical Oncology.* 2014; 12: 174. https://doi.org/10.1186/1477-7819-12-174
12. Church TR, *et al.* Interaction of CYP1B1, cigarette-smoke carcinogen metabolism, and lung cancer risk. *International Journal of Molecular Epidemiology and Genetics.* 2010; 1 (4): 295–309.
13. Ma Q, *et al.* Machine learning reveals impacts of smoking on gene profiles of different cell types in lung. *Life.* 2024; 14 (4): 502. https://doi.org/10.3390/life14040502
14. Irimie AI, *et al.* Differential effect of smoking on gene expression in head and neck cancer patients. *International Journal of Environmental Research and Public Health.* 2018; 15 (7): 1558. https://doi.org/10.3390/ijerph15071558
15. Zanetti F, *et al.* Comparative systems toxicology analysis of cigarette smoke and aerosol from a candidate modified risk tobacco product in organotypic human gingival epithelial cultures: A 3-day repeated exposure study. *Food and Chemical Toxicology.* 2017; 101: 15–35. https://doi.org/10.1016/j.fct.2016.12.027
16. Rostami MR, *et al.* Smoking shifts human small airway epithelium club cells toward a lesser differentiated population. *NPJ Genomic Medicine.* 2021; 6: 73. https://doi.org/10.1038/s41525-021-00237-1
17. Chang J-M, *et al.* The alteration of CTNNBIP1 in lung cancer. *International Journal of Molecular Sciences.* 2019; 20 (22): 5684. https://doi.org/10.3390/ijms20225684
18. Bol GM, *et al.* DDX3, a potential target for cancer treatment. *Molecular Cancer.* 2015; 14: 188. https://doi.org/10.1186/s12943-015-0461-7
19. Eriksson AM, Emini N, Harbo HF & Berge T. Is DEXI a multiple sclerosis susceptibility gene? *International Journal of Molecular Sciences.* 2025; 26 (3): 1175. https://doi.org/10.3390/ijms26031175
20. Davison LJ, *et al.* Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Human Molecular Genetics.* 2012; 21 (2): 322–333. https://doi.org/10.1093/hmg/ddr468

21. Dasari VK, *et al.* Expression analysis of Y chromosome genes in human prostate cancer. *The Journal of Urology.* 2001; 165 (4): 1335–1341. https://doi.org/10.1097/00005392-200104000-00080, https://doi.org/10.1016/S0022-5347(01)69895-1
22. Chen Y, *et al.* Network analysis of differentially expressed smoking-associated mRNAs, lncRNAs and miRNAs reveals key regulators in smoking-associated lung cancer. *Experimental and Therapeutic Medicine.* 2018; 16 (6): 4991–5002. https://doi.org/10.3892/etm.2018.6891
23. Yue W, *et al.* Frequent inactivation of RAMP2, EFEMP1 and Dutt1 in lung cancer by promoter hypermethylation. *Clinical Cancer Research.* 2007; 13 (15 Pt 1): 4336–4344. https://doi.org/10.1158/1078-0432.CCR-07-0015
24. Wang X, *et al.* Role of downregulated ADARB1 in lung squamous cell carcinoma. *Molecular Medicine Reports.* 2020; 21 (3): 1517–1526. https://doi.org/10.3892/mmr.2020.10958
25. Takizawa M, *et al.* Decrease in ADAR1 expression by exposure to cigarette smoke enhances susceptibility to oxidative stress. *Toxicology Letters.* 2020; 331: 22–32. https://doi.org/10.1016/j.toxlet.2020.05.019
26. Tu S-H, *et al.* Protein phosphatase Mg^{2+}/Mn^{2+} dependent 1F promotes smoking-induced breast cancer by inactivating phosphorylated-p53-induced signals. *Oncotarget.* 2016; 7 (47): 77516–77531. https://doi.org/10.18632/oncotarget.12717
27. Qiu Z, *et al.* SLIT3 deficiency promotes non-small cell lung cancer progression by modulating UBE2C/WNT signaling. *Open Life Sciences.* 2024; 19 (1): 20220956. https://doi.org/10.1515/biol-2022-0956
28. García-González J, *et al.* Identification of slit3 as a locus affecting nicotine preference in zebrafish and human smoking behaviour. *eLife.* 2020; 9: e51295. https://doi.org/10.7554/eLife.51295
29. Sun H, Zhu R, Guo X, Zhao P, *et al*. Exosome miR-101-3p derived from bone marrow mesenchymal stem cells promotes radiotherapy sensitivity in non-small cell lung cancer by regulating DNA damage repair and autophagy levels through EZH2. *Pathology – Research and Practice.* 2024; 256: 155271. https://doi.org/10.1016/j.prp.2024.155271
30. Walser T, *et al.* Smoking and lung cancer: The role of inflammation. *Proceedings of the American Thoracic Society.* 2008; 5 (8): 811–815. https://doi.org/10.1513/pats.200809-100TH
31. Nebert DW, Dalton TP, Okey AB & Gonzalez FJ. Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *Journal of Biological Chemistry.* 2004; 279: 23847–23850. https://doi.org/10.1074/jbc.R400004200
32. Tanaka M, Koyama T, Doi K, *et al.* The endothelial adrenomedullin-RAMP2 system regulates vascular integrity and suppresses tumour metastasis. *Cardiovascular Research.* 2016; 111 (4): 398–409. https://doi.org/10.1093/cvr/cvw166
33. Zhang Y, Li L, Mendoza JJ, Wang D, *et al*. Advances in A-to-I RNA editing in cancer. *Molecular Cancer.* 2024; 23: 280. https://doi.org/10.1186/s12943-024-02194-6
34. Yang J, Nie J, Ma X, *et al.* Targeting PI3K in cancer: Mechanisms and advances in clinical trials. *Molecular Cancer.* 2019; 18: 26. https://doi.org/10.1186/s12943-019-0954-x
35. Tsay JJ, *et al.* Aryl hydrocarbon receptor and lung cancer. *Anticancer Research.* 2013; 33 (4): 1247–1256.
36. Croft D, O'Kelly G, Wu G, Haw R, *et al*. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research.* 2011; 39 (Database issue): D691–D697. https://doi.org/10.1093/nar/gkq1018
37. Hu Y, *et al.* Exposure to tobacco smoking induces a subset of activated tumor-resident Tregs in non-small cell lung cancer. *Translational Oncology.* 2022; 15 (1): 101261. https://doi.org/10.1016/j.tranon.2021.101261
38. Wang R, *et al.* Multi-omics analysis of the effects of smoking on human tumors. *Frontiers in Molecular Biosciences.* 2021; 8: 704910. https://doi.org/10.3389/fmolb.2021.704910
39. NCI Staff. (2018, January 23). *Study finds biological differences in lung tumors of African Americans and Whites.* Cancer Currents Blog, National Cancer Institute. https://www.cancer.gov/news-events/cancer-currents-blog/2018/lung-cancer-biologic-differences-race
40. Hernández D, *et al.* Survival and comorbidities in lung cancer patients: Evidence from administrative claims data in Germany. *Oncology Research.* 2023; 30 (4): 173–185. https://doi.org/10.32604/or.2022.027262