Gene Expression Profiling of Tobacco Exposure in Buccal Tissue

Alberto Catalano – Network-based Data Analysis – 16/06/2025

Abstract

Tobacco **smoking** is a leading cause of preventable morbidity and mortality, with widespread biological effects including those on the **oral epithelium**, the first point of contact with tobacco smoke. This study investigated the transcriptomic differences in buccal mucosal tissues between **smokers** and **never-smokers** using **microarray data** from 79 human subjects. After rigorous quality control and feature selection, 803 genes were retained for downstream analysis. Unsupervised methods, including PCA, k-means, and hierarchical clustering, revealed a strong molecular separation between the two groups. Supervised machine learning models—Random Forest (RF), Linear Discriminant Analysis (LDA), and LASSO logistic regression—demonstrated high classification accuracies, with LASSO outperforming others slightly. A network-based approach using SCUDO and Spinglass community detection confirmed these findings. Functional enrichment of the most predictive genes identified pathways associated with xenobiotic metabolism, inflammation, and carcinogenesis, notably implicating the JAK-STAT signaling pathway, cytochrome P450 enzymes, and oxidative stress-related genes. Together, these results demonstrate that **smoking induces a distinct and detectable transcriptomic signature in the oral epithelium**.

1 Introduction

Tobacco smoking remains a global public health crisis, unequivocally established as the leading preventable cause of morbidity and mortality worldwide [1]. Its negative effects extend far beyond respiratory diseases, implicating it as a primary factor in several malignancies, including those of the oral cavity, lung, esophagus, and bladder, alongside cardiovascular and chronic inflammatory conditions [2]. The complex mixture of thousands of chemical compounds in tobacco smoke, including numerous known carcinogens, initiates a cascade of molecular and cellular damage upon contact with biological tissues [3]. The oral epithelium, being the first line of defense and direct point of contact for these toxicants in smokers, is particularly vulnerable [4]. Understanding the initial molecular perturbations in this accessible tissue is critical for elucidating the pathways leading to oral cancer and other smoking-related pathologies.

2 Methods

2.1 Dataset and Pre-processing

We analyzed the dataset GSE17913, which contains microarray-derived transcriptomic profiles from human buccal mucosal biopsies collected to investigate the effects of smoking. The data originate from 80 participants: 40 current smokers with a significant exposure history (>15 pack-years) and 40 never-smokers (<100 lifetime cigarettes), who were carefully age- and gender-matched. Following quality control that identified one outlier sample from a non-smoker, the final dataset comprised 79 samples. Analyses were performed in R, using a fixed *seed (123)* to ensure reproducibility. The initial gene expression matrix, containing 54,675 features, was subjected to a two-step feature selection pipeline to reduce dimensionality and enrich for biologically relevant signals. First, we performed variance-based filtering to retain the 10000—an approach commonly used to remove low-informative probes [5, 6]. Second, from this subset, we conducted independent t-tests between the smoker and never-smoker classes, selecting genes with a p-value < 0.1 [7]. This procedure resulted in a final feature set of 803 genes.

2.2 Machine learning analysis

To identify sample groupings and predictive biomarkers, we used both unsupervised and supervised machine learning approaches. Principal Component Analysis (PCA), k-means, and hierarchical clustering (Ward.D2 linkage on Euclidean distances [8, 9]) were employed for exploratory analysis. Optimal cluster number (k) was determined via silhouette analysis across a range of k between 2 and 10 [10]. Supervised modeling was conducted using a 75/25% train-test split. A Random Forest (RF) classifier with 500 trees was optimized via 10-fold cross-validation, using the accuracy as the performance metric [11]. Features were ranked based on Mean Decrease Accuracy. Additionally, we implemented Linear Discriminant Analysis (LDA) with equal prior probabilities (0.5) and a LASSO model with 10-fold CV-tuned lambda [12, 13]. Performance was evaluated on the test set using confusion matrices and ROC curves [14]. LASSO's embedded feature selection also provided a ranked list of influential predictors based on their non-zero coefficient magnitudes. Since all method were resampled in the same way, a direct accuracy comparison between them was also performed. We also constructed a sample similarity network using SCUDO, a gene signature-based approach that captures global transcriptomic similarity [15]. Optimal hyperparameters (nTop and nBottom) were selected via 10-fold cross-validation. The resulting network was reduced to its largest connected component, and Spinglass community detection was used to partition samples into two robust topological clusters [16].

2.3 Functional and Network-based enrichment Analysis

To functionally interpret the features identified by machine learning models, we used multiple gene set enrichment approaches. The top 200 genes ranked by Random Forest importance scores were analyzed using g:Profiler and DAVID. These tools complement each other by querying distinct annotation sources and applying different statistical frameworks. g:Profiler interrogated GO, KEGG, Reactome, and other pathway databases, with significance assessed via the g:SCS multiple testing correction (adjusted $p<0.05$) [17]. DAVID provided annotation clustering based on GO terms, KEGG pathways, UniProt keywords, and protein domains, using a modified Fisher's exact test (EASE score) to determine enrichment. For the 100 LASSO-selected features, we applied active-subnetwork enrichment using the Pathfinder framework, which integrates differential statistics with interaction networks to identify relevant subnetworks. Enrichment was performed against KEGG and Reactome databases, clustered by gene-set overlap, and visualized via term-gene interaction graphs. We also conducted protein-protein interaction (PPI) analysis using the STRING database (v12). Starting from selected gene symbols, we built an interaction network using STRING's PPI evidence. This network was clustered using k-means ($k = 4$) to define functional gene modules, and STRING's annotation tools characterized these clusters.

3 Results

3.1 Dimensionality Reduction Reveals Smoking-Associated Variation

Following the feature selection and scaling pipeline detailed in the *Methods*, the resulting expression matrix of 803 genes was used for all subsequent analyses. The normalized expression distribution for all samples is displayed in Fig S1. We initiated our investigation with PCA to conduct an unsupervised exploration of the major sources of variation. When performed on the pre-processed dataset of 803 selected genes, PCA revealed an interesting pattern: the first principal component (PC1), accounting for 24.1% of the variance, drove a clear separation trend between the two groups (Fig 1). Specifically, never-smoker samples predominantly clustered towards the negative side of the PC1 axis, while current smokers were largely distributed towards the positive side. While PC2, which explained an additional 9.6% of the variance, did not contribute to this group separation, the strong segregation along PC1 demonstrates that a systematic molecular signature is the primary driver of variation within the selected gene set. To underscore the critical importance of our feature selection pipeline, we performed the same analysis on the initial, unprocessed data containing all 54,675 features. In contrast, this analysis showed no discernible separation between the two groups (Fig 2).

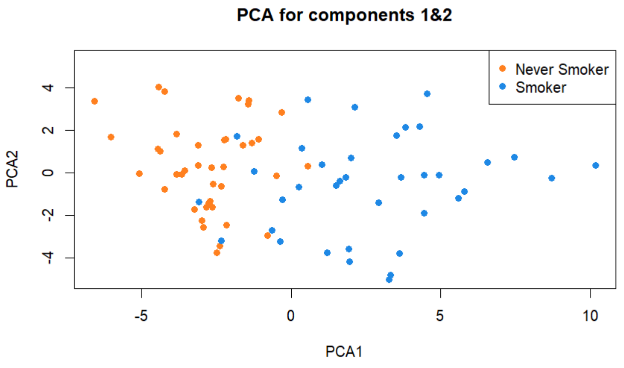


Figure 1: *PCA on the 803 gene filtered dataset reveals clear separation between smokers and never-smokers along PC1, which accounts for 24.1% of the variance, indicating a strong smoking-associated transcriptomic signature.*

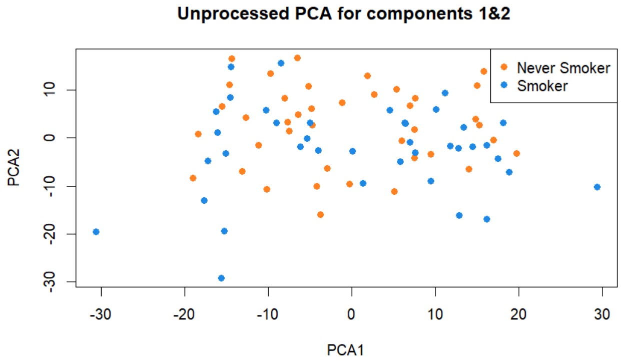


Figure 2: *PCA using all 54,675 features shows no discernible separation between smokers and never-smokers, highlighting the necessity of feature selection for uncovering biologically meaningful signals.*

The smoker and never-smoking samples appeared as a single, indistinguishable cloud, with no observable structure related to smoking status along either of the top principal components. This comparison demonstrates that our pre-processing strategy, was essential for removing non-informative genes and amplifying the signals associated with smoking that was otherwise completely obscured by noise and technical variation in the raw data.

3.2 Unsupervised Clustering Validates Class Separation

To quantitatively assess the sample groupings suggested by PCA, we applied k-means clustering to formally partition the dataset into two groups (k=2). This unsupervised approach successfully segregated the samples into two clusters containing 50 and 29 individuals, respectively. When these clusters were visualized on the principal component axes, their composition showed a remarkable alignment with the known smoking statuses. Cluster 1 (orange) was overwhelmingly composed of 'Never Smoker' samples, where Cluster 2 (blue) almost exclusively contained 'Smoker' samples, with minor overlap confined to a few samples near the boundary along PC1. This clustering achieved by k-means validates the presence of a strong, distinct molecular profile separating the two groups (Fig 3).

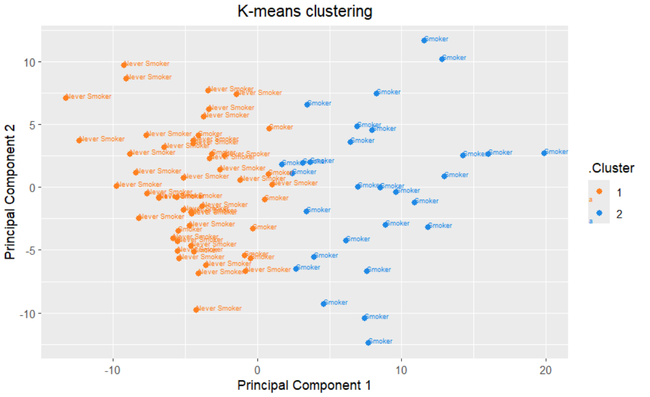


Figure 3: *K-means clustering (k=2) on the selected gene set aligns closely with smoking status, supporting the presence of a robust molecular distinction between the two classes.*

To further investigate the intrinsic structure of the data, we performed hierarchical clustering. The resulting dendrogram revealed a clear, primary division that partitioned the samples into two main branches (Fig 4).

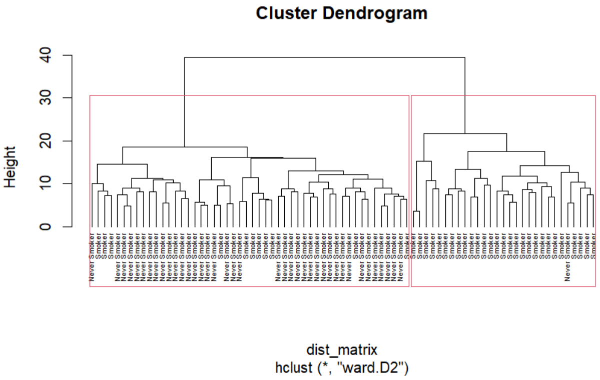


Figure 4: *Hierarchical clustering using Ward.D2 linkage further confirms two dominant sample clusters, largely corresponding to smoker and never-smoker groups, validating the transcriptomic stratification.*

This two-cluster structure was quantitatively supported by a silhouette analysis, which identified k=2 as the optimal partitioning by yielding the highest average silhouette score (Fig S2). The composition of these two clusters showed a strong but imperfect, correspondence with the known biological groups. The first major branch was predominantly composed of 'Never Smoker' samples but also contained several 'Smoker' individuals. Conversely, the second branch consisted almost exclusively of 'Smoker' samples, with only one 'Never Smoker' misclassified within this group. This result demonstrates that the smoking status is the principal factor driving sample organization, though the presence of misclassified samples indicates a degree of molecular overlap between the classes.

3.3 Supervised Classification Models Achieve High Predictive Accuracy

To develop a predictive model, we trained a Random Forest (RF) classifier using 10-fold cross-validation (Fig S3, S4) and extracted feature importance scores to rank the most discriminatory genes. A heatmap visualizing the expression patterns of the top 25 most important genes was then generated (Fig 5).

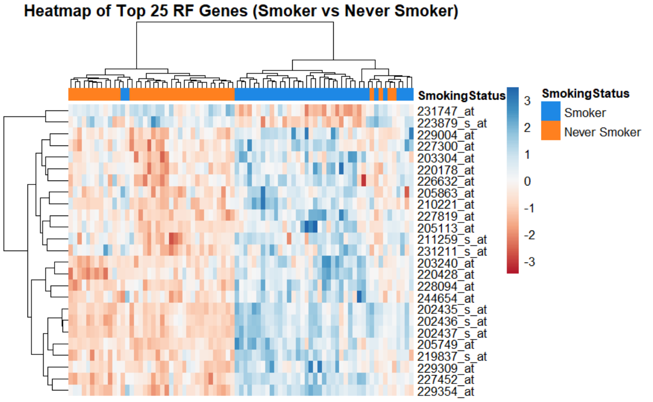


Figure 5: *Expression patterns of the top 25 RF-ranked genes distinguish smoker and never-smoker samples, with two co-regulated gene blocks reflecting opposing expression trends across groups.*

The intrinsic dendrogram based on this 25-gene signature effectively clusters the samples into their respective 'Smoker' and 'Never Smoker' classes, demonstrating the high predictive power of this concise feature set. The analysis also revealed two distinct blocks of co-regulated genes. The first block was consistently down-regulated in 'Never Smokers' and up-regulated in 'Smokers'. The second exhibited the opposite pattern. This strong bidirectional expression pattern underscores that an interesting set of genes, with systematically altered expression patterns, drives the accurate classification of smoking status.

We next employed Linear Discriminant Analysis (LDA) to build a classifier and formally assess the separation between the classes. The resulting linear discriminant successfully separated the training samples, projecting the 'Never Smoker' and 'Smoker' groups into distinct, non-overlapping distributions along a single axis (Fig 6).

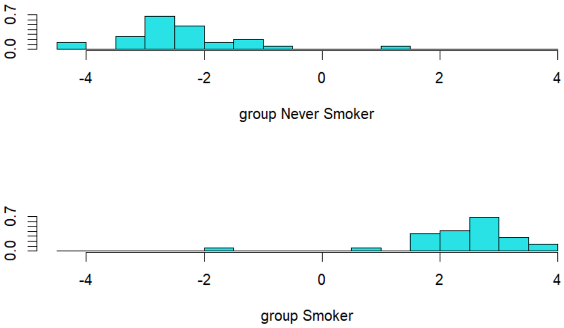


Figure 6: *LDA separates smoker and never-smoker samples along a single discriminant axis, revealing distinct class distributions in the training set.*

To evaluate the model's predictive power on unseen data, we assessed its performance on the hold-out test set. The model demonstrated excellent classification capability, achieving an Area Under the Curve (AUC) of 0.97 in the ROC analysis (Fig 7).

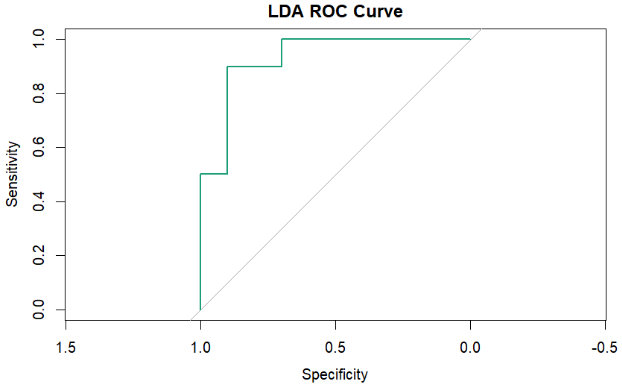


Figure 7: *ROC curve for LDA classifier evaluated on the test set, achieving an AUC of 0.97 and confirming excellent predictive accuracy.*

This high performance was confirmed by the confusion matrix, which showed that the model correctly classified 19 out of 20 test set samples, with an accuracy of 95%. The clear separation of the test set samples along the discriminant axis is further illustrated in Fig S5. These results confirm that a simple linear classifier is sufficient to accurately distinguish between the two groups.

To simultaneously perform classification and feature selection, we implemented a LASSO-regularized logistic regression model. The optimal regularization parameter, lambda (λ), was determined via 10-fold cross-validation on the training data, selecting the value that minimized the mean binomial deviance (Fig S6). When evaluated on the independent test set, the resulting model demonstrated excellent predictive power, achieving an AUC of 0.97 and correctly classifying 19 of 20 samples (Fig 8).

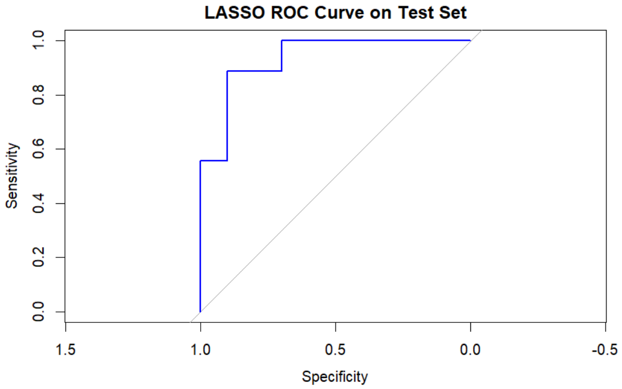


Figure 8: *ROC analysis of the LASSO logistic regression model on the test set shows high predictive performance (AUC = 0.97), highlighting its effectiveness for sparse genomic data.*

A key outcome of this approach was the generation of a model, from which we extracted a ranked list of the most influential gene predictors based on their non-zero coefficient magnitudes.

To formally compare the robustness of our supervised methods, we benchmarked the performance of LASSO, Random Forest, and LDA using an identical 10-fold cross-validation framework. The results showed that while all three models performed exceptionally well, LASSO achieved the highest mean accuracy (approximately 0.95), slightly outperforming Random Forest (0.93) and LDA (0.92) as seen in Fig 9.

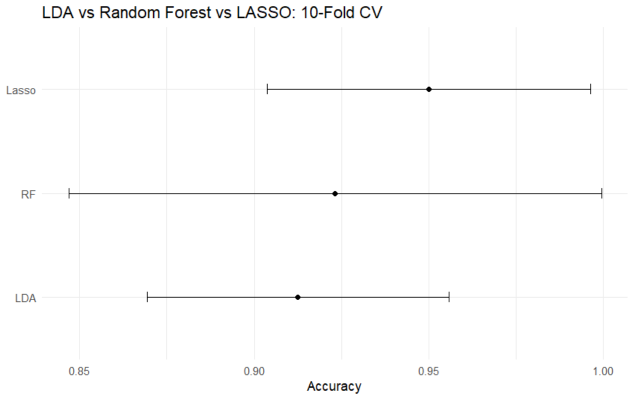


Figure 9: *Cross-validation results comparing LASSO, Random Forest, and LDA models. LASSO achieved the highest classification accuracy, followed closely by RF and LDA.*

This indicates that all three models effectively capture the underlying biological signal, with the penalized regression approach showing a marginal advantage in predictive accuracy in this context.

3.4 Network-Based Sample Profiling Confirms Group Separation

Finally, we employed the SCUDO algorithm to construct a sample similarity network. Following hyperparameter tuning via cross-validation, a final model was trained on the entire dataset using the optimal *nTop* and *nBottom* parameters. The resulting network displayed a distinct topology where samples from the 'Smoker' and 'Never Smoker' classes occupied separate, yet interconnected, regions (Fig 10).

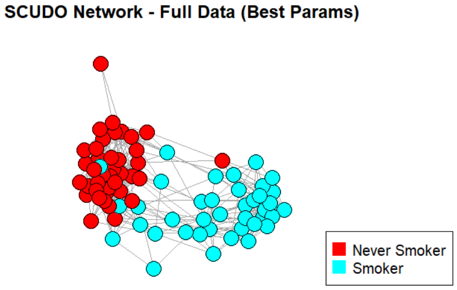


Figure 10: *Sample similarity network from SCUDO analysis shows topological segregation of smoker and never-smoker samples, indicating consistent expression-based grouping.*

To quantitatively partition this structure, we applied the Spinglass community detection algorithm to the network's largest connected component. This method successfully identified two primary communities that demonstrated a high degree of concordance with the ground-truth biological labels (Fig 11).

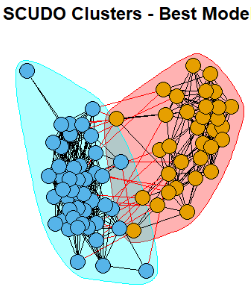
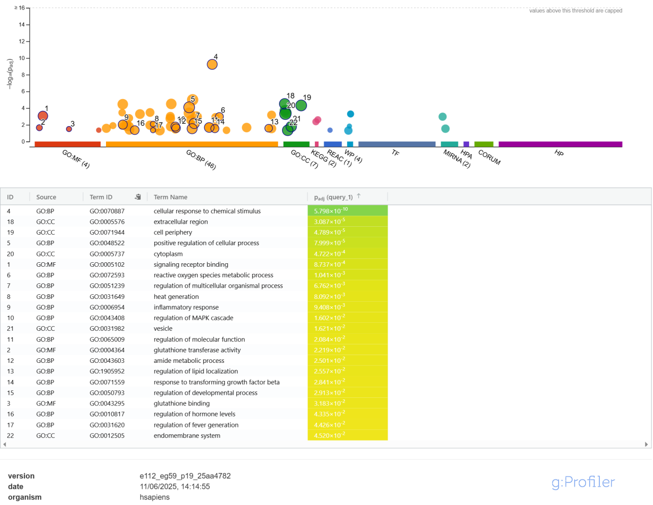


Figure 11: *Community detection on the largest network component identifies two clear communities with strong concordance to smoking status, validating molecular substructure.*

This network-based approach further validates that the smoking-associated gene expression signature is the dominant organizing principle in the dataset, effectively separating the two classes based on their molecular profiles.

3.5 Functional Enrichment Links Gene Signatures to Smoking-Related Pathways

To elucidate the biological functions of the predictive gene signature, we focused on the top 200 genes identified by the Random Forest (RF) model, which we selected due to its ability to provide interpretable feature importance scores (e.g., Mean Decrease in Accuracy or Gini). These quality metrics allowed us to rank genes based on their predictive contribution to the classification task. Functional enrichment analysis was performed on this gene set using both g:Profiler and DAVID, two complementary annotation tools. Since both approaches yielded consistent results, we present here the output from g:Profiler for clarity and visualization (Fig 12). The first lines of the DAVID results can be seen in Fig S7. The most substantial enrichments were observed within the Gene Ontology (GO) domain, particularly in the 'Biological Process' (GO:BP) category, which contained the most numerous and most statistically significant terms. Significant enrichments were also found for 'Cellular Component' (GO:CC) terms. Beyond the GO database, the analysis identified significant enrichment in established canonical pathway databases, including KEGG, Reactome, and WikiPathways.



ID	Source	Term ID	Term Name	p_{adj} (query_1) ↑
4	GO:BP	GO:0070887	cellular response to chemical stimulus	5.798×10^{-10}
18	GO:CC	GO:0005576	extracellular region	3.087×10^{-5}
19	GO:CC	GO:0071944	cell periphery	4.789×10^{-5}
5	GO:BP	GO:0048522	positive regulation of cellular process	7.999×10^{-5}
20	GO:CC	GO:0005737	cytoplasm	4.722×10^{-4}
1	GO:MF	GO:0005102	signaling receptor binding	8.737×10^{-4}
6	GO:BP	GO:0072593	reactive oxygen species metabolic process	1.041×10^{-3}
7	GO:BP	GO:0051239	regulation of multicellular organismal process	6.762×10^{-3}
8	GO:BP	GO:0031649	heat generation	8.092×10^{-3}
9	GO:BP	GO:0006954	inflammatory response	9.408×10^{-3}
10	GO:BP	GO:0043408	regulation of MAPK cascade	1.602×10^{-2}
21	GO:CC	GO:0031982	vesicle	1.621×10^{-2}
11	GO:BP	GO:0065009	regulation of molecular function	2.084×10^{-2}
2	GO:MF	GO:0004364	glutathione transferase activity	2.219×10^{-2}
12	GO:BP	GO:0043603	amide metabolic process	2.501×10^{-2}
13	GO:BP	GO:1905952	regulation of lipid localization	2.557×10^{-2}
14	GO:BP	GO:0071559	response to transforming growth factor beta	2.841×10^{-2}
15	GO:BP	GO:0050793	regulation of developmental process	2.913×10^{-2}
3	GO:MF	GO:0043295	glutathione binding	3.183×10^{-2}
16	GO:BP	GO:0010817	regulation of hormone levels	4.335×10^{-2}
17	GO:BP	GO:0031620	regulation of fever generation	4.426×10^{-2}
22	GO:CC	GO:0012505	endomembrane system	4.520×10^{-2}

version e112_eg59_p19_25aa4782
date 11/06/2025, 14:14:55
organism hsapiens

Figure 12: *Over-representation analysis of top 200 RF-selected genes highlights significant GO and pathway enrichment, implicating biological processes associated with smoking effects.*

Collectively, these results demonstrate that the genes most critical for classifying smoking status are strongly implicated in specific, coordinated biological processes and signaling pathways, providing a functional context for the predictive power of the RF model.

To elucidate the biological functions of the top 100-gene signature derived from the LASSO model, we performed active-subnetwork-based pathway analysis using both Pathfinder and STRING. The Pathfinder method integrates statistical evidence with biological network topology and identified several significantly enriched pathways relevant to the physiological impact of smoking. These include the JAK-STAT signaling pathway, cytokine-cytokine receptor interaction, and chemical carcinogenesis related to reactive oxygen species, as shown in (Fig 13).

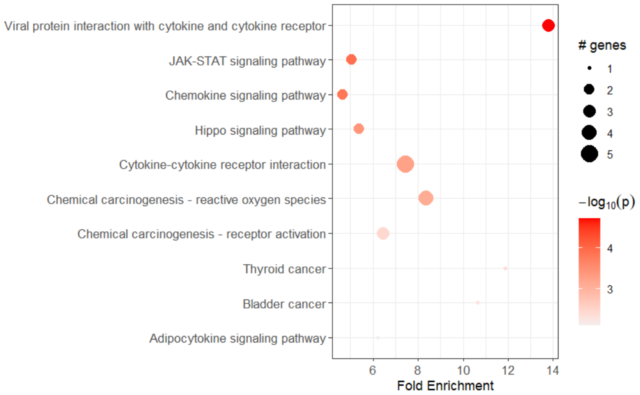


Figure 13: *KEGG enrichment of top 100 LASSO-selected genes using Pathfinder. Highlighted pathways are associated with immune response, signaling, and chemical carcinogenesis.*

To better interpret these results, enriched KEGG pathways were organized into functional clusters based on gene-set overlap (Fig 14).

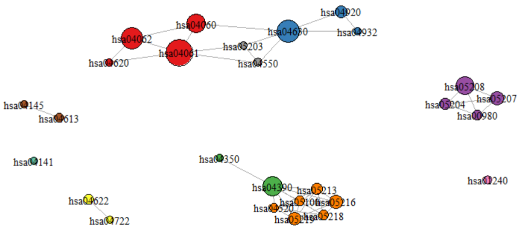


Figure 14: *Clustered KEGG pathways based on gene overlap. Red: immune and inflammatory signaling; orange: cancer-related pathways; green: apoptosis/stress; blue: signal transduction; purple: metabolism.*

Each node represents a pathway in KEGG, with size indicating statistical significance and edges denoting shared genes. Color-coded modules highlight distinct biological themes: red for immune and inflammatory signaling, orange for cancer-related pathways, green for apoptosis and cellular stress responses, blue for general signal transduction, and purple for metabolic processes. A complementary term-gene interaction network (Fig 15) mapped predictive genes to their enriched pathways, offering further insight into their functional roles. Notably, genes such as CYP1A1, CYP1B1, and GSTM3—downregulated in smokers—were central to carcinogenesis-related modules, while IL6ST and other immune mediators appeared in immune signaling clusters.

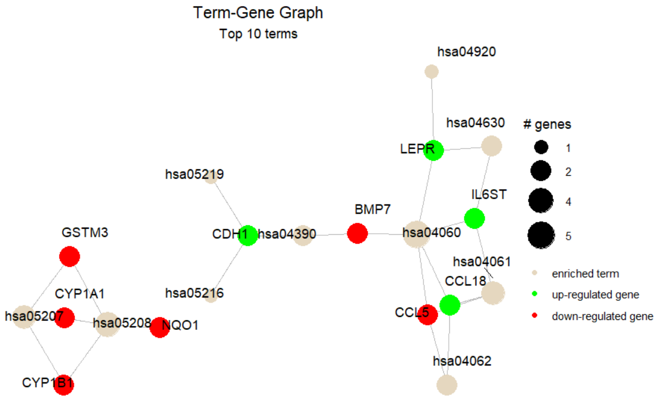


Figure 15: *Network linking predictive genes to enriched pathways. Genes such as CYP1A1, CYP1B1, GSTM3, and IL6ST act as central nodes within key biological modules.*

Additional validation using the Reactome database confirmed the involvement of xenobiotic metabolism pathways, including glucuronidation and cytochrome P450 activity. These results provide a mechanistic link between the selected gene features and smoking-associated biological disruptions (Fig S8, S9 and S10).

Protein-protein interaction analysis using STRING (Fig 16) further contextualized the gene signature within biological networks. The network revealed tightly interconnected modules of genes involved in immune regulation, epithelial differentiation, and detoxification processes, underscoring the functional coherence of the LASSO-selected gene set.

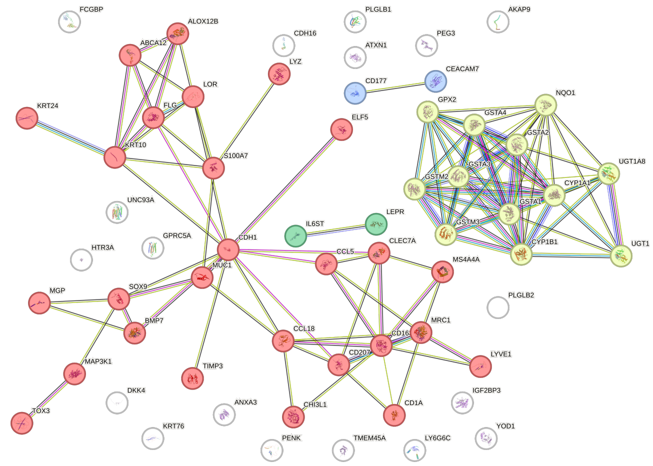


Figure 16: *STRING network analysis of LASSO-selected genes showing functional protein associations, grouped into modules related to inflammation, epithelial function, and detoxification*

4 Conclusion

In this study, we conducted a comprehensive transcriptomic and machine learning-based analysis to investigate the effects of tobacco smoking on the buccal mucosal gene expression profile. Utilizing the GSE17913 dataset, we demonstrated that smoking induces a strong and detectable molecular signature within the oral epithelium, a tissue directly exposed to tobacco smoke. Dimensionality reduction and unsupervised clustering techniques (PCA, k-means, and hierarchical clustering) revealed a clear separation between smokers and never-smokers, validating the effectiveness of our feature selection pipeline [6, 7, 9]. Supervised classifiers, including Random Forest, LDA, and LASSO, achieved high predictive performance with AUC values of up to 0.97, emphasizing the robustness of transcriptomic features in classifying smoking status [11, 13, 14]. Among these, LASSO emerged as the best-performing method, offering not only high accuracy but also a compact set of highly informative genes. Network-based approaches using SCUDO further confirmed the molecular stratification of the classes, highlighting topological coherence among expression profiles of similar biological classes [15, 16]. Pathway enrichment and network-based analyses revealed that the selected genes are involved in key biological processes disrupted by tobacco exposure [3, 4, 17]. These include immune regulation, xenobiotic metabolism, oxidative stress response, and epithelial signaling [2, 3, 12]. Downregulated genes such as CYP1A1, CYP1B1, and GSTM3 were central in carcinogenesis-related pathways, while upregulated markers like IL6ST were enriched in immune and inflammatory modules, particularly the JAK-STAT signaling cascade. STRING-based protein interaction analysis supported these results, revealing tightly connected subnetworks that reinforce the functional coherence of the selected signature. Overall, this integrative framework combining statistical modeling, machine learning, and multi-omic interpretation provides a mechanistic understanding of how smoking perturbs the molecular landscape of the oral epithelium.

Supplementary Figures

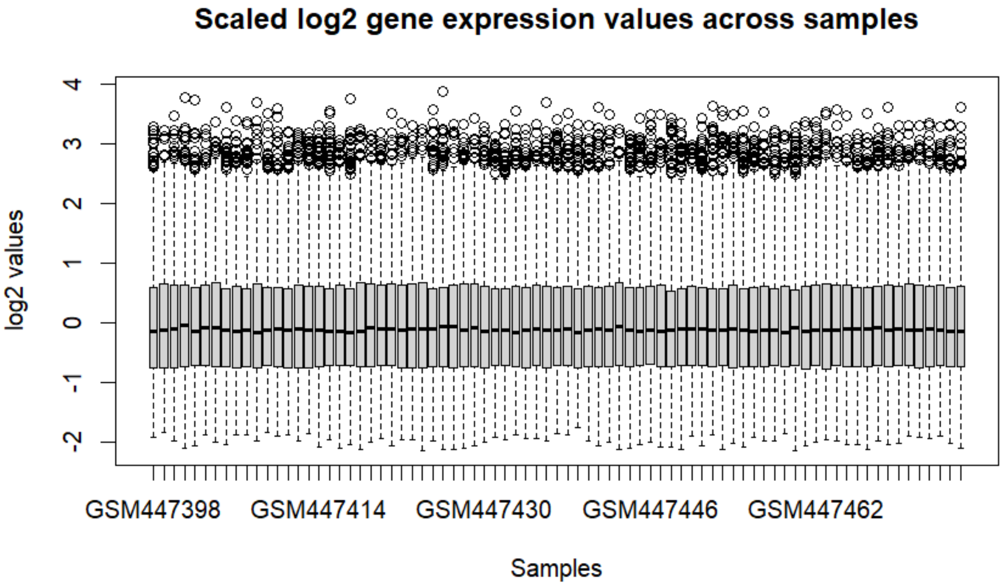


Fig. S1. *Distribution of normalized expression across all 79 samples post-feature selection, ensuring comparability and data quality for downstream analyses.*

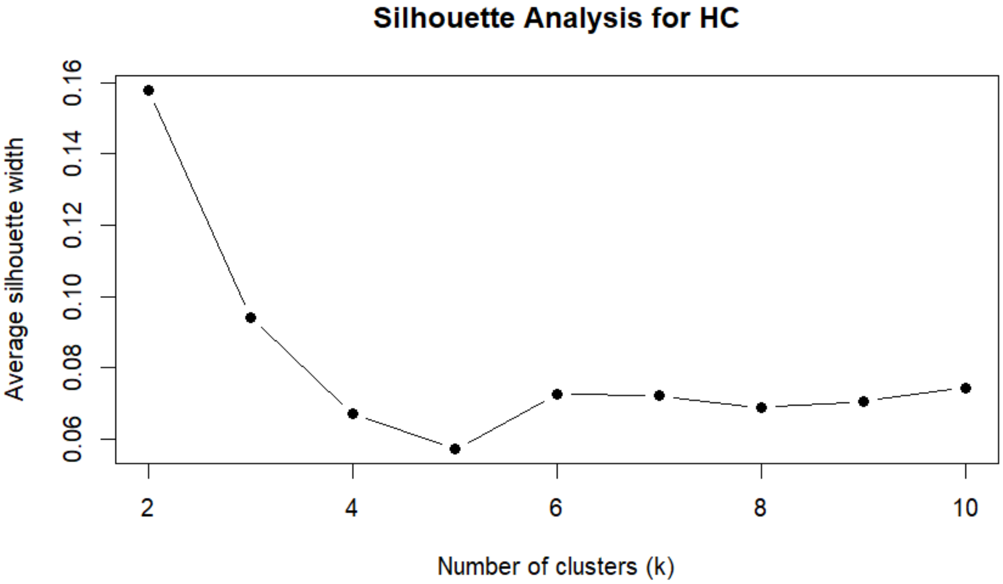


Fig. S2. *Silhouette width analysis identifies k=2 as the optimal number of clusters for both k-means and hierarchical clustering methods.*

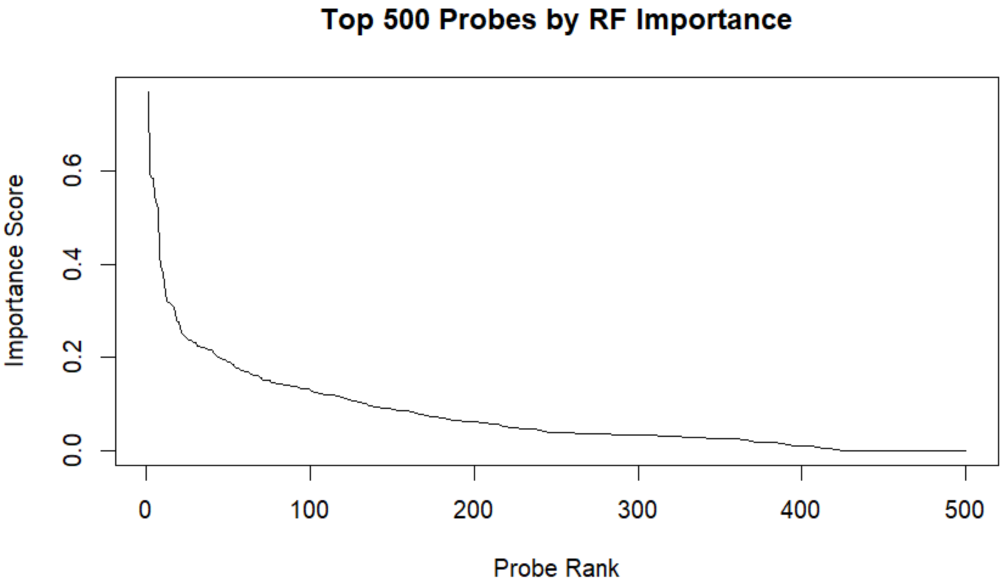


Fig. S3.*Silhouette width analysis identifies k=2 as the optimal number of clusters for both k-means and hierarchical clustering methods.*

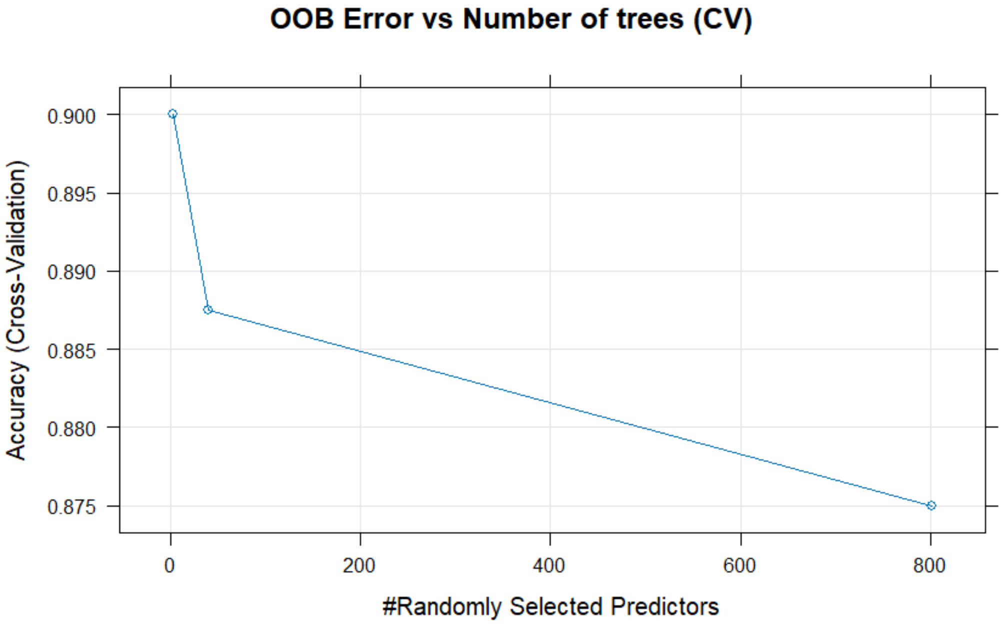


Fig. S4. *Ranked importance scores of genes derived from RF model, illustrating key contributors to smoking status classification..*

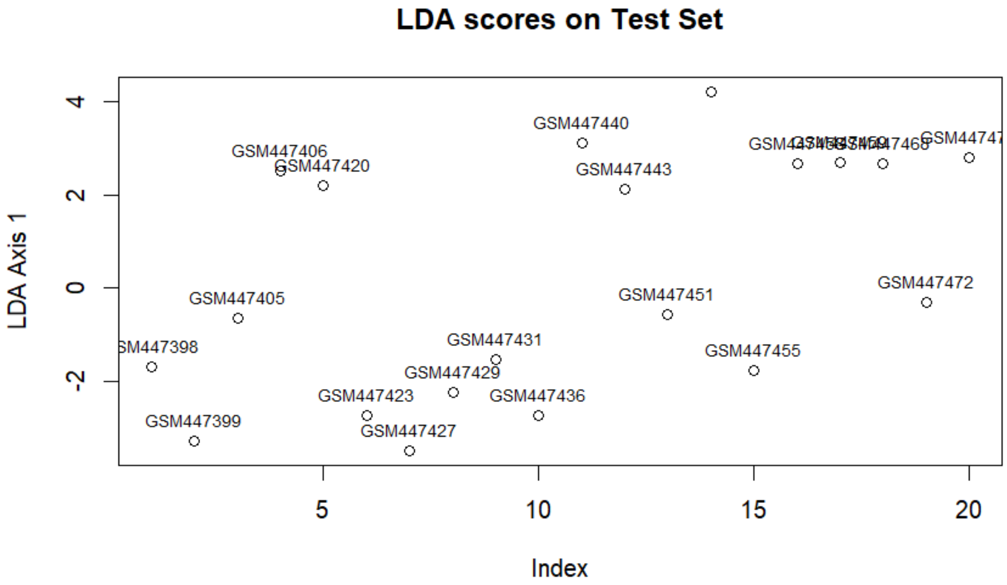


Fig. S5. *Test samples projected onto LDA discriminant axis, confirming strong separation between smoker and never-smoker groups.*

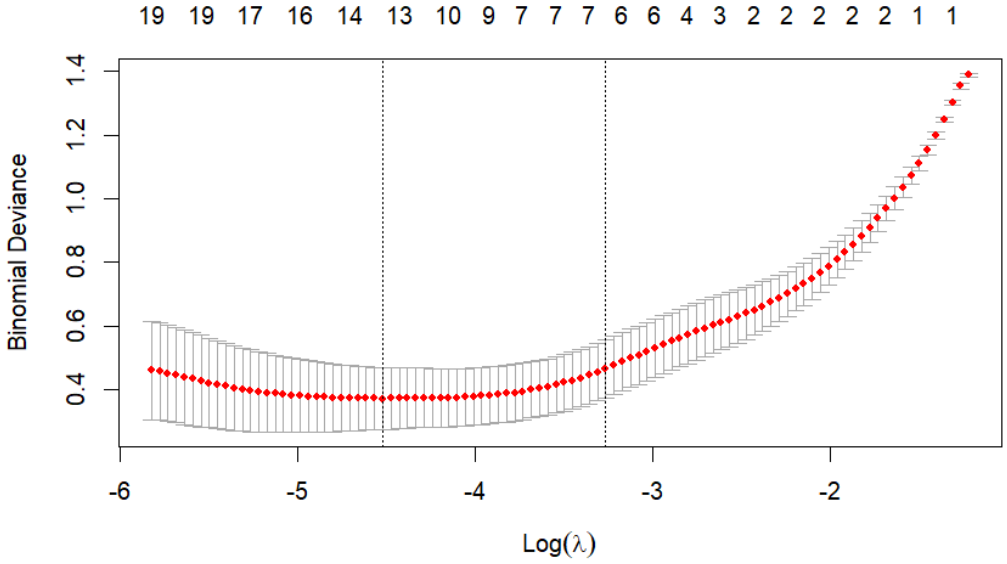
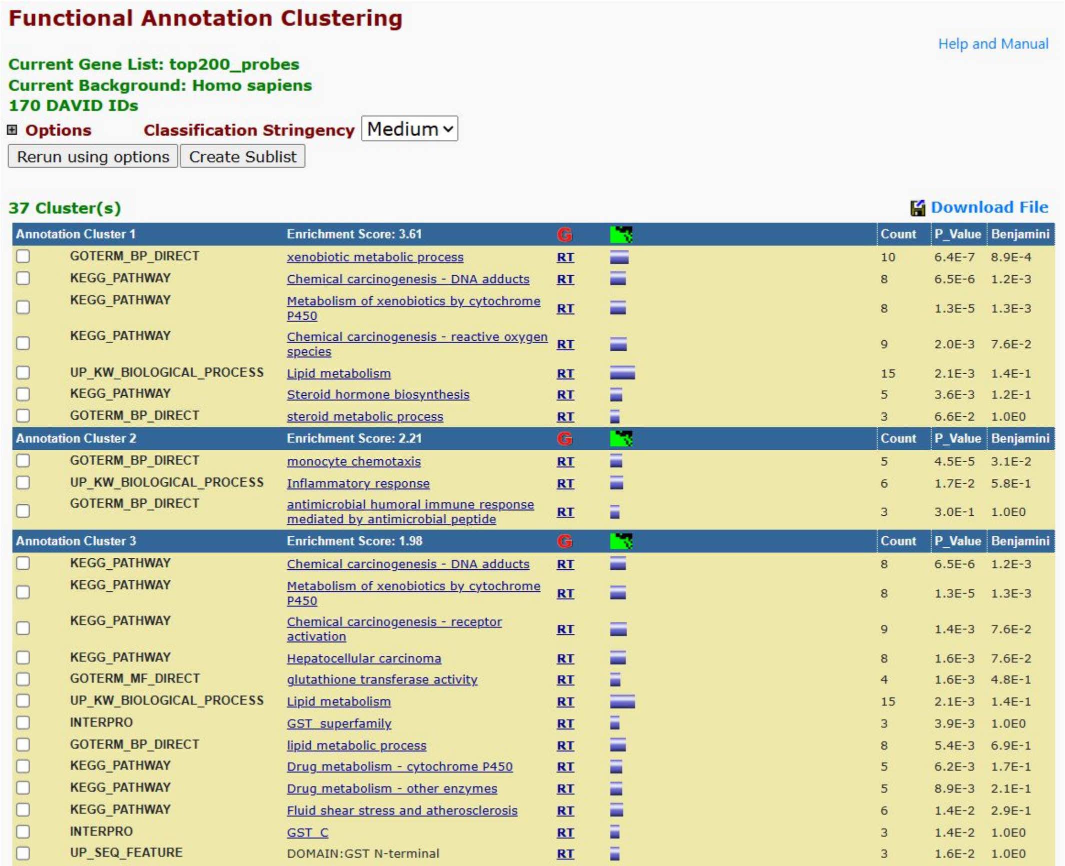


Fig. S6. *Plot of cross-validated deviance across different values, identifying the optimal regularization parameter for the LASSO model.*



Functional Annotation Clustering

Current Gene List: top200_probes
Current Background: Homo sapiens
170 DAVID IDs

⊞ **Options** **Classification Stringency** Medium

37 Cluster(s)

Annotation Cluster 1		Enrichment Score: 3.61	G		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	xenobiotic metabolic process	RT		10	6.4E-7	8.9E-4
☐	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT		8	6.5E-6	1.2E-3
☐	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		8	1.3E-5	1.3E-3
☐	KEGG_PATHWAY	Chemical carcinogenesis - reactive oxygen species	RT		9	2.0E-3	7.6E-2
☐	UP_KW_BIOLOGICAL_PROCESS	Lipid metabolism	RT		15	2.1E-3	1.4E-1
☐	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		5	3.6E-3	1.2E-1
☐	GOTERM_BP_DIRECT	steroid metabolic process	RT		3	6.6E-2	1.0E0
Annotation Cluster 2		Enrichment Score: 2.21	G		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	monocyte chemotaxis	RT		5	4.5E-5	3.1E-2
☐	UP_KW_BIOLOGICAL_PROCESS	Inflammatory response	RT		6	1.7E-2	5.8E-1
☐	GOTERM_BP_DIRECT	antimicrobial humoral immune response mediated by antimicrobial peptide	RT		3	3.0E-1	1.0E0
Annotation Cluster 3		Enrichment Score: 1.98	G		Count	P_Value	Benjamini
☐	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT		8	6.5E-6	1.2E-3
☐	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		8	1.3E-5	1.3E-3
☐	KEGG_PATHWAY	Chemical carcinogenesis - receptor activation	RT		9	1.4E-3	7.6E-2
☐	KEGG_PATHWAY	Hepatocellular carcinoma	RT		8	1.6E-3	7.6E-2
☐	GOTERM_MF_DIRECT	glutathione transferase activity	RT		4	1.6E-3	4.8E-1
☐	UP_KW_BIOLOGICAL_PROCESS	Lipid metabolism	RT		15	2.1E-3	1.4E-1
☐	INTERPRO	GST_superfamily	RT		3	3.9E-3	1.0E0
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		8	5.4E-3	6.9E-1
☐	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT		5	6.2E-3	1.7E-1
☐	KEGG_PATHWAY	Drug metabolism - other enzymes	RT		5	8.9E-3	2.1E-1
☐	KEGG_PATHWAY	Fluid shear stress and atherosclerosis	RT		6	1.4E-2	2.9E-1
☐	INTERPRO	GST_C	RT		3	1.4E-2	1.0E0
☐	UP_SEQ_FEATURE	DOMAIN:GST N-terminal	RT		3	1.6E-2	1.0E0

__Fig. S7.__ Visualization of DAVID top results from the functional annotation

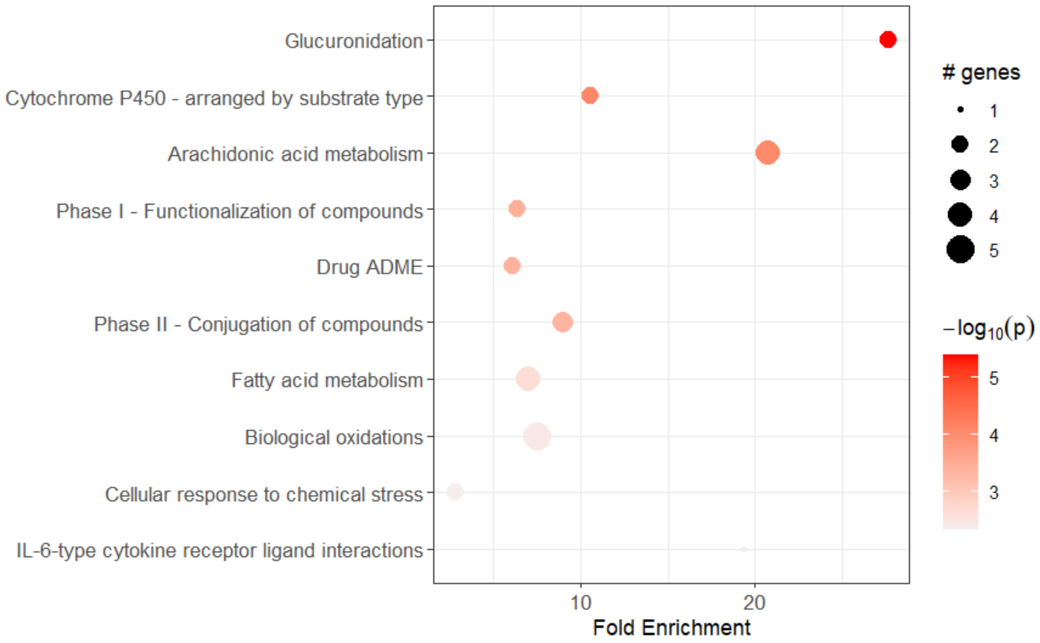


__Fig. S8.__ Reactome analysis of LASSO-selected genes shows significant enrichment in xenobiotic metabolism and detoxification pathways.

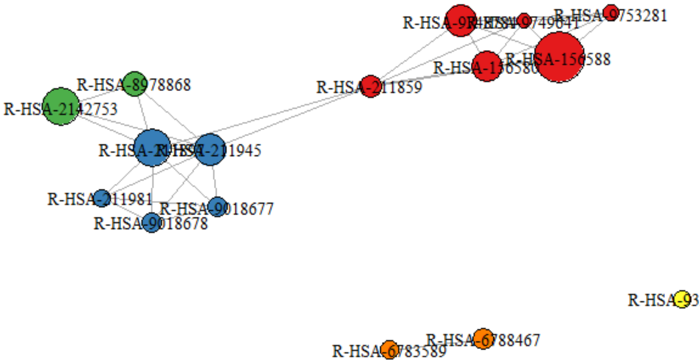


Fig. S9. *Clustered Reactome pathways grouped by shared genes, representing key functional domains disrupted by smoking.*

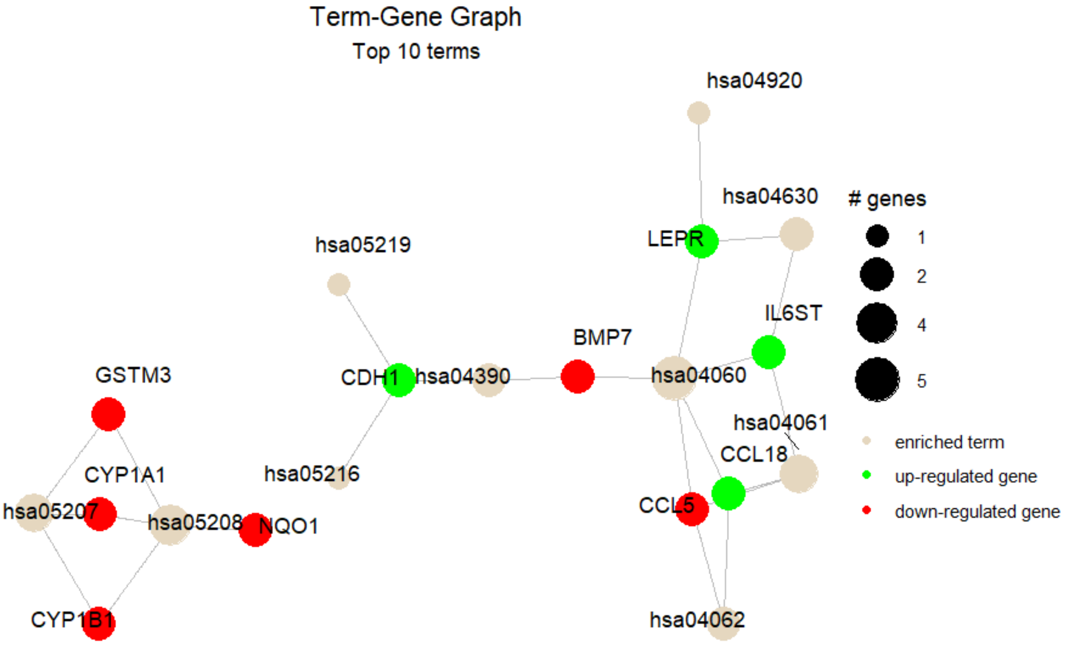


Fig. S10. *Visualization of Reactome pathway-gene relationships; central genes like CYP1A1 and GSTM3 underscore roles in chemical carcinogenesis.*

References

[1] GBD 2019 Tobacco Collaborators. “Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019”. In: *The Lancet* 397.10292 (2021), pp. 2337–2360. DOI: 10.1016/S0140-6736(21)01169-7.

[2] Stephen S. Hecht. “Lung carcinogenesis by tobacco smoke”. In: *International Journal of Cancer* 131.12 (2012), pp. 2724–2732. DOI: 10.1002/ijc.27816.

[3] Gerd P. Pfeifer et al. “Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers”. In: *Oncogene* 21.48 (2002), pp. 7435–7451. DOI: 10.1038/sj.onc.1205803.

[4] César Rivera. “Essentials of oral cancer”. In: *International Journal of Clinical and Experimental Pathology* 8.9 (2015), pp. 11884–11894.

[5] B. M. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data”. In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[6] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. “Independent filtering increases detection power for high-throughput experiments”. In: *PNAS* 107.21 (2010), pp. 9546–9551.

[7] Gordon K. Smyth. “Linear models and empirical Bayes methods for assessing differential expression in microarray experiments”. In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), Article3.

[8] Michael B. Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *PNAS* 95.25 (1998), pp. 14863–14868.

[9] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 2009.

[10] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[11] Ramón Díaz-Uriarte and Sara Alvarez De Andres. “Gene selection and classification of microarray data using random forest”. In: *BMC Bioinformatics* 7 (2006), p. 3.

[12] Robert Tibshirani. “Regression shrinkage and selection via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22.

[14] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874.

[15] Chiara Pastrello et al. “SCUDO: a tool for signature-based clustering of expression profiles”. In: *Nucleic Acids Research* 39.Web Server issue (2011), W111–W115.

[16] Jörg Reichardt and Stefan Bornholdt. “Statistical mechanics of community detection”. In: *Physical Review E* 74.1 (2006), p. 016110.

[17] Jüri Reimand et al. “Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap”. In: *Nature Protocols* 14.2 (2019), pp. 482–517.