JOURNAL OF
EMERGING INVESTIGATORS

Article

Predicting smoking status based on RNA sequencing data

Kevin Yang[1], Julian Stanley[2]
[1] Cranbrook Kingswood Upper School, Bloomfield Hills, Michigan
[2] Massachusetts Institute of Technology, Cambridge, Massachusetts

SUMMARY

Given an association between nicotine addiction and gene expression, we hypothesized that expression of genes commonly associated with smoking status would have variable expression between smokers and non-smokers. To test whether gene expression varies between smokers and non-smokers, we analyzed two publicly-available datasets that profiled RNA gene expression from brain (nucleus accumbens) and lung tissue taken from patients identified as smokers or non-smokers. We discovered statistically significant differences in expression of dozens of genes between smokers and non-smokers. To test whether gene expression can be used to predict whether a patient is a smoker or non-smoker, we used gene expression as the training data for a logistic regression or random forest classification model. The random forest classifier trained on lung tissue data showed the most robust results, with area under curve (AUC) values consistently between 0.82 and 0.93. Both models trained on nucleus accumbens data had poorer performance, with AUC values consistently between 0.65 and 0.7 when using random forest. These results suggest gene expression can be used to predict smoking status using traditional machine learning models. Additionally, based on our random forest model, we proposed *KCNJ3* and *TXLNGY* as two candidate markers of smoking status. These findings, coupled with other genes identified in this study, present promising avenues for advancing applications related to the genetic foundation of smoking-related characteristics.

INTRODUCTION

Nicotine addiction is an issue gaining increasing attention due to the introduction of electronic cigarettes and their influence on adolescents (1). While electronic cigarettes have far less than the 7000+ harmful chemicals present in traditional cigarette smoke, advertisements for these devices primarily center around a younger audience, introducing nicotine usage beyond just older generations (2,3). Users of electronic cigarettes also often still smoke traditional cigarettes: a 2014 study found that the percentage of e-cigarette users who continued to smoke traditional cigarettes was 93% in the United States, 83% in France, and 60% in the United Kingdom (4).

Nicotine primarily targets the nicotinic acetylcholine receptors, or nAChRs, in the brain thereby stimulating the release of neurotransmitters, primarily dopamine. The increase in these neurotransmitters then prompts the formation of oxygen radicals and a reduction of the antioxidant capacity of the lungs. These factors cause point mutations of the DNA which alter lung growth and function (5, 6). Especially when a single cell undergoes multiple mutations, they can lead to nonfunctional tumor-suppressor genes and cancer (7). Moreover, the general use of nicotine "hijacks" the brain's reward system and desensitizes the nucleus accumbens (the brain tissue involved in reward and pleasure), decreasing one's control over dangerous behavior (8).

Genetic factors also play a role in the development of nicotine dependency (9). The expression level of certain genes can contribute to becoming addicted and even assist medical professionals in determining the most effective treatment method (10). Prior studies surrounding the use of nicotine products focused on identifying genes of interest with respect to nicotine usage. Such studies tended to focus on identifying genes related to nicotine addiction rather than deriving a method or function to identify nicotine usage or potential risk for nicotine addiction in patients. For example, *SERPINA1, CHRNB3, CHRNA6, CHRNA5, CHRNA3, CHRNB4, DNMT3B, NOL4L*, and *CHRNA4* were all found to be associated with smokers via a genome-wide association study (GWAS) in healthy patients (11). These genes are largely responsible for inhibition of relevant enzymes related to nicotinic acetylcholine receptor subunits. Genetic variations of *SERPINA1* have been found to be associated with the pathogenesis of chronic obstructive pulmonary disease (COPD), to which smoking is a known contributor (12). Additionally, mutations in *CHRNB3*, *CHRNA6*, *CHRNA5*, *CHRNA3*, *CHRNB4*, *DNMT3B*, *NOL4L*, and *CHRNA4* have all been found to influence COPD (13). However, while GWAS studies identify correlations, they have been criticized because such correlations often lack direct biological relevance to disease (14). It remains unclear whether these genetic associations with smoking also extend to the level of mRNA expression, which has a more direct effect on protein levels.

We hypothesized that expression of genes commonly associated with smoking status would have variable expression between smokers and non-smokers. To address this question, we analyzed RNA sequencing (RNA-seq) data from studies of gene expression in the nucleus accumbens and the lungs of smokers and non-smokers. In this study, the random forest classifier trained on lung tissue data showed the most robust results, with area under curve (AUC) values consistently between 0.82 and 0.93. Based on this model, we determined *KCNJ3* and *TXLNGY* as two candidate markers of smoking status.

https://doi.org/10.59720/23-099

RESULTS

To test whether gene expression varies between smokers and non-smokers, we analyzed two publicly-available datasets that profiled RNA gene expression from nucleus accumbens (brain) and lung tissue taken from patients identified as smokers or non-smokers (23,24). First, we tested whether genes known to be associated with smoking status had different levels of expression across smokers and non-smokers. We found that expression of one of the primary genetic indicators for nicotine dependence, *SERPINA1*, showed a clear difference between the distribution of case and control points in both nucleus accumbens and lung samples. We found that non-smokers had a higher average expression of *SERPINA1* than smokers (p = 0.01, Student's t-test), suggesting that the regulation of *SERPINA1* may be an indicator of a patient's smoking status. However, other smoking-associated genes, such as *NOL4L*, did not show significant differential expression in either dataset (p = 0.3, Student's t-test) (**Figure 1**).

We hypothesized that we could predict smoking status based on the genes expressed differentially between smokers and non-smokers. To find genes with large expression differences, we calculated the fold change gene expression for all genes in each dataset (**Tables 1 and 2**). A fold change value of 1 indicates no difference between the two groups, while a value greater than 1 indicates an increase in the variable in the experimental group. We found significantly different genes in the lung tissue dataset compared to the genes found from the nucleus accumbens dataset, in particular, the presence of *SERPINA1*.

A GLMNET logistic model was then generated, which resulted in a consistent AUC output of 0.61-0.70 (**Figure 2**). Glmnet is a package that fits generalized linear models

Genes	Fold Change
SNORA38	3.531952634
SNORA60	1.788363573
RNA5SP145	1.744378786
PIK3R2	1.534884362
SNORD11-20	1.522911307

Table 1: Top five genes ranked descending by fold change values in nucleus accumbens dataset. Log10-transformed transcripts per million (TPM) gene expression across each of the three groups. N=50 case (smoker) samples, N=171 control (non-smoker), N=2 indeterminant smoking status. Fold change>1 indicates differential expression between smoker and non-smoker patients.

Genes	Fold Change
XIST	2.3771
TSIX	2.2682
Cyorf15B	1.9481
RPS4Y1	1.8637
Cyorf15A	1.8484

Table 2. Top 5 genes ranked descending by fold change value in lung tissue dataset. N=39 case (smoker) samples, N=34 control (non-smoker), N=4 indeterminant smoking status.

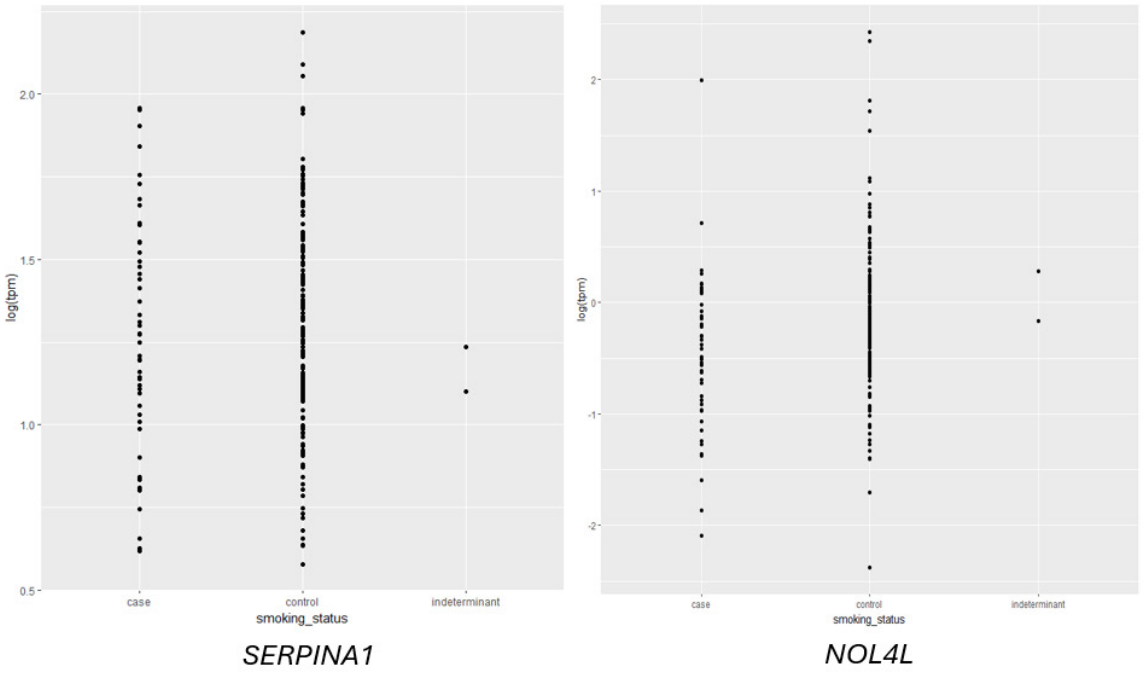
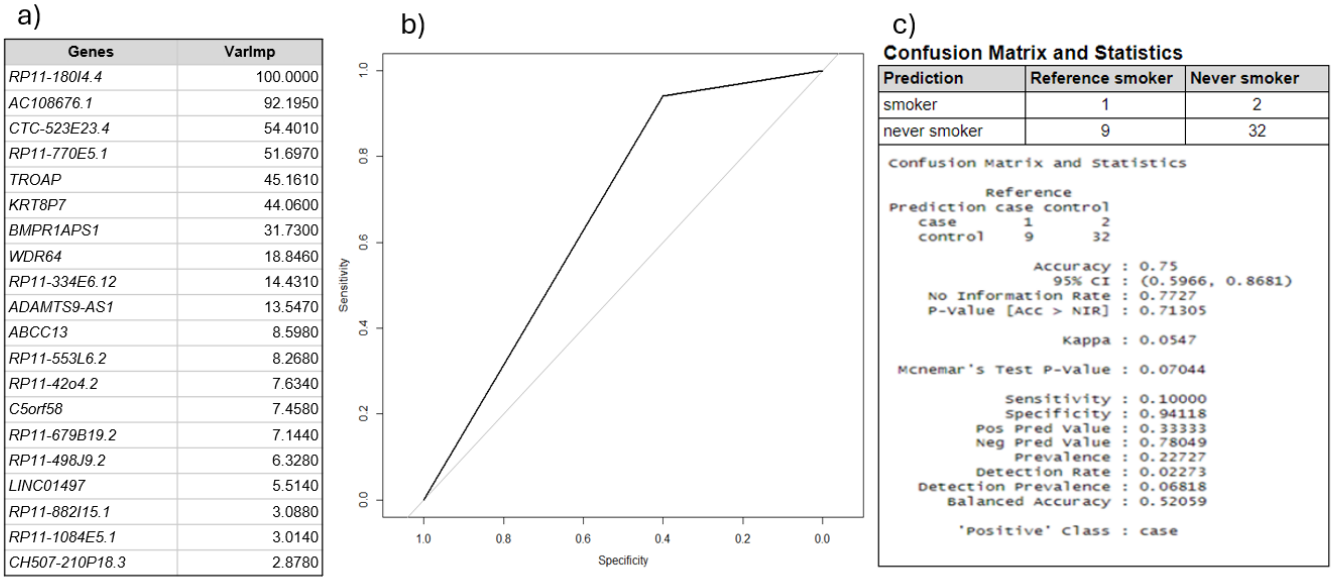


Figure 1: Expression of *SERPINA1* and *NOL4L* in smokers, non-smokers, and patients with indeterminant smoking status from the nucleus accumbens dataset. Log10-transformed transcripts per million (TPM) gene expression across each of the three groups. TPM values were calculated from raw quantified RNA-seq values. N=50 case (smoker) samples, N=171 control (non-smoker), N=2 indeterminant smoking status. p=0.01 between case and control (student's t-test) *SERPINA1*, p=0.3 between case and control (student's t-test) NOL4L.

https://doi.org/10.59720/23-099



Genes	VarImp
RP11-180I4.4	100.0000
AC108676.1	92.1950
CTC-523E23.4	54.4010
RP11-770E5.1	51.6970
TROAP	45.1610
KRT8P7	44.0600
BMPR1APS1	31.7300
WDR64	18.8460
RP11-334E6.12	14.4310
ADAMTS9-AS1	13.5470
ABCC13	8.5980
RP11-553L6.2	8.2680
RP11-42o4.2	7.6340
C5orf58	7.4580
RP11-679B19.2	7.1440
RP11-498J9.2	6.3280
LINC01497	5.5140
RP11-882I15.1	3.0880
RP11-1084E5.1	3.0140
CH507-210P18.3	2.8780

Confusion Matrix and Statistics

Prediction	Reference smoker	Never smoker
smoker	1	2
never smoker	9	32

```
Confusion Matrix and Statistics

          Reference
Prediction case control
   case      1       2
   control   9      32

               Accuracy : 0.75
                 95% CI : (0.5966, 0.8681)
    No Information Rate : 0.7727
    P-Value [Acc > NIR] : 0.71305

                  Kappa : 0.0547

 Mcnemar's Test P-Value : 0.07044

            Sensitivity : 0.10000
            Specificity : 0.94118
         Pos Pred Value : 0.33333
         Neg Pred Value : 0.78049
             Prevalence : 0.22727
         Detection Rate : 0.02273
   Detection Prevalence : 0.06818
      Balanced Accuracy : 0.52059

       'Positive' Class : case
```

Figure 2: GLMNET prediction model for nucleus accumbens dataset. (a) List of genetic markers ordered by variable importance in creating the prediction, (b) plot of AUC for the prediction, and (c) confusion matrix. N=50 case (smoker) samples, N=171 control (non-smoker), N=2 indeterminant smoking status. AUC = 0.61-0.70.

based on maximum likelihood. The AUC output evaluates how well a logistic regression model classifies positive and negative outcomes, ranging from 0.5 to 1, the larger the better. In this case, an output between 0.61 and 0.70 indicates a low ability to discriminate between smoker and nonsmoker patients in the dataset. We created a confusion matrix using the top 50 highest fold change genes (data not shown). Variable importance was then found, depicting the genes that were most significant in the prediction. However, in an attempt to provide a more accurate prediction, rather than using GLMNET logistic regression as a classifier, we used a random forest classifier as the predictor for both datasets. The random forest classifier is robust to outliers because they get averaged out by the aggregation of multiple tree outputs. Moreover, the nucleus accumbens dataset can be considered an imbalanced dataset, where the dataset is not half smokers and half non-smokers but has many more data points from the non-smoker class. Random forest classification is a robust algorithm that can handle imbalanced datasets, where one class is much rarer than the others. The result from the random forest model was a similar AUC of 0.65-0.70 (**Figure 3**).

Compared to the nucleus accumbens dataset, there was a clear increase in accuracy, and the predictions were not skewed towards either smoking status. Using RNA-seq from the lung sample dataset as input, the GLMNET model achieved AUC values consistently between 0.79 and 0.92 (**Figure 4**). The random forest classifier achieved similar levels of accuracy and AUC values, roughly 0.82 to 0.93 (**Figure 5**). These results suggest that the lung tissue dataset is a much better predictor of nicotine usage compared to the nucleus accumbens dataset.

Interestingly, the two most significant genes in the GLMNET and random forest models of the lung tissue dataset, *KCNJ3* and *TXLNGY* respectively, play very different functional roles but were both assigned very high significance.

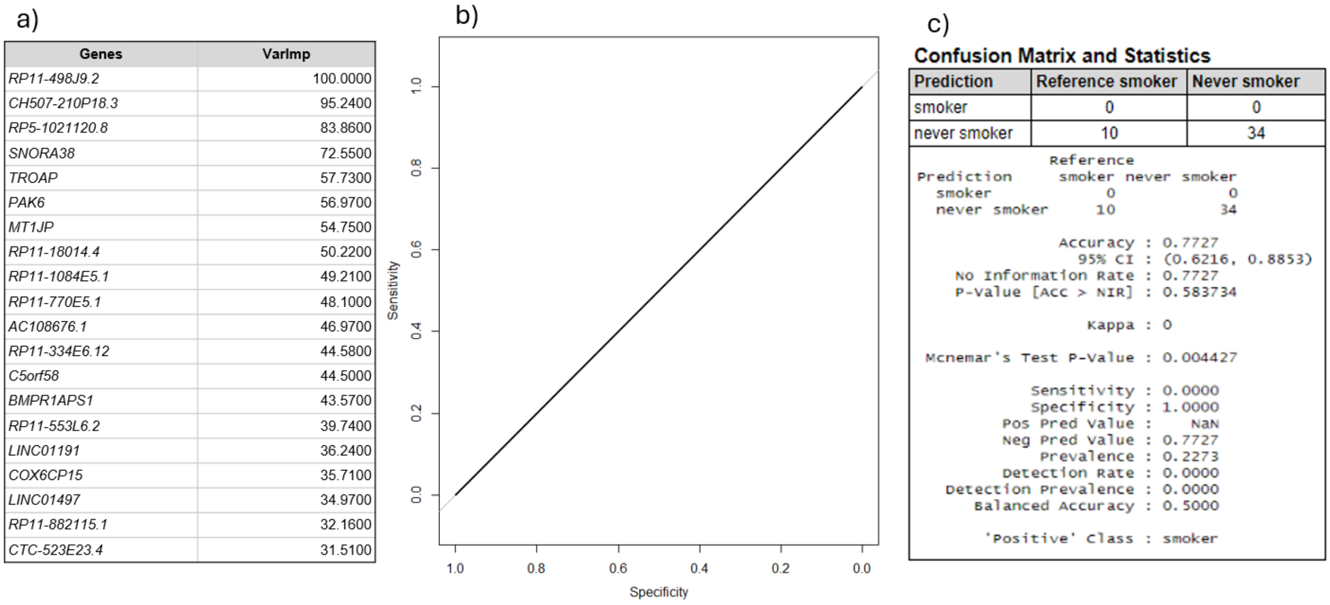
DISCUSSION

In this study, we sought to determine whether genes known to be associated with smoking status also had differential gene expression levels across smokers and non-smokers. We postulated that normalized gene expression data could be used to predict whether someone is a smoker or non-smoker. To make this prediction, we used logistic regression and random forest models, which both can be trained on large data with modest computational resources. When used on the nucleus accumbens dataset, the output resulted in an AUC between 0.61-0.70 between the two models and did not lead to more interpretable results. Meanwhile, both models of the lung tissue dataset performed relatively well, the random forest model performed marginally better. Random forest does not assume a linear relationship between predictors and the response. This is especially beneficial when dealing with high-dimensional data, such as that of this dataset, in which the relationships might not be linear. Moreover, random forest's capability to handle irrelevant or redundant features without significant drops in efficiency allows much better output for higher-dimensional scenarios where many features might not be relevant to the outcome.

Expanding our analysis, it is important to underscore the context of accurate predictions within this project. Notably, a significant proportion of accurate predictions originated from non-smoker individuals. This observation casts doubt on the feasibility of accurately determining smoking status based on gene expression data from the nucleus accumbens dataset. However, we acknowledge that other studies have established a strong link between gene expression and smoking status with a high degree of accuracy (15).

In the lung tissue dataset, we found that some genes with statistically significant expression differences between smokers and non-smokers were absent from the nucleus accumbens dataset, such as *SERPINA1*. This offers a much more informative result, as many other genes with similar or even higher fold change values may be possible indicators of

https://doi.org/10.59720/23-099



Genes	VarImp
RP11-498J9.2	100.0000
CH507-210P18.3	95.2400
RP5-1021120.8	83.8600
SNORA38	72.5500
TROAP	57.7300
PAK6	56.9700
MT1JP	54.7500
RP11-18014.4	50.2200
RP11-1084E5.1	49.2100
RP11-770E5.1	48.1000
AC108676.1	46.9700
RP11-334E6.12	44.5800
C5orf58	44.5000
BMPR1APS1	43.5700
RP11-553L6.2	39.7400
LINC01191	36.2400
COX6CP15	35.7100
LINC01497	34.9700
RP11-882115.1	32.1600
CTC-523E23.4	31.5100

Confusion Matrix and Statistics

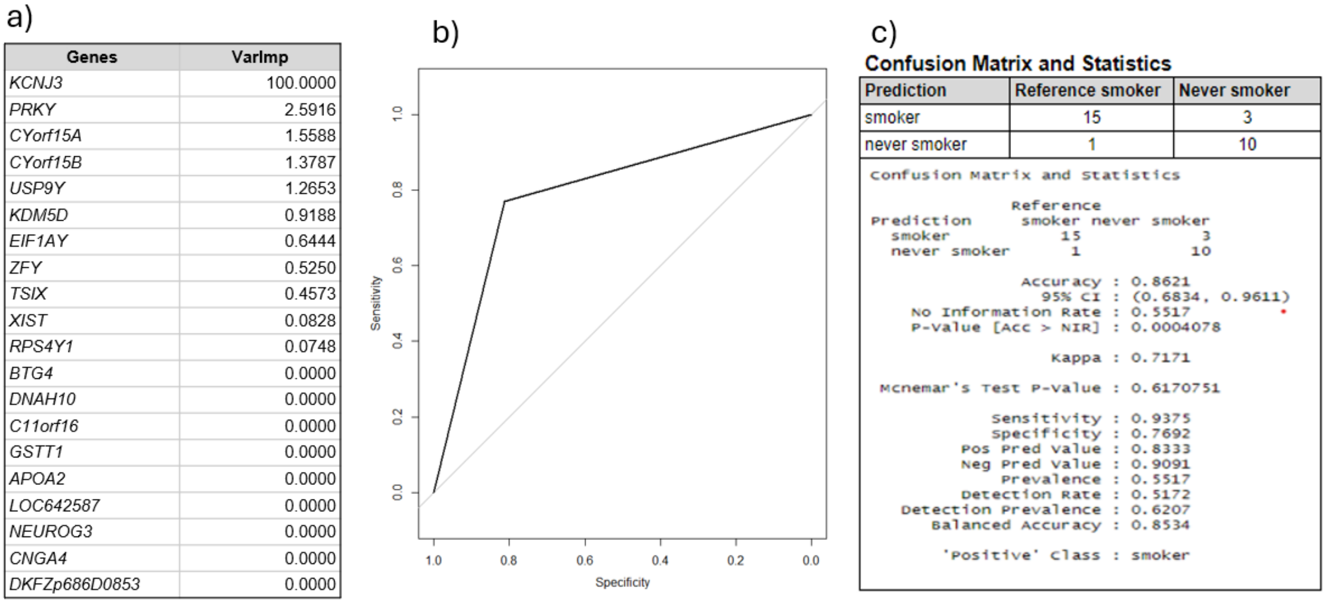
Prediction	Reference smoker	Never smoker
smoker	0	0
never smoker	10	34

```
                  Reference
Prediction         smoker never smoker
  smoker                0            0
  never smoker         10           34

               Accuracy : 0.7727
                 95% CI : (0.6216, 0.8853)
    No Information Rate : 0.7727
    P-Value [Acc > NIR] : 0.583734

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7727
             Prevalence : 0.2273
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : smoker
```

Figure 3: Random forest prediction model for nucleus accumbens dataset. (a) List of genetic markers ordered by variable importance in creating the prediction, (b) plot of AUC for the prediction, and (c) confusion matrix. N=50 case (smoker) samples, N=171 control (non-smoker), N=2 indeterminant smoking status. AUC = 0.65-0.70.

smoking status in patients.

For the secondary analysis of both datasets, logistic regression was used as a means to determine additional genes that could be predictors of smoking status. In the nucleus accumbens dataset, gene expression is not as strong a predictor of smoking status, but the lung tissue dataset shows much higher accuracy. One possible reason for this discrepancy is the difference between sample sources. The first dataset took samples from the nucleus accumbens part of the brain while the second took lung tissue samples. Nicotine intake is primarily done through smoking e-cigarettes or normal cigarettes, with nicotine directly entering the lungs and soon after entering the bloodstream as well (16). While the alveoli in the lungs are exposed to the full amount of nicotine and other carcinogens present in a nicotine product, the brain only comes into contact with these chemicals through the bloodstream, which must pass through much of the body and the blood-brain barrier before reaching the brain. This much longer travel path may function to decrease the number of absorbed carcinogens that reach the brain and lower their impact in comparison to the lungs, resulting in more genetic mutations in lung samples compared to brain



Genes	VarImp
KCNJ3	100.0000
PRKY	2.5916
CYorf15A	1.5588
CYorf15B	1.3787
USP9Y	1.2653
KDM5D	0.9188
EIF1AY	0.6444
ZFY	0.5250
TSIX	0.4573
XIST	0.0828
RPS4Y1	0.0748
BTG4	0.0000
DNAH10	0.0000
C11orf16	0.0000
GSTT1	0.0000
APOA2	0.0000
LOC642587	0.0000
NEUROG3	0.0000
CNGA4	0.0000
DKFZp686D0853	0.0000

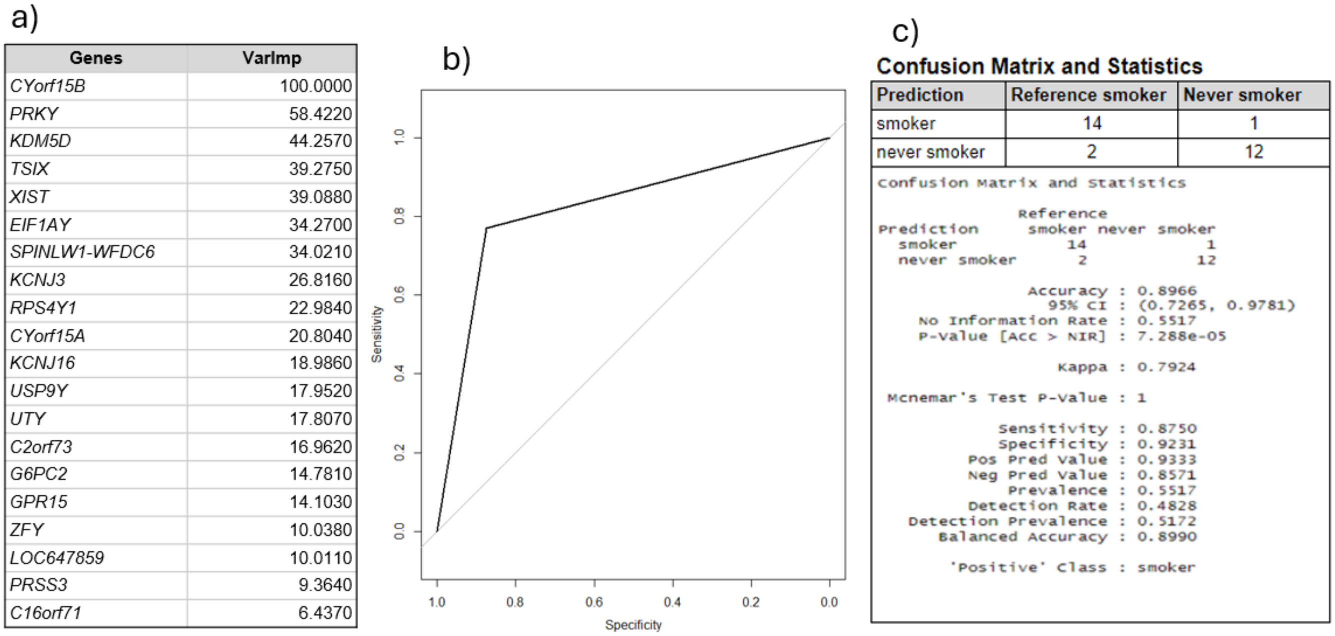
Confusion Matrix and Statistics

Prediction	Reference smoker	Never smoker
smoker	15	3
never smoker	1	10

```
Confusion Matrix and Statistics

                  Reference
Prediction         smoker never smoker
  smoker               15            3
  never smoker          1           10

               Accuracy : 0.8621
                 95% CI : (0.6834, 0.9611)
    No Information Rate : 0.5517
    P-Value [Acc > NIR] : 0.0004078

                  Kappa : 0.7171

 Mcnemar's Test P-Value : 0.6170751

            Sensitivity : 0.9375
            Specificity : 0.7692
         Pos Pred Value : 0.8333
         Neg Pred Value : 0.9091
             Prevalence : 0.5517
         Detection Rate : 0.5172
   Detection Prevalence : 0.6207
      Balanced Accuracy : 0.8534

       'Positive' Class : smoker
```

Figure 4. GLMNET prediction model for lung tissue dataset. (a) List of genetic markers ordered by variable importance in creating the prediction, (b) plot of AUC for the prediction, and (c) confusion matrix. N=39 case (smoker) samples, N=34 control (non-smoker), N=4 indeterminant smoking status. AUC = 0.79-0.92.

https://doi.org/10.59720/23-099



Genes	VarImp
CYorf15B	100.0000
PRKY	58.4220
KDM5D	44.2570
TSIX	39.2750
XIST	39.0880
EIF1AY	34.2700
SPINLW1-WFDC6	34.0210
KCNJ3	26.8160
RPS4Y1	22.9840
CYorf15A	20.8040
KCNJ16	18.9860
USP9Y	17.9520
UTY	17.8070
C2orf73	16.9620
G6PC2	14.7810
GPR15	14.1030
ZFY	10.0380
LOC647859	10.0110
PRSS3	9.3640
C16orf71	6.4370

Confusion Matrix and Statistics

Prediction	Reference smoker	Never smoker
smoker	14	1
never smoker	2	12

```
Confusion Matrix and Statistics

                 Reference
Prediction        smoker never smoker
   smoker             14            1
   never smoker        2           12

               Accuracy : 0.8966
                 95% CI : (0.7265, 0.9781)
    No Information Rate : 0.5517
    P-Value [Acc > NIR] : 7.288e-05

                  Kappa : 0.7924

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8750
            Specificity : 0.9231
         Pos Pred Value : 0.9333
         Neg Pred Value : 0.8571
             Prevalence : 0.5517
         Detection Rate : 0.4828
   Detection Prevalence : 0.5172
      Balanced Accuracy : 0.8990

       'Positive' Class : smoker
```

Figure 5. Random forest prediction model for lung tissue dataset. (a) List of genetic markers ordered by variable importance in creating the prediction, (b) plot of AUC for the prediction, and (c) confusion matrix. N=39 case (smoker) samples, N=34 control (non-smoker), N=4 indeterminant smoking status. AUC = 0.82-0.93.

samples. Additionally, as longer time of exposure correlates with mutation rate, both the amount of time and number of cigarettes smoked will influence gene expression (17).

Another reason for this discrepancy could lie in the data of the patients in the nucleus accumbens dataset. Compared to the lung tissue dataset, it is possible that the nucleus accumbens dataset was noisier, and many data points could potentially have skewed the result. We indeed found the data to be quite noisy, which can make it difficult to interpret general correlations simply due to the high variability in expression values (**Figures 1 and 2**). Of course, it is entirely possible that there is just more natural variation in gene expression in one tissue type compared to another. From the lung tissue dataset, smoking status is predicted with a high enough accuracy for many of the genes used in the model to be considered good indicators, all displaying high differential expression between smoker and non-smoker patients.

Other factors could also impact a gene's significance in logistic regression. In the case of gene expression, sex is a significant factor that was not considered in this study. For example, three of the five genes with the highest fold change values from the lung tissue dataset are Y-linked genes (*PRKY*, *TSIX*, and *CYORF15B)* (18). Despite all being genes found to be very relevant in both the GLMNET and random forest models, they cannot be fully considered as genes that can act as predictors for nicotine addiction in patients. Located solely on the Y chromosome, females would not express these genes, generating false significance in the prediction model. However, these genes may still be important, as a previous study has indicated that nicotine dependency may be partially sex-based (19). It would be extremely beneficial for future research on this topic to consider biological sex and possibly create two separate prediction models, which would help eliminate some confounding variables and test this hypothesis.

In this paper, we determined *KCNJ3* and *TXLNGY* as two candidate markers of smoking status. *KCNJ3* encodes a G-protein-activated potassium channel that plays an important role in regulating cell function in the heart and brain (20). These G-protein-gated inwardly rectifying potassium (GIRK) channels have been previously found to correlate with addiction, epilepsy, and other mental disorders (21). *CYORF15B*, more commonly known as *TXLNGY*, primarily enables syntaxin binding, a biological process that plays a role in the growth of neurons during brain development. Other genes performing the same function such as *SYN1A* have been found to interact with dopamine transporters, a known effect of nicotine in the brain (22). A better understanding of the genetic basis of smoking status will aid in identifying smoking-associated risk factors and the development of smoking cessation drugs.

MATERIALS AND METHODS

Data acquisition and cleaning

Nucleus accumbens RNA-seq data was retrieved from GSE171936 (23). It includes data from 223 deceased individuals: 50 current cigarette smokers, 171 nonsmokers, and 2 individuals with undetermined smoking status. Lung tissue RNA-seq data was retrieved from GSE40419 (24). It includes data from 87 lung adenocarcinomas and 77 adjacent normal tissues. For the purposes of this study, we only utilized the normal tissue samples, as the adenocarcinoma samples do not affect smoking status. Among the normal tissue, there are 39 smoking patients, 34 non-smoking patients, and 4 patients of undetermined smoking status. Moreover, in the data, the smoking status considered whether a patient was a current smoker or a smoker at some point in their life, but the groups were assembled as one smoker category for ease of processing.

Preliminary Analysis

First, R was used to sort through the dataset and identify genes that were either overexpressed or underexpressed in smokers compared to nonsmokers. ggplot2 was used to compare the transcripts per million (TPM) or reads per million (RPM) values of smokers and nonsmokers to identify significantly differentially expressed genes (using a Student's t-test threshold of $p<0.05$ for significance).

An Excel table containing the fold change values (sorted in descending order) was created by comparing the case-median expression and control-median expression of each gene (taking median rather than mean expression decreases the influence of outliers).

Predictive modeling

R Studio was utilized to perform the data training and modeling. Gene expression values with an absolute $\log_2$ fold change greater than one as input to GLMNET logistic regression or random forest classifiers were used to try to predict smoking status from gene expression. Training data sets were used as input to GLMNET or random forest classifiers, using 5-fold cross validation via the 'caret' package (25). To better visualize the results, a confusion matrix was created using its provided function in R.

Received: May 14, 2023
Accepted: September 29, 2023
Published: August 30, 2024

REFERENCES

1. Lin, Crystal et al. "Nicotine Dependence from Different E-Cigarette Devices and Combustible Cigarettes among US Adolescent and Young Adult Users." *International Journal of Environmental Research and Public Health*, vol. 19, no. 10, 11 May 2022, p. 5846, https://doi.org/10.3390/ijerph19105846.
2. "Harms of Cigarette Smoking and Health Benefits of Quitting." *National Cancer Institute*, 19 Dec. 2017, www.cancer.gov/about-cancer/causes-prevention/risk/tobacco/cessation-fact-sheet. Accessed 13 May 2023.
3. "Tobacco: Industry Tactics to Attract Younger Generations." *World Health Organization*, 25 Mar. 2020, www.who.int/news-room/questions-and-answers/item/tobacco-industry-tactics-to-attract-younger-generations.
4. Glantz, Stanton A., and David W. Bareham. "E-Cigarettes: Use, Effects on Smoking, Risks, and Policy Implications." *Annual Review of Public Health*, vol. 39, no. 1, 2018, pp. 215-35, https://doi.org/abs/10.1146/annurev-publhealth-040617-013757.
5. Maritz, Gert S. "Nicotine and Lung Development." *Birth Defects Research. Part C, Embryo Today: Reviews*, vol. 84, no. 1, 2008, pp. 45-53. https://doi.org/10.1002/bdrc.20116.
6. Benowitz, Neal L. "Pharmacology of Nicotine: Addiction, Smoking-Induced Disease, and Therapeutics." *Annual Review of Pharmacology and Toxicology*, vol. 49, 2009, pp. 57-71. https://doi.org/10.1146/annurev.pharmtox.48.113006.094742.
7. "How Does Smoking Cause Cancer?" *Cancer Research UK*, 19 Mar. 2021, www.cancerresearchuk.org/about-cancer/causes-of-cancer/smoking-and-cancer/how-does-smoking-cause-cancer#:~:text=Chemicals%20from%20cigarettes%20damage%20DNA,time%20that%20leads%20to%20cancer.
8. "Nicotine Addiction Overview Unit 1: The Brain." *Stanford Medicine*, med.stanford.edu/tobaccopreventiontoolkit-old/nicotine-addiction/ NicotineAddictionUnit1.html#:~:text=The%20Reward%20Pathway,cause%20a%20release%20of%20dopamine.
9. Mackillop, James et al. "The Role of Genetics in Nicotine Dependence: Mapping the Pathways from Genome to Syndrome." *Current Cardiovascular Risk Reports*, vol. 4, no. 6, 2010, pp. 446-453. https://doi.org/10.1007/s12170-010-0132-6.
10. Price, M. "Genes Matter in Addiction." *Monitor on Psychology*, vol. 39, no. 6, June 2008, https://www.apa.org/monitor/2008/06/genes-addict.
11. Quach, B.C., et al. "Expanding the Genetic Architecture of Nicotine Dependence and Its Shared Genetics with Multiple Traits." *Nature Communications*, vol. 11, 2020, p. 5562. https://doi.org/10.1038/s41467-020-19265-z.
12. Rotondo, J.C., et al. "Methylation of SERPINA1 Gene Promoter May Predict Chronic Obstructive Pulmonary Disease in Patients Affected by Acute Coronary Syndrome." *Clinical Epigenetics*, vol. 13, 2021, p. 79. https://doi.org/10.1186/s13148-021-01066-w.
13. Quach, B.C., et al. "Expanding the Genetic Architecture of Nicotine Dependence and Its Shared Genetics with Multiple Traits." *Nature Communications*, vol. 11, 2020, p. 5562. https://doi.org/10.1038/s41467-020-19265-z.
14. Tam, Vivian et al. "Benefits and Limitations of Genome-wide Association Studies." *Nature Reviews. Genetics*, vol. 20, no. 8, 2019, pp. 467-484. https://doi.org/10.1038/s41576-019-0127-1.
15. Hastie, Trevor, et al. "An Introduction to glmnet." *Stanford*, 27 Mar. 2023, glmnet.stanford.edu/articles/glmnet.html#:~:text=The%20algorithm%20is%20extremely%20fast,and%20relaxed%20lasso%20regression%20models.
16. Eaton DL, Kwan LY, Stratton K, editors. *Public Health Consequences of E-Cigarettes*. Washington (DC): National Academies Press (US), 2018 Jan 23. 4, Nicotine. Available from: www.ncbi.nlm.nih.gov/books/NBK507191/.
17. Tsai, Pei-Chien et al. "Smoking Induces Coordinated DNA Methylation and Gene Expression Changes in Adipose Tissue with Consequences for Metabolic Health." *Clinical Epigenetics*, vol. 10, no. 1, 20 Oct. 2018, https://doi.org/10.1186/s13148-018-0558-0.
18. Rozen, Steve, et al. "Remarkably Little Variation in Proteins Encoded by the Y Chromosome's Single-Copy Genes, Impling Effective Purifying Selection." American Journal of Human Genetics, vol. 85, no. 6, 2009, pp. 923-928. https://doi.org/10.1016/j.ajhg.2009.11.011
19. Kozlova, A., et al. "Sex-Specific Nicotine Sensitization and Imprinting of Self-Administration in Rats Inform GWAS Findings on Human Addiction Phenotypes." *Neuropsychopharmacology*, vol. 46, 2021, pp. 1746–1756. https://doi.org/10.1038/s41386-021-01027-0.
20. Hancock D. "Genome-wide RNA-sequencing differences in nucleus accumbens of smokers vs. nonsmokers." Gene Expression Omnibus, 14 Apr. 2021, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171936.
21. Seo J, Ju YS, Lee W. "The Transcriptional Landscape and Mutational Profile of Lung Adenocarcinoma." Gene Expression Omnibus, 6 Sep. 2012, www.ncbi.nlm.nih.gov/

geo/query/acc.cgi?acc=ERP001058.

22. Kuhn, Max. “Building Predictive Models in R Using the caret Package.” *Journal of Statistical Software*, vol. 28, no. 5, 2008, pp. 1–26. https://doi.org/10.18637/jss.v028.i05.
23. Benowitz, Neal L et al. “Nicotine Chemistry, Metabolism, Kinetics and Biomarkers.” *Handbook of Experimental Pharmacology*, vol. 192, 2009, pp. 29-60. https://doi.org/10.1007/978-3-540-69248-5_2.
24. Kano, H., et al. “Structural Mechanism Underlying G Protein Family-Specific Regulation of G Protein-Gated Inwardly Rectifying Potassium Channel.” *Nature Communications*, vol. 10, 2019, p. 2008. https://doi.org/10.1038/s41467-019-10038-x.
25. “KCNJ3 Potassium Inwardly Rectifying Channel Subfamily J Member 3 [Homo sapiens (human)].” *National Center for Biotechnology Information*, 29 Mar. 2023, www.ncbi.nlm.nih.gov/gene/3760#summary.