

Gene expression levels from scRNA-Seq between smokers and non-smokers in COVID-19 Patients

Andy Hsu, Kathy Khong and Phillip Tellier

Abstract

Muus et al. investigated the correlation between ACE2, TMPRSS2, and CTSL expression and smoking in COVID-19 patients. As an extension to this study, this paper investigated the predictability of smoking habits using scRNA-Seq gene expressions of COVID-19 patients. Topic modelling, logistic regression, and random forests revealed that CTSC, CTSL may be good predictors of smoking behaviours. It is recommended that the authors of the original study investigate other potential genes' expressions with smoking in COVID-19 patients and not limit their analysis to ACE2, TMPRSS2, and CTSL.

Keywords: Machine Learning, Topic Modelling, Random Forest, Logistic Regression, Single-cell RNA Sequencing, Non-negative Matrix Factorization, COVID-19, Smoking

1 Introduction

With the COVID-19 pandemic causing millions of deaths, much research has been conducted to identify potential risk factors associated with severe COVID-19 symptoms in hopes to reduce mortality. One such factor, smoking, has been found to increase the expression of angiotensin-converting enzyme 2 (ACE2) in the lungs. ACE2 is a receptor for SARS-CoV-2, the virus responsible for COVID-19 and hence it is suspected there is an influence of smoking on COVID-19. Unfortunately, this connection remains unclear. There have been conflicting findings regarding the relationship between ACE2 levels and COVID-19 infection risk (Leung et al., 2020), (Gheware et al., 2022). In this paper, using the data set provided by Muus et al, the analysis of a previously conducted experiment will be extended (Muus et al., 2020). Namely, with the gene expression provided, the fit of different models on the data will be investigated in hopes of examining the impact of smoking on COVID-19 patients. The findings could contribute to a better understanding of the complex relationship between smoking, COVID-19, and other risk factors, ultimately aiding in the development of effective prevention and treatment strategies.

Muus et al. (2020) conducted an integrated analysis of 107 single-cell and single-nucleus RNA sequencing studies on various tissues. This included 22 studies on tissue relating to the respiratory system. The researchers conducted differential analysis on ACE2 viral receptor, type II transmembrane serine proteases (TMPRSS2), and cathepsin L (CTSL); these are genes that have previously been found to be important for the uptake of SARS-CoV-2 into the cell. By calculating the correlation with various risk factors such as age, sex, and smoking status, they investigated how these risk factors affected the aforementioned gene expressions (Muus et al., 2020). This further allowed the researchers to identify gene expression programs of cells that are susceptible to SARS-CoV-2 infection. The gene program was compared across different cell types, organs, and species. In an experiment that investigated smoking, Muus et al. (2020) reported an upregulation of ACE2, TMPRSS2, and CTSL associated with smokers. However, Muus et al. (2020) did not investigate how gene expression could be a potential predictor for smoking status. Hence, the current study aims to expand on the findings by investigating the levels of gene expressions as a predictor for smokers and non-smokers in COVID-19 patients.

Given that smoking is a known risk factor for respiratory infections such as COVID-19 (He, Sun, Ding, & Wang, 2021), understanding how smoking affects the expression levels of these genes is critical in understanding the mechanisms by which smoking increases the risk of severe COVID-19. For instance, there are reasons as to why Muus et al. (2020) investigated ACE2, TMPRSS2, and CTSL. In a pathological sense, ACE2 is a receptor used by SARS-CoV-2 to infiltrate the cells. On the other hand, TMPRSS2 and CTSL are proteases responsible for cleaving the spike protein on SARS-CoV-2 (Chua et al., 2020). The cleavage of the spike protein is also a necessity for viral entry into host cells. These highlight the importance of gene

expression levels and certain changes can provide insight into the severity of the disease and the impact of smoking on the immune response to the virus.

2 Methods

To establish a baseline model, `DummyClassifier` from the Scikit-learn library with a "most frequent" strategy was used to predict the mode of the classes. This approach predicts the class label, in this case, the smoker status that has the most samples in the training set. We then calculated the accuracy of this model to be used as a reference for comparison.

To obtain a preliminary visualization of the data, the existing UMAP data provided by the authors was plotted, grouping by cell types, as well as smoking conditions.

Topic modelling was then performed to identify co-expressed gene pathways between samples using non-negative matrix factorization (NMF) from the Scikit-learn library. For hyperparameter tuning, the number of topics k was chosen based on the reconstruction error for various k values. The mean squared error (MSE) metric from Scikit-learn's mean squared error function was used and an elbow curve was plotted using the `matplotlib.pyplot` library (Fig. A2). The "elbow method" was used to determine the optimal value of k , which was found to be 8.

The Wilcoxon rank sum test from the `scipy.stats` library was used to investigate significant differences between smokers and non-smokers for each topic identified from the NMF analysis. The `ranksums` function was used to calculate the test statistic and p-value for each pair of topics at a significance level of 0.05. A pair of bar graphs using `matplotlib.pyplot` were then plotted to visually compare the mean weights for each of the 8 topics between smoker and non-smoker classes.

Next, a logistic regression model was fitted to the balanced data set using Scikit-learn's `LogisticRegression` method. `LogisticRegression` was first fitted on the whole data set, then on each cell category. Using cross-validation and L2 regularization, hyperparameters were optimized on each of the models generated according to cell categories. The best three models were found and the f1 score, coefficients for features, and intercept were obtained. Unfortunately, Scikit-learn did not provide a method for obtaining the statistics. To obtain the statistics, the package `statsmodel` was utilized. Using `statsmodel`, a binomial generalized linear model was fitted on the top 3 models to check the significance of the coefficients.

To explore how a general machine learning model could fit the data and possibly capture non-monotonic trends, three different random forest based models were fitted to the balanced data using `RandomForestClassifier` from Scikit-learn. Each model used five-fold cross-validation to decide the best maximum depth for its random trees (Fig. A3). The number of random trees for each model was set to be as many as was runnable in a reasonable time. The three models were: **1.** A single random forest trained with 100 trees on all training data **2.** An ensemble of random forests (10 trees each) where each forest is trained on data under a particular cell type annotation given by [Muus et al. \(2020\)](#) (cell-type based random forest ensemble). Cross-validation was run to select max depth for each forest separately. Each prediction used one forest for the cell's type. **3.** An ensemble of random forests for each NMF topic (topic based random forest ensemble). Each random forest had 20 trees. Each forest was trained on a random sample of examples where the selection probability of each example was weighted by its topic proportion for that forest topic. After using L1 normalization to make the topic proportions of each example sum to 1, prediction happened by taking the weighted average of predictions of all models, weighted by the topic proportions. Again, cross-validation selected the max-depth for each forest separately.

Feature importances were calculated for each random forest in all three random forest based models using the mean decrease in Gini impurity.

3 Results

The baseline Dummy Classifier achieved an accuracy of 0.498. The UMAP data showed a total of 17 clusters. When splitting between the conditions, it can be observed these clusters are co-localized between the cells except for the epithelial AT2 cells (Fig. 1).

When performing NMF, eight topics representing sets of genes that are possibly co-expressed and involved in similar pathways were identified (Table 1). Using the Wilcoxon Rank Sum test, pairwise comparisons of each topic between the two groups (smokers versus non-smokers) revealed a significant difference ($p \leq 0.05$) for each topic (Fig. 2.1). When examining the H matrix, we identified the top three genes that

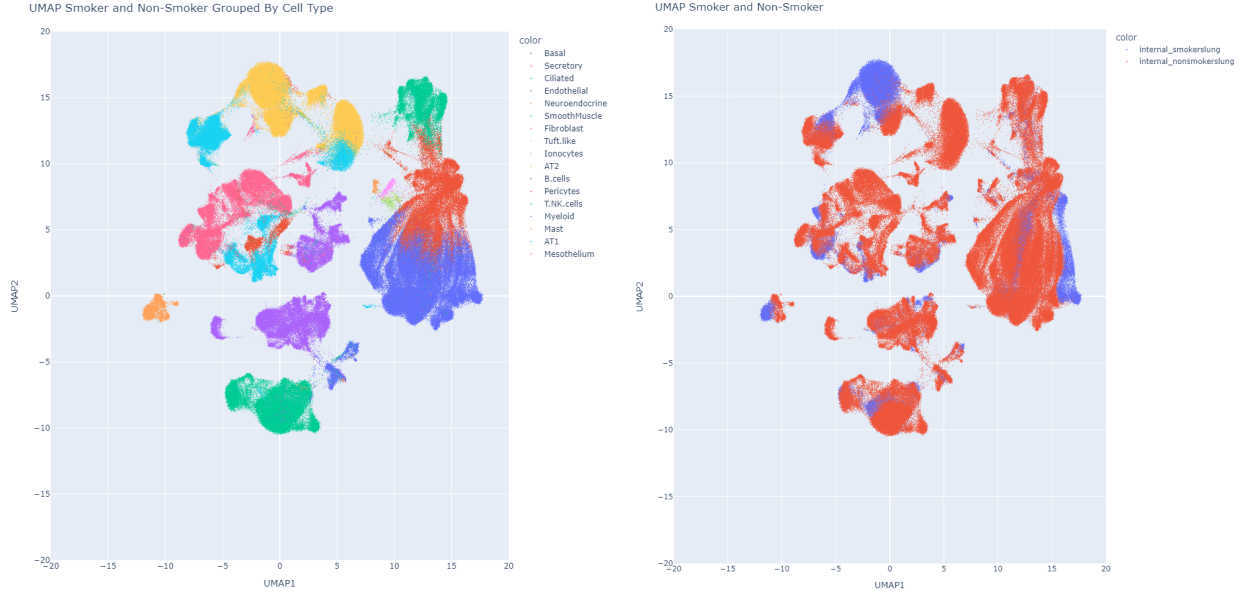


Fig. 1 Co-localization of cells can be observed except for that of epithelial AT2 cells. **1.** The UMAP data of smoker and non-smoker cells grouped by cell type (left). **2.** The UMAP data of smokers and non-smokers grouped by smoking status (right).

contributed the most weight to each topic (Fig. 2.2). Notably, the gene IL6ST was present in multiple topics (#2, #4, #5 and #6). CTSC was identified in both topics #4 and #8. C3, CTSC, C1R, CTSL, IL6 are the genes that contribute the most weight to topics #1, #2, #3, #4 and #5 respectively.

Table 1 Topic results from topic modelling using NMF

Topic	Genes	Topic	Genes
#1	C3, CFI, C2, CTSE, TMPRSS2, MBP, PCSK7	#5	IL6, IL6ST, PCSK5, CFI, PCSK7, FURIN, C3
#2	CTSC, IL6ST, PCSK7, PCSK5, PCSK1, C3, ACE2	#6	IL6ST, C2, PCSK7, TMPRSS2, MBP, CFI, FURIN
#3	C1R, IL6ST, PCSK5, CFI, C3, CTSC, PCSK1	#7	CTSS, C2, TMPRSS2, MBP, IL6R, PCSK5, FURIN
#4	CTSL, CTSC, IL6ST, C3, PCSK1, CTSS, MAG	#8	CTSB, MBP, PCSK7, FURIN, IL6R, PCSK5, MYRF

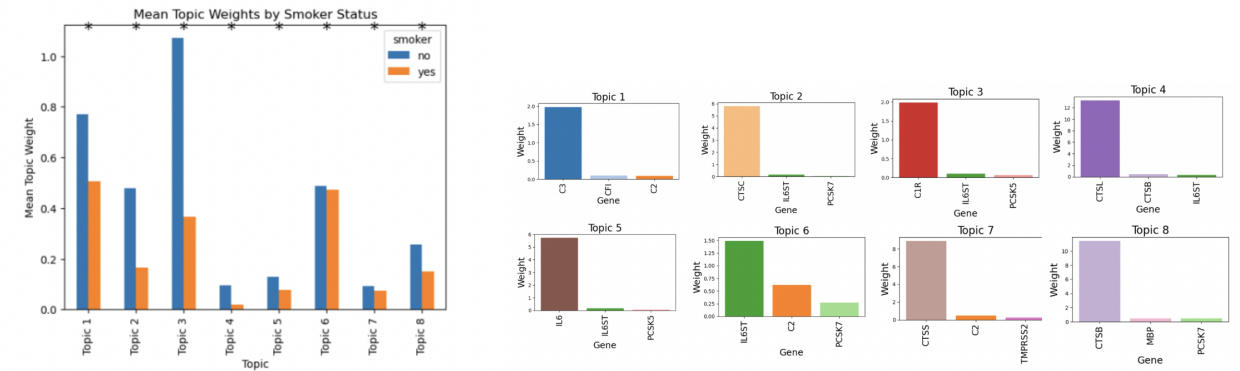


Fig. 2 **1.** Mean weights of topics 1-8 between smokers and non-smokers. Asterisks(*) show significant differences (left) **2.** Top 3 genes that contribute the most weight to each topic (right).

For the logistic regression model, the top three models grouped by cell types are epithelial AT2 cells, stromal fibroblasts, and stromal pericytes, which has testing scores of 0.7656, 0.7240, and 0.6635, respectively (Table 2). The top three coefficients for epithelial AT2 cells model are CTSC, CTSC, C1R (Table 2). The top three coefficients for the stromal fibroblasts model are CTSL, CTSC, C1R (Table 2). Lastly, the stromal pericytes model has CTSC, PCSK7, and CTSL as its top three coefficients (Table 2). All coefficients have

been validated with statsmodel, and all are significant to $\alpha = 0.05$ (Fig. A4). Note the intercepts are not significant (Fig. A4).

Table 2 Testing score, intercepts, and top three coefficients of top three logistic regression model from sklearn

Model	Test %	Intercept	Coefficients		
Epithelial (AT2)	76.48	-0.5907	-1.6304 (CTSC)	-1.1883 (CSTB)	-0.5009 (C1R)
Stromal (Fibroblast)	72.51	-1.0492	-2.2451 (CTSL)	-2.1979 (CTSC)	0.6803 (C1R)
Stromal (Pericyte)	66.35	-0.6252	-2.9893 (CTSC)	-1.004 (CTSL)	0.9539 (PCSK7)

The three random forest models performed worse than the best logistic regression model on the dataset as a whole (68.87%, 66.08%, and 59.53% test accuracy versus 76.56% respectively). The cell type based random forest ensemble had the best test accuracy (68.87%). It was only marginally more accurate than the single random forest (66.08%), and much more accurate than the topic based random forest ensemble (59.53%). Interestingly, the random tree for AT2 cells in the cell type based random forest ensemble had 92.9% accuracy on its testing subset and outperformed the AT2 cell type logistic regression model having 76.48% accuracy.

Table 3 Accuracies of Machine learning models to predict smoker or non-smoker status of cells from lung scRNA-seq data

Model	Test %	Train %	Hyper-parameters
logistic regression with discrete cell types	76.56	76.86	C = 1
single random forest	66.08	78.24	max_depth= 32, n_estimators= 100
random forest with discrete cell types	68.87	85.06	max_depth ¹ = 20 – 40, n_estimators= 10(each)
random forest with topic modeling	59.53	74.05	max_depth ¹ = 30 – 40, n_estimators= 20(each), k = 8

¹Separate cross-validation for each random forest lead to a separate max depth for each forest, hence there is a range of depths

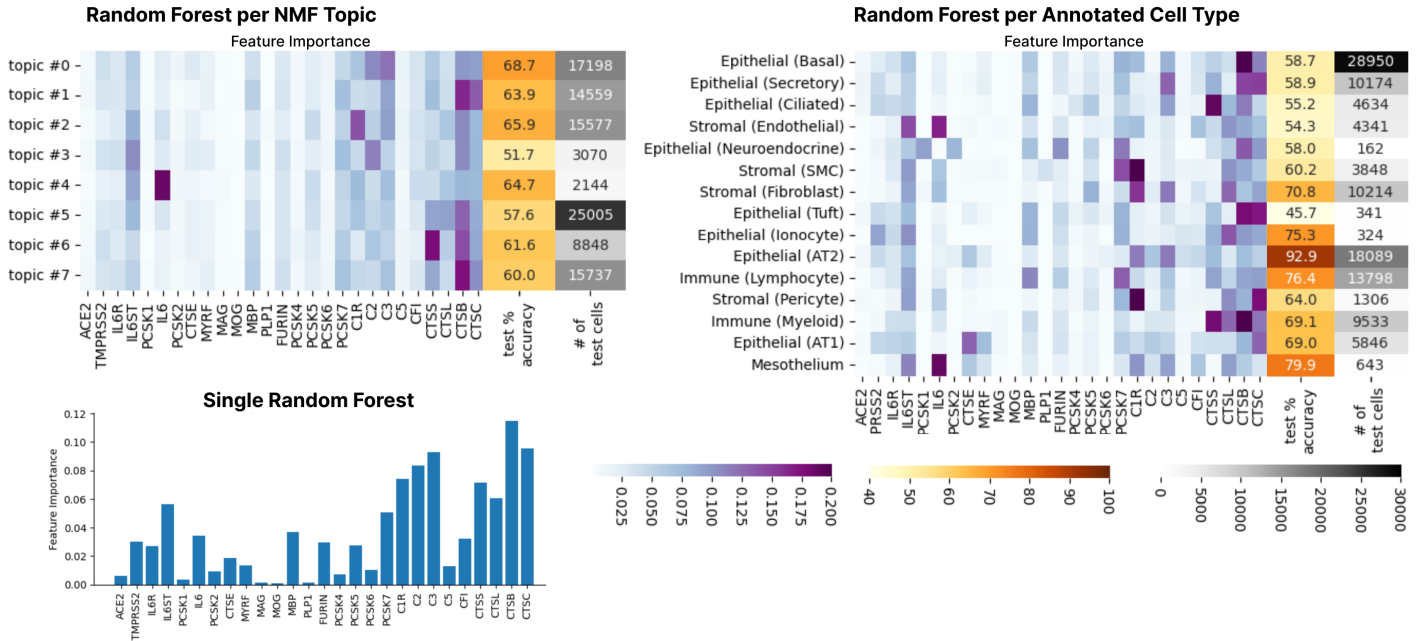


Fig. 3 Feature importances calculated by the mean decrease in Gini impurity for random forests in three forest based models predicting whether cells are from smokers or non-smokers based on gene expression: **1.** A single random forest trained with 100 trees on all training data **2.** A cell type based random forest ensemble with 10 trees for each forest **3.** A NMF topic based random forest ensemble with 20 trees for each forest

Examining feature importances from the single random forest shows 4/5 of the proteins from the cathepsin family of proteases are the most important in making predictions (Fig. 4.1), namely CTSS, CTSL, CTSE, and CTSC, but not CTSE. The next highest importance come from 3/4 complement proteins, namely C1R,

C2, C3, but not C5. IL6ST and PCSK7 are next. The random tree for AT2 cells in the cell type based random forest ensemble had high feature importances for many complement proteins (C1R, C3) and cathepsin proteases (CTSS, CTSB) as well (Fig. 4.2).

4 Discussion

From the UMAP’s perspective, it is interesting to observe that most cells exhibit a high degree of co-localization except for the epithelial AT2 cells and how this is aligning with our highest accuracy models. Interestingly, other articles have reported that AT2 cells are important in lung repair and immune response (Olajuyin, Zhang, & Ji, 2019).

From the results, we showed that our NMF-based topic modelling approach can identify distinct sets of genes that are potentially involved in specific biological pathways related to smoking. The significant differences ($p < 0.05$) between smokers and non-smokers suggest that smoking may have a differential effect on the expression of genes involved in these specific pathways. Smoking has been linked to increased levels of pro-inflammatory cytokines like IL-6. It is possible that smoking-induced changes in the immune system could lead to a downregulation of these particular pathways, resulting in lower mean weights for these topics among smokers (Chen, Cowan, Hasday, Vogel, & Medvedev, 2007). Moreover, smoking can cause changes in DNA methylation that increase the risk of developing lung cancer. The specific changes observed in DNA methylation were found to be linked to the C3 protein, highlighting the potential impact of smoking-induced changes on health. (Zeilinger et al., 2013). Taken together, our findings may provide insights into potential mechanisms for the observed differences in mean topic weights between smokers and non-smokers, suggesting that DNA methylation may be a crucial molecular mechanism underlying the effects of smoking on immune function and inflammation.

When examining the top three gene contributions for each topic, it is interesting to note that IL6ST was present in multiple topics, indicating its potential involvement in multiple pathways affected the pathogenesis of SARS-CoV-2. IL6ST is a receptor for the cytokine interleukin-6 (IL-6), which is a key mediator of the inflammatory response. Del Valle et al. (2020) have shown that IL-6 is a strong predictor for patient survival at time of hospitalization. IL-6’s role in patient survival may be due its involvement in the pathogenesis of cytokine storm. Cytokine storm is a sudden, heightened immune response that can lead to tissue damage, organ failure, and death (Del Valle et al., 2020).

Additionally, the gene CTSB was identified in both topics 4 and 8 suggesting a significant contribution across the topics. Previous findings have shown that CTSB is highly expressed in lung adenocarcinoma after SARS-CoV-2 infection and is also positively associated with proinflammatory protein expression (Ding et al., 2022). These findings suggest CTSB may play an important role in the hyper-inflammatory response of COVID-19 patients. Furthermore, C1R is involved in the activation of the C1 complex which is responsible for identifying and binding to foreign invaders, such as viruses, and triggering a cascade of biological pathways that ultimately leads to their destruction (Uhlen et al., 2010). Survival analysis findings have revealed a significant correlation between high expression of C1R and pathways involved in COVID-19 (Wang et al., 2022). It is important to note that topic modelling is an unsupervised machine learning model, and as such, it cannot establish causal relationships between variables. While the NMF-based approach has identified distinct sets of genes potentially involved in specific biological pathways affected by smoking, further research is required to establish causality and underlying mechanisms. Nonetheless, the findings provide valuable insights and hypotheses for future investigation into the effects of smoking on gene expression and immune function.

In the logistic regression and random forest based models, there were a few genes repeatedly observed to be of importance in the top models. These genes were not limited to, but included CTSC, CTSL, and C1R. Indeed, aligning with the literature, CTSC has been found to be a cause of immune-related diseases such as cancer and Papillon-Lefèvre syndrome (Korkmaz et al., 2018). Some even suggested CTSC inhibition may be a therapeutic target for cancer (Korkmaz et al., 2021). It is also not surprising that CTSL has similar functions. Previous research suggested that CTSL is upregulated in cancer and also a potential therapeutic target (Sudhan & Siemann, 2015). As well, C1R is a gene that encodes a protein important for the regulation of the innate immune system (Sim, 1981). From our perspective, it is interesting to observe that all the genes correlated in prediction have immune function implications, which could potentially be caused by the upregulation of the immune system due to smoking.

One interesting aspect to note is that the top three models of logistic regression did not include any immune cells. While unsure why this may be the case, it is suspected that the acute heightened immune response due to COVID infections may mask the predictability of immune cells on smoking behaviours. On the other hand, the chronically elevated proliferation of epithelial cells of smoker’s lungs compared to

non-smoker’s lungs remained a predictive factor (Thorley & Tetley, 2007). In contrast to logistic regression, the cell type based random forest ensemble had the third highest accuracy (76.43%) of all cell types for the lymphocyte immune subpopulation and the sixth highest accuracy for myeloid immune subpopulation (69.11%). Hence, the random forest lymphocyte submodel competed with the most accurate submodel of the logistic regression ensemble at 76.48% for the AT2 cell submodel on its subpopulation. It is possible that there are non-monotonic trends present in the immune cell gene expression that are better predicted by random forests since logistic regression can only model monotonic trends. Further experimentation using smokers and non-smokers in a healthy population may be a potential avenue to confirm this hypothesis. The logistic regression model has its limitation in that it only considers the data fitting a sigmoid curve. Nevertheless, there could still be other avenues for future studies such as incorporating other features provided with the data such as age.

5 Conclusion

After fitting various models, our results indicated that several immune regulatory genes (CTSB, CTSC, CTSL) are predictive of smoking and non-smoking behaviours in COVID-19 patients. This result is surprising in contrast with Muus et al’s result. As mentioned before, Muus et al. investigated ACE2, TMPRSS2, and CTSL, yet only CTSL appeared aligned with our findings. Thus, it is suggested that the author of the original paper should extend to potentially other gene expression analyses.

Appendix A

The GitHub repository to the code: <https://github.com/ptellier/covid-smoker-ml>

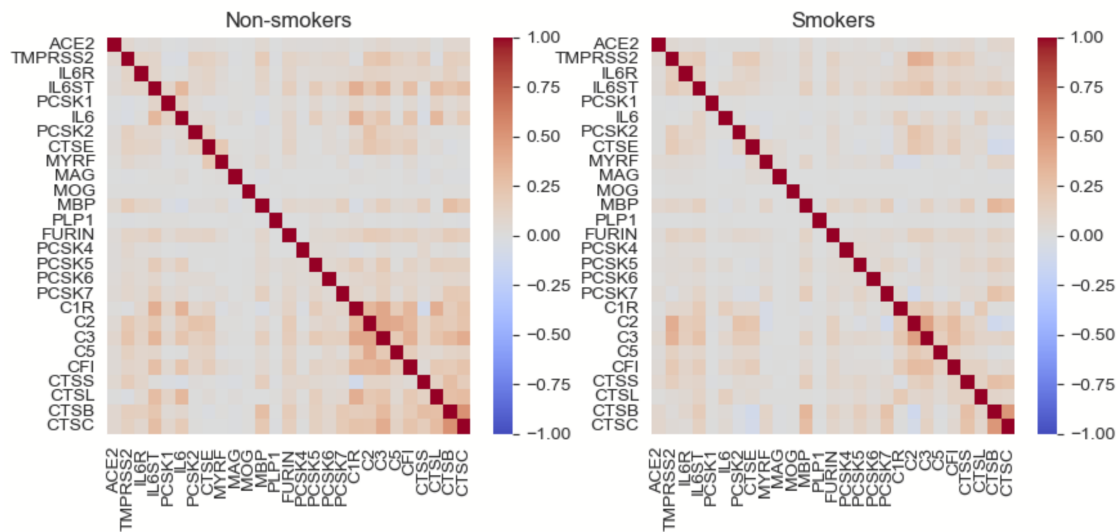


Fig. A1 Spearman Correlation from scRNA-seq data of every gene pair in the 27 most differentially expressed genes as selected by muus et al.

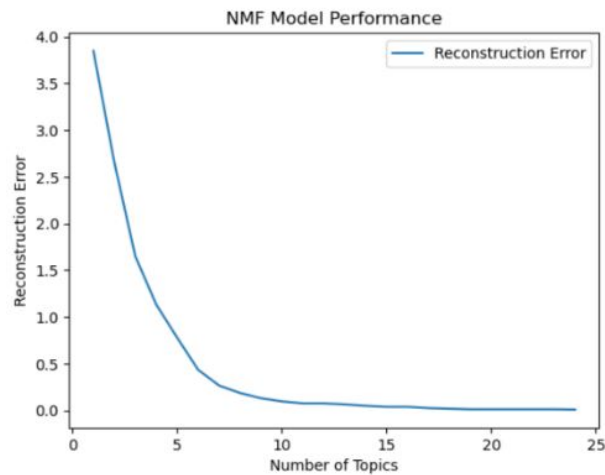


Fig. A2 NMF model performance gauged using the “elbow method” on the change in reconstruction error with number of topics

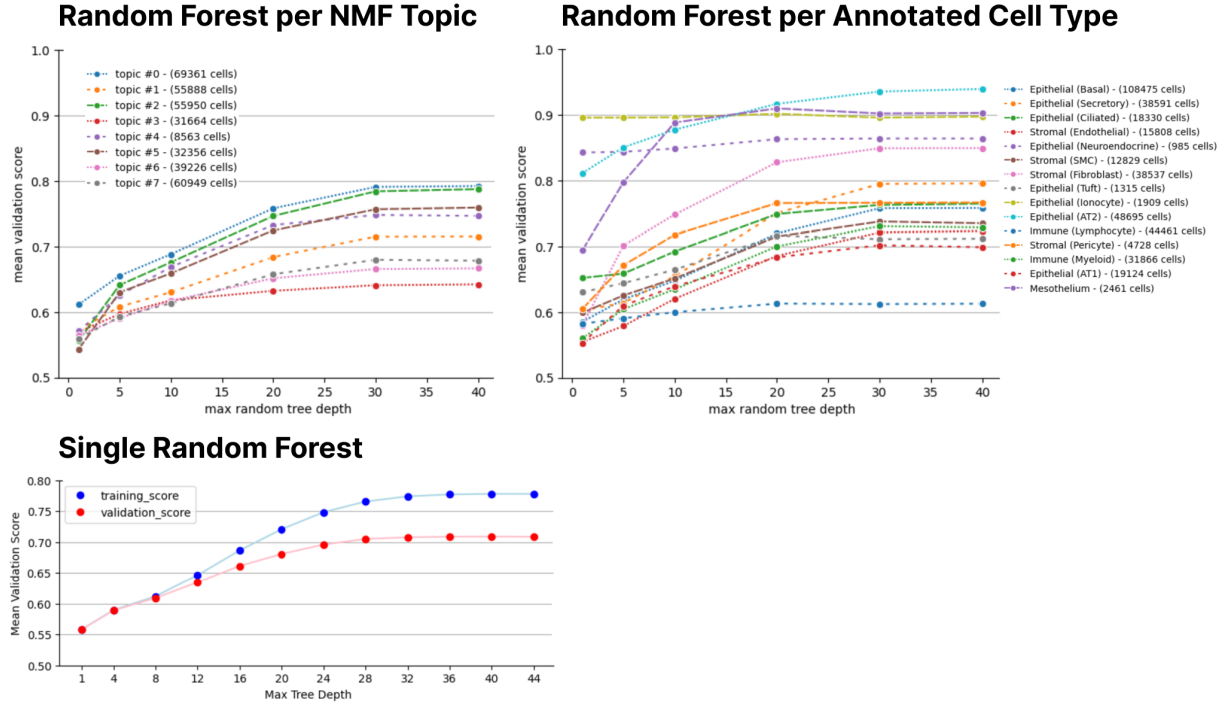


Fig. A3 Five-fold cross validation to decide the best maximum depth for random trees in the random forests of three models predicting whether cells are from smokers or non-smokers based on gene expression: **1.** A single random forest trained with 100 trees on all training data **2.** A cell type based random forest ensemble with 10 trees for each forest **3.** A NMF topic based random forest ensemble with 20 trees for each forest

Generalized Linear Model Regression Results									
Dep. Variable:	y	No. Observations:	44592						
Model:	GLM	DF Residuals:	44564						
Model Family:	Binomial	DF Model:	27						
Link Function:	Logit	Scale:	1.00000						
Method:	IRLS	Log-Likelihood:	-21315.						
Date:	Mon, 24 Apr 2023	Deviance:	42628.						
Time:	12:00:44	Pseudo R-squ. (CS):	0.3496						
No. Iterations:	20								
Covariance Type:	nonrobust								
coef	std err	z	P> z	[0.025	0.975]				
x1	-0.0218	0.013	-1.651	0.099	-0.048	0.004			
x2	0.2355	0.034	17.429	0.000	0.209	0.262			
x3	0.0858	0.013	6.527	0.000	0.060	0.112			
x4	0.1019	0.014	11.241	0.000	0.114	0.190			
x5	-0.0200	0.016	-1.943	0.055	-0.060	0.002			
x6	-0.0666	0.012	-5.704	0.000	-0.089	-0.044			
x7	-0.2093	0.013	-15.772	0.000	-0.235	-0.183			
x8	0.0458	0.016	2.925	0.003	0.015	0.077			
x9	-0.1208	0.012	-10.334	0.000	-0.144	-0.098			
x10	-0.0315	0.011	-2.805	0.005	-0.054	-0.009			
x11	-0.0002	0.013	-0.014	0.989	-0.026	0.025			
x12	0.0949	0.014	6.937	0.000	0.066	0.122			
x13	-0.1285	0.014	-9.001	0.000	-0.157	-0.100			
x14	-0.1113	0.014	-8.121	0.000	-0.139	-0.085			
x15	0.0200	0.012	1.672	0.095	-0.003	0.043			
x16	-0.0429	0.013	-3.275	0.001	-0.068	-0.017			
x17	-0.1871	0.016	-11.846	0.000	-0.218	-0.156			
x18	-0.1579	0.014	-11.134	0.000	-0.186	-0.130			
x19	-0.5000	0.021	-23.428	0.000	-0.543	-0.459			
x20	0.1551	0.016	9.572	0.000	0.123	0.187			
x21	-0.0739	0.017	-4.335	0.000	-0.107	-0.040			
x22	-0.4036	0.020	-20.998	0.000	-0.444	-0.363			
x23	0.1363	0.015	13.408	0.000	0.108	0.225			
x24	0.2957	0.016	18.835	0.000	0.265	0.327			
x25	-0.0126	0.016	-0.787	0.431	-0.044	0.019			
x26	-1.1883	0.027	-44.725	0.000	-1.240	-1.136			
x27	-1.6304	0.032	-51.467	0.000	-1.693	-1.568			
const	-0.5912	0.676	-0.875	0.382	-1.915	0.733			

Generalized Linear Model Regression Results									
Dep. Variable:	y	No. Observations:	40622						
Model:	GLM	DF Residuals:	40590						
Model Family:	Binomial	DF Model:	27						
Link Function:	Logit	Scale:	1.00000						
Method:	IRLS	Log-Likelihood:	-21231.						
Date:	Mon, 24 Apr 2023	Deviance:	42462.						
Time:	12:00:42	Pseudo R-squ. (CS):	0.2894						
No. Iterations:	24								
Covariance Type:	nonrobust								
coef	std err	z	P> z	[0.025	0.975]				
x1	-0.0140	0.012	-1.139	0.255	-0.038	0.010			
x2	0.1814	0.017	10.768	0.000	0.148	0.214			
x3	0.0043	0.014	14.142	0.000	0.178	0.233			
x4	-0.6351	0.031	-20.598	0.000	-0.696	-0.575			
x5	-0.2340	0.036	-6.535	0.000	-0.304	-0.164			
x6	0.3815	0.021	18.022	0.000	0.340	0.423			
x7	0.0066	0.011	0.578	0.563	-0.016	0.029			
x8	0.0273	0.013	2.092	0.036	0.002	0.053			
x9	0.0453	0.013	3.487	0.000	0.020	0.071			
x10	-1.8132	0.029	-62.537	0.000	-1.872	-1.754			
x11	0.0310	0.014	2.262	0.024	0.004	0.058			
x12	0.1095	0.016	7.013	0.000	0.079	0.140			
x13	-0.0914	0.024	-3.840	0.000	-0.138	-0.045			
x14	-0.1451	0.019	-7.632	0.000	-0.181	-0.109			
x15	0.0011	0.011	0.100	0.921	-0.021	0.023			
x16	-0.2718	0.029	-9.402	0.000	-0.328	-0.215			
x17	-0.0964	0.021	-4.635	0.000	-0.137	-0.056			
x18	0.2209	0.015	14.710	0.000	0.191	0.250			
x19	0.0804	0.025	26.972	0.000	0.031	0.130			
x20	-0.0418	0.019	-2.263	0.024	-0.082	0.006			
x21	-0.4512	0.022	-20.666	0.000	-0.494	-0.409			
x22	0.0565	0.014	4.046	0.000	0.029	0.084			
x23	-0.2087	0.022	-9.483	0.000	-0.252	-0.166			
x24	0.1429	0.019	7.521	0.000	0.106	0.180			
x25	-2.2456	0.064	-34.929	0.000	-2.372	-2.120			
x26	-0.1935	0.023	-8.387	0.000	-0.239	-0.148			
x27	-2.1576	0.091	-24.043	0.000	-2.377	-1.938			
const	-1.0931	0.483	-2.263	0.024	-1.988	0.812			

Generalized Linear Model Regression Results									
Dep. Variable:	y	No. Observations:	4892						
Model:	GLM	DF Residuals:	4864						
Model Family:	Binomial	DF Model:	27						
Link Function:	Logit	Scale:	1.00000						
Method:	IRLS	Log-Likelihood:	-2783.4						
Date:	Mon, 24 Apr 2023	Deviance:	5566.9						
Time:	12:00:46	Pseudo R-squ. (CS):	0.2199						
No. Iterations:	23								
Covariance Type:	nonrobust								
coef	std err	z	P> z	[0.025	0.975]				
x1	0.0207	0.034	0.777	0.437	-0.041	0.084			
x2	0.0120	0.041	0.290	0.772	-0.069	0.093			
x3	0.0653	0.050	1.319	0.187	-0.032	0.162			
x4	-0.4587	0.054	-8.520	0.000	-0.564	-0.353			
x5	-2.0549	0.071	-28.901	0.000	-2.195	-1.914			
x6	0.1083	0.043	2.544	0.011	0.025	0.192			
x7	0.0436	0.041	1.069	0.285	-0.036	0.123			
x8	0.1448	0.038	3.822	0.000	0.070	0.218			
x9	0.0953	0.042	2.275	0.023	0.013	0.177			
x10	-0.3127	0.030	-10.400	0.000	-0.371	-0.254			
x11	-0.0218	0.030	-0.729	0.466	-0.080	0.037			
x12	0.0005	0.051	0.010	0.992	-0.099	0.100			
x13	-0.3703	0.030	-12.350	0.000	-0.430	-0.310			
x14	0.0349	0.038	0.927	0.354	-0.039	0.109			
x15	-0.1506	0.049	-3.045	0.002	-0.247	-0.054			
x16	-0.1279	0.037	-3.440	0.001	-0.201	-0.055			
x17	-0.1916	0.049	-3.912	0.000	-0.289	-0.093			
x18	0.0539	0.078	0.691	0.488	-0.100	0.208			
x19	0.7228	0.062	11.632	0.000	0.601	0.845			
x20	-0.0315	0.053	-0.596	0.551	-0.135	0.072			
x21	0.0506	0.045	1.133	0.257	-0.037	0.138			
x22	-0.4205	0.077	-5.452	0.000	-0.572	-0.269			
x23	0.0289	0.037	0.789	0.430	-0.045	0.101			
x24	-0.0315	0.053	-0.596	0.551	-0.135	0.072			
x25	0.0715	0.058	1.243	0.214	-0.041	0.184			
x26	-1.0036	0.098	-10.201	0.000	-1.196	-0.811			
x27	0.1660	0.055	3.044	0.002	0.059	0.273			
const	-2.9886	0.212	-14.073	0.000	-3.405	-2.572			

Fig. A4 Statsmodel binomial generalized linear model for the top 3 models reported by sklearn. **1.** Left, the result for epithelial AT2 cells, x27 denotes CTSL, x26 denotes CTSC, x19 denotes C1R. **2.** Middle, the result for stromal fibroblast cells, x25 denotes CTSL, x27 denotes CTSC, x19 denotes C1R. **3.** Right, the result for stromal pericyte cells, x27 denotes CTSC, x18 denotes PCSK7, x25 denotes CTSL.

References

- Chen, H., Cowan, M.J., Hasday, J.D., Vogel, S.N., Medvedev, A.E. (2007, 11). Tobacco Smoking Inhibits Expression of Proinflammatory Cytokines and Activation of IL-1R-Associated Kinase, p38, and NF-B in Alveolar Macrophages Stimulated with TLR2 and TLR4 Agonists1. *The Journal of Immunology*, 179(9), 6097-6106, <https://doi.org/10.4049/jimmunol.179.9.6097> Retrieved from <https://doi.org/10.4049/jimmunol.179.9.6097> <https://journals.aai.org/jimmunol/article-pdf/179/9/6097/1255711/zim02107006097.pdf>
- Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., ... Eils, R. (2020, Aug 01). Covid-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nature Biotechnology*, 38(8), 970-979, <https://doi.org/10.1038/s41587-020-0602-4> Retrieved from <https://doi.org/10.1038/s41587-020-0602-4>
- Del Valle, D.M., Kim-Schulze, S., Huang, H.-H., Beckmann, N.D., Nirenberg, S., Wang, B., ... Gnajatic, S. (2020, Oct 01). An inflammatory cytokine signature predicts covid-19 severity and survival. *Nature Medicine*, 26(10), 1636-1643, <https://doi.org/10.1038/s41591-020-1051-9> Retrieved from <https://doi.org/10.1038/s41591-020-1051-9>
- Ding, X., Ye, N., Qiu, M., Guo, H., Li, J., Zhou, X., ... Li, J. (2022). Cathepsin b is a potential therapeutic target for coronavirus disease 2019 patients with lung adenocarcinoma. *Chemico-Biological Interactions*, 353, 109796, <https://doi.org/https://doi.org/10.1016/j.cbi.2022.109796> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0009279722000011>
- Gheware, A., Ray, A., Rana, D., Bajpai, P., Nambirajan, A., Arulselvi, S., ... Jain, D. (2022, Mar 08). Ace2 protein expression in lung tissues of severe covid-19 infection. *Scientific Reports*, 12(1), 4058, <https://doi.org/10.1038/s41598-022-07918-6> Retrieved from <https://doi.org/10.1038/s41598-022-07918-6>
- He, Y., Sun, J., Ding, X., Wang, Q. (2021, Feb.). Mechanisms in which smoking increases the risk of covid-19 infection: A narrative review. *Iranian Journal of Public Health*, 50(3), 431-437, <https://doi.org/10.18502/ijph.v50i3.5582> Retrieved from <https://ijph.tums.ac.ir/index.php/ijph/article/view/22376>
- Korkmaz, B., Caughey, G.H., Chapple, I., Gauthier, F., Hirschfeld, J., Jenne, D.E., ... Thakker, N.S. (2018). Therapeutic targeting of cathepsin c: from pathophysiology to treatment. *Pharmacology Therapeutics*, 190, 202-236, <https://doi.org/https://doi.org/10.1016/j.pharmthera.2018.05.011> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0163725818300913>
- Korkmaz, B., Lamort, A.-S., Domain, R., Beauvillain, C., Geldon, A., Önder Yildirim, A., ... Kettritz, R. (2021). Cathepsin c inhibition as a potential treatment strategy in cancer. *Biochemical Pharmacology*, 194, 114803, <https://doi.org/https://doi.org/10.1016/j.bcp.2021.114803> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0006295221004196>
- Leung, J.M., Yang, C.X., Tam, A., Shaipanich, T., Hackett, T.-L., Singhera, G.K., ... Sin, D.D. (2020). Ace-2 expression in the small airway epithelia of smokers and copd patients: implications for covid-19. *European Respiratory Journal*, 55(5), , <https://doi.org/10.1183/13993003.00688-2020> Retrieved from <https://erj.ersjournals.com/content/55/5/2000688> <https://erj.ersjournals.com/content/55/5/2000688.full.pdf>
- Muus, C., Luecken, M.D., Eraslan, G., Waghray, A., Heimberg, G., Sikkema, L., ... Network, T.H.C.A.L.B. (2020). Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of sars-cov-2 viral entry and highlights inflammatory programs in putative target cells. *bioRxiv*, , <https://doi.org/10.1101/2020.04.19.049254> Retrieved from <https://www.biorxiv.org/content/early/2020/04/21/2020.04.19.049254> <https://www.biorxiv.org/content/early/2020/04/21/2020.04.19.049254.full.pdf>

- Olajuyin, A.M., Zhang, X., Ji, H.-L. (2019, Feb 08). Alveolar type 2 progenitor cells for lung injury repair. *Cell Death Discovery*, 5(1), 63, <https://doi.org/10.1038/s41420-019-0147-9> Retrieved from <https://doi.org/10.1038/s41420-019-0147-9>
- Sim, R. (1981). [4] the human complement system serine proteases c1r and c1s and their proenzymes. *Proteolytic enzymes, part c* (Vol. 80, p. 26-42). Academic Press. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0076687981800067>
- Sudhan, D.R., & Siemann, D.W. (2015). Cathepsin l targeting in cancer treatment. *Pharmacology Therapeutics*, 155, 105-116, <https://doi.org/10.1016/j.pharmthera.2015.08.007> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0163725815001655>
- Thorley, A.J., & Tetley, T.D. (2007). Pulmonary epithelium, cigarette smoke, and chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*, 2(4), 409-428,
- Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., ... Ponten, F. (2010, Dec 01). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, 28(12), 1248-1250, <https://doi.org/10.1038/nbt1210-1248> Retrieved from <https://doi.org/10.1038/nbt1210-1248>
- Wang, X., Yang, G., Wang, Q., Zhao, Y., Ding, K., Ji, C., ... Li, S. (2022, Jun 18). C1r, ccl2, and tnfrsf1a genes in coronavirus disease-covid-19 pathway serve as novel molecular biomarkers of gbm prognosis and immune infiltration. *Disease Markers*, 2022, 8602068, <https://doi.org/10.1155/2022/8602068> Retrieved from <https://doi.org/10.1155/2022/8602068>
- Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., ... Illig, T. (2013, 05). Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PLOS ONE*, 8(5), 1-14, <https://doi.org/10.1371/journal.pone.0063812> Retrieved from <https://doi.org/10.1371/journal.pone.0063812>