Early detection of lung cancer using RNA transcriptomes derived from nasal airway epithelial cells (GitHub: bit.ly/2BdiDZZ)

Sean Friedowitz (sfriedo) and Kevin J Hou (kjhou)

(Dated: December 13, 2018)

We have developed a classifier which uses a combination of clinical data and the nasal cell RNA transcriptome — the expression levels of genes in nasal cells — to predict if a patient will develop lung cancer. Such classifiers allow for the early detection of lung cancer using non-invasive nasal swabs, which has enormous potential to improve detection and treatment. We trained the classifier on a labeled dataset produced by the AEGIS clinical trials,[1] which provides the expression levels of $\approx$ 32,000 genes for 484 patients with suspect lung cancer, 298 of which were diagnosed with cancer within a few months of the study. T-statistic based filter feature selection and PCA have been applied to reduce the large dimensionality of this feature set. Using reduced features, we have constructed a logistic regression classifier which achieves 88% test set recall and 59% specificity, which is comparable to current work in the field.[2,3]

I. INTRODUCTION AND RELATED WORK

Lung cancer is the second most common type of cancer, with roughly half a million new cases being diagnosed every year in the United States.[4] Patient age and smoking status are the greatest contributing risk factors to developing lung cancer, with smoking status posing a significantly greater risk to developing the disease.[4] Although the success rate for curing early stage lung cancer is remarkably high, lung cancer is still by far the leading cause of cancer-related death in adults.[4] Indeed, cases where the disease is diagnosed before advancing to a non-curable stage yield very favorable prognoses, placing a high demand for diagnostic tools capable of routinely and accurately detecting the presence of lung cancer in patients, as well as risk factors for cancer development.[5]

Genome screening methods show great promise for the early detection of lung cancer, as well as a variety of other cancers. These methods analyze the genetic material within target cells — either DNA or transcribed mRNA — and use the information to characterize physiological state and indicate disease pathology. In the last decade, these methods have been revolutionized by the advent of next-generation RNA sequencing techniques, which have enabled the collection of vast amounts of genetic sequence data.[6] These massively parallel sequencing techniques have made it possible to obtain gene expression profiles which characterize cellular states in relation to various diseases, genetic mutations, environmental conditions, etc. This data is typically reported as an "RNA transcriptome", a collection of expression vectors for cell samples in which each vector entry is the expression level of a particular gene.

There have been a number of medical studies investigating the use of genetic data as a biomarker for the presence of cancer. One of the earliest studies gathered gene expression data from non-cancerous, bronchial endothelial cells taken from 129 chronic smokers at a single medical center, and followed these patients through until final cancer diagnosis.[7] This study identified an 80-gene biomarker corresponding to the most over- and under-expressed genes between cancerous and non-cancerous patients. This 80-gene biomarker (i.e. a subset of the feature space) was integrated into a simple weighted voting algorithm[7,8] to predict the presence of cancer. The authors were able to achieve a remarkable 83% classification sensitivity with this approach, but did not explore more sophisticated classification models.

Recently, the AEGIS (Airway Epithelium Gene Expression in the Diagnosis of Lung Cancer) clinical trials have substantially expanded the amount of RNA transcriptome data available for lung cancer classification. In these clinical trials, gene expression data was collected from hundreds of at-risk smokers across 28 medical centers in the US and Canada.[1] With this expanded set of patient data, researchers have developed classifiers based on RNA transcriptomes from both bronchial[3] and nasal[2] cells. However, much like previous work, both of these studies rely on a simplistic weighted-voting classification algorithm. Although these studies have been able to achieve a high true positive classification rate, neither model was able to achieve high specificity, i.e. neither model achieved definitive classification of patients without risk of developing cancer.[1–3]

The expansive datasets obtained through the AEGIS study are freely available online through the NCBI Gene Expression Omnibus (GEO).[9] We have obtained the AEGIS dataset derived from nasal epithelial cell samples taken from 484 chronic smokers.[10] This dataset contains expression levels for roughly 33,000 genes for each of the patients, as well as auxiliary clinical factors relevant to their cancer diagnosis. Nasal cell samples were collected prior to patients' final diagnosis. This diagnosis serves as the 'label' for the presence of cancer in each patient.

In this work, we seek to build a predictive classifier for the presence of lung cancer. As input, our predictor uses the RNA transcriptome (i.e. expression levels of all genes) derived from the nasal epithelial cells of a patient suspected to have lung cancer. We then use filter feature selection to select a subset of features (i.e. specific

TABLE I. Clinical data for patients in the AEGIS study.

CHARACTERISTIC	NO CANCER (N = 186)	CANCER (N = 298)
SEX – NO.		
MALE	119	186
FEMALE	67	112
AGE – MEDIAN	57.6	61.3
SMOKING STATUS – NO.		
FORMER	117	189
CURRENT	69	109
TIME SINCE QUITTING – NO.		
> 15 YEARS	75	101
< 15 YEARS	111	197
YEARS TOBACCO USE – MEDIAN	25.0	37.5
LESION SIZE – NO.		
ILL DEFINED	44	16
< 3 CM	99	96
> 3 CM	43	186

genes) for use in classification, and apply PCA to further reduce the dimensionality of the selected features. The reduced feature set is then used in a logistic regression for classification. As output, our predictor returns either a positive (1) or negative (0) label indicating the presence of cancer.

The remainder of this report is organized as follows. In Section II, we describe the training dataset, and detail our methodology for feature selection. In Section III, we describe our use of PCA to further reduce the dimensionality of the selected features, and detail the logistic regression classifier that we have implemented. In Section IV, we present the results of our classifier trained on the AEGIS dataset, and compare our results to the best reported in the literature. Finally, in Section V, we discuss areas of improvement in our classification scheme, and provide potential next steps for this project.

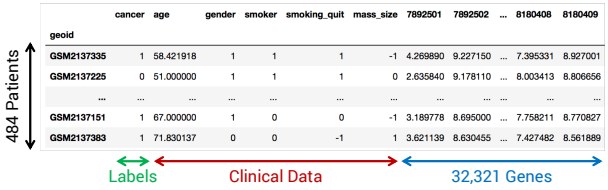
II. DATASET AND FEATURES

We have trained our classifier on the AEGIS nasal epithelial cell dataset. This data has been made publicly available on the NCBI Gene Expression Omnibus as Series GSE80796.[10] The dataset consists of clinical data and RNA transcriptomes from 484 chronic smokers suspected of having lung cancer due to the observation of lesion(s) in their lung tissue.

Of the 484 patients, 298 were ultimately diagnosed with cancer. The clinical data consisted of sex, age, smoking status (current/former), time since quitting smoking, years of tobacco usage, and lesion size. Clinical information is summarized in Table 1.

Each RNA transcriptome was taken from a sample of nasal epithelial cells prior to the official diagnosis of cancer. For each patient, the transcriptome consists of an array of numerical expression levels for approximately 32,000 RNA transcripts in the cell sample, measured using an Affymetrix Gene 1.0 ST Microarray.[11]

TABLE II. Pre-processed data to be used in classifier.



geoid	cancer	age	gender	smoker	smoking_quit	mass_size	7892501	7892502	...	8180408	8180409
GSM2137335	1	58.421918	1	1	1	-1	4.269890	9.227150	...	7.395331	8.927001
GSM2137225	0	51.000000	1	1	1	0	2.635840	9.178110	...	8.003413	8.806656
...	...	...	...	...	...	...	...	...	...	...	...
GSM2137151	1	67.000000	1	0	0	-1	3.189778	8.695000	...	7.758211	8.770827
GSM2137383	1	71.830137	0	0	-1	1	3.621139	8.630455	...	7.427482	8.561889

We pre-processed this data by combining clinical data with RNA transcriptome data. For each non-numerical clinical feature, we assigned a numerical value to each possible value (e.g. former smoker = 0, current smoker = 1). Each RNA transcriptome was converted into a n-dimensional feature vector of expression levels ($n = 32,321$). The clinical and transcriptome feature vectors were then combined to form our working clinico-genetic data matrix. The pre-processed data is visually depicted in Table II.

After pre-processing, the dataset was split 90%-10% into a training set (438 samples) and a test set (49 samples), with the large training split due to the small number of overall samples. This dataset is very high-dimensional ($n \approx 32{,}321$) with only a small number of examples ($m = 484$). As such, there is an extreme danger of overfitting. Importantly, we note that only a small number of features (gene expression levels) are expected to be indicative of the presence of lung cancer. Many genes encode vital biological functions which have consistent expression levels regardless of cell type, and are not useful for distinguishing between cancerous and non-cancer prone cells. Thus, we select a subset of the RNA transcriptome for use in classification.

The goal of our feature selection methods was to identify genes whose expression levels correlated with the two class distinctions. In this work, we tested three feature selection techniques: filtering based on a variant of a 'T-score',[2,8] filtering based on an ANOVA f-statistic, and feature selection using an ensemble of decision trees. Ultimately, we used the filter 'T-score' method for our classifier as it provided the best performance.

Filter feature selection was a natural starting point for this dataset. We found in our progress report that the computational cost of forward feature search was prohibitive due to the large number of features. Thus, we drew upon previous work in the field and used a T-score[2,8] to rank features. For each feature, we compute an independent-samples t-statistic using the positive (cancer) and negative (no cancer) populations as our two groups. For feature x_i, the T-score is,

$$T(x_i) = \frac{|\mu_{0,i} - \mu_{1,i}|}{s_{0,i} + s_{1,i}}.$$

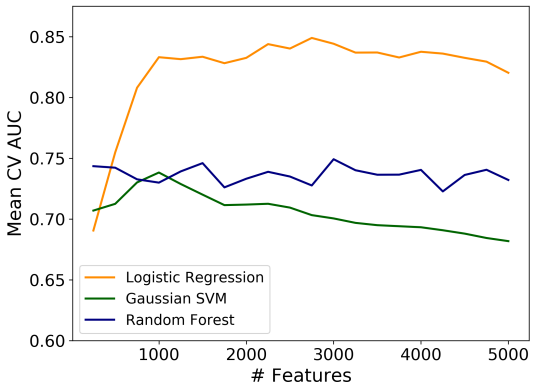


FIG. 1. Mean AUC obtained from 10-fold cross-validation on the training set for three different models, where features were ranked according to 'T-score' values.

In the preceding equation, $\mu_{0,i}$ and $s_{0,i}$ are the mean and standard deviation of feature x_i among the negative population; $\mu_{1,i}$ and $s_{1,i}$ are the same for the positive population. The T-statistic defined above is the established standard for transcriptome classification in the medical field.[8] Intuitively, a high T-score for feature x_i indicates that it is differentially expressed among the two populations. Feature selection was performed by selecting the top k features with the highest T-score.

We also attempted feature selection using an ensemble of decision trees. In this approach, we fit an ensemble of decision trees on different random sub-samples of the dataset, and use the average prediction value of the ensemble to select features. This was done using `sklearn`'s built-in `ExtraTreesClassifier` using the Gini loss for each decision step, which provides rankings of "feature importance" normalized for the ensemble of trees. A set number of k features can then be selected from the fit model based on ranked importance.

To identify features using these approaches, we fit logistic regression, gaussian SVM, and random forest models using the highest ranked k features from the dataset as input. In this selection step, we selected k to optimize the mean, cross-validated area-under-curve (AUC for sensitivity vs. specificity) for each of the classifiers on the training set. We found, using 10-fold cross-validation, that $k \approx 3500$ resulted in the best training AUC for 'T-score' selection with a logistic regression model (Fig. 1). SVM and random forest classifiers achieved substantially lower mean AUC values during cross-validation across the entire range of k, with the SVM showing a steadily decreasing AUC as k was increased.

Among the three feature-selection methods we attempted, 'T-score' feature selection coupled with logistic regression produced the best mean AUC. Ranking features according to the ANOVA f-statistic resulted in similar model performance as the 'T-score' method, but ultimately the 'T-score' method produced better results. We were unable to obtain good results using the decision-tree based feature selector. This is likely a result of the randomness inherent in this ensemble method; a large number of random trees are need to sample the large number of features in our dataset. We found that using large ensembles reintroduced the computational burden we sought to avoid in filter feature selection. Thus, we use the T-score selection technique with $k \approx 3500$ as input for our final clasifier.

TABLE III. Functions of notable genes from feature selection.

PROBE ID	GENE	FUNCTION
8091385	CP	Related to iron transport across cell membrane, plays role in lung development
7978123	PSME2	Proteosome subunit, related to cell cycle, mitotic, and RET signaling pathways
8146092	IDO1	Catalyzes tryptophan metabolism, acts as a suppressor for anti-tumor immunity
8115147	CD74	CD74 molecule, chaperones antigen presentation in immune response

We define our initial predictor as the set of approximately 3500 features obtained from the feature selection procedure detailed above. This predictor contains only one piece of clinical data, namely the lesion size. All other features correspond to gene expression levels in the transcriptome. Broadly speaking, most of the genes in our predictor are linked to various cellular metabolic pathways, and many play vital roles in immune signaling and response. For example, many of the genes encode proteins which play active roles in the major histocompatibility complex, which is a series of surface proteins present on cellular membranes that interact with immune cells and antigen presentation. Some genes also encode transcription factors and non-coding RNA (ncRNA), which control the rate of gene expression. Notable genes are highlighted in Table III.

III. METHODS

To derive our final classifier, we begin with the top $k = 3500$ features derived from the 'T-score' ranking and cross-validation trials described in Section II. We use a logistic regression as our final classification algorithm, as it performed the best out of the three models (logistic, SVM, random forest) tested in feature selection. Although feature selection reduces the size of the feature space greatly, we are still left with 3500 features, roughly seven times the number of samples in our dataset. To avoid overfitting and to enable our classifier to generalize, we have utilized PCA to further reduce the dimension of the selected features.

We conducted PCA using `sklearn`, which has built-in functions for normalizing the data (i.e. setting the mean and variance of each feature to zero and one). We

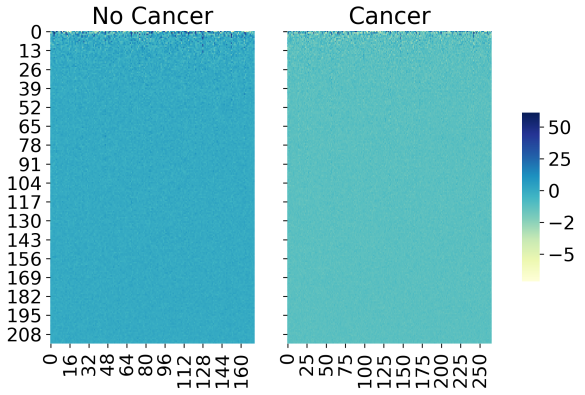


FIG. 2. PCA heatmap using 215 PCs. There is obvious contrast between the heatmaps for different classes. Horizontal columns represent each patient sample, while vertical colmns represent the 215 PCs.

found that 90% of the variance in the data was captured within the first 215 principal components. This is a far more reasonable feature space for our roughly 400-sample dataset. Visualization of the 215-dimensional, PCA-reduced dataset in Fig. 2 shows a clear distinction between the positive and negative labels, indicating that our selected features are good indicators for cancer presence. For the rest of this paper, we use this set of 215 PCA-reduced, selected features for classification.

As discussed previously, these 215 components were then used as input for our final logistic regression classifier. We have attempted to optimize the logistic regression classifier with the PCA-reduced dataset. Due to the relatively small sample size, even a simple logistic regression can be prone to overfitting. Additionally, there is a slight class imbalance in the AEGIS dataset ($n_0 = 186$, $n_1 = 298$), making our model more prone to generalization error. To improve on these points, we utilized `sklearn`'s built-in method `GridSearchCV`, which performs cross-validation across a range of hyperparameters, to find the parameter set that optimizes a given cross-validation metric. Using this method, we performed 10-fold cross-validation using the AUC of sensitivity vs. specificity as the scoring metric, and determined that the optimal parameters are an inverse regularization constant of $C = 0.05$, and balance class weights for the positive and negative samples.

IV. RESULTS AND DISCUSSION

We found that the optimal classifier derived from our feature selection, PCA, and parameter tuning cross-validation tests was a standard logistic regression model. This clinical-genomic model uses a mix of 3500 clinical and genomic features from the original dataset, reduced to a set of 215 principal components which capture 90% of the data variance. We believe that logistic regression outperformed the other tested models — Gaussian SVM and random forest — due to its simplicity. Given the small number of samples and large feature space in the AEGIS dataset, more expressive models are prone to overfitting. These models have higher variance and do not generalize well to unseen data. Although the high variance in these models could potentially be reduced with intricate and exhaustive parameter turning, we opted to use the more simple logistic regression model as our final classifier.

TABLE IV. Test set results for the final clinical vs. clinical-genomic logistic regression classifiers.

	Clinical	Clinical-Genomic
Accuracy	65.3%	77.6%
Precision	71.4%	80.0%
Recall	78.1%	87.5%
Specificity	41.2%	58.8%

To better assess the results of the final classifier, we have also developed a simple "clinical" logistic regression model, which takes as input only clinical and none of the genomic features from the dataset (e.g. smoking status, lesion size, age). Using this "clinical" classifier as a reference allows us to clearly understand the additional information provided by the RNA transcriptome. Specifically, this comparison enables us to quantify the extent to which RNA transcriptome data enhances a clinical diagnosis. Final results for our clinical-genomic model are reported in Table IV, along with a comparison to the bare clinical model.

Within the context of lung cancer identification, the test set **recall** is by far the most important metric. The recall, or true positive rate, is the percentage of patients with cancer which were accurately classified as having cancer. The recall is optimized in opposition to the specificity, which is the fraction of patients the classifier accurately predicts do not have cancer. From the standpoint of medical diagnosis, some amount of false positives are acceptable. Diagnostic techniques like the classifier we have developed typically serve as screening or early-detection tools, which may motivate a doctor to conduct a more accurate, more invasive procedure.

The results in Table IV indicate that the addition of genomic data enriches clinical diagnosis significantly. In addition to improvements in overall accuracy and precision, test set recall is improved to 87.5%, meaning the classifier is able to correctly detect lung cancer in nearly 90% of affected patients. These results are comparable to the current best work in the field; the most recently published study with the AEGIS nasal dataset reports a recall of 91% and a specificity of 52%.[2]

The relative performance between the clinical and clinical-genomic models can be seen in the receiver oper-

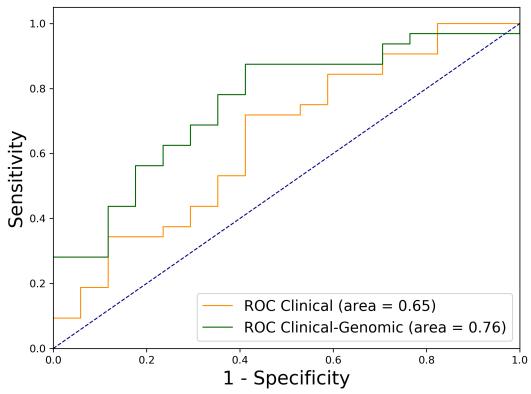


FIG. 3. Test set ROC curves for clinical and clinical-genomic logistic regressions after model optimization. The clinical-genomic model was trained on the final 215 PCs derived from 3500 selected features.

ating characteristic (ROC) curves shown in Fig. 3. This curve highlights the diagnostic capabilities of each classifier by plotting the sensitivity (true positive rate) versus the specificity (true negative rate), at various threshold levels for distinguishing between each class. An ideal classifier would achieve 100% sensitivity and specificity at the same time, and result in a vertical line coming out of the origin on the x-axis, while a random guess would yield a diagonal line, as shown in the figure. Classifier performance is quantified by the area under the curve (AUC) of this plot, with an AUC of unity correpsonding to the ideal case. Clearly, including the genomic features within our classifier increases the AUC of our model from 0.65 to 0.76, which again falls within the error bounds of the best classifiers reported in literature.[2]

Most work in this field typically leverages a large amount of domain knowledge. For example, many groups utilize internal databases of genes implicated in various cancers.[7] This allows for additional 'weighting' of genes during feature selection. We have found in this report that with careful feature selection and model tuning, comparable results can be attained without the incorporation of domain knowledge. Although certainly not a new idea in the field, this result also indicates that exploratory studies of these datasets have the potential to select genes which are directly linked to the onset of various cancers. Notable genes obtained from feature selection can then be targeted for further study, to investigate their connection to specific biochemical processes that lead to the onset and propagation of cancer.

V. CONCLUSION/FUTURE WORK

Lung cancer is one of the leading causes of cancer-related deaths across the world, largely owing to the difficulty of early detection. Since traditional detection methods often rely on invasive procedures such as bronchoscopy, there is a need to develop non-invasive screening methods, particularly for populations susceptible to the disease (e.g. smokers). In this work, we have developed a classification model which uses RNA transcriptome data derived from nasal endothelial cells to predict the presence of cancer among smokers suspected to have the disease.

Our training dataset consists of samples from 484 patients in the AEGIS clinical trials, where 298 were ultimately diagnosed with lung cancer. Features in the dataset consist of expression levels for roughly 32,000 genes (the RNA transcriptome), as well as clinical factors such as age, smoking status, and size of lesions in lung tissue.

Owing to the large dimensionality of this data and the relatively small number of samples (patients), we sought to filter out those genes which were most relevant to cancer prediction. Using filter based feature selection based on a t-statistic between the two class means, we identified 3500 features that yielded the best performance with a logistic regression classifier during cross-validation testing on the training set. These features included the clinical factor of lung lesion size, as well as an assortment of genes, many of which encoded biological functions related to metabolic activity and immune signaling. PCA was then applied to reduce the dimensionality of this feature set to a more reasonable 215 PCs, which captured 90% of the data variance.

A final logistic regression classifier was trained on this PCA-reduced set of selected features. Logistic regression outperformed more sophisticated SVM and random forest algorithms in cross-validation, likely due to the simplicity of the model. Due to the small sample size and high dimensionality of this data, more intricate non-linear models appeared susceptible to overfitting and did not generalize well during cross-validation, or to the test set. Our best logistic regression classifier achieved 88% sensitivity and 58% specificity on the test set, which is comparable to the results reported in the literature,[2] and is a significant improvement over the 78% sensitivity and 41% specificity obtained for a classifier trained only on clinical features available in the dataset.

A natural next step for this line of work would be to spend time carefully optimizing more expressive classifiers than logistic regression, such as a random-forest or SVM model. As identified previously, great care is required to prevent more expressive models from overfitting limited training data. However, we expect that even better performance is possible with these models. Additionally, the AEGIS clinical trials have produced data for other cell samples, such as from bronchial (lung) cells. A potential second future work would be to apply our approach to different cell samples, and to compare selected features and final performance across these datasets.

VI. CONTRIBUTIONS

Sean Friedowitz conducted data-preprocessing and wrote the feature and model selection code. Kevin Hou and Sean Friedowitz both worked on model design, optimization, and final analysis. Kevin Hou wrote code to identify specific genes from selected features. Both team members contributed equally to literature review, creating the final poster, and writing this report.

VII. REFERENCES

[1] G. A. Silvestri, A. Vachani, D. Whitney, M. Elashoff, K. P. Smith, J. S. Ferguson, E. Parsons, N. Mitra, J. Brody, M. E. Lenburg, and A. Spira, "A bronchial genomic classifier for the diagnostic evaluation of lung cancer," New England Journal of Medicine **373**, 243–251 (2015).

[2] J. F. Perez-Rogers, J. Gerrein, C. Anderlind, G. Liu, S. Zhang, Y. Alekseyev, K. P. Smith, D. Whitney, W. E. Johnson, D. A. Elashoff, S. M. Dubinett, J. Brody, A. Spira, and M. E. L. and, "Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection," JNCI: Journal of the National Cancer Institute **109** (2017), 10.1093/jnci/djw327.

[3] D. H. Whitney, M. R. Elashoff, K. Porta-Smith, A. C. Gower, A. Vachani, J. S. Ferguson, G. A. Silvestri, J. S. Brody, M. E. Lenburg, and A. Spira, "Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy," BMC Medical Genomics **8** (2015), 10.1186/s12920-015-0091-3.

[4] American Cancer Society, "Key statistics for lung cancer," `https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html` (), accessed: 2018-11-18.

[5] American Cancer Society, "Lung cancer prevention and early detection," `https://www.cancer.org/cancer/lung-cancer/prevention-and-early-detection.html` (), accessed: 2018-11-18.

[6] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: a revolutionary tool for transcriptomics," Nature Reviews Genetics **10**, 57–63 (2009).

[7] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. E. Lenburg, and J. S. Brody, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," Nature Medicine **13**, 361–366 (2007).

[8] T. R. Golub, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science **286**, 531–537 (1999).

[9] National Center for Biotechnology Information, "NCBI Gene Expression Omnibus (GEO)," `https://www.ncbi.nlm.nih.gov/gds` (), accessed: 2018-11-18.

[10] National Center for Biotechnology Information, "Gene expression profiling of nasal epithelial cells in current and former smokers with and without lung cancer," `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80796` (), accessed: 2018-11-26.

[11] Affymetrix, "Affymetrix human gene probeset," `http://www.affymetrix.com/Auth/analysis/downloads/na36/wtgene/HuGene-1_1-st-v1.na36.hg19.probeset.csv.zip` (2018), accessed: 2018-12-06.