Pineda et al. BMC Cancer (2016) 16:184
DOI 10.1186/s12885-016-2223-3

BMC Cancer

RESEARCH ARTICLE Open Access

On Predicting lung cancer subtypes using 'omic' data from tumor and tumor-adjacent histologically-normal tissue

Arturo López Pineda[1*], Henry Ato Ogoe[1], Jeya Balaji Balasubramanian[1], Claudia Rangel Escareño[2], Shyam Visweswaran[1], James Gordon Herman[3] and Vanathi Gopalakrishnan[1]

Abstract

Background: Adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the most prevalent histological types among lung cancers. Distinguishing between these subtypes is critically important because they have different implications for prognosis and treatment. Normally, histopathological analyses are used to distinguish between the two, where the tissue samples are collected based on small endoscopic samples or needle aspirations. However, the lack of cell architecture in these small tissue samples hampers the process of distinguishing between the two subtypes. Molecular profiling can also be used to discriminate between the two lung cancer subtypes, on condition that the biopsy is composed of at least 50 % of tumor cells. However, for some cases, the tissue composition of a biopsy might be a mix of tumor and tumor-adjacent histologically normal tissue (TAHN). When this happens, a new biopsy is required, with associated cost, risks and discomfort to the patient. To avoid this problem, we hypothesize that a computational method can distinguish between lung cancer subtypes given tumor and TAHN tissue.

Methods: Using publicly available datasets for gene expression and DNA methylation, we applied four classification tasks, depending on the possible combinations of tumor and TAHN tissue. First, we used a feature selector (ReliefF/Limma) to select relevant variables, which were then used to build a simple naïve Bayes classification model. Then, we evaluated the classification performance of our models by measuring the area under the receiver operating characteristic curve (AUC). Finally, we analyzed the relevance of the selected genes using hierarchical clustering and IPA® software for gene functional analysis.

Results: All Bayesian models achieved high classification performance (AUC > 0.94), which were confirmed by hierarchical cluster analysis. From the genes selected, 25 (93 %) were found to be related to cancer (19 were associated with ADC or SCC), confirming the biological relevance of our method.

Conclusions: The results from this study confirm that computational methods using tumor and TAHN tissue can serve as a prognostic tool for lung cancer subtype classification. Our study complements results from other studies where TAHN tissue has been used as prognostic tool for prostate cancer. The clinical implications of this finding could greatly benefit lung cancer patients.

Keywords: Bayes Theorem, Adenocarcinoma of Lung, Squamous Cell Carcinoma, DNA Methylation

* Correspondence: arl68@pitt.edu
[1]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Boulevard, 15206 Pittsburgh, PA, USA
Full list of author information is available at the end of the article

Background

Lung cancer is the leading cause of human cancer death in both sexes in the United States. In 2014, there was an estimate of 224,210 new cases, while 159,260 patients were estimated to have died from the disease [1]. Cigarette smoking is the main risk factor for the development of lung cancer [2]. While smoking has been proven to have a high correlation with epigenetic changes in the DNA [3], other behavioral and environmental factors might also be recorded by changes in the epigenetics of the DNA (i.e. passive smoking, air pollution, occupational exposure, alcohol consumption, poor diet, low physical activity).

Adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the most common histological subtypes among all lung cancers. Both of them are a form of cancer that develops in the epithelial cells (carcinoma), and belong to the category of non-small cell lung cancer. Lung ADC develops in the glands that secrete products into the bloodstream or some other cavity in the body –the mucus secreting glands in the lungs. Most lung ADC arise in the outer, or peripheral, areas of the lung [4]. In contrast, lung SCC develops in flat surface covering cells. Squamous cells allow trans-membrane movement, like filtration and diffusion, for example the exchange of air in the alveoli of lungs. Squamous cells can also serve as boundary and protection of various organs. Most lung squamous cell cancers frequently arise in the central chest area in the bronchi [5].

The diagnosis of early stage lung cancer involves the use of imaging techniques, followed by a biopsy for pathology analysis [6]. Initially, screening of lung cancer is done using chest x-ray, or low-dose computed tomography. The American Cancer Society recommends screening to patients between the ages of 55–74 years old who are smokers or who quit smoking within the past 15 years [7]. Imaging techniques are not foolproof, so further analyses are usually required to make final diagnostic decisions. For instance, a cytological analysis is still required to confirm the imaging analysis [8]. In addition, tissue samples, albeit small, are often obtained during a needle aspiration biopsy or a bronchoscopy biopsy. The lack of tissue architecture in these small tissue specimens limits the pathologic analysis under a microscope [9].

Several studies have shown that molecular profiling of lung carcinoma is a viable tool for disease diagnosis [10], and prognosis [11]. What is more, distinguishing between ADC and SCC has significant clinical implications – both can have different treatment regimens. In this era of precision medicine, molecular characterization can be crucially important in the selection of an effective drug regimen. Potentially, patients can be subjected to drug regimens that are beneficial and/or harmful. Four possibilities summarize this situation: when a drug 1) has both therapeutic and adverse effects, 2) has only therapeutic effects (no adverse effects), 3) has adverse but no therapeutic effects, and 4) has no adverse nor therapeutic effects. Treatment safety and efficacy outcomes are important reasons of concern and the main reason for tumor subtyping [12]. Furthermore, ADC and SCC have distinct progression rate and progression free survival, which determines the selection of treatment [13].

The molecular mechanisms of ADC and SCC are considerably different. The standard molecular testing for lung cancer is to check for mutations of two molecules: epidermal growth factor receptor (EGFR) and rearrangement of anaplastic lymphoma kinase (ALK). Each protein has mutations that lead to the development of lung cancer. However, EGFR is found to be mutated only in around 10 % of tumors [14]. Similarly, ALK mutation occurs only in 6 % of tumors [15]. Although some drugs target EGFR and ALK positive tumors with therapeutic benefits for the patient, 75 % of lung tumors do not possess these molecular alterations [16]. The high sensitivity and low specificity of these diagnostic molecules is a motivation to research into new diagnostic models.

DNA methylation is an emerging diagnostic technology to measure the epigenetic changes in the DNA, characterized by the addition of a methyl group in regions of the DNA known by having CpG islands. Traditionally, gene expression has been used as a prognostic biomarker for lung carcinoma, and differentially expressed genes between lung cancer subtypes have been found [17]. However, it has been suggested that DNA methylation signatures of cancer should also be considered as a potential diagnostic biomarker of the disease [18]. Distinct DNA methylation signatures exist between ADC and SCC [19], and also between tumor tissue and normal surrounding tissue [20]. Since DNA methylation plays a significant role in the regulation of gene expression [21], there is an added value of investigating both data types.

Computational modeling methods, such as Bayesian classifiers, have been used successfully to model the complexity of genomic data. A study by Chang and Ramoni [22], yielded very high classification performance (accuracy = 0.95) to distinguish between lung tumor ADC and lung tumor SCC. Despite these results, the study still has open questions that are significant for the cause of precision medicine. For instance, selecting appropriate tissue samples to maximize microarray analysis is a big challenge. Inadequate biopsies can cause misdiagnosis and delay appropriate treatment [23]. In some cases, the amount of tissue available in the biopsy might not be enough to make a diagnosis from pathology and characterize the DNA changes in the cancer cells.

A major challenge of our study is the lack of tissue availability in public datasets. Typically, a biopsy tissue represents a very small portion of the lung. In spite of ultrasound guidance, it is easy to miss a small focal

malignancy, and end up retrieving tumor-adjacent histologically-normal tissue (TAHN) along with Tumor tissue. In those cases, the biopsy is discarded if it cannot retrieve more than 50 % of tumor tissue [9]. The patient would have to undergo a new procedure to obtain another biopsy. Thus, it is worth exploring computational alternatives for classifying lung cancer subtypes given a small biopsy sample and a mix of TAHN and tumor tissue.

Our goal in this work was to test whether computational modeling can be a viable approach to accurately differentiate between lung cancer subtypes, given molecular profiles of tumor tissue using DNA methylation data. Specifically, we tested the hypothesis that "Bayesian modeling is sufficient to classify lung cancer subtypes, regardless of the tissue sample being tumor or tumor-adjacent." In this paper, we evaluated the ability of a Bayesian classifier to accurately differentiate lung cancer subtypes using real lung cancer molecular profiling data sets that are also publicly available.

Methods

Datasets

To test our hypothesis, we extracted datasets containing gene expression and DNA methylation beta values from the Cancer Genome Atlas (TCGA) data portal for lung adenocarcinoma (LUAD [24]) and lung squamous cell carcinoma (LUSC, [25]). Additionally, we also used the gene expression dataset of lung adenocarcinoma patients, described by Landi et al. [26], GEO accession number GDS3257. Table 1 describes the characteristics of the samples we used for this study. For each dataset, it provides information on the type of 'omic' data type, source of data, assay platform, including number of features (i.e. genes or DNA methylation sites), and the number of sample distribution – that is, tumor tissue (T and TAHN) – within each subtype, where available. The formatted TCGA dataset used in this study, along with sample IDs, are provided in Additional file 1 ($TAHN_{ADC}$ vs. $Tumor_{ADC}$ in gene expression), Additional file 2 ($TAHN_{SCC}$ vs. $Tumor_{SCC}$ in gene expression), and Additional file 3 ($TAHN_{ADC}$ vs. $Tumor_{ADC}$ in methylation). The annotations from TCGA to identify these samples are provided in Additional file 4 (Appendix A).

Table 1 Datasets and sample distributions

Dataset Source	Tissue type	ADC	SCC
GEO: GDS3257 (gene expression)	Tumor	58	***
	TAHN	49	***
TCGA: LUAD+LUSC (gene expression)	Tumor	32	153
	TAHN	***	***
TCGA: LUAD+LUSC (DNA methylation)	Tumor	65	132
	TAHN	24	27

See challenge in Background on lack of TAHN tissue availability (***). GEO gene expression platform: Affymetrix Human Genome U133A Array (22,283 features), TCGA gene expression platform: Agilent 244 K Custom Gene Expression (17,814 features). TCGA methylation platform: Illumina Infinium HumanMethylation 27 k (27,578 features)

Experimental design

We followed a supervised classification process on 10-fold cross-validation. That is, for each fold we partitioned the dataset into training and test, where the former contains 90 % of the samples, while the latter contains the remaining 10 %. We ensured that each partition maintains the same class distribution as the whole dataset (stratified). In each fold, we analyzed the datasets using the experimental design as illustrated in Fig. 1. According to the design, there are four main components, namely, a) Feature Selection, b) Discretization, c) Model Building and d) Evaluation. We additionally perform Gene Functional Analysis, and apply Clustering methods to better understand the characteristics of the features chosen by this framework. Below, we explain each component in detail.

Feature selection

High-throughput platforms, such as gene expression and methylation microarrays, generate high-dimensional data that is typically very complex for analysis. Feature selection is a machine learning pre-processing step that tries to find a subset of the original variables (also called features or attributes) that are highly associated with the target class variable (i.e. phenotype, like a disease state). We used the ReliefF algorithm [27] to rank all variables and select the top scoring ones. ReliefF is a multivariate filter algorithm that estimates how well a given variable can distinguish the target class given the instances that are near to each other. The initial number of variables (17,814 in gene expression, and 27,578 in methylation) is reduced to the top 30 scoring variables. In previous studies [28], it has been reported that 30 is a sufficient number of genes to create computational classification models. With this number of genes, the classification models created would have a good trade-off between relevance and complexity of the model.

Similarly, we also selected the differentially expressed (DE) genes and differentially methylated (DM) probe sites from each dataset using Limma, which is an R-language package for the analysis of microarray data [29]. Limma uses a t-statistic to rank genes in order of evidence for differential expression. It first fits linear models for each gene (lmFit), and then it uses empirical Bayes (eBayes) moderation to adjust the standard error of the models by borrowing information from the rest of the genes (average variance across all genes). This method is very effective in finding differentially expressed (DE) genes in microarray data, however with methylation datasets it has not been

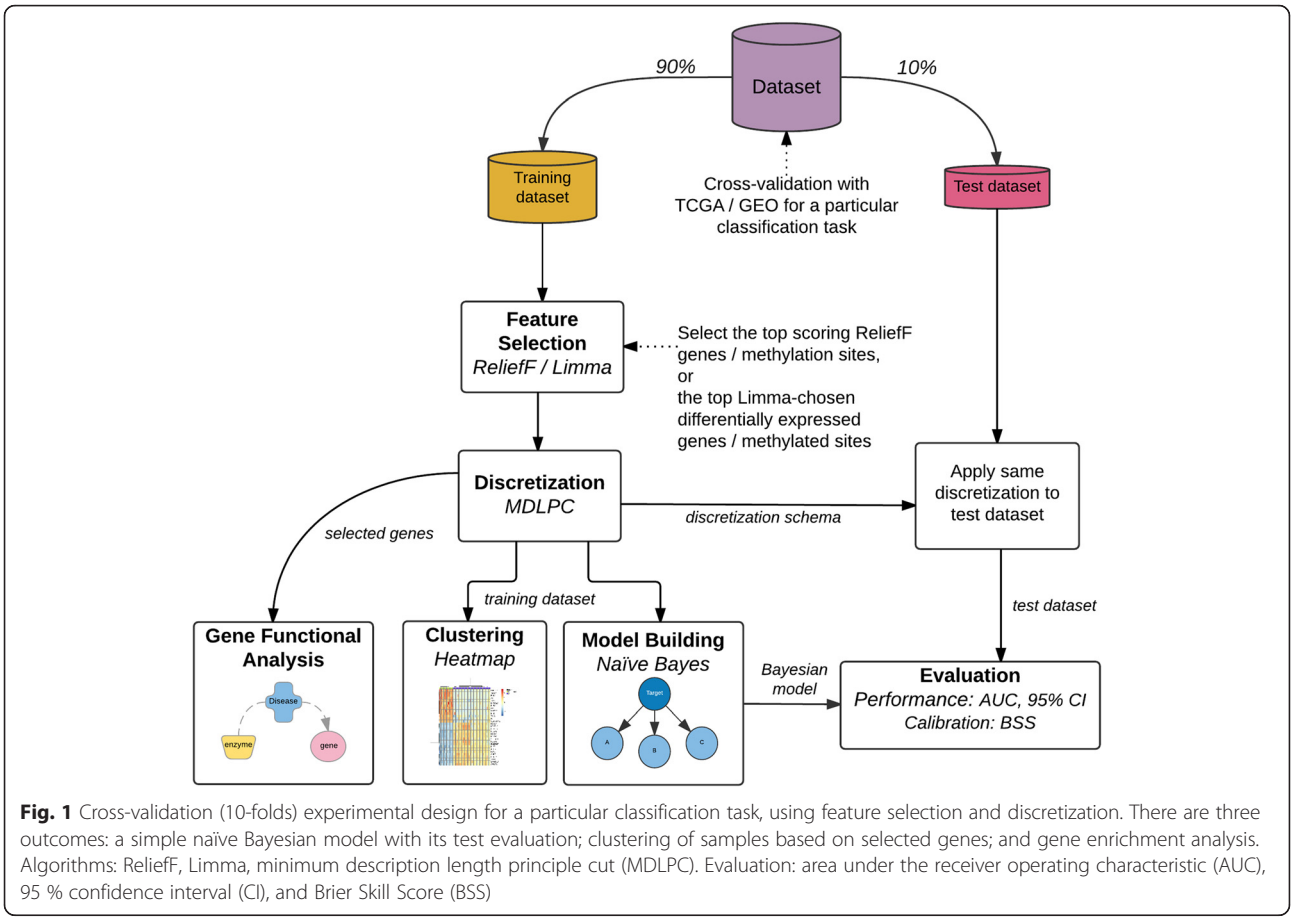

Fig. 1 Cross-validation (10-folds) experimental design for a particular classification task, using feature selection and discretization. There are three outcomes: a simple naïve Bayesian model with its test evaluation; clustering of samples based on selected genes; and gene enrichment analysis. Algorithms: ReliefF, Limma, minimum description length principle cut (MDLPC). Evaluation: area under the receiver operating characteristic (AUC), 95 % confidence interval (CI), and Brier Skill Score (BSS)

equally successful [30]. The output of finding the DE genes and DM probe sites with *Limma* can be seen as a feature selection method (or ranked list). Similarly to the ReliefF selection, we selected the top 30 most DE genes and DM probe sites (based on $\log_2$-fold change) to build a classifier for comparison with ReliefF. The output of the resulting classifiers was evaluated using the area under the receiver operating characteristic curve (AUC) performance metric in the test datasets.

Discretization

Most 'omic' data such as gene expression and methylation are represented with continuous values. However, many machine learning algorithms are designed to only handle discrete (categorical) data, using nominal variables, while real-world applications, like 'omic' data analysis, typically involves continuous-valued variables. Discretization, the process of transforming continuous values into discrete ones, has been shown to improve the performance of machine learning classifiers [31]. To discretize the variables, we used the Fayyad and Irani's minimum description length principle cut (MDLPC) [32]. This method, which is widely used in the machine learning community, applies a supervised greedy search strategy to recursively find the minimal number of cut-points in each variable that minimizes the entropy of the resulting subintervals.

For continuous methylation values ranging from 0 to 1, three possible strategies for discretization can occur. The first strategy is when a fixed cut-point is determined arbitrarily for all variables (for example, choosing > 0.5 methylated, while ≤ 0.5 could refer to unmethylated). The second strategy, when an expert-based discretization is made for all variables (i.e. unmethylated < 0.1, partially methylated between 0.1 and 0.8, and methylated > 0.8 [33]). The third strategy is when a supervised discretization method creates independent cut-points for each variable. For the first and second strategies, the same discretization scheme (i.e. same number of intervals or cut-points) is used for all variables. However, this approach is suboptimal for a classification task. For instance, when using MDLPC we observed that the methylation site cg19782598 was discretized into two categories: methylated (>0.86) and unmethylated (≤0.86); while methylation site cg11693019 was discretized into three categories: methylated (>0.76), partially methylated (between 0.76 and 0.47), and unmethylated (<0.47). Thus, supervised discretization could help identify appropriate cut-points for each variable, as opposed to the others, which naïvely assume the same cut-points for variables.

Clustering

In computational genomics, heatmaps are used to graphically show the level of expression that a selected group of genes have in a cohort of patient samples. A heatmap can also be built with methylation intensity values. We build heatmaps from the genes selected by Limma and ReliefF to further validate the results obtained with these feature selection methods. The clusters are a visual representation of the class discrimination ability of the genes selected.

The order in which genes (rows) and samples (columns) are ordered in the heatmap matrix is often based on an agglomerative hierarchical clustering. We used the Minkowski measure to calculate the pairwise distances between elements, and then aggregated the closest elements in clusters using the Ward linkage calculation of distances between clusters. This combination of Minkowski distance and Ward linkage has been shown to perform well in biomedical and synthetic datasets [34].

Gene functional analysis

We also performed Gene Functional Analysis using QIAGEN's Ingenuity® Pathway Analysis tool (IPA®, QIAGEN Redwood City, www.ingenuity.com) to gain insight into the biological role of the genes selected by our framework. First, all gene symbols selected were used as input for the IPA platform, which will search for correlations between these genes and functions or pathways in their curated literature. A p-value is computed using Fisher's right-tailed exact test for the gene list to a function/pathway it may be associated with. The p-values indicate the likelihood of association between the gene set (as selected by ReliefF) and a specific function (set of genes associated with a function) to have occurred due to random chance alone. A p-value of less than 0.05 is considered to be significantly better than random chance. Methylation probe sites were mapped into their corresponding gene symbols that they methylate.

Model building

In the machine learning literature, a classifier is a computational model that can differentiate between two (or more) states of disease. Bayesian networks [35] are particularly useful classifiers that are very popular in the classification of biomedical data. A Bayesian network (BN) is a probabilistic graphical representation of random variables (nodes) and probabilistic dependencies among them (arcs). Once a Bayesian network is learned, the structure and conditional probability tables can be used to calculate the posterior probabilities for a new case to be a member of a given class, i.e. the probabilities of a new case being ADC given the BN and the data. P(ADC = True|BN, data). A special case of BN is the naïve Bayesian classifier (NB), which assumes a strong conditional independence among the variables. In a NB structure, the target node (i.e. class variable) is the parent for all other features, and there are no arcs among those children nodes. The child nodes are independent given the parent, which facilitates the calculation of posterior probabilities by substituting the joint probability with the product of their probabilities. NBs have been shown to predict poorly in high-dimensional genomic datasets [36], but it is expected that the use of a feature selection method (ReliefF or Limma) will improve the NB classification performance. Moreover, its simplicity makes it a powerful tool to be considered in a biomedical classification framework, while giving us insights into the baseline performance on a given dataset.

Evaluation

We evaluated the NB classifiers using the area under the receiver operating characteristic (AUC), which is a measurement of the area created by plotting the performance of a classifier for the true positive rate versus the false positive rate. When presented with a test dataset, the Bayesian network calculates a posterior probability for every case, and a threshold is used to assign the class for the new cases. The curve is constructed by varying the threshold to which the probability is considered for class determination. Also, the 95 % confidence interval (C.I.) of the AUC was calculated using DeLong's method for variance estimation [37].

AUC (equivalent to c-statistic) is a useful measurement of the ability of models to discriminate between two (or more) classes [38]. Calibration deals with agreement between observed outcomes and predictions. For this purpose, we used the Brier Skill Score (BSS) [39] creates an index between −1 and 1 that provides information as of how far away the results of any classifier are in relation to the unskilled classifier. The unskilled classifier is one that only considers the distribution of data. A classifier with a positive BSS would therefore be skilled and unbiased.

Results

We investigated four classification tasks depending on the tissue type. These tasks test our hypothesis that the TAHN tissue has distinct genomic signatures that can differentiate among non-small cell lung cancer subtypes. We describe the classification tasks as follows:

1. $TAHN_{ADC}$ vs. $Tumor_{ADC}$, and $\mathrm{TAHN_{SCC}}$ vs $\mathrm{Tumor_{SCC}}$, searches for molecular differences between tumor tissue and TAHN tissue. These tasks are only applied to one lung cancer subtype at a time, either adenocarcinoma or squamous cell carcinoma patients;
2. $Tumor_{ADC}$ vs. $Tumor_{SCC}$, which searches for molecular differences between subtypes using only Tumor tissue;

3. $TAHN_{ADC}$ vs. $TAHN_{SCC}$, which searches for molecular differences between subtypes using only TAHN tissue; and
4. $TAHN\text{-}Tumor_{ADC}$ vs. $TAHN\text{-}Tumor_{SCC}$, which searches for molecular differences between subtypes using both TAHN and Tumor tissue.

The classification performance for every naïve Bayes classifier was calculated by averaging the AUCs over all folds from the experimental design illustrated in Fig. 1. Table 2 shows results for the classification tasks, including 95 % confidence interval (C.I.) and Brier Skill Score (BSS) as a calibration measurement. Contingency tables for these models can be seen in Additional file 4 (Appendix B).

All classification tasks achieved high predictive performances with AUC values higher than 0.8. For these datasets, the classification performance was similar between the NB classifiers created after applying ReliefF and Limma as feature selection methods. Limma is a popular method, among the genomics community, for the selection of differentially expressed genes, but it is not used as a feature selection method by the machine learning community. In contrast, ReliefF is a popular method among machine learning studies but not widely used in genomic studies. Figure 2 shows heatmaps and clusters for each classification task with the methylation probe sites selected using ReliefF.

We analyzed the genes found by ReliefF in the classification task of $TAHN\text{-}Tumor_{ADC}$ vs $TAHN\text{-}Tumor_{SCC}$ using IPA®. The results of the IPA® core analysis show a significant association between ReliefF-selected genes and the following diseases: cancer (25 out of 27) connective tissue disorder (13 out of 27), dermatological diseases and conditions (13 out of 27). Interestingly, the ReliefF-selected genes (19 out of 27) are associated with either adenocarcinoma (16 genes), squamous-cell carcinoma (4 genes) or carcinoma of the lung (4 genes). The list of genes and their associations can be seen in Table 3. Using these interesting 19 genes, we generated a gene interaction network to graphically visualize the relationships between genes and the disease class (adenocarcinoma, squamous-cell carcinoma and carcinoma of the lung). The network is illustrated in Fig. 3.

Discussion

Evaluation of classifiers

The classification performance for all models is high (AUC≥0.81), with positive calibration (BSS > 0). This positive calibration is a good indication that the models will perform well for other cases, and that they were not biased by the distribution of the data.

In the classification task of $TAHN_{ADC}$ vs. $Tumor_{ADC}$, the naïve Bayesian model created obtained high predictive performances (AUC≥0.99 with ReliefF, and≥0.81 with Limma). The classification task $TAHN_{SCC}$ vs. $Tumor_{SCC}$ also obtained high predictive performances (≥0.99 with both feature selection methods). The molecular differences between TAHN and tumor tissue show distinctive signatures regardless of 'omic' dataset, feature selection method or lung cancer subtype. The results for these classification tasks were as expected since the tissue architecture between TAHN and Tumor is recognizable under a microscope if enough tissue samples are provided. They also could be achieved with the relatively small number of normal tissues available for analysis, since these normal tissues are very homogenous in expression and methylation features.

In the classification task of $Tumor_{ADC}$ vs. $Tumor_{SCC}$ the predictive performance was very high (AUC≥0.89, for gene expression, and≥0.89 with methylation). Previous studies for the same classification task also show a similar classification performance. For example, Ben-Hamo et al. [40] correctly classified 85 %, using linear models. Meanwhile, Cai et al. [10] obtained an accuracy of 86 % using ensemble methods; Li et al. [41] achieved an AUC of 0.98 using Support Vector Machines; and Zhang et al. [42] achieved AUCs of 0.89 using naïve Bayesian models. Similarly, the study by

Table 2 AUC classification performance for different classification tasks

Classification Task	Omic	Feature selection with ReliefF			Feature selection with Limma		
		AUC	95 % C.I.	BSS	AUC	95 % C.I.	BSS
$TAHN_{ADC}$ vs. $Tumor_{ADC}$	G	0.99	0.97–1.0	0.89	0.94	0.82–1.0	0.73
	M	1.0	1.0–1.0	0.99	0.81	0.58–0.97	0.17
$TAHN_{SCC}$ vs. $Tumor_{SCC}$	M	1.0	0.99–1.0	0.94	0.99	0.96–1.0	0.66
$Tumor_{ADC}$ vs. $Tumor_{SCC}$	G	0.89	0.83–0.96	0.29	0.90	0.89–0.9	0.81
	M	0.97	0.94–0.99	0.71	0.89	0.74–1.0	0.38
$TAHN_{ADC}$ vs. $TAHN_{SCC}$	M	1.0	1.0–1.0	0.92	1.0	1.0–1.0	0.99
$TAHN\text{-}Tumor_{ADC}$ vs. $TAHN\text{-}Tumor_{SCC}$	M	0.92	0.89–0.95	0.42	0.94	0.87–1.0	0.56

G: gene expression, M: DNA methylation. The Brier Skill Score is a measurement of calibration of the classifier. A positive value on the BSS means that the classifier is well calibrated. A baseline classification is the work by Chang and Ramoni [22] which obtained an accuracy of 0.95 in the classification task $Tumor_{ADC}$ vs. $Tumor_{SCC}$

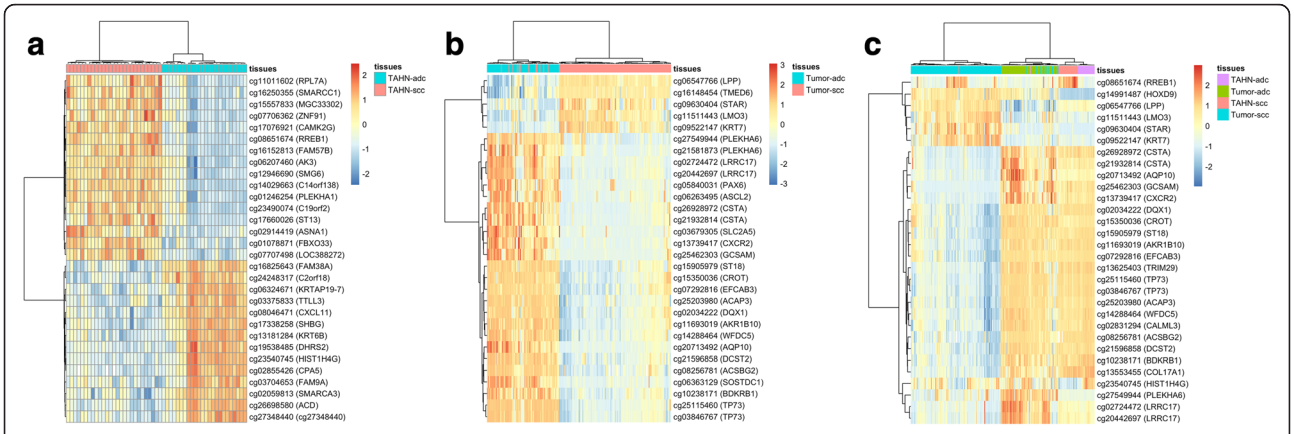

Fig. 2 Heatmaps for classification task **a** $TAHN_{ADC}$ vs. $TAHN_{SCC}$, **b** $Tumor_{ADC}$ vs. $Tumor_{SCC}$ and **c** TAHN-$Tumor_{ADC}$ vs. TAHN-$Tumor_{SCC}$ using the ReliefF feature selection algorithm. In the vertical axis the corresponding methylation site and gene symbol (in parenthesis) are shown. Some methylation sites do not lie in a particular gene, therefore, no symbol is provided. When multiple methylation sites are selected for the same gene, these sites should have similar methylation intensity, for it to be included. In the horizontal axis, a color-coded representation of the tissue samples is provided. Two distinct groups are observed in all three heatmaps. Cluster purity (accuracy by classification using clustering) for each task is calculated to be 1.0, 0.94 and 0.85 respectively

Chang and Ramoni [22] achieved an accuracy of 0.95, using naïve Bayesian models. It is worth noting that none of these studies used methylation datasets and they fail to clearly recognize the importance of TAHN tissue for classification.

The classification task of $TAHN_{ADC}$ vs. $TAHN_{SCC}$ also had very high evaluation performances ($AUC = 1$). This high performance means that all samples were correctly classified. We hypothesize that an explanation of this excellent result can be attributed to the distinctive epigenetic differences between lung tissues. We did not evaluate the gene expression in this classification task due to the lack of an available dataset. To the best of our knowledge reporting of TAHN tissue in public repositories is still an open challenge that should be addressed to improve experimental designs of other studies. A study by Haaland et al. [43], showed that there are differentially expressed genes between TAHN tissues in prostate cancer. In our study, we investigate DNA methylation data to indicate that the same differences could also be found in lung cancer TAHN tissues, and we hypothesize that the use of TAHN tissues might also help in the classification performance of other cancer types.

The classification task of TAHN-$Tumor_{ADC}$ vs. TAHN-$Tumor_{SCC}$ is a novel approach, where a mix of tissue types are used to classify between lung cancer subtypes. The noise introduced by mixing tissue types is overcome by our experimental design, which was able to obtain a very good classification performance ($AUC \geq 0.92$). Despite, the 'noisy' tissue samples, a simple naïve Bayesian classifier can accurately classify between lung cancer subtypes. This classification performance is confirmed by the heatmap analysis in Fig. 2c, where the tumor tissue of ADC creates a distinct cluster, while the remaining samples cluster together in three distinct subclusters. Furthermore, our Gene Functional Analysis using IPA® shows strong associations to cancer pathways, with 19 genes found to be associated with adenocarcinoma, squamous-cell carcinoma and carcinoma of the lung. Out of these 19 genes we found 4 genes associated specifically with lung cancer subtypes: AKR1B10, AQP10, CXCR2, TP73.

The value of using TAHN tissue for classification

Lung cancer patients could benefit with a potentially novel approach for subtyping. The diagnosis of adenocarcinoma vs. squamous cell carcinoma is routinely accomplished using histology supplemented by immunohistochemistry (TTF-1 and p63/p40). It is therefore not likely that our approach would change this practice, which is well established, quick and inexpensive. Rather, we suggest that the use of epigenomic changes could help in the small number of tumors which remain difficult to classify. However, the primary importance of our work may be in providing additional understanding of the origins of squamous cell and adenocarcinomas, which suggest that these phenotypes are associated with, or perhaps even derived from, different epigenomic phenotypes. Epigenomic alterations, in the form of DNA methylation, prevent the binding of transcription machinery, resulting in gene silencing [44]. Moreover, DNA methylation signatures are different between tissue types and between tumors and normal surrounding tissue [20]. In our study, tumor-adjacent histologically normal tissue samples were used to classify lung cancer subtypes with excellent results. This classification performance was achieved when no tumor samples were involved ($TAHN_{ADC}$ vs. $TAHN_{SCC}$), and when a mix of tissue was used (TAHN-$Tumor_{ADC}$ vs. TAHN-$Tumor_{SCC}$). The high

Table 3 Genes selected for the classification task of TAHN-Tumor$_{ADC}$ Vs. TAHN-Tumor$_{SCC}$

Gene Symbol	Gene Name	Known Literature Evidence to Cancer
ST18	suppression of tumorigenicity 18, zinc finger	Yes [45]
CSTA	cystatin A (stefin A)	Yes [45, 46]
LPP	LIM domain containing preferred translocation partner in lipoma	Yes [45]
CROT	carnitine O-octanoyltransferase	Yes [45]
BDKRB1	bradykinin receptor B1	Yes [47]
AKR1B10	aldo-keto reductase family 1, member B10 (aldose reductase)	Yes [48]
TP73	tumor protein p73	Yes [49–51]
EFCAB3	EF-hand calcium binding domain 3	Yes
RREB1	ras responsive element binding protein 1	Yes [45]
HIST1H4G	histone cluster 1, H4g	No
STAR	steroidogenic acute regulatory protein	Yes
ACSBG2	acyl-CoA synthetase bubblegum family member 2	Yes [45]
DQX1	DEAQ box RNA-dependent ATPase 1	Yes [45]
AQP10	aquaporin 10	Yes [45]
PLEKHA6	pleckstrin homology domain containing, family A member 6	Yes [52, 53]
GCSAM	germinal center-associated, signaling and motility	No
WFDC5	WAP four-disulfide core domain 5	Yes
KRT7	keratin 7, type II	Yes [54]
DCST2	DC-STAMP domain containing 2	Yes [45]
CALML3	calmodulin-like 3	Yes
ACAP3	ArfGAP with coiled-coil, ankyrin repeat and PH domains 3	Yes
LRRC17	leucine rich repeat containing 17	Yes [45]
TRIM29	tripartite motif containing 29	Yes [55]
CXCR2	chemokine (C-X-C motif) receptor 2	Yes [45, 56, 57]
HOXD9	homeobox D9	Yes [58]
COL17A1	collagen, type XVII, alpha 1	Yes [45]
LMO3	LIM domain only 3 (rhombotin-like 2)	Yes

The list of genes is ordered by their ranks, as selected by ReliefF for the classification task of TAHN-Tumor$_{ADC}$ Vs. TAHN-Tumor$_{SCC}$. The Entrez gene symbol, and the gene name are listed in the first two columns respectively. The 'Known Literature Evidence to Cancer' indicates if links to cancer were detected by the IPA® software. Citations are provided to literature indicating links to adenocarcinoma, squamous-cell carcinoma and carcinoma in lung

AUC results are an indication of the diagnostic potential of this technology.

Limitations and future work

Our study had some limitations, which include the following: 1) There were a limited number of tumor-adjacent histologically normal tissue samples used. However, the homogeneity of these normal tissues we observed suggests that additional normal tissues would not improve the classifier. 2) The resulting classifiers were not validated in another dataset outside of TCGA lung samples. 3) Each 'omic' classifier is independent of one another. In the future, we would like to explore data integration models in a multi-omic approach. 4) The classification problem of discriminating cancer subtypes of adenocarcinoma and squamous cell carcinoma could also be explored in a pan-cancer analysis, to validate the same finding seen in our study of lung cancer subtypes. 5) Due to the challenge of data availability, in this study we did not analyze biopsies with varying percentages of tumor and TAHN tissue (mixed biopsies). Instead, we took relatively 'pure' biopsies of either tumor or TAHN to classify between lung cancer subtypes. A future study could consider the molecular classification or discovery of cancer given a mixture of tumor and TAHN tissue. For example, an analysis of 'omic' data from cancerous and non-cancerous tumor tissues, as well as TAHN tissue for both types of tumors, might be performed in the same way as presented in this manuscript.

Conclusions

In this paper, we addressed the issue of lung cancer subtyping using DNA methylation data from TAHN tissue, which is a novel strategy for classification of non-small

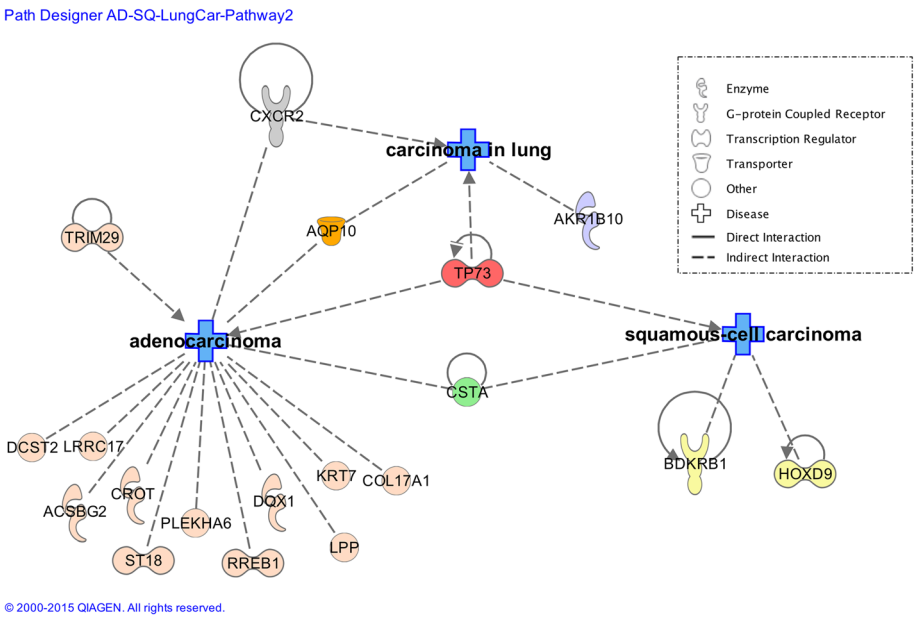

Fig. 3 Gene interaction network generated by the IPA® software. It shows an analysis of the genes found by ReliefF in the classification task TAHN-Tumor$_{ADC}$ vs TAHN-Tumor$_{SCC}$. Three diseases are being shown (carcinoma of the lung, adenocarcinoma and squamous cell carcinoma), and the selected genes from our analysis were connected to these diseases via literature evidence that indicates: direct interactions (straight line), or indirect interactions (dashed line). Some of those interactions have arrow-heads indicating causation (e.g. BDKRB1). An arrow-head with a bar (i.e., TP73) indicates inhibition

cell lung cancer samples. This study demonstrated that using computational Bayesian modeling, it is possible to discover the molecular differences between tumor and tumor-adjacent tissue of lung cancer patients. This discovery will allow clinicians to use the available biopsy material without worrying about its tissue composition, yielding in less invasive diagnostic procedures for the patient. We hope that our results will encourage researchers to also make use of TAHN tissue samples generated in their laboratories for predictive modeling and make this data available for public use. As more data becomes available, our models can be further improved, and future discoveries could be made in other cancers.

Availability of supporting data

The datasets used in this study are publicly available from The Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/) in datasets LUAD and LUSC; and also from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), accession number GDS3257. The formatted datasets used in this study, along with sample IDs, are provided in Additional file 1 (TAHN$_{ADC}$ vs. Tumor$_{ADC}$ in gene expression), Additional file 2 (TAHN$_{SCC}$ vs. Tumor$_{SCC}$ in gene expression), and Additional file 3 (TAHN$_{ADC}$ vs. Tumor$_{ADC}$ in methylation). The annotations from TCGA to identify these samples are provided in Additional file 4 (Appendix A).

Additional files

Additional file 1: Formatted TCGA dataset used in this study, along with sample IDs for classification task TAHN$_{ADC}$ vs.Tumor$_{ADC}$ in gene expression. (CSV 5182 kb)

Additional file 2: Formatted TCGA dataset used in this study, along with sample IDs for classification task TAHN$_{SCC}$ vs.Tumor$_{SCC}$ in gene expression. (CSV 24086 kb)

Additional file 3: Formatted TCGA dataset used in this study, along with sample IDs for classification task TAHN$_{ADC}$ vs.Tumor$_{ADC}$ in DNA methylation. (CSV 41150 kb)

Additional file 4: Appendix A shows the Cancer Genome Atlas annotations to identify the types of samples used in this study.Appendix B shows additional performance measures for the models described. (DOCX 106 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

ALP, SV and VG designed the study. ALP, HAO and JBB performed the analysis of the data. CRE and JGH provided interpretation of the results. ALP drafted the manuscript, and all authors contributed critically, read, revised and approved the final version.

Acknowledgements

The research reported in this publication was supported in part by the following grants: National Cancer Institute (USA): P50CA90440; National Library of Medicine (USA): R01LM010950 and R01LM012095, training grant 5T15LM007059-26; National Institute of General Medical Sciences (USA): R01GM100387; The International Fulbright Science and Technology Award (USA): 15101109; Mexican National Council of Science and Technology (CONACyT, Mexico): scholarship 213941.

Author details
[1]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Boulevard, 15206 Pittsburgh, PA, USA. [2]Department of Computational Genomics, National Institute of Genomic Medicine, Periferico Sur No. 4809, Col. Arenal Tepepan, Tlalpan 14610Mexico City, Mexico. [3]Division of Hematology/Oncology, Department of Medicine, University of Pittsburgh School of Medicine, UPMC Cancer Pavilion, 5150 Centre Avenue, 15232 Pittsburgh, PA, USA.

Received: 13 August 2015 Accepted: 28 February 2016
Published online: 04 March 2016

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64:9–29.
2. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc. 2008;83:584–94.
3. Yao H, Rahman I. Current concepts on the role of inflammation in COPD and lung cancer. Curr Opin Pharmacol. 2009;9:375–83.
4. College of American Pathologists. Lung Adenocarcinoma. 2011. p. 1–2.
5. College of American Pathologists. Lung Squamous Cell Carcinoma. 2011. p. 1–2.
6. Cagle PT. The new American Cancer Society Lung Cancer Screening guidelines and the role of the pathologist. Arch Pathol. 2013;137:451.
7. Wender R, Fontham ETH, Barrera E, Colditz GA, Church TR, Ettinger DS, Etzioni R, Flowers CR, Gazelle GS, Kelsey DK, LaMonte SJ, Michaelson JS, Oeffinger KC, Shih Y-CT, Sullivan DC, Travis W, Walter L, Wolf AMD, Brawley OW, Smith RA. American Cancer Society lung cancer screening guidelines. CA Cancer J Clin. 2013;63:107–17.
8. Stamatis G. Staging of lung cancer: the role of noninvasive, minimally invasive and invasive techniques. Eur Respir J. 2015;46(2):521–31. ERJ–01267–2014.
9. Dooms C, Vliegen L, Vander Borght S, Yserbyt J, Hantson I, Verbeken E, Wauters E, Nackaerts K, Ninane V, Vansteenkiste J, Vandenberghe P. Suitability of small bronchoscopic tumour specimens for lung cancer genotyping. Respiration. 2014;88:371–7.
10. Cai Z, Xu D, Zhang Q, Zhang J, Ngai S-M, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. Mol Biosyst. 2014;11(3):791–800.
11. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102:464–74.
12. Langer CJ, Besse B, Gualberto A, Brambilla E, Soria J-C. The evolving role of histology in the management of advanced non-small-cell lung cancer. J Clin Oncol. 2010;28:5311–20.
13. Chiu C-H, Chou T-Y, Chiang C-L, Tsai C-M. Should EGFR mutations be tested in advanced lung squamous cell carcinomas to guide frontline treatment? Cancer Chemother Pharmacol. 2014;74:661–5.
14. Dacic S, Shuai Y, Yousem S, Ohori P, Nikiforova M. Clinicopathological predictors of EGFR/KRAS mutational status in primary lung adenocarcinomas. Mod Pathol. 2010;23:159–68.
15. Soda MM, Choi YLY, Enomoto MM, Takada SS, Yamashita YY, Ishikawa SS, Fujiwara S-IS, Watanabe HH, Kurashina KK, Hatanaka HH, Bando MM, Ohno SS, Ishikawa YY, Aburatani HH, Niki TT, Sohara YY, Sugiyama YY, Mano HH. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6.
16. Richer AL, Friel JM, Carson VM, Inge LJ, Whitsett TG. Genomic profiling toward precision medicine in non-small cell lung cancer: getting beyond EGFR. Pharmgenomics Pers Med. 2015;8:63–79.
17. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Rosell R, Fárez-Vidal ME. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. Int J Cancer. 2011;129:355–64.
18. Pfeifer GP, Rauch TA. DNA methylation patterns in lung carcinomas. Semin Cancer Biol. 2009;19:181–7.
19. Rauch TA, Wang Z, Wu X, Kernstine KH, Riggs AD, Pfeifer GP. DNA methylation biomarkers for lung cancer. Tumor Biol. 2012;33:287–96.
20. Szyf M. DNA methylation signatures for breast cancer classification and prognosis. Genome Med. 2012;4:26.
21. Phillips T. The role of methylation in gene expression. Nat Educ. 2008;1(1):116. http://www.nature.com/scitable/topicpage/the-role-ofmethylation-in-gene-expression-1070
22. Chang H-H, Ramoni MF. Transcriptional network classifiers. BMC Bioinformatics. 2009;10 Suppl 9:S1.
23. Guimarães MD, Hochhegger B, Benveniste MFK, Odisio BC, Gross JL, Zurstrassen CE, Tyng CC, Bitencourt AGV, Marchiori E. Improving CT-guided transthoracic biopsy of mediastinal lesions by diffusion-weighted magnetic resonance imaging. Clinics (Sao Paulo). 2014;69:787–91.
24. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
25. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
26. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, Murphy SE, Yang P, Pesatori AC, Consonni D, Bertazzi PA, Wacholder S, Shih JH, Caporaso NE, Jen J. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One. 2008;3:e1651.
27. Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Appl Intell. 1997;7:39–55.
28. Dudoit S, Fridlyand J, Speed TP. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. J Am Stat Assoc. 2002;97:77–87.
29. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
30. Buhule OD, Minster RL, Hawley NL, Medvedovic M, Sun G, Viali S, Deka R, McGarvey ST, Weeks DE. Stratified randomization controls better for batch effects in 450 K methylation analysis: a cautionary tale. Front Genet. 2014;5:354.
31. Garcia S, Luengo J, Sáez JA, López V, Herrera F. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng. 2013;25:734–50.
32. Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.
33. Capra JA, Kostka D. Modeling DNA methylation dynamics with approaches from phylogenetics. Bioinformatics. 2014;30:i408–14.
34. Lee A, Willcox B. Minkowski generalizations of Ward's method in hierarchical clustering. J Classif. 2014;31:194–218.
35. Neapolitan RE. Probabilistic Reasoning in Expert Systems. 2012.
36. Jiang X, Cai B, Xue D, Lu X, Cooper GF, Neapolitan RE. A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. J Am Med Inform Assoc. 2014;21:e312–9.
37. DeLong ERE, DeLong DMD, Clarke-Pearson DLD. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
38. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol. 2012;12:82.
39. Wilks DS. Statistical Methods in the Atmospheric Sciences, 3rd Edition from Daniel Wilks. ISBN-9780123850225, Printbook, Release Date: 2011 Academic Press; 2011; 284–287. http://store.elsevier.com/Statistical-Methods-in-the-Atmospheric-Sciences/Daniel-Wilks/isbn-9780123850225/
40. Ben-Hamo R, Boue S, Martin F, Talikka M, Efroni S. Classification of lung adenocarcinoma and squamous cell carcinoma samples based on their gene expression profile in the sbv IMPROVER Diagnostic Signature Challenge. Systemsbiomedicine. 2013;1:68–77.
41. Li J, Li D, Wei X, Su Y. In silico comparative genomic analysis of two non-small cell lung cancer subtypes and their potentials for cancer classification. Cancer Genomics Proteomics. 2014;11:303–10.
42. Zhang A, Wang C, Wang S, Li L, Liu Z, Tian S. Visualization-aided classification ensembles discriminate lung adenocarcinoma and squamous cell carcinoma samples using their gene expression profiles. PLoS One. 2014;9:e110052.
43. Haaland CM, Heaphy CM, Butler KS, Fischer EG, Griffith JK, Bisoffi M. Differential gene expression in tumor adjacent histologically normal prostatic tissue indicates field cancerization. Int J Oncol. 2009;35:537–46.
44. Brzeziańska E, Dutkowska A, Antczak A. The significance of epigenetic alterations in lung carcinogenesis. Mol Biol Rep. 2013;40:309–25.
45. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(Database issue):D805–11.
46. Costea DE, Hills A, Osman AH, Thurlow J, Kalna G, Huang X, Murillo CP, Parajuli H, Suliman S, Kulasekara KK, Johannessen AC, Partridge M. Identification of two distinct carcinoma-associated fibroblast subtypes with

differential tumor-promoting abilities in oral squamous cell carcinoma. Cancer Res. 2013;73:3888–901.
47. Dlamini Z, Bhoola KD. Upregulation of tissue kallikrein, kinin B1 receptor, and kinin B2 receptor in mast and giant cells infiltrating oesophageal squamous cell carcinoma. J Clin Pathol. 2005;58:915–22.
48. Kim B, Lee HJ, Choi HY, Shin Y, Nam S, Seo G, Son D-S, Jo J, Kim J, Lee J, Kim J, Kim K, Lee S. Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. Cancer Res. 2007;67:7431–8.
49. Flores ER, Sengupta S, Miller JB, Newman JJ, Bronson R, Crowley D, Yang A, McKeon F, Jacks T. Tumor predisposition in mice mutant for p63 and p73: Evidence for broader tumor suppressor functions for the p53 family. Cancer Cell. 2005;7:363–73.
50. Lu H, Yang X, Duggal P, Allen CT, Yan B, Cohen J, Nottingham L, Romano R-A, Sinha S, King KE, Weinberg WC, Chen Z, Van Waes C. TNF-alpha Promotes c-REL/Delta Np63 alpha Interaction and TAp73 Dissociation from Key Genes That Mediate Growth Arrest and Apoptosis in Head and Neck Cancer. Cancer Res. 2011;71:6867–77.
51. Tomasini R, Tsuchihara K, Wilhelm M, Fujitani M, Rufini A, Cheung CC, Khan F, Itie-Youten A, Wakeham A, Tsao M-S, Iovanna JL, Squire J, Jurisica I, Kaplan D, Melino G, Jurisicova A, Mak TW. TAp73 knockout shows genomic instability with infertility and tumor suppressor functions. Genes Dev. 2008; 22:2677–91.
52. The Cancer Genome Atlas Research Network, Getz G, Saksena G, Zhang J, Zhang H, Shukla S, Lawrence MS, Sivachenko A, Stojanov P, Jing R, Park PJ, Chin L, Chan TA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
53. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, Chaudhuri S, Guan Y, Janakiraman V, Jaiswal BS, Guillory J, Ha C, Dijkgraaf GJP, Stinson J, Gnad F, Huntley MA, Degenhardt JD, Haverty PM, Bourgon R, Wang W, Koeppen H, Gentleman R, Starr TK, Zhang Z, Largaespada DA, Wu TD, de Sauvage FJ. Recurrent R-spondin fusions in colon cancer. Nature. 2012;488:660–4.
54. Laurell H, Bouisson M, Berthelemy P, Rochaix P, Dejean S, Besse P, Susini C, Pradayrol L, Vaysse N, Buscail L. Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation with gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples. World J Gastroenterol. 2006;12:3344–51.
55. Wang L, Yang H, Abel EV, Ney GM, Palmbos PL, Bednar F, Zhang Y, Leflein J, Waghray M, Owens S, Wilkinson JE, Prasad J, Ljungman M, Rhim AD, di Magliano MP, Simeone DM. ATDC induces an invasive switch in KRAS-induced pancreatic tumorigenesis. Genes Dev. 2015;29:171–83.
56. Raghuwanshi SK, Nasser MW, Chen X, Strieter RM, Richardson RM. Depletion of beta-arrestin-2 promotes tumor growth and angiogenesis in a murine model of lung cancer. J Immunol. 2008;180:5699–706.
57. Raghuwanshi SK, Smith N, Rivers EJ, Thomas AJ, Sutton N, Hu Y, Mukhopadhyay S, Chen XL, Leung T, Richardson RM. G protein-coupled receptor kinase 6 deficiency promotes angiogenesis, tumor progression, and metastasis. J Immunol. 2013;190:5329–36.
58. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, Zhao M, Alves MVO, Chang K, Drummond J, Cortez E, Xie T-X, Di Zhang, Chung W, Issa J-PJ, Zweidler-McKay PA, Wu X, El-Naggar AK, Weinstein JN, Wang J, Muzny DM, Gibbs RA, Wheeler DA, Myers JN, Frederick MJ. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 2013;3:770–81.