Small Patient Datasets Reveal Genetic Drivers of Non-Small Cell Lung Cancer Subtypes using a Novel Machine Learning Approach

Moses Cook [1], Bessi Qorri [2], Amruth Baskar [3,4], Jalal Ziauddin [3], Luca Pani [5,6,7,8], Shashi Bushan Yenkanchi [3], Joseph Geraci [3,9,10] *

1 Department of Medical Biophysics, University of Toronto, Toronto ON, Canada
2 Department of Biomedical and Molecular Sciences, Queen's University, Kingston ON, Canada
3 NetraMark Corp, Toronto, ON, Canada
4 Faculty of Mathematics, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo ON, Canada
5 Department of Psychiatry and Behavioral Sciences, Leonard M. Miller School of Medicine, University of Miami, FL, USA
6 Nurosene Inc., Toronto, ON, Canada
7 Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy
8 VeraSci, Durhan, NC, USA
9 Department of Pathology and Molecular Medicine, Queen's University, Kingston ON, Canada
10 The Centre for Biotechnology and Genomics Medicine, Medical College of Georgia, USA
** Author to whom correspondence should be addressed.*

Abstract: There are many small datasets of significant value in the medical space that are being underutilized. Due to the heterogeneity of complex disorders found in oncology, systems capable of discovering patient subpopulations while elucidating etiologies is of great value as it can indicate leads for innovative drug discovery and development. Here, we report on a machine intelligence-based study that utilized a combination of two small non-small cell lung cancer (NSCLC) datasets consisting of 58 samples of adenocarcinoma (ADC) and squamous cell carcinoma (SCC) and 45 samples from the gene expression analysis of human lung cancer and control samples series (GSE18842). Utilizing a novel machine learning approach, we were able to uncover subpopulations of ADC and SCC while simultaneously extracting which genes, in combination, were significantly involved in defining the subpopulations. An interactive hypothesis-generating interface designed to work with machine learning methods allowed us to explore the hypotheses generated by the unsupervised components of the system. Using these methods, we were able to uncover genes implicated by other methods and accurately discover known subpopulations without being asked, such as different levels of aggressiveness within the SCC and ADC subtypes. Furthermore, *PIGX* was a novel gene implicated in this study that warrants further study due to its role in breast cancer proliferation. Here we demonstrate the ability to learn from small datasets and reveal well-established properties of NSCLC. These machine learning techniques can reveal the driving factors behind subpopulations of patients altering the approach to drug discovery and development by making precision medicine a reality.

Keywords: Machine learning; genetic subtypes; disease heterogeneity; augmented intelligence; machine intelligence; artificial intelligence; squamous cell carcinoma; adenocarcinoma

1. Introduction

The collection of transcriptomic data is expensive, resulting in datasets with a small number of sample numbers (in the hundreds) but thousands of variables. As a result, several techniques that are making significant strides in the imaging space, such as deep neural networks, are not suitable for these data, as a large number of samples are required. Furthermore, the heterogeneity of the patient population and the complexity of diseases found in oncology requires going beyond the labels. The development of techniques that can explain the driving variables behind patient subpopulations is tremendously valuable in identifying and developing novel therapeutic agents – this is particularly relevant for mapping out heterogeneous diseases such as lung cancer.

Lung cancer is the leading cause of cancer mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for 85% of all lung cancers [1]. NSCLC can be divided into three histological subtypes with distinct phenotypes and prognoses: adenocarcinoma (ADC), squamous cell carcinoma (SCC) and large cell carcinoma (LCC) [2,3]. The histological differences across these subtypes suggest that distinct molecular mechanisms are underlying the observed phenotypic differences. Although the differential gene expressions across NSCLC subtypes have been of increasing interest, the therapeutic implications on how these pathways interact, is only more recently being investigated [4]. The remarkable degree of genetic variability within each histological subtype only highlights the importance of molecular biology and genotyping for NSCLC [5,6].

Fortunately, machine learning (ML) advancements have served as promising tools for stratifying NSCLC, predicting transcriptional mutations based on histological slides or discriminating NSCLC subtype through genomic expression levels. The bulk of ML efforts have focused on image analysis for predicting the stage of NSCLC [7-10]. However, the growing body of evidence highlighting the molecular abnormalities that underlie the genomic subtypes of NSCLC can train ML algorithms to identify novel biomarkers for NSCLC, moving towards precision medicine [11-13]. For instance, previous reports have identified that ADC is associated with increased expression of genes related to protein transport and cell junction, while SCC is associated with increased expression of genes related to cell division and DNA replication [14]. An analysis of gene expression profiles between ADC and SCC using machine learning algorithms has been previously reported, identifying several genes including *CSTA, TP63, SERPINB13, CLCA2, BICD2, PERP, FAT2, BNC1, ATP11B, FAM83B, KRT5, PARD6G,* and *PKP1* that were differentially expressed in ADC and SCC [15].

To consolidate of progress in NSCLC classification using ML, a review of recent work using imaging and genetic biomarkers has been performed. This review highlights the foundation by which our methodology was built on, while contextualizing the need for an ML framework that can further identify NSCLC genetic drivers using small datasets when labelled and/or big data is unavailable.

Review of Previous Work

Discriminating Tumor vs. Normal Tissue

ML models are highly efficient and well suited for image classification, outpacing human efforts [16]. More specifically, tumor size is a prognostic marker for patient survival, highlighting the necessity for automatic pattern analysis of tumor boundary detection [17]. Several ML methods have been proposed for this application, most notably using convolutional neural networks (CNNs). For example, a CNN-based approach has been used to identify specific features associated with patient prognosis based on histological slides of lung ADC [18]. The utility of CNNs to classify cancer vs. non-cancerous tissue has been corroborated as CNNs were found to outperform traditional non-deep learning methods, such as Gray-level Co-occurrence Matrix texture analysis with a support vector machine (SVM) classifier [19]. However, this model had a relatively low Area Under the Curve (AUC) of the Receiver Operator Curve (ROC) compared to other ML methods, owing to the complex morphometry of histological image data.

Discriminating ADC vs. SCC vs. Normal Tissue

ADC and SCC account for approximately 40% and 25% of all lung cancer cases, respectively, with treatment strategies for the two subtypes differing significantly. Several methods have been developed to automate patient diagnosis based on both radiomic and histological images. For example, an automated GLCM-SVM binary classifier method was used to classify hybrid positron emission tomography (PET)/computed tomography (CT) images of ADC and SCC patient tumors [20]. They distinguished ADC and SCC patient samples with an AUC of 0.89 based on colour and texture features. Similarly, another group trained a hybrid genetic algorithm and SVM model on CT images to parse the heterogeneity of NSCLC subtypes with a classification success rate of 96.2%. CT images were used in a more recent study to classify NSLSC subtypes using a hybrid minimum redundancy maximum relevance algorithm and SVM method to attain an AUC of 0.655 [21].

On the microscopic scale, one study used histologically stained images of SCC and ADC patients for subtype classification using a novel CNN, PathCNN, which was able to differentiate normal and pathological tissue but found AUCs for distinguishing the two subtypes to be 0.93 [22]. Interestingly, microRNA data was the foundation of an investigation of the two NSCLC subtypes were distinguished using microRNA sequencing data from a publicly available dataset. They used a decision tree algorithm, where the two nodes corresponding to two microRNA thresholds to distinguish the two subtypes, obtaining an AUC of 0.916 [23]. These findings demonstrate the complexity of NSCLC subtyping at the tumor level; however, subtyping at the morphological level can impose additional challenges.

Morphometry Subtype Prediction

NSCLC subtypes present remarkable structural heterogeneity. The World Health Organization (WHO) has released guidelines on diagnosing the morphological subtypes of ADC: lepidic, acinar, papillary, micropapillary, and solid [24]. However, there is a high level of inter-variability that exists between classifications. Furthermore, a single patient can present with a range of morphological subtypes. As a result, a careful diagnosis must be made as each subtype imparts a distinct set of characteristics, including therapeutic options and prognosis [25,26]. The Cohen's kappa score (κ) can assess statistical reliability as no ground truth exists for classifying

the predominant morphological pattern. Thus, quantitative methods that can minimize variability in diagnosis are paramount in delivering targeted care. Using a CNN on histological slides of ADC, researchers classified tumors into one of the five subtypes, with a κ score of 0.41–0.60 [27]. Similarly, a de-novo trained CNN was used to characterize patches of the ADC morphological subtypes on histological slides [28].

Mutation Subtype Prediction

Understanding the etiology of NSCLC subtypes is essential to elucidate the mechanisms underpinning the morphological heterogeneity of NSCLC. Therefore, it comes as no surprise that there is a remarkable degree of genetic heterogeneity even within subtypes. For example, 20% of ADC tumours have been shown to have *EGFR* mutations and have subsequently become a druggable target [29,30]. As a result, efforts have been focused on using ML and histological or radiological images to predict the mutation status of NSCLC tumours. Using a CNN has enabled researchers to predict the six most common ADC mutations, such as *EGFR* and *KRAS*, based on histological slides [31]. This network correctly classified *EGFR* and *KRAS* mutations with AUCs of 0.826 and 0.733, respectively. Comparable results were reported for *TP53* mutations with an AUC of 0.76 using the same CNN architecture [32].

Research efforts have also begun to use non-invasive radiomic data at the mutational level. Gradient Tree Boosting outperformed a Random Forest model to distinguish the *EGFR* mutation status in PET/CT images of ADC and SCC patients with an AUC of 0.659 [33]. One study employed PET, CT and genomics data to test a spectrum of ML algorithms to predict *EGFR* and *KRAS* mutation status in a cohort of NSCLC patients with maximal AUCs of 0.82 and 0.83, respectively, using a stochastic gradient descent classifier [34]. Using a CNN trained on tumor morphology and RNA sequencing data, researchers have been able to classify ADC and SCC transcriptomic subtypes with AUCs of 0.771-0.892 and 0.7, respectively [35]. These results showcase the potential for ML methods to predict NSCLC transcriptional subtypes.

NSCLC Stratification Based on Gene Expression

ML also has applications in classifying the genetic expression profiles of ADC and SCC. In attempts to predict overall survival (OS) in NSCLC patients, investigators used a deep learning network, trained on clinical prognostic factors and microarray data to predict the probability of 5-year survival after the first treatment with an AUC of 0.8163 [36].

NSCLC progression through the various pathologic stages has been shown to decrease overall 5-year survival [37]. Monotonically expressed genes have been hypothesized to give rise to stage progression and correlate to survival risk levels. Using a feature selection algorithm of microarray data accessed from the GEO of both ADC and SCC patients across NSCLC stages, no monotonically expressed genes were found to correlate to ADC or SCC stage [38]. However, a handful of genes were found to correlate with risk level for the ADC subtype, suggesting that ML-identified gene signatures might be useful for patient prognosis.

Conversely, ML efforts can also be focused on discovering biomarkers for early tumor development. For instance, one group used a semi-restricted Boltzmann ML algorithm using clinical data, including OS and tumor stage, and a feature selection technique to derive the genes driving early cancer development and predicting tumor stage [39]. These genes may serve as therapeutic targets for early-stage NSCLC and warrant further investigation. In previous efforts in this field, groups used a Monte Carlo feature selection to build genetic profiles of ADC and SCC tumour samples obtained from the GEO and applied an SVM classifier to create a list of

optimal genes that distinguish the two subtypes [40]. Their classifier derived a list of 13 differentially expressed genes in the two subtypes, including *CLCA2*, a member of the chloride channel family.

In another study, a k-means clustering method was used to classify genetic subtypes of ADC. Healthy and ADC tissue was then classified using an SVM followed by input into a self-organizing map neural network. The neurons in the output layer were categorized using a hierarchical clustering method to divide ADC tumours into four genetic subtypes [41]. In addition to performing a survival analysis on the four subtypes, two subtypes were found to have high expression levels of immune-related genes, suggesting the crucial role of immune dysregulation in ADC development.

Here, using a novel set of ML tools designed to learn from patient datasets to analyze gene expression data derived from ADC and SCC NSCLC patients, we were able to identify novel driving genes that distinguish these two broad subtypes. ML with statistical modelling tailored for small datasets has shown promise in showcasing disease heterogeneity[42]. Because large datasets are critical for contemporary machine learning methods, such as CNNs, there is a need for alternative techniques when data banks are insufficient to train the model. As such, our ML and statistical framework will allow for discovery of the non-linear ways in which groups of genes may interact to drive disease heterogeneity. This framework is designed for small datasets, which presents as a novel way of hypothesizing genetic subpopulations that may result in pathanogenesis. Our findings support genes previously reported to distinguish ADC and SCC subtypes; however, the novelty of this work lies in the machine's ability to discover previously unknown subpopulations that are defined by several genes at a time. These findings shed light on the different mechanisms at play within these subtypes as well as highlight novel potential therapeutic interventions.

2. Materials and Methods

Datasets

The dataset consisted of 58 samples of ADC and SCC (GSE10245) and 45 samples of human lung cancer and controls (GSE18842) to obtain a total of 103 samples. Only GSE10245 was used when analyzing gene expression levels for discriminating differences between sex as this data was omitted from GSE18842. Genetic expression levels denote relative RMA-calculated signal intensity [43]. Bar plots represent the mean expression level and error bars represent the standard deviation of the pooled data from each probe ID.

Machine Intelligence

In this study, we used a proprietary tool to organize the resulting models from several well-known machine learning methods to explore NSCLC genetic heterogeneity within a small dataset. This organizational technique was used to extract insights from models that could then be compared with statistical methods suitable for small data. An interactive hypothesis-generating interface was used such that human interaction could facilitate the analysis of different models [44,45]. This methodology allows the user to explore hypotheses generated by the unsupervised clustering methods of the system. For the work reported in this paper, we only utilized the following process, coupled with a proprietary tool of organizing the resulting models:

1) Feature selection was performed via standard univariate variable reduction methods and ensemble trees (Random Forest) through cross-validation [46,47]. The only dependent variables used were ADC vs SCC.
2) Principal components were utilized as a linear unsupervised clustering method to reveal obvious subpopulation structures.
3) The loadings from the principal components were utilized to reduce the variables.
4) Using the t-SNE [48] and UMAP [49] algorithms, we were able to extract subpopulations.
5) We then collected the sample IDs from the clusters formed from these two clustering models, systematically compared each group with the others, and then applied statistical methods to determine differentially expressed gene candidates.
6) A proprietary mathematical system was utilized to capture the models created up to this point in order to create maps. The advantage of using the NetraAI system is that it is easier to explore the subpopulations found to extract precisely which genes are most significant. In order to determine the significance of a gene, a standard Student t-test was used when two subpopulations were compared, and if more than two subpopulations were compared, an ANOVA was used.

Clustering was performed via principal components, t-SNE, and UMAP and these were the basis of the maps found in this paper. Some proprietary algorithms were used to organize the resulting clustering models, in addition to the random forest models, so that we were able to explore the models interactively to derive a deeper understanding of the driving genes behind the sub-clusters [44]. The NetraAI system goes beyond these capabilities, but we did not utilize these proprietary methods to maintain academic standards. By allowing ourselves to use the proprietary organization methods provided by the NetraAI, we were able to identify subpopulations that we could compare with statistical methods suitable for a dataset with so few samples and avoid overfitting that often comes with utilizing machine learning methods with small datasets.

3. Results

3.1 Machine learning identifies differentially expressed genes from a small NSCLC dataset

Using the ADC and SCC tumor gene expression data, our algorithm was able to generate a map distinguishing SCC (blue) and ADC subjects (red) (Figure 1). The genes that were found to have driven this distinction were *DSC3*, *VSNL1*, *SLC6A10P*, *IRF6*, *DST*, *CLCA2*, *DSG3*, *LPCAT1* and *PIGX*. Previous studies have reported on differentially expressed genes in ADC and SCC. Here, we identified 17 genes that discriminate between SCC and ADC (Table 1). It is noteworthy that 16 of the 17 genes we identified have been previously reported to be differentially expressed in SCC and ADC, validating our methods. Interestingly, we found genes associated with gap junctions and tight junctions to be strong driving forces differentiating SCC and ADC. It is worth mentioning that *PIGX* was the only gene identified that has not been previously associated with NSCLC. Although there have been reports that *PIGX* promotes cancer cell proliferation by suppressing *EHD2* and *ZIC1*, this warrants further investigation [50].

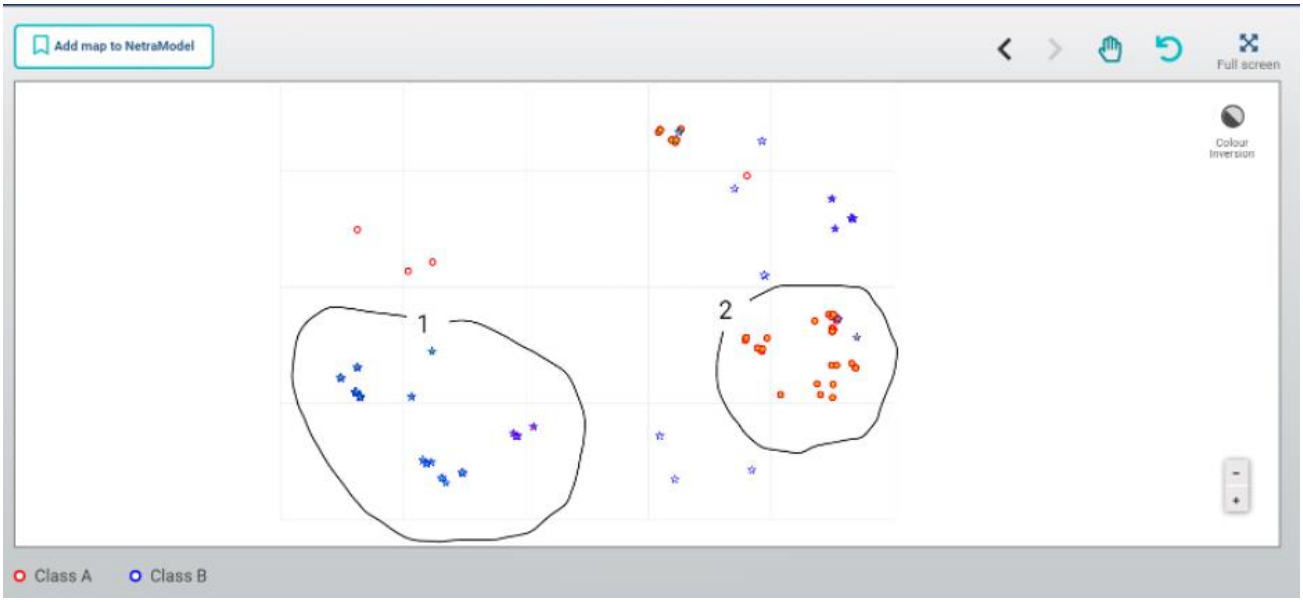

Figure 1. Algorithm-generated map of NSCLC patients stratified into SCC and ADC. *SCC (blue) and ADC subjects (red) delineated by user-defined segmentation of unsupervised patient map. The encircled groups were distinguished by DSC3, VSNL1, SLC6A10P, IRF6, DST, CLCA2, DSG3, LPCAT1 and PIGX.*

Table 1 Genes identified as discriminating between squamous cell carcinoma and adenocarcinoma.

Gene Symbol	Gene Name	Description	Reference
DSC3	Desmocollin-3	Ca^{2+}-dependent glycoprotein involved in cell adherence	[51]
VSNL1	Visinin-like protein 1	Neuronal Ca^{2+} sensor protein; tumor suppressor gene	[52,53]
IRF6	Interferon regulatory transcription factor 6	Transcription factor	[54]
DST	Dystonin	Cell adhesion	[55]
CLCA2	Chloride channel accessory 2	Cell adhesion; tumor suppressor	[56]
PIGX	Phosphatidylinositol glycan anchor biosynthesis, class X	Tumor suppressor	
DSG3	Desmoglein 3	Cell adhesion	[57-59]

LPCAT1	Lysophosphatidylcholine acyltransferase 1	Cancer progression and metastasis	[60,61]
GJB5/CX31.1	Gap junction protein beta 5	Intracellular communication, gap junction protein	[62-64]
SLC16A1	Solute carrier family 16 member 1	Cell metabolism	[65]
BNC1	Zinc finger protein basonuclin-1	Keratinocyte proliferation	[66]
GBP6	Guanylate binding protein family member 6		[56]
SLC6A10P	Solute carrier family 6 member 10	Neurotransmitter transporter; pseudogene of SLC6A8	[67]
KRT5	Keratin 5	Cytoskeleton and structural support	[68]
TRIM29	Tripartite motif-containing 29	Migration and invasion	[69]
KRT17	Keratin 17	Cytoskeleton and structural support	[70]
CGN	Cingulin	Tight junction	[64]

3.2 ADC and SCC are associated with distinct cellular adhesion molecules

Reports of SCC being characterized by the upregulation of desmosome and gap junction genes and ADC characterized by the upregulation of tight junction genes suggest that NSCLC subtypes are associated with a distinct set of adhesion molecules [64]. Here, we found that SCC was associated with cell adhesion marker *DSC3*, and ADC was associated with tight junction marker *CGN* (Figure 2). We identified two probes corresponding to *DSC3*, 206032_at and 206033_s_at. There was a statistically significant association of both *DSC3* probes with SCC ($p < 0.0001$) (Figure 2A). Interestingly, the elevated expression of *DSC3* was associated with males; however, this was not statistically significant ($p = 0.062$ for 206032_at and $p = 0.077$ for 206033_s_at) (Figure 2B). In contrast, the two probes corresponding to *CGN*, 223232_s_at and 223233_s_at were significantly associated with ADC ($p < 0.0001$) (Figure 2C). The *CGN* probes were significantly associated with females ($p = 0.014$) (Figure 2D). The variability of adhesion

molecule expression across sex warrants further investigation to elucidate the details of the correlation and advance towards gender related precision medicine.

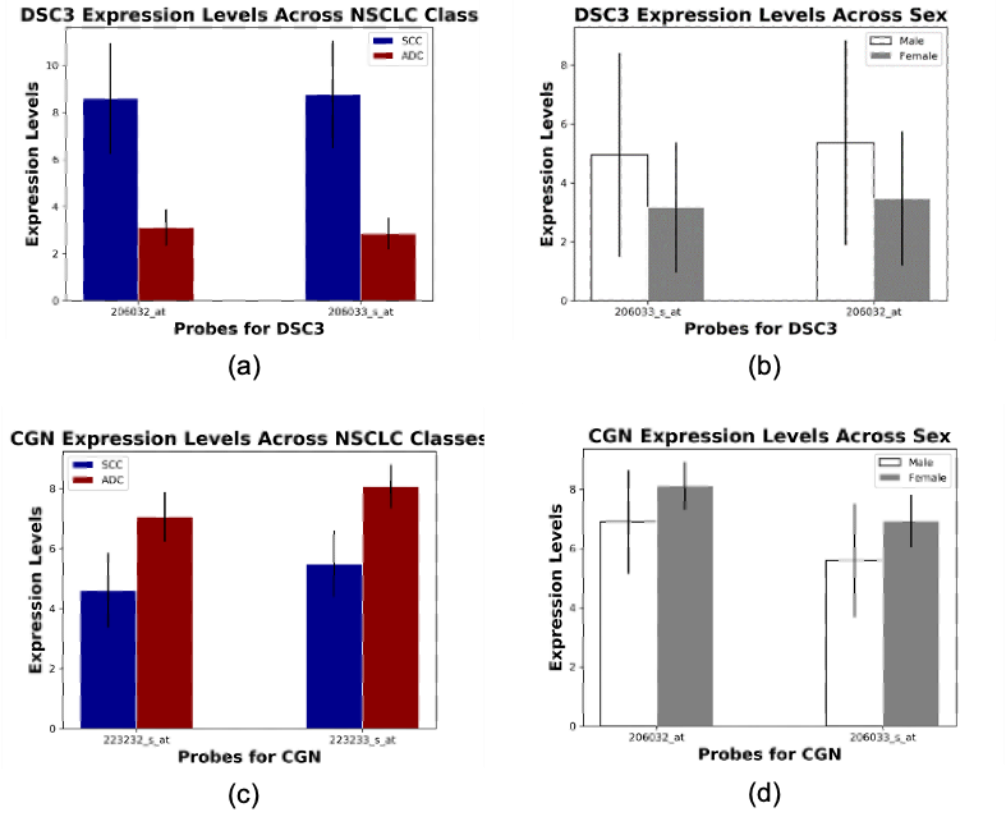

Figure 2. Differential expression of DSC3 and CGN in NSCLC. *(A) Expression level of DSC3 probes (mean ± SD), 206032_at and 206033_s_at, in SCC and ADC. (B) DSC3 probes, 206032_at and 206033_s_at, expression levels in males and females. (C) Expression level of CGN probes, 223232_s_at and 223233_s_at, in SCC and ADC. (D) CGN probes, 223232_s_at and 223233_s_at, expression levels in males and females. Abbreviations: NSCLC, non-small cell lung cancer; SD, standard deviation; SCC, squamous cell carcinoma; ADC, adenocarcinoma.*

3.3 SLC6A10P is a key driver of a more aggressive ADC subtype

Elevated *SLC6A10P* was significantly associated with two subgroups of ADC ($p<0.0001$) (Figure 3), in line with previous reports [54,67]. Interestingly, increased expression of the pseudogene *SLC6A10P* in ADC has been associated with increased metastatic risk and reported to be a significant predictor of poor clinical outcome [67]. Our ML methodology was able to reveal subpopulations of ADC subjects that are uniquely classified by *SLC6A10P* ($p = 1.3\times10^{-9}$), in an unsupervised way. This demonstrates the potential power of machine intelligence to reveal aetiologies within complex diseases, even when a small number of samples are present. However, the methods must be used to reveal subpopulations that can then be compared using appropriate statistical methods suitable for comparing small groups.

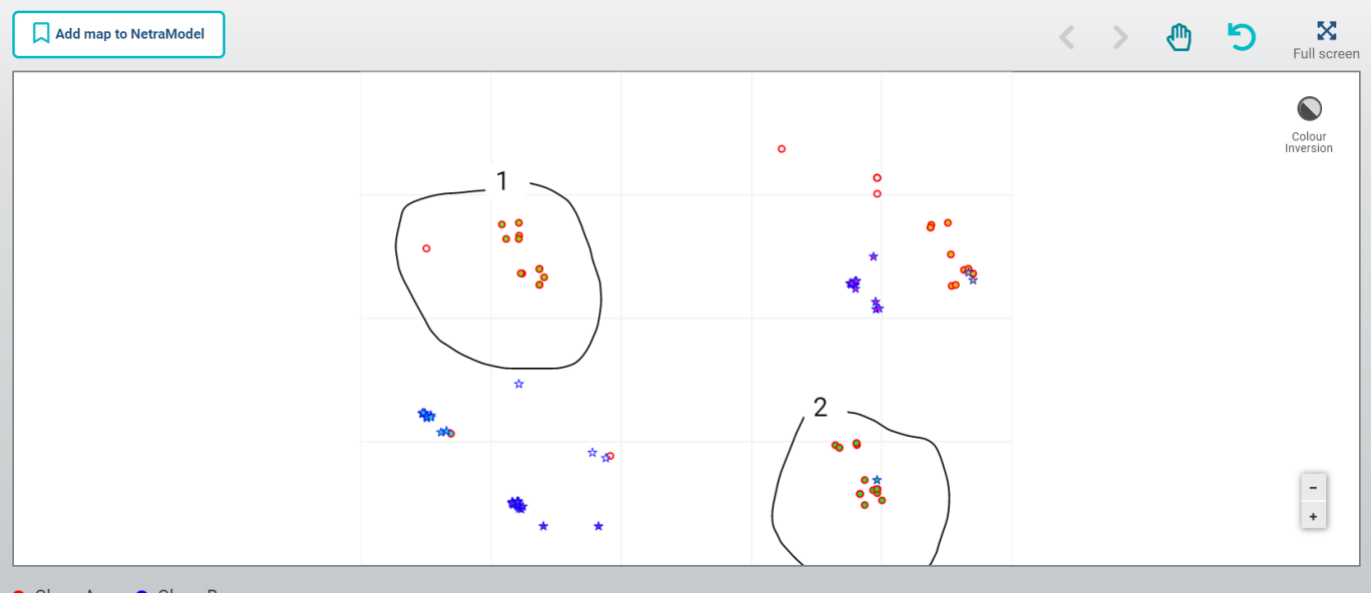

Figure 3. Algorithm-generated map of ADC subject subgroups. *ADC subject (red) subgroups delineated by user-defined segmentation of unsupervised patient map. The NetraAI tool organized the resulted models in order to reveal that the two encircled ADC patient subgroups were driven by SLC6A10P. Abbreviations: ADC, adenocarcinoma.*

3.4 IRF6 and CLCA2 drive unique subpopulations of SCC

Consistent with previous reports, we found two distinct subpopulations of SCC were found to be driven by *IRF6* and *CLCA2* (Figure 4) [54,56]. *IRF6* and *CLCA2* expression levels were higher in SCC than ADC ($p<0.0001$) (Figure 4B and 4C). The significance value between the *CLCA2* and *IRF6* probes in the two encircled SCC groups were evaluated to be 4.4×10^{-10}, 5.8×10^{-7}, 9.3×10^{-11} and 0.046 for 206164_at, 206165_s_at, 206166_s_at and 1552477_a_at probes, respectively.

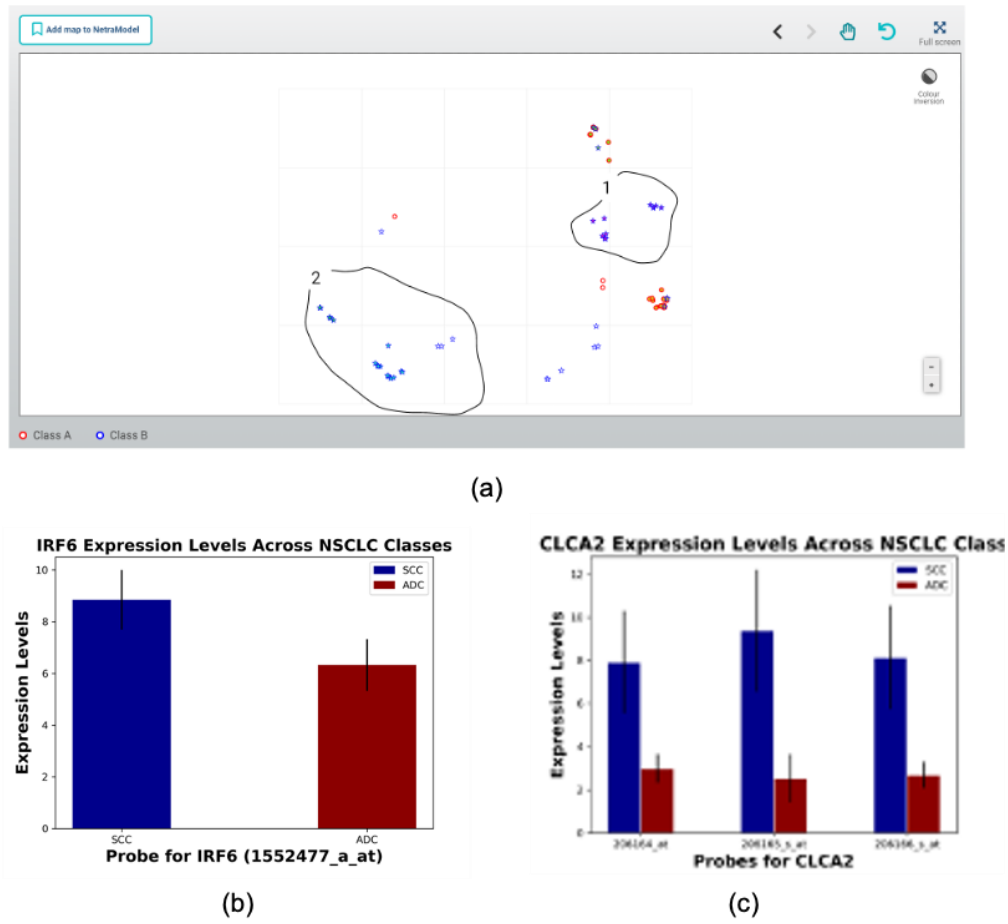

Figure 4. Differential expression of IRF6 and CLCA2 in NSCLC. *(A) Algorithm hypothesized that encircled patient SCC subgroups were driven by IRF6 and CLCA2. (B) Expression level of the IRF6 probe (mean ± SD), 1552477_a_at in SCC and ADC. (C) Expression level of CLCA2 probes 206164_at, 206165_s_at, and 206166_s_at in SCC and ADC (mean ± SD). Abbreviations: NSCLC, non-small cell lung cancer; SCC, squamous cell carcinoma; ADC, adenocarcinoma; SD, standard deviation.*

4. Discussion

This study highlights the genetic heterogeneity within NSCLC subtypes. Using a small dataset, we were able to identify a set of 17 genes that distinguish between SCC and ADC (Table1). Within these 17 genes, several have been previously reported to be associated either with NSCLC or a specific NSCLC, validating our ML approach. These findings were inline with previous reports on SCC genes being associated with the organization and assembly of cell and gap junctions, glutathione conjugation and the redox stress response, ECM organization and collagen-related proteins, interferon and cytokine signaling, and HLA downregulation and ADC genes associated with ECM organization proteins and complement, interferon and cytokine signaling, and collagen-related genes and proteins for ECM organization [62]. Another study identified epidermis development, cell division, and epithelial cell differentiation as the most common categories characterizing SCC, and cell adhesion enrichment, biological adhesion, and coagulation for ADC [63]. However, some of the genes we identified have not been previously associated with NSCLC or a specific subtype and represent areas that warrant greater

investigation for the advancement of precision medicine in NSCLC. Below, the genes of interest found using our methodology are highlighted in the context of previous findings in NSCLC.

The first of the previously reported NSCLC-associated genes we identified, *DSC3,* plays a role in epidermal morphology and keratinocyte proliferation [51]. There are several studies that report on *DSC3* distinguishing ADC and SCC, with a higher expression in SCC [71-74]. Interestingly, there has been a report on the association between *DSC3* and tumor suppressor activity in NSCLC mediated by inhibition of *EGFR* [75]. However, there remain contradictory associations with *DSC3* and prognosis, with elevated levels associated with increased metastatic risk in melanoma and better prognosis in lung and colon cancer [73]. This suggests that the same molecule may have differential effects in the tumor microenvironment (TME), which presents as an interesting field of research to understand how *DSC3* expression correlates with NSCLC subtypes depending on where they originate in the lung.

VSNL1 codes for the calcium-sensor protein VILIP1. Lower *VSNL1* expression has been correlated with poor clinical outcomes in NSCLC patients [52]. VILIP-1 has been reported to be decreased or undetectable in aggressive and invasive SCC, while less aggressive SCC displayed VILIP-1 expression [52]. There is evidence linking decreased VILIP1 expression to increased cell motility and malignancy, suggesting that VSNL1 downregulation promotes SCC tumor invasiveness [76].

Although a direct role of *IRF6* in lung cancer has not been identified, studies suggest that *IRF6* is a crucial regulator of the cell cycle, promoting progression to the G_0 state and allowing for uncontrolled cell proliferation [51]. Decreased *IRF6* expression has been associated with poor prognosis of gastric cancer and increased invasiveness of breast cancer [77,78]. Interestingly, SLC6A10P was the single gene that we found to drive two specific subtypes of ADC. SLC6A10P was previously found to be a marker for aggressive ADC [67], and recently, involved in the Notch signaling pathway [79]. Our findings suggest that SLC6A10P warrants further investigation as a genetic biomarker in the context the ADC patient subpopulation. *DST* and *DSC3* have been increasingly reported to be highly expressed in both ADC and SCC. Overexpression of these desmosomal genes is associated with increased $CD8^+$ T-cell infiltration in ADC [73].

CLCA2 has been implicated as a negative regulator of cancer cell migration [80]. In the lung, *CLCA2* has been reported to be highly expressed in SCC, suggesting that it may serve as a diagnostic marker to differentiate SCC from ADC. Female patients with *CLCA2*-negative SCC exhibited significantly poorer prognoses [56].

DSG3 has been reported to play a role in SCC and has been used as a sensitive and specific marker for SCC and is an effective discriminator between SCC and ADC [57,58]. Higher *DSG3* expression correlated with lower survival in SCC [59]. DSG3 and KRT5 have been reported to be downregulated in AC [68].

LPCAT1 has recently been shown to be overexpressed in lung SCC and associated with decreased OS [60]. In lung ADC, gene overexpression was associated with higher probabilities of ADC metastasis and poor clinical outcomes [61].

There is evidence that supports the role of cell adhesion proteins in both ADC and SCC. However, *GJB5* has been implicated in SCC mechanisms and is associated with gap junctions [62]. It is not surprising that there is a higher expression of *GJB5* in SCC as it is primarily associated with gap junctions (Figure 2) [63,64]. *GJB5* (gap junction protein beta 5 or protein-coding gene: Cx31.1) is involved in intercellular communication related to epidermal differentiation and environmental sensing. Cx31.1 was found to be downregulated in NSCLC

with expression levels reversely related to metastatic potential, suggesting it inhibits malignant properties of NSCLC cell lines. Cx31.1 is colocalized with LC3-II (autophagy marker light chain 3) and acts like a tumor suppressor as it plays a role in the regulation of cell proliferation, cell differentiation, tissue development and apoptosis [63].

TRIM29 has been shown to be upregulated in NSCLC, and may be a marker for tumour aggressiveness [81]. It was been further associated with poorer histological grade and clinical and outcomes in SCC [69]. It was been suggested this may be due to the inhibition of p53 via TRIM29 [82].

KRT17 overexpression has been associated with both subtypes of NSCLC, but was significantly correlated to more advanced tumour grade, lymph node metastatic potential, and overall survival in ADC [70].

BCN1 has been reported to be hypermethylated in NSCLC tissue [66]. Furthermore, decreased expression of *BNC1* has been observed in other carcinomas [83]. Aberrant *BNC1* and *BNC2* expression contribute to tumor progression [84].

Reports of upregulation of desmosomes and gap junctions in SCC and tight junctions in ADC suggest that SCC and ADC are characterized by a distinct set of adhesion molecules [64]. Here, we found that ADC was identified by *CGN* and SCC by *DSC3* (Figure 2). *CGN* (cingulin) is involved in the organization of tight junctions and is downregulated in SCC [64]. In contrast, ADC has been reported to be characterized by tight junctions, while SCC is characterized by gap junctions.

In addition to the 17 genes identified differentially expressed genes in ADC and SCC, *PTGFRN* (prostaglandin F2 receptor negative regulator; CD315) was also found to be associated with ADC. *PTGFRN* has been reported to be associated with worse survival in glioblastoma, while inhibition has been associated with decreased proliferation and tumor growth [85,86]. *PTGFRN* inhibits the binding of prostaglandin F2α to its receptor. Interestingly, there are reports that *PTGFRN* is associated with small cell lung cancer; however, the role remains unknown [87,88].

IRF6 and CLCA2 have previously been implicated in lung SCC [54,56]. CLCA2 in particular was highlighted to differentiate ADC and SCC. Furthermore, SCC was expressed was correlated to tumour grade upon histological characterization. In particular, *CLCA2* negative samples were associated with poorly differentiated tumours [56].

Males have been reported to have a significantly poorer NSCLC prognosis compared to women, shifting efforts towards sex-based approaches to diagnosis, prognosis, and therapeutic interventions [89,90]. Additionally, estrogens have been associated with increased risk in ADC in women despite equal expression of estrogen receptors α and β, however, the role remains unclear [91]. While there are several reports on the sex-based differences in cancer mechanisms, including differences in metabolism, immunity, and angiogenesis, differences in *CGN* and *DSC3* expression have not been previously reported, to the best of our knowledge [92]. Gap junction proteins, also known as connexins, serves as channels that connect the interior of adjacent cells, facilitating intracellular homeostasis and coordination of activities via second messengers [93]. Desmosomes primarily provide mechanical strength via a structural network. In contrast, tight junctions form a barrier around the cell, regulating permeability of the paracellular space [94,95]. These molecules play critical roles in epithelial-to-mesenchymal transition (EMT), a process involved in cancer metastasis. Though no sex-based differences have been reported, this presents as a unique field of research, as there may be different druggable targets for men and women.

Finally, the phosphatidylinositol glycan anchor biosynthesis class gene, *PIGX*, was found to be a driver of ADC and SCC differentiation in several instances (Figure 1). Little is known about the role of *PIGX* in NSCLC. However, it has been noted that *PIGX* has a proliferative role when expressed in breast cancer cells [50]. In addition, authors found higher *PIGX* expression was associated with shorter recurrence-free survival. This suggests that this gene plays a role in NSCLC that warrants further study to determine if it is a druggable target or a biomarker.

5. Conclusions

Genetic data sets are expensive to acquire, and therefore, many of them are small, *i.e.*, containing few samples. Machine intelligence methods are becoming popular, but a major problem with machine learning, especially powerful methods like ensemble trees and deep neural networks, is that they require thousands of samples to create robust predictive models capable of generalizing. This is a real challenge within the medical sciences, and so research programs have started to explore the ability of machine intelligence to make an impact on small datasets.

The approach utilized here to derive the insights relied on the ability for certain machine learning methods to create hypotheses about subpopulations of patients, and then to statistically test the driving variables of these subgroups of patients. In this way, we utilize machine learning to derive potential insights and then utilize statistical methods that are suitable for small data to evaluate differential expression. In order to create robust predictive models with machine intelligence, one requires large data sets, but here we utilized the ability for some of these methods to create hypotheses instead, and then use methods appropriate for small data to test these hypotheses. This bidirectional attack allowed us to derive insights from these small data that have been previously validated and to derive a new potential role for the gene *PIGX* in NSCLC.

References

1. Ridge, C.A.; McErlean, A.M.; Ginsberg, M.S. Epidemiology of lung cancer. In Proceedings of Seminars in interventional radiology; p. 93.
2. Thomas, A.; Liu, S.V.; Subramaniam, D.S.; Giaccone, G. Refining the treatment of NSCLC according to histological and molecular subtypes. *Nature reviews Clinical oncology* **2015**, *12*, 511.
3. Lawrence, M.S.; Stojanov, P.; Mermel, C.H.; Robinson, J.T.; Garraway, L.A.; Golub, T.R.; Meyerson, M.; Gabriel, S.B.; Lander, E.S.; Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **2014**, *505*, 495-501.
4. Pikor, L.A.; Ramnarine, V.R.; Lam, S.; Lam, W.L. Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung cancer* **2013**, *82*, 179-189.
5. Manegold, C. Treatment algorithm in 2014 for advanced non-small cell lung cancer: therapy selection by tumour histology and molecular biology. *Advances in medical sciences* **2014**, *59*, 308-313.
6. Carnio, S.; Novello, S.; Bironzo, P.; Scagliotti, G.V. Moving from histological subtyping to molecular characterization: new treatment opportunities in advanced non-small-cell lung cancer. *Expert review of anticancer therapy* **2014**, *14*, 1495-1513.
7. Yu, L.; Tao, G.; Zhu, L.; Wang, G.; Li, Z.; Ye, J.; Chen, Q. Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. *BMC cancer* **2019**, *19*, 1-12.
8. Tau, N.; Stundzia, A.; Yasufuku, K.; Hussey, D.; Metser, U. Convolutional neural networks in predicting nodal and distant metastatic potential of newly diagnosed non–small cell lung cancer on FDG PET images. *American Journal of Roentgenology* **2020**, *215*, 192-197.
9. Kriegsmann, M.; Haag, C.; Weis, C.-A.; Steinbuss, G.; Warth, A.; Zgorzelski, C.; Muley, T.; Winter, H.; Eichhorn, M.E.; Eichhorn, F. Deep Learning for the Classification of Small-Cell and Non-Small-Cell Lung Cancer. *Cancers* **2020**, *12*, 1604.
10. Mu, W.; Jiang, L.; Zhang, J.; Shi, Y.; Gray, J.E.; Tunali, I.; Gao, C.; Sun, Y.; Tian, J.; Zhao, X. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nature communications* **2020**, *11*, 1-11.
11. Rabbani, M.; Kanevsky, J.; Kafi, K.; Chandelier, F.; Giles, F.J. Role of artificial intelligence in the care of patients with nonsmall cell lung cancer. *European journal of clinical investigation* **2018**, *48*, e12901.
12. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214-218.
13. Podolsky, M.D.; Barchuk, A.A.; Kuznetcov, V.I.; Gusarova, N.F.; Gaidukov, V.S.; Tarakanov, S.A. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention* **2016**, *17*, 835-838.
14. 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*, 303-310.
15. Yuan, F.; Lu, L.; Zou, Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta Molecular Basis of Disease* **2020**, *1866*, 165822.

16. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. *J Med Syst* **2018**, *42*, 226, doi:10.1007/s10916-018-1088-1.
17. Balagurunathan, Y.; Kumar, V.; Gu, Y.; Kim, J.; Wang, H.; Liu, Y.; Goldgof, D.B.; Hall, L.O.; Korn, R.; Zhao, B., et al. Test-retest reproducibility analysis of lung CT image features. *Journal of Digital Imaging* **2014**, *27*, 805-823, doi:10.1007/s10278-014-9716-x.
18. Wang, S.; Chen, A.; Yang, L.; Cai, L.; Xie, Y.; Fujimoto, J.; Gazdar, A.; Xiao, G. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific Reports* **2018**, *8*, 10393, doi:10.1038/s41598-018-27707-4.
19. Li, Z.; Hu, Z.; Xu, J.; Tan, T.; Chen, H.; Duan, Z.; Liu, P.; Tang, J.; Cai, G.; Ouyang, Q., et al. Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study. *arXiv:1803.05471 [cs]* **2018**.
20. Ma, Y.; Feng, W.; Wu, Z.; Liu, M.; Zhang, F.; Liang, Z.; Cui, C.; Huang, J.; Li, X.; Guo, X. Intra-tumoural heterogeneity characterization through texture and colour analysis for differentiation of non-small cell lung carcinoma subtypes. *Physics in Medicine and Biology* **2018**, *63*, 165018, doi:10.1088/1361-6560/aad648.
21. E, L.; Lu, L.; Li, L.; Yang, H.; Schwartz, L.H.; Zhao, B. Radiomics for Classification of Lung Cancer Histological Subtypes Based on Nonenhanced Computed Tomography. *Academic Radiology* **2019**, *26*, 1245-1252, doi:10.1016/j.acra.2018.10.013.
22. Bilaloglu, S.; Wu, J.; Fierro, E.; Sanchez, R.D.; Ocampo, P.S.; Razavian, N.; Coudray, N.; Tsirigos, A. Efficient pan-cancer whole-slide image classification and outlier detection using convolutional neural networks. *bioRxiv* **2019**, 10.1101/633123, 633123, doi:10.1101/633123.
23. Sherafatian, M.; Arjmand, F. Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data. *Oncology Letters* **2019**, *18*, 2125-2131, doi:10.3892/ol.2019.10462.
24. Travis, W.D.; Brambilla, E.; Burke, A.P.; Marx, A.; Nicholson, A.G. Introduction to The 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart. *Journal of Thoracic Oncology* **2015**, *10*, 1240-1242, doi:10.1097/JTO.0000000000000663.
25. Tsao, M.-S.; Marguet, S.; Le Teuff, G.; Lantuejoul, S.; Shepherd, F.A.; Seymour, L.; Kratzke, R.; Graziano, S.L.; Popper, H.H.; Rosell, R., et al. Subtype Classification of Lung Adenocarcinoma Predicts Benefit From Adjuvant Chemotherapy in Patients Undergoing Complete Resection. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **2015**, *33*, 3439-3446, doi:10.1200/JCO.2014.58.8335.
26. Hung, J.-J.; Yeh, Y.-C.; Jeng, W.-J.; Wu, K.-J.; Huang, B.-S.; Wu, Y.-C.; Chou, T.-Y.; Hsu, W.-H. Predictive value of the international association for the study of lung cancer/American Thoracic Society/European Respiratory Society classification of lung adenocarcinoma in tumor recurrence and patient survival. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **2014**, *32*, 2357-2364, doi:10.1200/JCO.2013.50.1049.
27. Wei, J.W.; Tafe, L.J.; Linnik, Y.A.; Vaickus, L.J.; Tomita, N.; Hassanpour, S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports* **2019**, *9*, 3358, doi:10.1038/s41598-019-40041-7.
28. Gertych, A.; Swiderska-Chadaj, Z.; Ma, Z.; Ing, N.; Markiewicz, T.; Cierniak, S.; Salemi, H.; Guzman, S.; Walts, A.E.; Knudsen, B.S. Convolutional neural networks can accurately

distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific Reports* **2019**, *9*, 1483, doi:10.1038/s41598-018-37638-9.
29. Terra, S.B.; Jang, J.S.; Bi, L.; Kipp, B.R.; Jen, J.; Yi, E.S.; Boland, J.M. Molecular characterization of pulmonary sarcomatoid carcinoma: analysis of 33 cases. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* **2016**, *29*, 824-831, doi:10.1038/modpathol.2016.89.
30. Pérez-Soler, R.; Chachoua, A.; Hammond, L.A.; Rowinsky, E.K.; Huberman, M.; Karp, D.; Rigas, J.; Clark, G.M.; Santabárbara, P.; Bonomi, P. Determinants of tumor response and survival with erlotinib in patients with non--small-cell lung cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **2004**, *22*, 3238-3247, doi:10.1200/JCO.2004.11.057.
31. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* **2018**, *24*, 1559-1567, doi:10.1038/s41591-018-0177-5.
32. Noorbakhsh, J.; Farahmand, S.; Soltanieh-ha, M.; Namburi, S.; Zarringhalam, K.; Chuang, J. Pan-cancer classifications of tumor histological images using deep learning. *bioRxiv* **2019**, 10.1101/715656, 715656, doi:10.1101/715656.
33. Koyasu, S.; Nishio, M.; Isoda, H.; Nakamoto, Y.; Togashi, K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. *Annals of Nuclear Medicine* **2020**, *34*, 49-57, doi:10.1007/s12149-019-01414-0.
34. Shiri, I.; Maleki, H.; Hajianfar, G.; Abdollahi, H.; Ashrafinia, S.; Hatt, M.; Oveisi, M.; Rahmim, A. Next Generation Radiogenomics Sequencing for Prediction of EGFR and KRAS Mutation Status in NSCLC Patients Using Multimodal Imaging and Machine Learning Approaches. *arXiv:1907.02121 [physics, q-bio]* **2019**, 10.1007/s11307-020-01487-8, doi:10.1007/s11307-020-01487-8.
35. Yu, K.-H.; Wang, F.; Berry, G.J.; Ré, C.; Altman, R.B.; Snyder, M.; Kohane, I.S. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. *Journal of the American Medical Informatics Association: JAMIA* **2020**, *27*, 757-769, doi:10.1093/jamia/ocz230.
36. Lai, Y.-H.; Chen, W.-N.; Hsu, T.-C.; Lin, C.; Tsao, Y.; Wu, S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports* **2020**, *10*, 4679, doi:10.1038/s41598-020-61588-w.
37. Goldstraw, P.; Chansky, K.; Crowley, J.; Rami-Porta, R.; Asamura, H.; Eberhardt, W.E.; Nicholson, A.G.; Groome, P.; Mitchell, A.; Bolejack, V., et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* **2016**, *11*, 39-51, doi:10.1016/j.jtho.2015.09.009.
38. Tian, S. Identification of monotonically differentially expressed genes for non-small cell lung cancer. *BMC Bioinformatics* **2019**, *20*, 177, doi:10.1186/s12859-019-2775-8.
39. Jin, T.; Talos, F.; Wang, D. ECMarker: Interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *bioRxiv* **2019**, 825414.

40. Yuan, F.; Lu, L.; Zou, Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2020**, *1866*, 165822, doi:10.1016/j.bbadis.2020.165822.
41. Hu, F.; Zhou, Y.; Wang, Q.; Yang, Z.; Shi, Y.; Chi, Q. Gene Expression Classification of Lung Adenocarcinoma into Molecular Subtypes. *IEEE/ACM transactions on computational biology and bioinformatics* **2020**, *17*, 1187-1197, doi:10.1109/TCBB.2019.2905553.
42. Robinson, G.A.; Peng, J.; Dönnes, P.; Coelewij, L.; Naja, M.; Radziszewska, A.; Wincup, C.; Peckham, H.; Isenberg, D.A.; Ioannou, Y., et al. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: patient stratification using a machine-learning approach. *Lancet Rheumatol* **2020**, *2*, e485-e496, doi:10.1016/S2665-9913(20)30168-5.
43. Wu, Z.; Irizarry, R.A.; Gentleman, R.; Martinez-Murillo, F.; Spencer, F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **2004**, *99*, 909-917, doi:10.1198/016214504000000683.
44. Qorri, B.; Tsay, M.; Agrawal, A.; Au, R.; Geraci, J. Using machine intelligence to uncover Alzheimer's disease progression heterogeneity. *Exploration of Medicine* **2020**, *1*, 377-395, doi:10.37349/emed.2020.00026.
45. Tsay, M.; Geraci, J.; Agrawal, A. *Next-Gen AI for Disease Definition, Patient Stratification, and Placebo Effect*; OSF Preprints: 2020/04/06/T02:51:09.502Z, 2020.
46. Lai, C.; Reinders, M.J.; van't Veer, L.J.; Wessels, L.F. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics* **2006**, *7*, 235, doi:10.1186/1471-2105-7-235.
47. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99*, 323-329, doi:10.1016/j.ygeno.2012.04.003.
48. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579--2605.
49. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.
50. Nakakido, M.; Tamura, K.; Chung, S.; Ueda, K.; Fujii, R.; Kiyotani, K.; Nakamura, Y. Phosphatidylinositol glycan anchor biosynthesis, class X containing complex promotes cancer cell proliferation through suppression of EHD2 and ZIC1, putative tumor suppressors. *International journal of oncology* **2016**, *49*, 868-876.
51. Sanchez-Palencia, A.; Gomez-Morales, M.; Gomez-Capilla, J.A.; Pedraza, V.; Boyero, L.; Rosell, R.; Fárez-Vidal, M.W. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International journal of cancer* **2011**, *129*, 355-364.
52. Fu, J.; Fong, K.; Bellacosa, A.; Ross, E.; Apostolou, S.; Bassi, D.E.; Jin, F.; Zhang, J.; Cairns, P.; De Caceres, I.I. VILIP-1 downregulation in non-small cell lung carcinomas: mechanisms and prediction of survival. *PLoS One* **2008**, *3*, e1698.
53. Gonzalez Guerrico, A.M.; Jaffer, Z.M.; Page, R.E.; Braunewell, K.-H.; Chernoff, J.; Klein-Szanto, A.J.P. Visinin-like protein-1 is a potent inhibitor of cell adhesion and migration in squamous carcinoma cells. *Oncogene* **2005**, *24*, 2307-2316, doi:10.1038/sj.onc.1208476.
54. Liu, Y.; Shao, G.; Yang, Z.; Lin, X.; Liu, X.; Qian, B.; Liu, Z. Interferon regulatory factor 6 correlates with the progression of non-small cell lung cancer and can be regulated by miR-320. *Journal of Pharmacy and Pharmacology* **2021**, 10.1093/jpp/rgab009, doi:10.1093/jpp/rgab009.

55. Chae, Y.K.; Choi, W.M.; Bae, W.H.; Anker, J.; Davis, A.A.; Agte, S.; Iams, W.T.; Cruz, M.; Matsangou, M.; Giles, F.J. Overexpression of adhesion molecules and barrier molecules is associated with differential infiltration of immune cells in non-small cell lung cancer. *Scientific Reports* **2018**, *8*, doi:10.1038/s41598-018-19454-3.
56. Shinmura, K.; Igarashi, H.; Kato, H.; Kawanishi, Y.; Inoue, Y.; Nakamura, S.; Ogawa, H.; Yamashita, T.; Kawase, A.; Funai, K., et al. CLCA2 as a Novel Immunohistochemical Marker for Differential Diagnosis of Squamous Cell Carcinoma from Adenocarcinoma of the Lung. *Disease Markers* **2014**.
57. Savci-Heijink, C.D.; Kosari, F.; Aubry, M.C.; Caron, B.L.; Sun, Z.; Yang, P.; Vasmatzis, G. The role of desmoglein-3 in the diagnosis of squamous cell carcinoma of the lung. *Am J Pathol* **2009**, *174*, 1629-1637, doi:10.2353/ajpath.2009.080778.
58. Fukuoka, J.; Dracheva, T.; Shih, J.H.; Hewitt, S.M.; Fujii, T.; Kishor, A.; Mann, F.; Shilo, K.; Franks, T.J.; Travis, W.D., et al. Desmoglein 3 as a prognostic factor in lung cancer. *Hum Pathol* **2007**, *38*, 276-283, doi:10.1016/j.humpath.2006.08.006.
59. Dong, Y.; Li, S.; Sun, X.; Wang, Y.; Lu, T.; Wo, Y.; Leng, X.; Kong, D.; Jiao, W. Desmoglein 3 and Keratin 14 for Distinguishing Between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. *Onco Targets Ther* **2020**, *13*, 11111-11124, doi:10.2147/OTT.S270398.
60. Liu, F.; Wu, Y.; Liu, J.; Ni, R.J.; Yang, A.G.; Bian, K.; Zhang, R. A miR-205-LPCAT1 axis contributes to proliferation and progression in multiple cancers. *Biochem Biophys Res Commun* **2020**, *527*, 474-480, doi:10.1016/j.bbrc.2020.04.071.
61. Wei, C.; Dong, X.; Lu, H.; Tong, F.; Chen, L.; Zhang, R.; Dong, J.; Hu, Y.; Wu, G. LPCAT1 promotes brain metastasis of lung adenocarcinoma by up-regulating PI3K/AKT/MYC pathway. *J Exp Clin Cancer Res* **2019**, *38*, 95, doi:10.1186/s13046-019-1092-4.
62. Lucchetta, M.; da Piedade, I.; Mounir, M.; Vabistsevits, M.; Terkelsen, T.; Papaleo, E. Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response. *BMC Cancer* **2019**, *19*, 824, doi:10.1186/s12885-019-5965-x.
63. Wang, T.; Zhang, L.; Tian, P.; Tian, S. Identification of differentially-expressed genes between early-stage adenocarcinoma and squamous cell carcinoma lung cancer using meta-analysis methods. *Oncology Letters* **2017**, *13*, 3314-3322.
64. Kuner, R.; Muley, T.; Meister, M.; Ruschhaupt, M.; Buness, A.; Xu, E.C.; Schnabel, P.; Warth, A.; Poustka, A.; Sültmann, H. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung cancer* **2009**, *63*, 32-38.
65. Liu, H.Y.; Lu, S.R.; Guo, Z.H.; Zhang, Z.S.; Ye, X.; Du, Q.; Li, H.; Wu, Q.; Yu, B.; Zhai, Q., et al. lncRNA SLC16A1-AS1 as a novel prognostic biomarker in non-small cell lung cancer. *J Investig Med* **2020**, *68*, 52-59, doi:10.1136/jim-2019-001080.
66. Shames, D.S.; Girard, L.; Gao, B.; Sato, M.; Lewis, C.M.; Shivapurkar, N.; Jiang, A.; Perou, C.M.; Kim, Y.H.; Pollack, J.R., et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med* **2006**, *3*, e486, doi:10.1371/journal.pmed.0030486.
67. Yuan, K.; Gao, Z.-J.; Yuan, W.-D.; Yuan, J.-Q.; Wang, Y. High expression of SLC6A10P contributes to poor prognosis in lung adenocarcinoma. *International journal of clinical & experimental pathology* **2018**, *11*, 720.
68. Xiao, J.; Lu, X.; Chen, X.; Zou, Y.; Liu, A.; Li, W.; He, B.; He, S.; Chen, Q. Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma. *Oncotarget* **2017**, *8*, 71759-71771, doi:10.18632/oncotarget.17606.

69. Zhou, Z.Y.; Yang, G.Y.; Zhou, J.; Yu, M.H. Significance of TRIM29 and β-catenin expression in non-small-cell lung cancer. *J Chin Med Assoc* **2012**, *75*, 269-274, doi:10.1016/j.jcma.2012.04.015.
70. Wang, Z.; Yang, M.Q.; Lei, L.; Fei, L.R.; Zheng, Y.W.; Huang, W.J.; Li, Z.H.; Liu, C.C.; Xu, H.T. Overexpression of KRT17 promotes proliferation and invasion of non-small cell lung cancer and indicates poor prognosis. *Cancer Manag Res* **2019**, *11*, 7485-7497, doi:10.2147/CMAR.S218926.
71. Warth, A.; Muley, T.; Herpel, E.; Meister, M.; Herth, F.J.; Schirmacher, P.; Weichert, W.; Hoffmann, H.; Schnabel, P.A. Large-scale comparative analyses of immunomarkers for diagnostic subtyping of non-small-cell lung cancer biopsies. *Histopathology* **2012**, *61*, 1017-1025.
72. Tsuta, K.; Tanabe, Y.; Yoshida, A.; Takahashi, F.; Maeshima, A.M.; Asamura, H.; Tsuda, H. Utility of 10 immunohistochemical markers including novel markers (desmocollin-3, glypican 3, S100A2, S100A7, and Sox-2) for differential diagnosis of squamous cell carcinoma from adenocarcinoma of the Lung. *Journal of Thoracic Oncology* **2011**, *6*, 1190-1199.
73. Chae, Y.K.; Choi, W.M.; Bae, W.H.; Anker, J.; Davis, A.A.; Agte, S.; Iams, W.T.; Cruz, M.; Matsangou, M.; Giles, F.J. Overexpression of adhesion molecules and barrier molecules is associated with differential infiltration of immune cells in non-small cell lung cancer. *Scientific reports* **2018**, *8*, 1-10.
74. Angulo, B.; Suarez-Gauthier, A.; Lopez-Rios, F.; Medina, P.; Conde, E.; Tang, M.; Soler, G.; Lopez-Encuentra, A.; Cigudosa, J.; Sanchez-Cespedes, M. Expression signatures in lung cancer reveal a profile for EGFR-mutant tumours and identify selective PIK3CA overexpression by gene amplification. *The Journal of pathology* **2008**, *214*, 347-356.
75. Cui, T.; Chen, Y.; Yang, L.; Knösel, T.; Huber, O.; Pacyna-Gengelbach, M.; Petersen, I. The p53 target gene desmocollin 3 acts as a novel tumor suppressor through inhibiting EGFR/ERK pathway in human lung cancer. *Carcinogenesis* **2012**, *33*, 2326-2333.
76. Guerrico, A.M.G.; Jaffer, Z.M.; Page, R.E.; Braunewell, K.-H.; Chernoff, J.; Klein-Szanto, A.J. Visinin-like protein-1 is a potent inhibitor of cell adhesion and migration in squamous carcinoma cells. *Oncogene* **2005**, *24*, 2307-2316.
77. Li, D.; Cheng, P.; Wang, J.; Qiu, X.; Zhang, X.; Xu, L.; Liu, Y.; Qin, S. IRF6 is directly regulated by ZEB1 and ELF3, and predicts a favorable prognosis in gastric cancer. *Frontiers in oncology* **2019**, *9*, 220.
78. Bailey, C.M.; Khalkhali-Ellis, Z.; Kondo, S.; Margaryan, N.V.; Seftor, R.E.; Wheaton, W.W.; Amir, S.; Pins, M.R.; Schutte, B.C.; Hendrix, M.J.J.J.o.B.C. Mammary serine protease inhibitor (Maspin) binds directly to interferon regulatory factor 6: identification of a novel serpin partnership. **2005**, *280*, 34210-34217.
79. Feng, Y.; Guo, X.; Tang, H. SLC6A8 is involved in the progression of non-small cell lung cancer through the Notch signaling pathway. *Ann Transl Med* **2021**, *9*, 264, doi:10.21037/atm-20-5984.
80. Sasaki, Y.; Koyama, R.; Maruyama, R.; Hirano, T.; Tamura, M.; Sugisaka, J.; Suzuki, H.; Idogawa, M.; Shinomura, Y.; Tokino, T. CLCA2, a target of the p53 family, negatively regulates cancer cell migration and invasion. *Cancer Biology & Therapy* **2012**, *13*, 1512-1521, doi:10.4161/cbt.22280.
81. Song, X.; Fu, C.; Yang, X.; Sun, D.; Zhang, X.; Zhang, J. Tripartite motif-containing 29 as a novel biomarker in non-small cell lung cancer. *Oncol Lett* **2015**, *10*, 2283-2288, doi:10.3892/ol.2015.3623.

82. Yuan, Z.; Villagra, A.; Peng, L.; Coppola, D.; Glozak, M.; Sotomayor, E.M.; Chen, J.; Lane, W.S.; Seto, E. The ATDC (TRIM29) protein binds p53 and antagonizes p53-mediated functions. *Mol Cell Biol* **2010**, *30*, 3004-3015, doi:10.1128/MCB.01023-09.
83. Wu, Y.; Zhang, X.; Liu, Y.; Lu, F.; Chen, X. Decreased Expression of BNC1 and BNC2 Is Associated with Genetic or Epigenetic Regulation in Hepatocellular Carcinoma. *Int J Mol Sci* **2016**, *17*, doi:10.3390/ijms17020153.
84. Wu, Y.; Zhang, X.; Liu, Y.; Lu, F.; Chen, X. Decreased expression of BNC1 and BNC2 is associated with genetic or epigenetic regulation in hepatocellular carcinoma. *International journal of molecular sciences* **2016**, *17*, 153.
85. Marquez, J.; Dong, J.; Dong, C.; Tian, C.; Serrero, G. Identification of Prostaglandin F2 Receptor Negative Regulator (PTGFRN) as an internalizable target in cancer cells for antibody-drug conjugate development. *Plos one* **2021**, *16*, e0246197.
86. Aguila, B.; Morris, A.B.; Spina, R.; Bar, E.; Schraner, J.; Vinkler, R.; Sohn, J.W.; Welford, S.M. The Ig superfamily protein PTGFRN coordinates survival signaling in glioblastoma multiforme. *Cancer letters* **2019**, *462*, 33-42.
87. George, J.; Lim, J.S.; Jang, S.J.; Cun, Y.; Ozretić, L.; Kong, G.; Leenders, F.; Lu, X.; Fernández-Cuesta, L.; Bosco, G. Comprehensive genomic profiles of small cell lung cancer. *Nature* **2015**, *524*, 47-53.
88. Krushkal, J.; Silvers, T.; Reinhold, W.C.; Sonkin, D.; Vural, S.; Connelly, J.; Varma, S.; Meltzer, P.S.; Kunkel, M.; Rapisarda, A.J.C.e. Epigenome-wide DNA methylation analysis of small cell lung cancer cell lines suggests potential chemotherapy targets. **2020**, *12*, 1-28.
89. Wainer, Z.; Wright, G.M.; Gough, K.; Daniels, M.G.; Russell, P.A.; Choong, P.; Conron, M.; Ball, D.; Solomon, B.J.C.l.c. Sex-dependent staging in non–small-cell lung cancer; analysis of the effect of sex differences in the eighth edition of the Tumor, Node, Metastases Staging System. **2018**, *19*, e933-e944.
90. Radkiewicz, C.; Dickman, P.W.; Johansson, A.L.V.; Wagenius, G.; Edgren, G.; Lambe, M. Sex and survival in non-small cell lung cancer: A nationwide cohort study. *PLoS ONE* **2019**, *14*, e0219206.
91. Ivanova, M.M.; Mazhawidza, W.; Dougherty, S.M.; Klinge, C.M. Sex differences in estrogen receptor subcellular location and activity in lung adenocarcinoma cells. *American Journal of Respiratory Cell and Molecular Biology* **2010**, *42*, 320-330.
92. Rubin, J.B.; Lagas, J.S.; Broestl, L.; Sponagel, J.; Rockwell, N.; Rhee, G.; Rosen, S.F.; Chen, S.; Klein, R.S.; Imoukhuede, P. Sex differences in cancer mechanisms. *Biology of Sex Differences* **2020**, *11*, 1-29.
93. Ruch, R. Gap Junctions and Connexins in Cancer Formation, Progression, and Therapy. *Cancers* **2020**, *12*, 3307, doi:10.3390/cancers12113307.
94. Ylermi, S. Tight junctions in lung cancer and lung metastasis: a review. *International journal of clinical & experimental pathology* **2012**, *5*, 126.
95. Bhat, A.A.; Uppada, S.; Achkar, I.W.; Hashem, S.; Yadav, S.K.; Shanmugakonar, M.; Al-Naemi, H.A.; Haris, M.; Uddin, S. Tight junction proteins and signaling pathways in cancer and inflammation: a functional crosstalk. *Frontiers in physiology* **2019**, *9*, 1942.