Exploration of Medicine

Open Exploration

Open Access Original Article

Small patient datasets reveal genetic drivers of non-small cell lung cancer subtypes using machine learning for hypothesis generation

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

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

[2]NetraMark, Toronto, ON M4P 2E5, Canada

[3]Faculty of Mathematics, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[4]Department of Psychiatry and Behavioral Sciences, Leonard M. Miller School of Medicine, University of Miami, Coral Gables, FL 33124, USA

[5]Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, 41121 Modena, Italy

[6]VeraSci, Durhan, NC 27707, USA

[7]Department of Pathology and Molecular Medicine, Queen's University, Kingston, ON K7L 3N6, Canada

[8]The Centre for Biotechnology and Genomics Medicine, Medical College of Georgia, Augusta University, Augusta, GA 30912, USA

[9]The Clarke Center for Human Imagination, University of California San Diego, La Jolla, CA 92093-0021, USA

***Correspondence:** Joseph Geraci, Department of Pathology and Molecular Medicine, Queen's University, Kingston ON, Canada. joseph.geraci@netramark.com

Academic Editor: Lindsay A. Farrer, Boston University School of Medicine, USA

Received: October 17, 2022 **Accepted:** March 1, 2023 **Published:** July 26, 2023

Cite this article: Cook M, Qorri B, Baskar A, Ziauddin J, Pani L, Yenkanchi S, et al. Small patient datasets reveal genetic drivers of non-small cell lung cancer subtypes using machine learning for hypothesis generation. Explor Med. 2023;4:428–40. https://doi.org/10.37349/emed.2023.00153

Abstract

Aim: Many small datasets of significant value exist 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 are of great value as they can indicate leads for innovative drug discovery and development.

Methods: Two small non-small cell lung cancer (NSCLC) datasets (GSE18842 and GSE10245) consisting of 58 samples of adenocarcinoma (ADC) and 45 samples of squamous cell carcinoma (SCC) were used in a machine intelligence framework to identify genetic biomarkers differentiating these two subtypes. Utilizing a set of standard machine learning (ML) methods, subpopulations of ADC and SCC were uncovered while simultaneously extracting which genes, in combination, were significantly involved in defining the subpopulations. A previously described interactive hypothesis-generating method designed to work with ML methods was employed to provide an alternative way of extracting the most important combination of variables to construct a new data set.

Results: Several genes were uncovered that were previously implicated by other methods. This framework accurately discovered known subpopulations, such as genetic drivers associated with differing levels of aggressiveness within the SCC and ADC subtypes. Furthermore, phyosphatidylinositol glycan anchor biosynthesis, class X (*PIGX*) was a novel gene implicated in this study that warrants further investigation due to its role in breast cancer proliferation.

Conclusions: The ability to learn from small datasets was highlighted and revealed well-established properties of NSCLC. This showcases the utility of ML techniques to reveal potential genes of interest, even from small datasets, shedding light on novel driving factors behind subpopulations of patients.

Keywords

Artificial intelligence, small datasets, genetic subtypes, disease heterogeneity, squamous cell carcinoma, adenocarcinoma

Introduction

The collection of transcriptomic data is expensive, resulting in datasets with a small number of sample sizes (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 datasets, as they require a large number of samples. 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 underlie 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 are only more recently being investigated [4]. The remarkable degree of genetic variability within each histological subtype further 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, and discriminating NSCLC subtypes through genomic expression levels. The bulk of ML efforts has 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 junctions, 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 ML has been previously reported, identifying several genes including cystatin-A (*CSTA*), tumor protein p63 (*TP63*), serpin family B member 13 (*SERPINB13*), chloride channel accessory 2 (*CLCA2*), bicaudal D cargo adaptor 2 (*BICD2*), P53 apoptosis effector related to PMP22 (*PERP*), FAT atypical cadherin 2 (*FAT2*), basonuclin 1 (*BNC1*), ATPase phospholipid transporting 11B (*ATP11B*), family with sequence similarity 83 member B (*FAM83B*), keratin 5 (*KRT5*), par-6 family cell polarity regulator gamma (*PARD6G*), and plakophilin 1 (*PKP1*) which were differentially expressed in ADC and SCC [15].

Other computational methods for discriminating genetic drivers of NSCLC have been previously investigated. A k-means clustering method was used to classify genetic subtypes of ADC [16]. Healthy and ADC tissue was then classified using a support vector machine 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. Two subtypes were found to have high expression levels of immune-related genes, suggesting the existence of heterogenous subpopulations of NSCLC. In another study, researchers also used hierarchal clustering of copy number variations to derive insights into NSCLC drug response [17].

Several ML frameworks have been previously developed specifically tailored for small datasets. A one-shot learning approach called CancerSiamese has been used to predict cancer types while highlighting several marker genes to predict metastatic or primary tumour signatures [18]. A second ML approach has shown promise in deriving insights into immune cell populations in a rare disease application [19].

In order to identify novel driving genes that distinguish these two broad subtypes, a combination of ML tools was designed to learn from patient datasets to analyze gene expression data derived from ADC and SCC NSCLC patients. Because large datasets are critical for most contemporary ML methods such as deep neural networks, there is a need for alternative techniques when data banks are insufficient to train the model. In addition, significant features found within small datasets may become diluted by more obvious statistical features and hence over-represented in large datasets. As such, ML methods must be carefully used and complemented by statistical methods that allow for the discovery of non-linear ways in which groups of genes may interact to drive disease heterogeneity. The methodology presented here is designed for small datasets—a novel way of hypothesizing genetic subpopulations that may result in pathogenesis. For example, the ML framework proposed here has been previously used on a small genetic dataset consisting of Alzheimer's disease brain samples [20]. Several genetic pathways associated with Alzheimer's disease were uncovered, suggesting that even with a small dataset, there exists a high degree of genetic complexity within pathophysiology. Similarly, the findings presented here support genes previously reported to distinguish ADC and SCC subtypes. However, the novelty of this work lies in the 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.

Materials and methods

Datasets

Two data sets were used, consisting of 40 samples of ADC and 18 samples of SCC (GSE10245) [21] and 14 samples of ADC and 32 samples of SCC (GSE18842) [22] to obtain a total of 104 samples (Table 1). 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 Robust Multi-Array Average-calculated signal intensity [23].

Table 1. Characteristics of datasets used to generate NSCLC hypotheses

Dataset	ADC/SCC samples	Male/female samples	Reference
GSE10245	40/18	14/44	[21]
GSE18842	14/32	N/A	[22]

Machine intelligence

The methodology was developed to organize the resulting models from several well-known ML methods to explore NSCLC genetic heterogeneity within a small dataset. The only proprietary method used for these results is a novel feature selection tool that is part of the NetraAI system which incorporates systems biology [20, 24, 25] and can help produce clustering diagrams as provided in this paper. This was used to create several reduced datasets with significantly fewer variables, e.g., less than a hundred. These reduced datasets are available upon request to encourage reproducibility and further research. The following algorithm was used based on standard methods to create models and insights (Figure 1). For the work reported in this paper, the following tailored process was utilized after performing the aforementioned feature reduction:

(1) First, variable importance was calculated via ensemble trees (Random Forest) through cross-validation [26, 27]. The 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 further.

(4) Using the t-SNE [28], HDBSCAN [29], and UMAP [30] algorithms, subpopulations were extracted.

(5) Sample IDs were collected from the clusters formed from these two clustering models, then systematically compared each group with the others, and applied statistical methods to determine differentially expressed gene candidates.

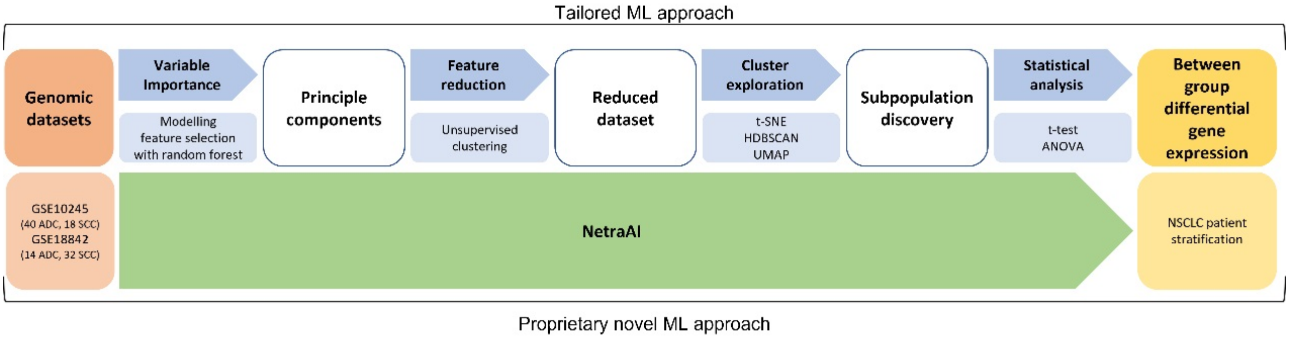

Figure 1. ML approach for small datasets. Using two NSCLC datasets, a tailored ML approach was used consisting of feature selection with random forest, unsupervised clustering, cluster exploration with t-SNE, HDBSCAN, UMAP, and statistical analysis to obtain between group differential gene expression for NSCLC patient stratification. These results were validated using the proprietary NetraAI which generates hypotheses across different groups of patients. t-SNE: t-distributed stochastic neighbor embedding; HDBSCAN: hierarchical density-based spatial clustering of applications with noise; UMAP: uniform manifold approximation and projection; ANOVA: analysis of variance

Clustering was performed via principal components, t-SNE, HDBSCAN, and UMAP as these were the basis of the maps found in this paper. These methods were used to organize the resulting clustering models, in addition to the random forest models, such that the models were capable of being explored interactively to derive a deeper understanding of the driving genes behind the subclusters [20].

A critical shortcoming of working with small data is that it is highly unlikely to represent the totality of the real-world phenomenon it represents, in this case, NSCLC. This means that creating reliable models that are meant to become biomarkers for the disorder is nearly impossible. However, what is possible is the discovery of a subpopulation of patients that all have a set of variables in common, in this case, gene expression. This cluster of patients can be regarded as a hypothesis and therefore classical statistics can be used to evaluate the significance of the findings. In this way, small data sets can be interrogated with the tailored process summarized in Figure 1 in order to extract potentially meaningful discoveries.

The methods here and those described in [20, 24, 25] are designed to address how to extract clear insights about subgroups of patients and their driving variables, while innovative methods found in [31, 32] are well suited to create models for predictions and decision making when sufficient data to do so is available.

Statistical analysis

Statistical analyses were implemented in order to determine significant differences in gene expression data. The following represents a summary of the statistical methods employed:

(1) Bar plot means values represent the mean expression level while error bars represent the standard deviation (SD) of the pooled data from each probe ID. Bar plot *P*-values were calculated using an unpaired *t*-test, where *P*-values < 0.05 were considered statistically significant.

(2) 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, ANOVA was used. The resulting clusters were plotted for the purpose of illustrating the findings.

(3) Bonferroni corrections were implemented whenever univariate statistics were utilized for feature selection and when initiation comparisons with ANOVA and *t*-tests.

Results

A tailored ML process identifies differentially expressed genes from a small NSCLC dataset

Using the ADC and SCC tumour gene expression data, this tailored ML approach for small datasets was able to help generate a map distinguishing SCC (blue) and ADC (red) subjects, Loop 1 and Loop 2, respectively (Figure 2). The key genes that were found to have driven this distinction were desmocollin-3 (*DSC3*), visinin-like protein 1 (*VSNL1*), solute carrier family 6 member 10 (*SLC6A10P*), interferon regulatory transcription factor 6 (*IRF6*), dystonin (*DST*), *CLCA2*, desmoglein 3 (*DSG3*), lysophosphatidylcholine acetyltransferase 1 (*LPCAT1*), cingulin (*CGN*), and phyosphatidylinositol glycan anchor biosynthesis, class X (*PIGX*). Of these, all genes except for *LPCAT1* were overexpressed in Loop 1, the SCC subjects. Meanwhile, Loop 2 consisting of ADC subjects was characterized by *LPCAT1* overexpression.

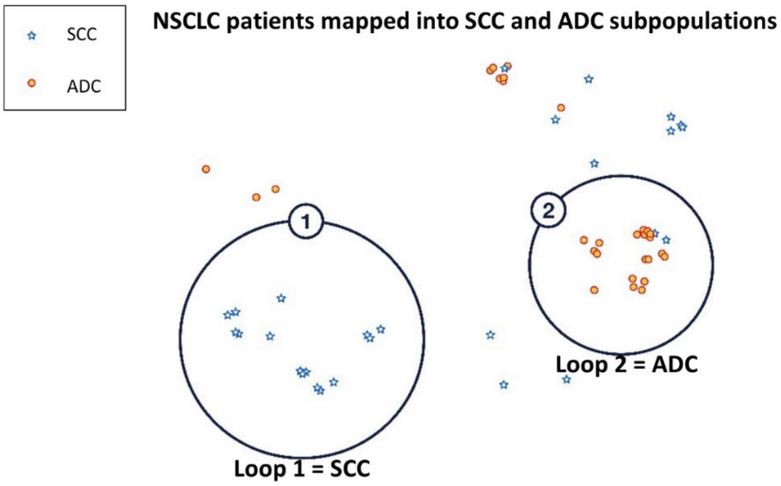

Figure 2. Stratification of NSCLC patients into SCC and ADC using NetraAI. Loop 1 consisting of SCC (blue) subjects and Loop 2 consisting of ADC (red) subjects were delineated by HDBSCAN. These subpopulations were identified by clustering methods that stratified patients due to the statistically significant differential expression of *DSC3*, *VSNL1*, *SLC6A10P*, *IRF6*, *DST*, *CLCA2*, *DSG3*, *LPCAT1*, and *PIGX* between the two loops

Collectively, in the analysis of these two datasets, total of 10 genes were identified that discriminate ADC and SCC patient populations. It is worth mentioning that 9 of the 10 genes identified have been previously reported to be differentially expressed in SCC and ADC (Table 2), further validating the methods used here. The novel gene identified that has not been previously associated with NSCLC populations at the time of this report is *PIGX*.

Table 2. Genes discriminating between SCC and ADC

Gene	Function	Upregulation (SCC/ADC)	Reference
DSC3	Ca^{2+}-dependent glycoprotein involved in cell adherence	SCC	[22]
VSNL1	Neuronal Ca^{2+} sensor protein; tumour suppressor gene	SCC	[33, 34]
IRF6	Transcription factor	SCC	[35]
DST	Cell adhesion	SCC	[36]
CLCA2	Cell adhesion; tumour suppressor	SCC	[37]
PIGX	Tumour suppressor	SCC	-
DSG3	Cell adhesion	SCC	[38–40]
LPCAT1	Cancer progression and metastasis	ADC	[41, 42]
SLC6A10	Neurotransmitter transporter; *pseudogene of *SLC6A8*	ADC	[43]
CGN	Tight junction	ADC	[21]

-: blank cell; Ca^{2+}: calcium ion; *: pseudogene of its parent gene *SLC6A8*

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 [21]. Here, SCC was found to be associated with cell adhesion marker *DSC3*, and ADC was associated with tight junction marker *CGN* (Figure 3). Specifically, two probes corresponding to *DSC3* were identified, 206032_at and 206033_s_at. There was a statistically significant association of both *DSC3* probes with SCC ($P < 0.0001$; Figure 3A). Interestingly, when looking at the dataset including sex, 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). In contrast, the two probes corresponding to *CGN*, 223232_s_at, and 223233_s_at were significantly associated with ADC ($P < 0.0001$; Figure 3B). In contrast, the *CGN* probes were significantly associated with females ($P = 0.014$). These results highlight a potential role of sex-based differences in NSCLC that warrant further investigation.

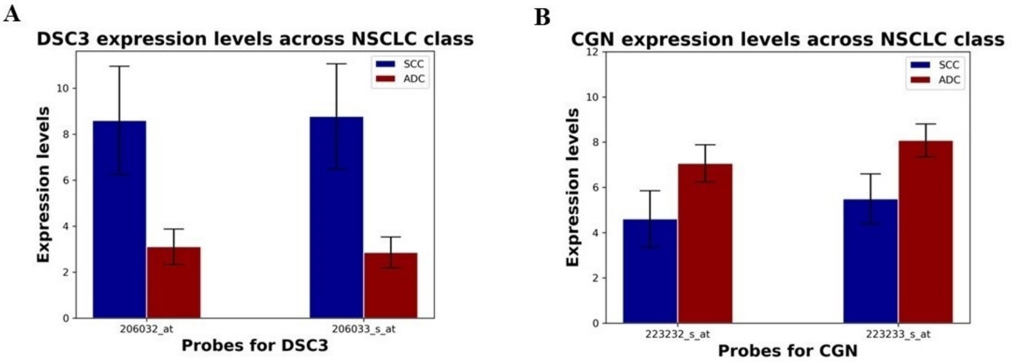

Figure 3. Differential expression of *DSC3* and *CGN* in SCC and ADC NSCLC patient subpopulations. (A) The expression levels of *DSC3* probes 206032_at and 206033_s_at (mean ± SD) in SCC and ADC subpopulations; (B) the expression levels of *CGN* probes 223232_s_at and 223233_s_at in SCC and ADC subpopulations

SLC6A10P may be a key driver of an ADC subtype

Further analysis of the two datasets revealed two distinct ADC patient subpopulations (Figure 4). These two loops were distinguished by *SLC6A10P*, with Loop 2 characterized by overexpression of *SLC6A10P* ($P = 1.3 \times 10^{-5}$). The association of *SLC6A10P* with ADC patients is in line with previous reports [35, 43]. However, 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 [43]. This suggests that within the ADC patient population there exist unique subpopulations that may be associated with increased aggressive and metastatic propensity.

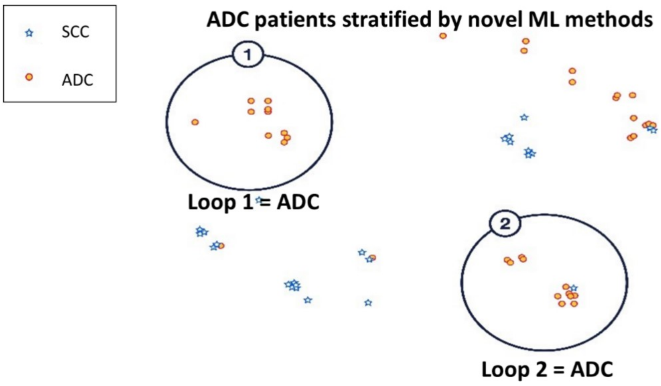

Figure 4. Semi-supervised clustering of ADC patient subpopulations using NetraAI. Analysis of the NSCLC patients revealed two distinct subpopulations of ADC (red) subjects delineated by HDBSCAN. Both Loop 1 and Loop 2 ADC subpopulations were identified by clustering methods that stratified patients due to statistically significant differential expression of *SLC6A10P* between the two Loops

IRF6 and *CLCA2* drive a unique SCC subpopulation

Not surprisingly, further analysis also revealed two distinct subpopulations of SCC driven by *IRF6* and *CLCA2* (Figure 5A), which have been previously associated with SCC [35, 37]. *IRF6* and *CLCA2* expression levels were higher in SCC than in ADC ($P < 0.0001$; Figure 5B and 5C). Here, Loop 2 was associated with a significantly higher expression of both *IRF6* and *CLCA2* compared to Loop 1. The significance value between the *CLCA2* and *IRF6* probes in the two encircled SCC groups were evaluated to be 4.4×10^{-7}, 5.8×10^{-3}, 9.3×10^{-7}, and 0.046 for the 206164_at, 206165_s_at, 206166_s_at and 1552477_a_at probes, respectively. Considering the strong association of both genes with one specific subpopulation of SCC patients, it highlights an avenue of research focusing on the pathways at play in the etiology of the disease as well as for the identification of novel drugs targeting their combined pathways.

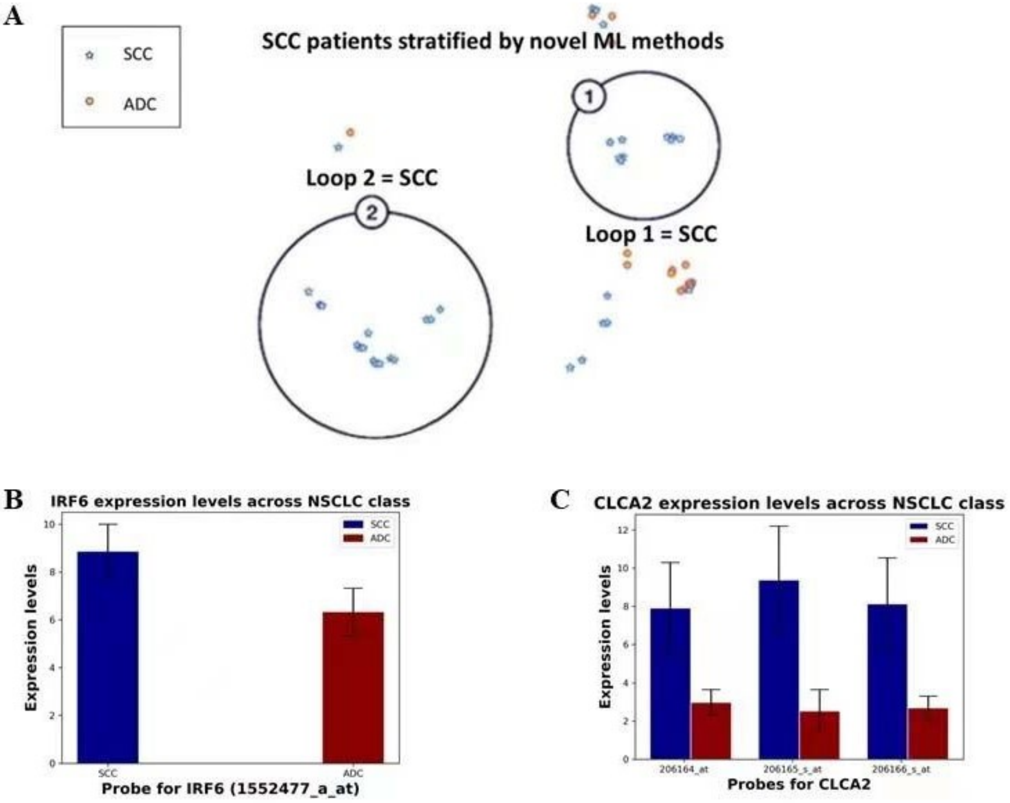

Figure 5. Semi-supervised clustering of SCC patients and differential expression of *IRF6* and *CLCA2*. (A) Analysis of NSCLC patients revealed two distinct subpopulations of SCC (blue) subjects delineated by HDBSCAN within NetraAI. Both Loop 1 and Loop 2 subpopulations were identified by clustering methods that stratified patients due to statistically significant differential expression of *IRF6* and *CLCA2*; (B) the expression levels of *IRF6* probe 1552477_a_at (mean ± SD) in SCC and ADC patient subpopulations; (C) the expression levels of *CLCA2* probes 206164_at, 206165_s_at, and 206166_s_at (mean ± SD) in SCC and ADC patient subpopulations

Discussion

Using publicly available NSCLC datasets with a suite of ML techniques appropriate for small datasets had an excellent signal for separating ADC and SCC. The main philosophy followed here is that for smaller datasets, where the patients are unlikely to reflect the distribution of patients in the totality of reality, one can allow ML methods to generate hypotheses about the population available in a small dataset. This allows a researcher to benefit from the power of statistics, in that they can test the hypothesis and derive some measure of confidence. Proprietary methods like the NetraAI empower this hypothesis testing paradigm, but the method described above is also capable of expressing hypotheses in the form of patient clusters.

Here, patient clusters were compared using statistical methods suitable for a dataset with so few samples in order to avoid overfitting that often comes with utilizing contemporary ML methods with small datasets. Further, the transparency of the driving factors is important so that human experience can be used to evaluate what is being 'discovered' by the machine.

This study highlights the genetic heterogeneity within NSCLC subtypes. Using this dataset, a set of 10 genes that distinguish ADC and SCC were identified (Table 2). Within these 10 genes, 9 have been previously reported to be associated either with NSCLC or a specific subtype of NSCLC, validating this ML approach. These findings were aligned 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 [44]. 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 [45]. However, some of the genes 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.

The first of the previously reported NSCLC-associated genes identified was *DSC3*, which plays a role in epidermal morphology and keratinocyte proliferation [22]. There are several studies that report on *DSC3* distinguishing ADC from SCC, with a higher expression in SCC [36, 46–48]. Notably, there has been a report on the association between *DSC3* and tumour suppressor activity in NSCLC mediated by inhibition of *EGFR* [49]. However, there remain contradictory associations between *DSC3* and prognosis, with elevated levels associated with increased metastatic risk in melanoma and better prognosis in lung and colon cancer [40]. This suggests that the same protein may have differential effects in the tumour microenvironment (TME), which presents an interesting field of research to understand how *DSC3* expression correlates with NSCLC subtypes depending on where they originate in the lung. 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 [21].

In the results presented here, ADC has been reported to be characterized by tight junctions and was identified by *CGN* and SCC has been characterized by gap junctions and was identified by *DSC3* (Figure 3). Males have been reported to have a significantly poorer NSCLC prognosis compared to females, shifting efforts towards sex-based approaches to diagnosis, prognosis, and therapeutic interventions [50, 51]. Additionally, estrogens have been associated with an increased risk of ADC in women despite equal expression of estrogen receptors α and β; however, the role remains unclear [52]. 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 [53]. Gap junction proteins, also known as connexins, serve as channels that connect the interior of adjacent cells, facilitating intracellular homeostasis and coordination of activities via second messengers [54]. Desmosomes primarily provide mechanical strength via a structural network. In contrast, tight junctions form a barrier around the cell, regulating the permeability of the paracellular space [55, 56]. These molecules play critical roles in epithelial-to-mesenchymal transition, a process involved in cancer metastasis. Aside from the current work relating CGN expression to females, no sex-based differences have been previously reported. This presents a unique field of research, as there may be different druggable targets for males and females. The variability of adhesion molecule expression across sex warrants further investigation to elucidate the details of the correlation and advance toward gender related precision medicine.

Interestingly, *SLC6A10P* was the single gene that was found to distinguish between two specific subpopulations of ADC. *SLC6A10P* was previously found to be a marker for aggressive ADC [43], and recently, implicated within the Notch signaling pathway [57]. These findings suggest that *SLC6A10P* warrants further investigation as a genetic biomarker in the context of the ADC patient subpopulation. This demonstrates the power of machine intelligence to reveal etiologies 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.

With respect to the SCC patient population, *CLCA2* and *IRF6* were found to distinguish between two distinct SCC subpopulations. *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 [37]. Furthermore, SCC expression was correlated with tumour grade upon histological characterization. In particular, *CLCA2*-negative samples were associated with poorly differentiated tumours [37].

Most noteworthy, phosphatidylinositol glycan anchor biosynthesis class gene *PIGX*, was the only gene identified that has not been previously associated with NSCLC. However, there have been reports that *PIGX* promotes cancer cell proliferation by suppressing *EHD2* and *ZIC1* in breast cancer [58]. The authors reported that *PIGX* expression was associated with shorter recurrence-free survival. In the present study, *PIGX* was found to be a driver of ADC and SCC differentiation, being overexpressed in SCC patients (Figure 2). As a novel gene associated with NSCLC or a specific subtype, this highlights an area that warrants further investigation for the advancement of precision medicine in NSCLC.

In order to create robust predictive models with machine intelligence, large datasets are required, but this study 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 method uncovered several genetic subtypes of ADC of SCC, including those driven by *SLC6A10P*, *CLCA2*, and *IRF6*, respectively. Furthermore, these data suggest that the expression levels of adhesion proteins encoded by *CGN* and *DSC3* may play a role in sex-based differences in NSCLC. Finally, this study uncovered a statistically significant driver of NSCLC heterogeneity, *PIGX*, which warrants further investigation.

This report highlights the use of a novel set of ML techniques that are appropriate for small datasets. The primary aim of using such techniques is to encourage other researchers to explore small datasets that are often otherwise skipped with ML as there may be hidden valuable information within them. Adopting these approaches, one can extract meaningful insights with the techniques described here to move closer toward precision medicine.

Abbreviations

ADC: adenocarcinoma

ANOVA: analysis of variance

CGN: cingulin

CLCA2: chloride channel accessory 2

DSC3: desmocollin-3

DSG3: desmoglein 3

DST: dystonin

HDBSCAN: hierarchical density-based spatial clustering of applications with noise

IRF6: interferon regulatory transcription factor 6

LPCAT1: lysophosphatidylcholine acetyltransferase 1

ML: machine learning

NSCLC: non-small cell lung cancer

PIGX: phyosphatidylinositol glycan anchor biosynthesis, class X

SCC: squamous cell carcinoma

SD: standard deviation

SLC6A10P: solute carrier family 6 member 10

t-SNE: t-distributed stochastic neighbor embedding

UMAP: uniform manifold approximation and projection

VSNL1: visinin-like protein 1

Declarations

Author contributions

JG: Conceptualization, Methodology, Resources, Writing—review & editing, Supervision, Project administration, Funding acquisition. MC: Conceptualization, Visualization. MC and BQ: Methodology, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing. JZ, SY and AB: Software. SY: Validation. LP and JG: Writing—review & editing. LP: Project administration.

Conflicts of interest

JG is a major shareholder of NetraMark Corp, where NetraMark is a technology company providing clinical trial support to pharmaceutical companies. LP has previously acted as a scientific consultant for AbbVie USA; Acadia USA; BCG Switzerland; Boehringer Ingelheim International GmbH; Compass Pathways; EDRA-Publishing, Italy; Ferrer Spain; Gedeon-Richter, Hungary; Inpeco SA, Switzerland; Johnson & Johnson USA; NeuroCog Trials USA; Novartis-Gene Therapies, Switzerland; Otsuka USA; Pfizer Global USA; PharmaMar Spain; Relmada Therapeutics USA; Takeda, USA; VeraSci, USA; Vifor Switzerland.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Data was obtained from publicly available datasets GSE10245 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10245 and GSE18842 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18842.

Funding

Part of this research was funded by NetraMark Corp in the form of salary for Dr. Joseph Geraci, and computational resources. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright

References

1. Ridge CA, McErlean AM, Ginsberg MS. Epidemiology of lung cancer. Semin Intervent Radiol. 2013;30:93–8.
2. Thomas A, Liu SV, Subramaniam DS, Giaccone G. Refining the treatment of NSCLC according to histological and molecular subtypes. Nat Rev Clin Oncol. 2015;12:511–26.

3. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505:495–501.
4. Pikor LA, Ramnarine VR, Lam S, Lam WL. Genetic alterations defining NSCLC subtypes and their therapeutic implications. Lung Cancer. 2013;82:179–89.
5. Manegold C. Treatment algorithm in 2014 for advanced non-small cell lung cancer: therapy selection by tumour histology and molecular biology. Adv Med Sci. 2014;59:308–13.
6. Carnio S, Novello S, Bironzo P, Scagliotti GV. Moving from histological subtyping to molecular characterization: new treatment opportunities in advanced non-small-cell lung cancer. Expert Rev Anticancer Ther. 2014;14:1495–513.
7. Yu L, Tao G, Zhu L, Wang G, Li Z, Ye J, et al. Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. BMC Cancer. 2019;19:464.
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. AJR Am J Roentgenol. 2020;215:192–7.
9. Kriegsmann M, Haag C, Weis CA, Steinbuss G, Warth A, Zgorzelski C, et al. Deep learning for the classification of small-cell and non-small-cell lung cancer. Cancers (Basel). 2020;12:1604.
10. Mu W, Jiang L, Zhang J, Shi Y, Gray JE, Tunali I, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat Commun. 2020;11:5228.
11. Rabbani M, Kanevsky J, Kafi K, Chandelier F, Giles FJ. Role of artificial intelligence in the care of patients with nonsmall cell lung cancer. Eur J Clin Invest. 2018;48:e12901.
12. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.
13. Podolsky MD, Barchuk AA, Kuznetcov VI, Gusarova NF, Gaidukov VS, Tarakanov SA. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. Asian Pac J Cancer Prev. 2016;17:835–8.
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;11:303–10.
15. Yuan F, Lu L, Zou Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochim Biophys Acta Mol Basis Dis. 2020;1866:165822.
16. Hu F, Zhou Y, Wang Q, Yang Z, Shi Y, Chi Q. Gene expression classification of lung adenocarcinoma into molecular subtypes. IEEE/ACM Trans Comput Biol Bioinform. 2020;17:1187–97.
17. Shen Y, Xiang Y, Huang X, Zhang Y, Yue Z. Pharmacogenomic cluster analysis of lung cancer cell lines provides insights into preclinical model selection in NSCLC. Interdiscip Sci. 2022;14:712–21.
18. Mostavi M, Chiu YC, Chen Y, Huang Y. CancerSiamese: one-shot learning for predicting primary and metastatic tumor types unseen during model training. BMC Bioinformatics. 2021;22:244.
19. Robinson GA, Peng J, Dönnes P, Coelewij L, Naja M, Radziszewska A, 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:e485–96.
20. Qorri B, Tsay M, Agrawal A, Au R, Geraci J. Using machine intelligence to uncover Alzheimer's disease progression heterogeneity. Explor Med. 2020;1:377–95.
21. Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu EC, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer. 2009;63:32–8.
22. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, et al. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. Int J Cancer. 2011;129:355–64.
23. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc. 2004;99:909–17.

24. Tsay M, Geraci J, Agrawal A. Next-gen AI for disease definition, patient stratification, and placebo effect. OSF Preprints [Preprint]. 2020 [cited 2023 Jan 21]. Available from: https://osf.io/pc7ak/

25. Choi J, Bodenstein DF, Geraci J, Andreazza AC. Evaluation of postmortem microarray data in bipolar disorder using traditional data comparison and artificial intelligence reveals novel gene targets. J Psychiatr Res. 2021;142:328–36.

26. Lai C, Reinders MJ, van't Veer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics. 2006;7:235.

27. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99:323–9.

28. van der Maaten L, Hinton G. Visualizing data using t-SNE. JMLR. 2008;9:2579–605.

29. McInnes L, Healy J. Accelerated hierarchical density based clustering. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). 2017 IEEE International Conference on Data Mining Workshops (ICDMW); 2017 Nov 18–21; New Orleans, LA, USA. IEEE; 2017. pp. 33–42.

30. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. J Open Source Softw. 2018;3:861.

31. Liu K, Chen Z, Wu J, Tan Y, Wang L, Yan Y, et al. Big medical data decision-making intelligent system exploiting fuzzy inference logic for prostate cancer in developing countries. IEEE Access. 2019;7:2348–63.

32. Zhou J, Khushi M, Moni MA, Uddin S, Poon SK. Lung cancer prediction using curriculum learning based deep neural networks. In: 2021 IEEE International Conference on Digital Health (ICDH). 2021 IEEE International Conference on Digital Health (ICDH); 2021 Sep 5–10;Chicago, IL, USA. IEEE; 2021. pp.11–8.

33. Fu J, Fong K, Bellacosa A, Ross E, Apostolou S, Bassi DE, et al. *VILIP-1* downregulation in non-small cell lung carcinomas: mechanisms and prediction of survival. PLoS One. 2008;3:e1698.

34. Gonzalez Guerrico AM, Jaffer ZM, Page RE, Braunewell KH, Chernoff J, Klein-Szanto AJ. Visinin-like protein-1 is a potent inhibitor of cell adhesion and migration in squamous carcinoma cells. Oncogene. 2005;24:2307–16.

35. Liu Y, Shao G, Yang Z, Lin X, Liu X, Qian B, et al. Interferon regulatory factor 6 correlates with the progression of non-small cell lung cancer and can be regulated by miR-320. J Pharm Pharmacol. 2021;73:682–91.

36. Chae YK, Choi WM, Bae WH, Anker J, Davis AA, Agte S, et al. Overexpression of adhesion molecules and barrier molecules is associated with differential infiltration of immune cells in non-small cell lung cancer. Sci Rep. 2018;8:1023.

37. Shinmura K, Igarashi H, Kato H, Kawanishi Y, Inoue Y, Nakamura S, et al. CLCA2 as a novel immunohistochemical marker for differential diagnosis of squamous cell carcinoma from adenocarcinoma of the lung. Dis Markers. 2014;2014:619273.

38. Savci-Heijink CD, Kosari F, Aubry MC, Caron BL, Sun Z, Yang P, et al. The role of desmoglein-3 in the diagnosis of squamous cell carcinoma of the lung. Am J Pathol. 2009;174:1629–37.

39. Fukuoka J, Dracheva T, Shih JH, Hewitt SM, Fujii T, Kishor A, et al. Desmoglein 3 as a prognostic factor in lung cancer. Hum Pathol. 2007;38:276–83.

40. Dong Y, Li S, Sun X, Wang Y, Lu T, Wo Y, et al. Desmoglein 3 and keratin 14 for distinguishing between lung adenocarcinoma and lung squamous cell carcinoma. Onco Targets Ther. 2020;13:11111–24.

41. Liu F, Wu Y, Liu J, Ni RJ, Yang AG, Bian K, et al. A miR-205-LPCAT1 axis contributes to proliferation and progression in multiple cancers. Biochem Biophys Res Commun. 2020;527:474–80.

42. Wei C, Dong X, Lu H, Tong F, Chen L, Zhang R, et al. LPCAT1 promotes brain metastasis of lung adenocarcinoma by up-regulating PI3K/AKT/MYC pathway. J Exp Clin Cancer Res. 2019;38:95.

43. Yuan K, Gao ZJ, Yuan WD, Yuan JQ, Wang Y. High expression of SLC6A10P contributes to poor prognosis in lung adenocarcinoma. Int J Clin Exp Pathol. 2018;11:720–6.

44. 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.
45. 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. Oncol Lett. 2017;13:3314–22.
46. Warth A, Muley T, Herpel E, Meister M, Herth FJ, Schirmacher P, et al. Large-scale comparative analyses of immunomarkers for diagnostic subtyping of non-small-cell lung cancer biopsies. Histopathology. 2012;61:1017–25.
47. Tsuta K, Tanabe Y, Yoshida A, Takahashi F, Maeshima AM, Asamura H, et al. 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. J Thorac Oncol. 2011;6:1190–9.
48. Angulo B, Suarez-Gauthier A, Lopez-Rios F, Medina PP, Conde E, Tang M, et al. Expression signatures in lung cancer reveal a profile for *EGFR*-mutant tumours and identify selective *PIK3CA* overexpression by gene amplification. J Pathol. 2008;214:347–56.
49. Cui T, Chen Y, Yang L, Knösel T, Huber O, Pacyna-Gengelbach M, et al. The p53 target gene desmocollin 3 acts as a novel tumor suppressor through inhibiting EGFR/ERK pathway in human lung cancer. Carcinogenesis. 2012;33:2326–33.
50. Wainer Z, Wright GM, Gough K, Daniels MG, Russell PA, Choong P, et al. 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. Clin Lung Cancer. 2018;19:e933–44.
51. Radkiewicz C, Dickman PW, Johansson ALV, Wagenius G, Edgren G, Lambe M. Sex and survival in non-small cell lung cancer: a nationwide cohort study. PLoS One. 2019;14:e0219206.
52. Ivanova MM, Mazhawidza W, Dougherty SM, Klinge CM. Sex differences in estrogen receptor subcellular location and activity in lung adenocarcinoma cells. Am J Respir Cell Mol Biol. 2010;42:320–30.
53. Rubin JB, Lagas JS, Broestl L, Sponagel J, Rockwell N, Rhee G, et al. Sex differences in cancer mechanisms. Biol Sex Differ. 2020;11:17.
54. Ruch R. Gap junctions and connexins in cancer formation, progression, and therapy. Cancers (Basel). 2020;12:3307.
55. Soini Y. Tight junctions in lung cancer and lung metastasis: a review. Int J Clin Exp Pathol. 2012;5:126–36.
56. Bhat AA, Uppada S, Achkar IW, Hashem S, Yadav SK, Shanmugakonar M, et al. Tight junction proteins and signaling pathways in cancer and inflammation: a functional crosstalk. Front Physiol. 2019;9:1942.
57. 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. Erratum in: Ann Transl Med. 2022;10:845.
58. Nakakido M, Tamura K, Chung S, Ueda K, Fujii R, Kiyotani K, et al. Phosphatidylinositol glycan anchor biosynthesis, class X containing complex promotes cancer cell proliferation through suppression of EHD2 and ZIC1, putative tumor suppressors. Int J Oncol. 2016;49:868–76.