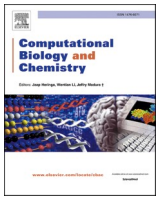
Computational Biology and Chemistry 119 (2025) 108560

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Identification of the differences in molecular networks between idiopathic pulmonary fibrosis and lung squamous cell carcinoma using machine learning

Yosui Nojima [a,b,*], Kenji Mizuguchi [a,c,**]

[a] Artificial Intelligence Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, 17-3 Senrioka-Shinmachi, Settu, Osaka 566-0002, Japan
[b] Center for Mathematical Modeling and Data Science, The University of Osaka, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
[c] Institute for Protein Research, The University of Osaka, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

ARTICLE INFO

Keywords:
Idiopathic pulmonary fibrosis
Lung squamous cell carcinoma
Machine learning
Molecular networks
RNA sequencing
Public data

ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is an independent risk factor for lung cancer, especially squamous cell carcinoma (SCC). The prognosis of patients with both IPF and SCC is poorer than that of patients with only IPF, and preventive measures against SCC in patients with IPF remain elusive. Understanding the distinct mechanisms that induce both diseases is crucial for mitigating SCC onset in patients with IPF. We developed highly accurate machine learning (ML) models to classify patients with IPF or SCC using public RNA sequencing data. To construct the ML models, a random restart technique was applied to the five algorithms. To identify the differentially expressed genes (DEGs) between IPF and SCC, feature importance was calculated in the classification models. Furthermore, we detected somatic mutations affecting gene expression using SCC data. The ML models identified *VCX2*, *TMPRSS11B*, *PRUNE2*, *PRG4*, *PZP*, *SCARA5*, *DES*, *HPSE2*, *HOXD11*, *S100A7A*, and *PLA2G2A* as DEGs. Somatic mutations were detected in four transcription factors, *BHLHE40*, *MYC*, *STAT1*, and *E2F4*, which regulate the expression of these 11 genes. Furthermore, a molecular network comprising four transcription factors and 11 downstream genes was discovered. This newly identified molecular network enhances our understanding of the distinct mechanisms underlying IPF and SCC onset, and provides new insights into preventing SCC complications in patients with IPF.

1. Introduction

Idiopathic pulmonary fibrosis (IPF), a common and prognostically unfavorable idiopathic interstitial pneumonia (IIP), is characterized by inflammatory and fibrotic changes in the lung parenchyma. IPF, constituting 55–60 % IIP cases, has no curative treatment, and the two recommended antifibrotic drugs only slow disease progression (Raghu et al., 2015). Briefly, 10–20 % IPF cases concurrently involve lung cancer (Ozawa et al., 2009), and the incidence of lung cancer in IPF cases is 7–14 times higher than in non-IPF cases (Matsushita et al., 1995; Turner-Warwick et al., 1980). Both conditions are more prevalent in old adults, males, and smokers, sharing risk factors, such as smoking and exposure to environmental or occupational hazards (Zaman and Lee, 2018). Additionally, IPF is an independent risk factor for lung cancer.

Patients with both IPF and lung cancer have a significantly worse prognosis than those with lung cancer or IPF (Tomassetti et al., 2015; Yoon et al., 2018). Squamous cell carcinoma (SCC) is the most common lung cancer type in patients with IPF, whereas adenocarcinoma (ADC) is the most common non-small-cell lung cancer (NSCLC) subtype in the general population (Ballester et al., 2019). Surgical intervention is a treatment option for patients with both IPF and lung cancer; however, the postoperative incidence of acute exacerbation (AE) of IPF is 10.1 % (Sato et al., 2014). Postoperative AE is associated with 35.6–43.9 % mortality, indicating poor prognosis (Sato et al., 2015, 2014). Thus, preventing lung cancer occurrence is crucial to avoid AE, but no preventive measures have been established for patients with both IPF and lung cancer. To prevent lung cancer in patients with IPF, it is essential to understand the molecular mechanisms by which lung cancer cells in

* Corresponding author at: Center for Mathematical Modeling and Data Science, The University of Osaka, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan.
** Corresponding author at: Institute for Protein Research, The University of Osaka, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan.
E-mail addresses: nojima.yosui.mmds@osaka-u.ac.jp (Y. Nojima), kenji@protein.osaka-u.ac.jp (K. Mizuguchi).

https://doi.org/10.1016/j.compbiolchem.2025.108560
Received 11 October 2024; Received in revised form 9 June 2025; Accepted 13 June 2025
Available online 19 June 2025

patients with both IPF and lung cancer arise from lung cells in patients with only IPF. This requires an understanding of the differences in the onset mechanisms of both diseases; however, such analyses remain underexplored.

One reason is that surgical intervention in patients with both IPF and lung cancer carries AE risk, making the collection of clinical samples ethically challenging. However, Hata *et al.* showed that the somatic mutation profiles were highly similar in SCC patients, regardless of the presence of IPF (Hata et al., 2021), suggesting that genetic information from the cancerous tissues of patients with both IPF and SCC could be substituted for that from patients with only SCC. Since data on IPF without lung cancer and lung cancer without IPF are relatively abundant in public databases, these datasets may be analyzed to understand the differences between the onset mechanisms of both diseases.

Classical statistics and machine learning (ML) differ in terms of computational tractability, with increasing variables per subject. Classical statistical modeling, designed for data with a few dozen input variables and small to moderate sample sizes, focuses on unobserved system aspects. Conversely, ML focuses on predictions using general-purpose learning algorithms to search for patterns in complex and unwieldy data (Bzdok, 2017; Bzdok et al., 2018, 2017); these methods can be effective even without a carefully controlled experimental design or complicated nonlinear interactions (Bzdok et al., 2018).

This study used various public RNA sequencing (RNA-Seq) data from the lungs of patients with IPF or SCC and employed a random restart technique to construct high-performance ML models for classifying IPF and SCC. These models highlight the differences in IPF and SCC onset mechanisms, contributing to a clear understanding of lung cancer onset and prevention mechanisms in patients with IPF.

2. Methods

2.1. Public RNA-Seq analysis and batch-effect correction

Public RNA-Seq datasets were obtained from the NCBI Sequence Read Archive (https://trace.ncbi.nlm.nih.gov/Traces/sra/; Table 1). The quality of FASTQ file data was confirmed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic version 0.36 (http://www.usadellab.org/cms/?page=trimmomatic) (Bolger et al., 2014) was used to trim the reads using the Illumina TruSeq adapter removal process (2:30:10) with the following parameters: LEADING, 20; TRAILING, 20; SLIDINGWINDOW, 4:20; and MINLEN, 25. The trimmed reads were mapped to the reference human genome (version GRCh38) available in Ensembl using HISAT2 version 2.1.0 (http://daehwankimlab.github.io/hisat2/) with default parameters (Kim et al., 2019). The resulting BAM files were input into featureCounts version 2.0.1 (http://subread.sourceforge.net/) (Liao et al., 2014) to generate read count data. Genes listed in Ensembl BioMart (https://www.ensembl.org/info/data/biomart/index.html) as protein-coding genes, long non-coding RNAs (lncRNAs), and microRNAs (miRNAs) were selected for further analysis using the biomaRt package (version 2.54.1) in R version 4.2.3 (https://www.r-project.org/). Count data were converted into transcripts per million (TPM) and subsequently transformed into $\log_2$ (TPM + 1).

Subsequently, the transformed values were fed into the ComBat_seq function of the sva package (version 3.46.0) in R to perform batch-effect corrections (BEC) among the datasets. The BEC outcomes were estimated using a uniform manifold approximation and projection (UMAP) via the umap function of the umap package (version 0.2.10.0). The UMAP visualization was performed using the ggplot function in the ggplot2 package (version 3.4.4). The normal samples after BEC were excluded from subsequent analyses.

2.2. ML model construction and feature importance

We selected the dataset with the largest sample size for each disease and designated it as the training data, whereas the remaining datasets were assigned as the test data. The training data were oversampled and downsampled using the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) function of the DMwR package (version 0.4.1) in R. Classification models were constructed using the training function of the caret package (version 6.0–94). The training samples were split into 8:2 training: validation data and trained using 5-fold cross-validation. The algorithms *k*-nearest neighbors (knn), support vector machines (SVM) with radial basis function kernel (svmRadial), SVM with linear kernel (svmLinear), eXtreme gradient boosting (xgbTree), and random forest (rf) were selected to construct the classification models because these algorithms are commonly used in supervised learning of RNA-Seq data (Liñares Blanco et al., 2019; Gakii et al., 2023) and high explainability. The hyperparameter settings for each ML algorithm used in the grid search are presented in Supplementary Table 1. A random restart was performed with seed values 1–2000 in increments of 1. The accuracy, area under the curve (AUC), and kappa value were calculated using the confusion matrix function of the caret package.

2.3. Permutation test

To evaluate the stability of the high-feature-importance genes, permutation tests were performed using R. For one or more genes, the values for each sample were shuffled among the samples, and the classification models were built as described above. A random restart was not performed in the permutation test, and the seed value with the highest performance from the original data was selected. Additionally, permutation tests were conducted only for the knn, svmLinear, and xgbTree models, which calculated the feature importance. *P*-values were calculated using the following formula:

$$Pvalue = \frac{b}{m}$$

where b is the number of permutations that yielded values equal to or greater than the original accuracy and m is the total number of permutations (100 times).

2.4. Somatic mutation analysis using RNA-Seq data

Somatic mutation analysis was conducted using the SRP114315 samples with reference to the GATK Best Practices workflow for RNA-Seq data (https://github.com/broadgsa/gatk/blob/master/doc_archive/methods/Calling_variants_in_RNAseq.md). The reads were trimmed as previously described and then mapped to the hs37d5 reference human genome from the 1000 Genome Project (https://www.internationalgenome.org/) using multi-sample 2-pass mapping in STAR version 2.7.0b (https://github.com/alexdobin/STAR?tab=readme-ov-file) (Dobin et al., 2013) with the outFilterMismatchNmax 2 option to generate BAM files. These files were input into the MarkDuplicates of Picard (version 2.21.8, https://broadinstitute.github.io/picard/) to identify and tag duplicate reads. The marked BAM files were sorted using Picard SortSam with the SORT_ORDER=coordinate option, input into SplitNCigarReads of GATK version 4.1.4.1 (https://gatk.broadinstitute.org/) to split reads that contained Ns in their cigar string, and then input into BaseRecalibrator of GATK to generate a recalibration table for Base Quality Score Recalibration (BQSR). The BQSR and BAM files were then inputted into ApplyBQSR of GATK to apply the BQSR results. The final BAM files were input into the GATK Mutect2 to call somatic mutations, generating variant call format (VCF) files for each sample. SnpEff version 4.3 t (https://pcingola.github.io/SnpEff/) was used for variant annotation and impact prediction in the VCF files, and variants with high or moderate impact were filtered through SnpSift (version 4.3 t). The selected VCF files were converted into mutation annotation format (MAF) files using vcf2maf (version v1.6.17). Finally, MAF files

Table 1
Public RNA sequencing data used in this study.

Accession ID	Disease type	No. of diseased samples	No. of normal samples	Data from	Ref.
GSE92592	IPF	20	19	U.S.A	(Schafer et al., 2017)
GSE99621	IPF	18	8	U.S.A	(Luzina et al., 2018)
GSE52463	IPF	8	7	U.S.A	(Nance et al., 2014)
GSE83717	IPF	6	5	Serbia	(Vukmirovic et al., 2017)
GSE138239	IPF	11	4	U.S.A	(Yin et al., 2020)
SRP114315	SCC	101	101	Korea	(Seo et al., 2018)
GSE81089	SCC	67	19	Sweden	(Mezheyeuski et al., 2018)

were integrated using the merge_mafs function of the maftools package (version 2.10.05) in R. When conducting somatic mutation analysis, whole genome sequencing or whole exome sequencing (WES) is commonly used because RNA-Seq data may have insufficient depth. Therefore, when using RNA-Seq, mutations detected in only few samples have low confidence. To address this issue, we focused on mutations detected in > 50 % samples.

2.5. Network construction

Network construction consisted of two steps. The first step detected 13,538 mutated genes in 101 MAF files. After filtering out genes with < 50 % mutation frequency across all samples, 727 genes remained. Of these, 26 genes were identified as transcription factor (TF)-encoding genes using the Human Transription Factors, a public database (http://humantfs.ccbr.utoronto.ca/index.php) (Lambert et al., 2018). In the second step, we examined the TFs regulating the expression of the 20 genes identified via feature importance using ChIP-Seq data from Encyclopedia of DNA Elements (ENCODE) (Consortium, 2011) downloaded from Harmonizome 3.0 (https://maayanlab.cloud/Harmonizome/). In total, 22,819 downstream genes were identified using ENCODE ChIP-Seq data.

2.6. Statistical analyses

Classical statistical methods were performed in order to identify the differentially expressed genes (DEGs). DEGs were determined using Welch's *t*-test and Storey's method with a threshold of FDR < 0.001 (Storey and Tibshirani, 2003). We defined the top 10 genes with the highest fold-change that were more highly expressed in IPF than in SCC, and the top 10 genes with the highest fold-change that were more highly expressed in SCC than in IPF as the DEGs. AUC values were calculated using the roc function of the pROC package in R. Spearman's rank correlation coefficient was performed using the cor function.

3. Results

3.1. BEC among multiple datasets

To detect differences in gene expression between IPF and SCC, we constructed ML models using public RNA-Seq datasets. We focused on protein-coding genes, lncRNAs, and miRNAs involved in lung cancer and IPF development (Ali et al., 2022; Hadjicharalambous and Lindsay, 2020; Schmitt and Chang, 2016; Wang et al., 2015, 2020). The ML model construction strategy is illustrated in Fig. 1. BEC was necessary because multiple datasets were deposited in different countries (Table 1). The pre-BEC training and test data were projected onto a

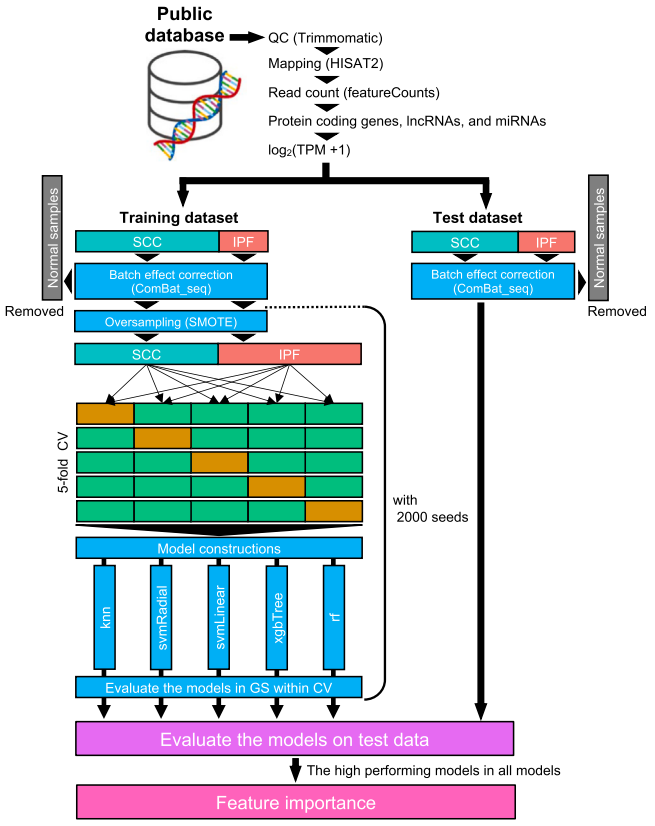


Fig. 1. Overview of machine learning (ML) model construction and validation using test data. RNA sequencing data for both the training and testing datasets were obtained from a public database. Subsequently, the data were analyzed and batch-effect corrected using Combat_seq. Training data imbalance was addressed using the synthetic minority oversampling technique (SMOTE), and models were constructed using five algorithms through 5-fold cross-validation and a grid search. From the SMOTE to grid search cross-validation, 2000 seed values were used. The model accuracies were evaluated using test data, and the feature importance of the highest-performing model was calculated. QC, quality control; GS, grid search; CV, cross-validation; SCC, squamous cell carcinoma; IPF, idiopathic pulmonary fibrosis.

UMAP embedding for each dataset (Supplementary Fig. 1), whereas the post-BEC training and test data were projected for each sample category with a distinct separation in the training data (Fig. 2).

3.2. ML model construction

Biomedical datasets for rare diseases, such as IPF, are often severely imbalanced, rendering most ML algorithms unsuitable (Mirza et al., 2019). To address this, we used SMOTE, a common tool when performing ML with RNA-Seq data (Chen et al., 2022; Mahin et al., 2022), to oversample IPF and downsample SCC in the training data. We then constructed classification models using five algorithms to classify IPF and SCC with 2000 different seed values based on the known impact of seed values on ML model performance (Goldberg, 2017). Optimal hyperparameters were searched during the learning process using a grid search and evaluated via 5-fold cross-validation (Supplementary Fig. 2). The model with the highest performance was selected for all seed values. Owing to test data imbalance, the final performance evaluations were conducted based on the kappa values. The kappa values for knn, svmLinear, and xgbTree were generally high, whereas those for svmRadial and rf were low (Fig. 3A). The highest kappa values for knn, svmRadial, svmLinear, xgbTree, and rf were 0.7670, 0.4909, 0.7630, 0.7855, and 0.6383, respectively (Table 2). The AUC and receiver operating characteristic (ROC) curves of each model with the highest

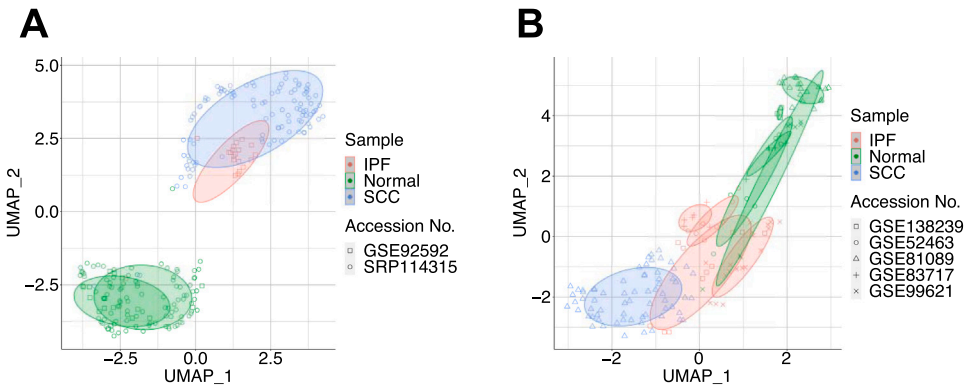


Fig. 2. Uniform manifold approximation and projection (UMAP) of training and test data after batch-effect correction. Plot colors indicate sample categories, whereas plot shapes indicate the accession numbers of publicly available RNA sequencing data. IPF: idiopathic pulmonary fibrosis; SCC: squamous cell carcinoma. **(A)** training and **(B)** test data.

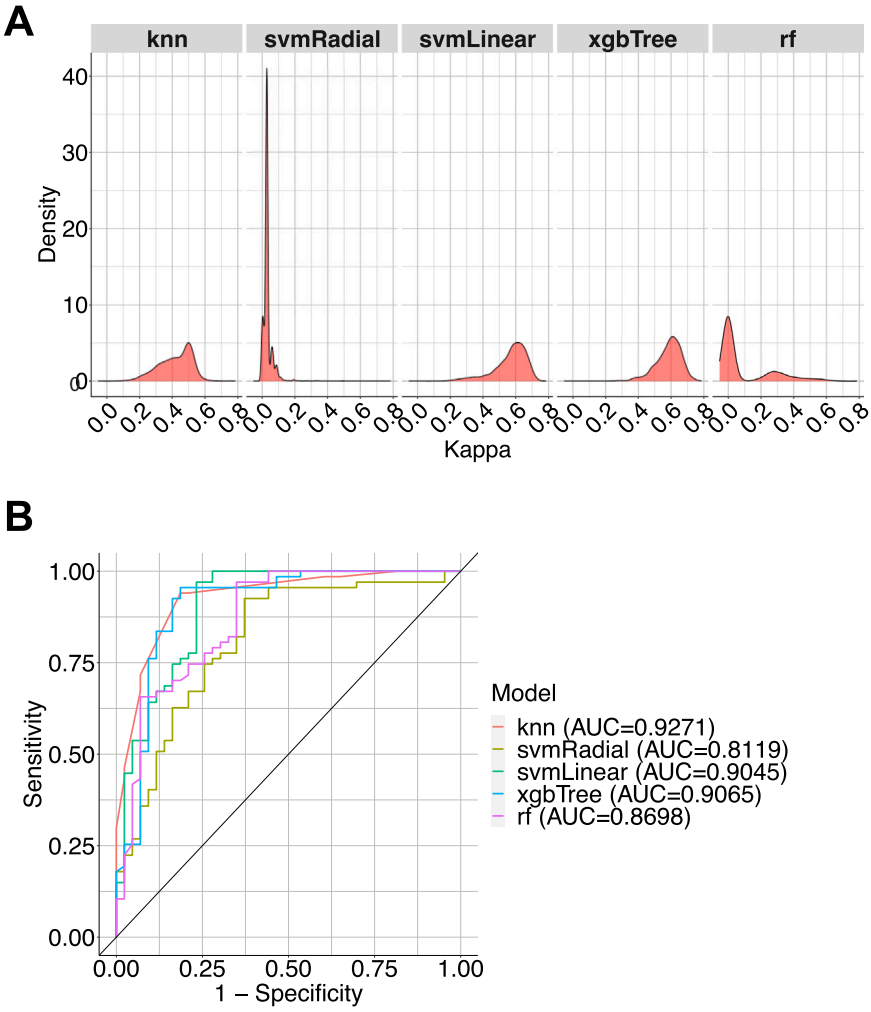


Fig. 3. Model performance in each algorithm validated using test data. **(A)** Kappa value density in each algorithm. **(B)** Receiver operating characteristic curve of the best-performing model in each algorithm.

kappa values are shown in Fig. 3B. The accuracy, kappa, and AUC in 5-fold cross-validation of each model with the highest kappa value were approximately 1 for each fold and each algorithm (Supplementary Fig. 3). According to Landis *et al.*, kappa value 0.61–0.80 is considered "substantial" and 0.81–1.00 is considered "almost perfect" (Landis and Koch, 1977). Thus, we selected the knn, svmLinear, and xgbTree models with the highest kappa values close to 0.8 for downstream analysis.

To evaluate the effect of BEC on model performance, the classification models were constructed using a learning process without applying BEC to the training data. Consequently, the accuracy decreased significantly (Supplementary Fig. 4).

Table 2
Kappa value statistics in each algorithm validated against test data.

	knn	svmRadial	svmLinear	xgbTree	rf
Max	0.7670	0.4909	0.7630	0.7855	0.6383
Mean	0.4172	0.03324	0.5658	0.5845	0.09426
Median	0.4361	0.02819	0.5871	0.5945	0
Min	0.09213	0	0.1669	0.2364	−0.05373

3.3. Feature importance

Significant genes were identified using the classification models (knn, svmLinear, and xgbTree) based on feature importance. The top 10 genes included melanoma-associated antigen 4 (*MAGEA4*) in all models; a class A scavenger receptor (*SCARA5*), a member of phospholipase A2 (*PLA2G2A*), *ENSG00000234638*, *HPSE2*, *ENSG00000279712*, desmin (*DES*), and a human homolog of the *Drosophila* prune gene (*PRUNE2*) in knn and svmLinear; and proteoglycan 4 (*PRG4*) in svmLinear and xgbTree. *ENSG00000250920* and pregnancy zone protein (*PZP*) were

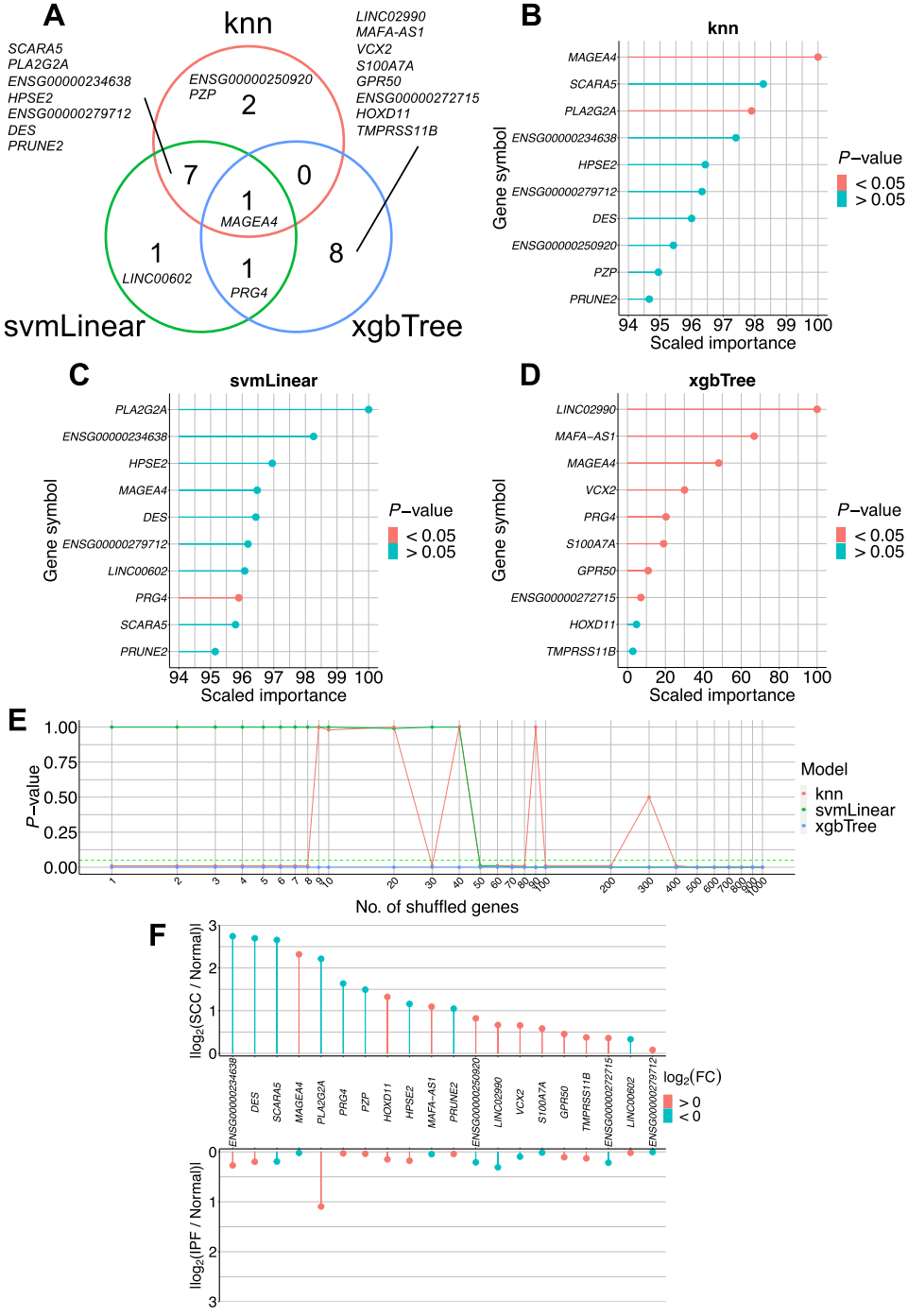


Fig. 4. Identification of significant genes based on feature importance. **(A)** Overlap among the top 10 genes based on feature importance in knn, svmLinear, and xgbTree. Top 10 genes based on feature importance in **(B)** knn, **(C)** svmLinear, and **(D)** xgbTree. **(E)** P-values obtained from the permutation test for each number of shuffled genes. The x-axis is shown on a logarithmic scale. **(F)** Log_2(FC) values of the top 20 differentially expressed genes between normal and IPF, or normal and SCC. Red and blue indicate positive and negative values, respectively. FC; fold change; SCC, squamous cell carcinoma; IPF, idiopathic pulmonary fibrosis.

exclusive to knn; *LINC00602* to svmLinear; and *LINC02990*, *MAFA-AS1*, *VCX2*, *S100A7A*, *GPR50*, *ENSG00000272715*, Homeobox D11 (*HOXD11*), and transmembrane serine protease (*TMPRSS11B*) to xgbTree (Fig. 4A). The scaled importance of the top 10 genes was > 90 in knn and svmLinear (Fig. 4B, C), and 2.83–100 in xgbTree (Fig. 4D). The *P*-values of some genes when performing the permutation test were not < 0.05 (Fig. 4B–D). However, when multiple genes were shuffled simultaneously in the descending order of feature importance, the model performance decreased below the original level at the point of shuffling the top 50 and 400 genes for svmLinear and knn, respectively (Fig. 4E). For xgbTree, performance fell below the original level when at least one top-ranked gene was shuffled (Fig. 4E). The expression levels of these 20 genes were substantially and slightly altered in SCC and IPF lung tissue, respectively, than in normal lung tissue (Fig. 4F).

3.4. Somatic mutation analysis of TFs regulating gene expression identified via feature importance

Intratumor heterogeneity refers to cellular variations within tumors (Ono et al., 2021), with unique genomic profiles even in adjacent tumors of the same patient; this phenomenon is also observed in SCC (Bruin et al., 2014). This makes it challenging to associate gene expression data from a tumor sample with somatic mutation data from an adjacent sample. Although the dataset with accession no. SRP114315 includes both RNA-Seq and WES data, we used the RNA-Seq data to identify genomic profiles causing gene expression differences, as the tumor tissue samples for RNA-Seq and WES were different (Seo et al., 2018). Genome-wide sequencing studies have detected a higher number of somatic mutations in NSCLC than in other cancers, indicating a significant role for gene mutations in NSCLC onset (Vogelstein et al., 2013). Therefore, we investigated somatic mutations affecting gene expression, focusing on mutations in TF-encoding genes that directly affect gene expression. The identification process shown in Fig. 5 consists of two steps. Variant classification after the first step revealed that frameshift insertions were the most common mutations, followed by missense mutations, with a median of 32 variants per sample (Fig. 6A, B). Twenty-six TFs with mutations in > 50 % samples are shown in Fig. 6C and Supplementary Fig. 5. After the second step, 13 downstream genes overlapped with the important feature genes detected by the ML models and chromatin immunoprecipitation sequencing (ChIP-Seq) data from the ENCODE, which were regulated by one or more of the 128 TFs (Fig. 7A, Supplementary Data 1).

The TFs *MYC*, *BHLHE40*, *STAT1*, and *E2F4* were identified as overlaps between the 26 TFs from the first step and the 128 TFs from the second step. All somatic mutations detected in the four TF genes are shown in Supplementary Data 2. In *MYC*, *BHLHE40*, and *STAT1*, Frame_Shift_Ins mutations were the most frequently detected, followed by Missense_Mutation (Fig. 6C and Supplementary Fig. 6A–C). The most somatic mutations in *E2F4* were In_Frame_Del (Fig. 6C and Supplementary Fig. 6D). The most common mutations in *BHLHE40* and *E2F4* were not located in domains (Supplementary Fig. 6A, D). In contrast, the most frequent mutations in *MYC* and *STAT1* were located in the HLH and STAT_alpha domains, respectively (Supplementary Fig. 6B, C). Less frequent mutations were scattered throughout the sequence (Supplementary Fig. 6A–D). Based on ENCODE data, the expression of 11 downstream genes was regulated by one or more of these four TFs (Fig. 7B). In addition, missense mutations in these four TFs were detected in at least one of the 14 samples in the TCGA-LUSC dataset (Supplementary Fig. 7).

4. Discussion

In this study, we constructed ML models to classify patients with IPF or SCC using RNA-Seq data from seven datasets. The benefits of merging gene expression datasets to improve the ML model performance are often undermined by batch effects, which are unwanted data variations caused by the technical differences across batches (Zhang et al., 2020). Therefore, minimizing these batch effects, particularly for training data, is crucial. The training data after BEC were distinctly distributed for each sample category, indicating successful correction of batch effects. Although the test data were more widely spread than the training data, the normal and diseased samples were separately distributed, indicating successful BEC. The model performance significantly decreased when BEC was not applied to the training data, indicating that BEC application to the training data significantly contributed to the model performance.

ML methods are particularly useful when dealing with data where the number of samples (n) is considerably greater than the number of features (p). However, the scenario is reversed in omics data, with the number of features (such as genes and proteins) significantly exceeding the number of samples, leading to a data structure expressed as $n << p$ (Teschendorff, 2019). This can result in overfitting when ML methods are applied to the omics data. Overfitting occurs when the derived predictive model fits a random variation in the data, which does not represent the true biological variation associated with the phenotype of interest (Simon et al., 2003; Teschendorff, 2019). Therefore, the performance of ML models should be validated using test data constructed independently from training data when conducting ML with omics data (Teschendorff, 2019). When training complex structures, employing

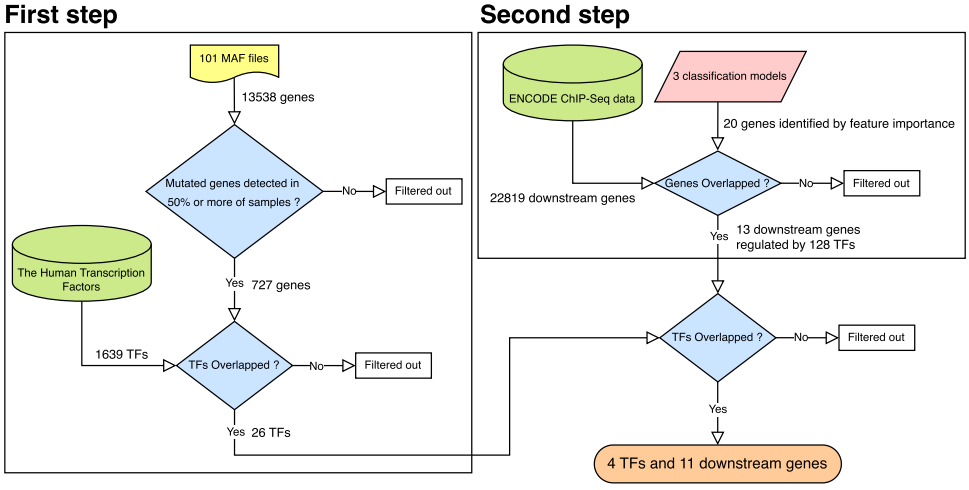


Fig. 5. Flowchart for the identification of somatic mutations in transcription factors (TFs) regulating the expression of genes identified via feature importance. Diagram was generated using draw.io (https://app.diagrams.net/). MAF, mutation annotation format; ENCODE, Encyclopedia of DNA Elements.

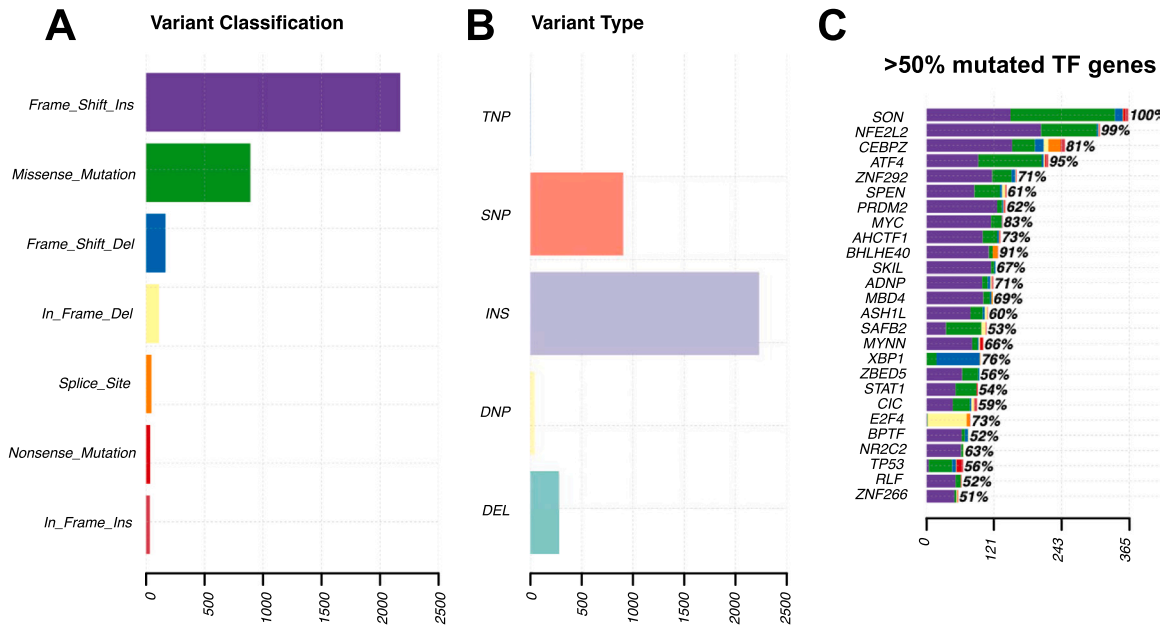


Fig. 6. Summary of somatic mutation analysis in squamous cell carcinoma (SCC) samples. Variant **(A)** classification and **(B)** type. In both cases, the horizontal axis represents the number of variants detected across all samples. **(C)** List of transcription factors (TFs) with a mutation in > 50 % samples. Horizontal axis represents the number of variants detected for each TF. Plots were constructed using the plotmafSummary function of the maftools package (version 2.10.05) in R. TNP, triple nucleotide polymorphism; SNP, Single nucleotide polymorphism; INS, Insertion; DNP, Double nucleotide polymorphism; DEL, Deletion.

various random seeds can result in different final solutions, each with its own accuracy (Goldberg, 2017). This method is known as random restart (Goldberg, 2017). In a random restart, the algorithm begins with several initial solutions (random initial states) and conducts an optimization process for each solution. In other words, the same problem is addressed multiple times under various initial conditions, and the optimal solution is selected based on the outcomes of each attempt. This method is beneficial as it mitigates the risk of local optimum entrapment, thus enabling an extensive solution search by testing multiple initial solutions. However, this could increase computational costs, necessitating efficient implementation. Therefore, the training process should be executed multiple times with varying random seeds when computational resources are available, followed by selecting the random seed that performs best on the development sets (Goldberg, 2017). In this study, the random restart technique was employed to select a high-performance classification model. The classification models exhibited high performance in 5-fold cross-validation, indicating potential overfitting. Therefore, independently constructed test data were used for the evaluation with 2000 seed values. The svmLinear, knn, and xgbTree classification models were selected based on their kappa values after validation against the test data. The kappa values for all the three models exceeded 0.75, which indicates substantial model performance (Landis and Koch, 1977). This demonstrates the feasibility of constructing high-performing ML models, even for omics data with $n << p$, when independently constructed test data and random restarts technique are available. As mentioned above, the classification model constructed in this study consistently exhibited high performance during the grid search, regardless of the hyperparameter conditions, indicating that random restarts are more critical and warrant greater consideration than model-specific hyperparameters in studies using ML for omics data with $n << p$. When the goal is to detect differences between two classes by extracting feature importance from the model, as in this study, selecting the seed value that yields the highest accuracy among those obtained through random restarts is effective in achieving a highly accurate model. In contrast, accuracy results may vary depending on the test data used, even with the same seed value. Therefore, the random restart technique may be inappropriate for constructing a general-purpose ML model.

Interestingly, the top 10 feature importance scores calculated based on the svmLinear and knn classification models were 90–100. In contrast, those calculated based on the xgbTree classification model showed a wide range of values from 2 to 100, indicating a high degree of variability. This suggests that the svmLinear and knn models make predictions using a larger geneset, whereas the xgbTree model relies on a smaller and more selective geneset. This was also supported by the results of permutation tests when multiple genes were shuffled. This difference likely stems from algorithmic variations. The top 10 genes with high feature importance in knn and svmLinear largely overlapped, unlike those in xgbTree. Despite this, the kappa values and accuracies of the three models were similar, suggesting that predictions could be made using two distinct gene lists. We also used classical statistical methods to identify the DEGs between patients with IPF and those with SCC. The gene lists differed from those selected based on feature importance in the ML models. This suggests that ML-based methods should not replace classical statistical approaches but it should be used in combination. However, when calculating the AUC for each DEG selected using classical statistical methods, the highest AUC was found to be 0.8671 (Supplementary Table 2). Therefore, except for svmRadial, the ML-based approach demonstrated a higher performance, suggesting its potential to provide more generalizable DEGs between IPF and SCC.

The feature importance based on the three classification models revealed 20 key genes for IPF and SCC classification. SCARA5 suppresses lung cancer cell proliferation, and increased methylation levels in *SCARA5* promoter region are associated with reduced gene expression in patients with SCC (Peng et al., 2021). PLA2G2A induces pyroptosis in alveolar epithelial cells, resulting cell death, thereby contributing to lung tissue destruction, and its gene and protein expressions are high in patients with IPF (Bauer et al., 2015; Jaiswal et al., 2023). In addition, *PLA2G2A* expression is decreased owing to somatic mutations in tumor suppressor genes (TSGs; *TP53*, *CDKN2A*, *PTEN*, *RB1*, and *BRCA1*) in the lung tissues of patients with SCC (Kim et al., 2021). Among the 101 patients with SCC, 78 had somatic mutations in at least one TSG (Supplementary Data 3). Thus, the PLA2G2A reduction in SCC may be caused by mutations in TSGs. Although the role of desmin in lung cancer and IPF is unclear, *DES* expressions are lower in the lung tissues of patients with SCC than in those with IPF (Fallahian et al., 2018). PRUNE2

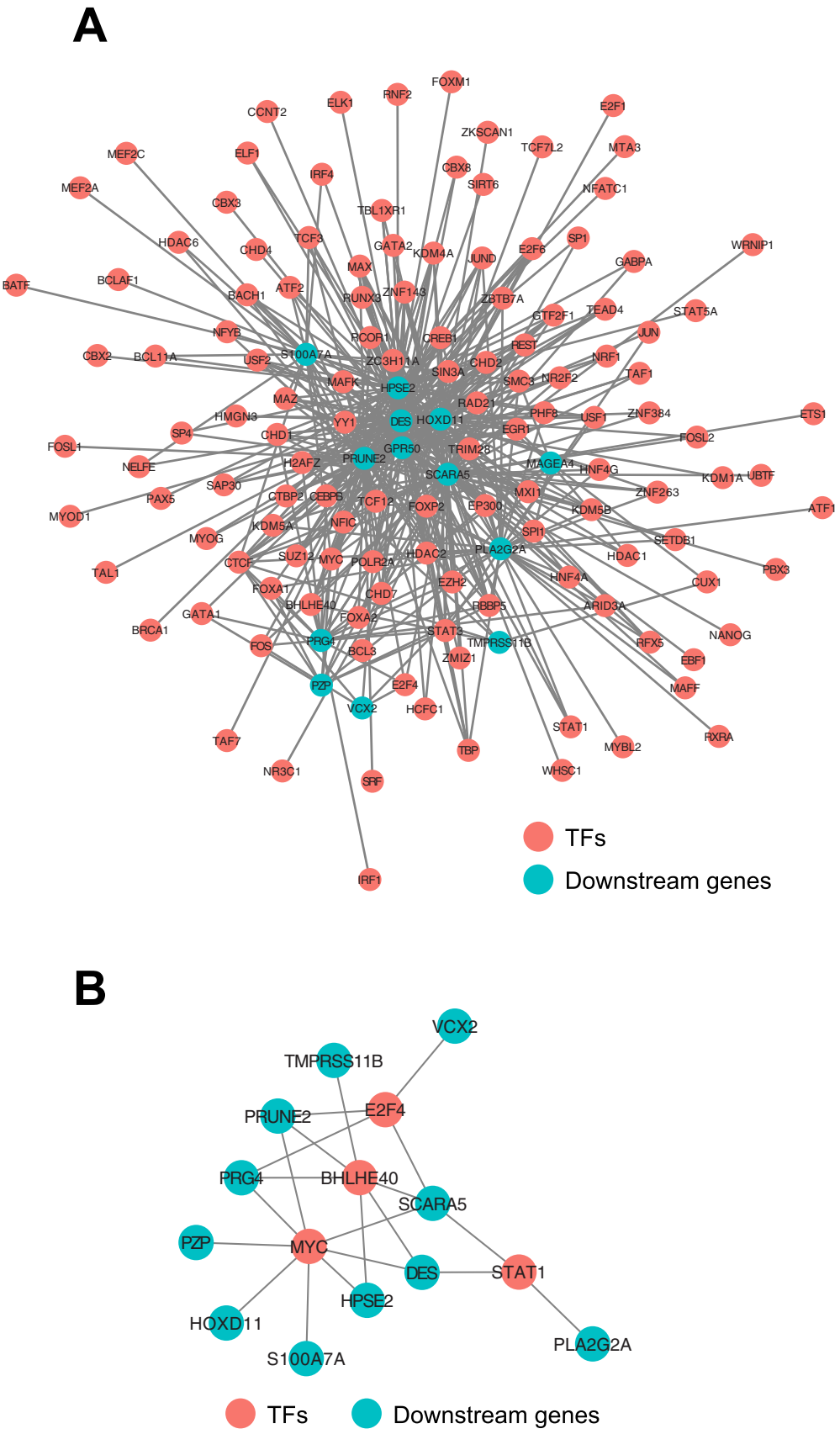


Fig. 7. Molecular networks consisting of transcription factors (TFs) with somatic mutations and the genes regulated by these TFs. **(A)** A network of 13 downstream genes identified based on the second step and 128 TFs regulating these genes. **(B)** A network of four TFs identified based on the first and second steps along with the genes regulated by these TFs according to feature importance. Red and blue nodes represent TFs and downstream genes, respectively. Connections (edges) indicate regulatory relationships between TFs and downstream genes. Networks were visualized using Cytoscape version 3.10.0 (https://cytoscape.org/).

suppresses colorectal and prostate cancer cell proliferation and invasion (Li et al., 2022; Salameh et al., 2015), and its expression is lower in SCC lung tissues than in normal lung tissues (Wu et al., 2020). PZP expression is low in tumor tissues; its downregulation is correlated with poor clinical outcomes in lung ADC and hepatocellular carcinoma, and its gene expression is lower in lung SCC tissues than in normal lung tissues (Chen et al., 2023; Su et al., 2020). PRG4 represses the genesis and metastasis of osteosarcoma (Zhang et al., 2024) and is correlated with longer survival in hepatocellular carcinoma (HCC) patients (Dituri et al., 2020). In addition, its gene expression is downregulated in SCC lung tissues than in normal lung tissues (Wu et al., 2020). HOXD11 promotes cell invasion and metastasis in penile squamous cell carcinoma (Tan et al., 2022), and its gene expression is higher in the lung tissues of patients with lung SCC than in normal lung tissue (Zhang et al., 2017). *MAFA-AS1*, a long non-coding RNA, is a candidate biomarker for poor prognosis in HCC, and its expression is high in the lung tissue of patients with SCC (Tian et al., 2023). TMPRSS11B promotes tumor transformation and growth in lung SCC (Updegraff et al., 2018). S100A7, a member of the S100 multigenic family, promotes lung ADC to squamous carcinoma transdifferentiation and is selectively expressed in lung SCC, but not lung ADC (Wang et al., 2017; Zhang et al., 2008). *MAGEA4* is uniquely expressed in SCC lung tissue, but not in ADC and normal lung tissues (Peikert et al., 2006). VCX2, identified as a cancer/testis antigen, is also expressed in SCC lung tissues (Taguchi et al., 2014). Although *HPSE2* and *GPR50* expression profiles in SCC and IPF remain to be elucidated, they play a role in suppressing tumorigenesis (Kayal et al., 2023; Saha et al., 2020). *LINC00602*, *LINC02990*, *ENSG00000272715*, *ENSG00000234638*, *ENSG00000279712*, and *ENSG00000250920* are listed as lncRNAs in the Ensembl Genome Database. However, despite their potential roles in carcinogenesis, no studies have investigated them. Therefore, further investigation is required to elucidate their roles in SCC and IPF development.

The expression profiles of the above-mentioned genes in IPF and SCC have been partially reported, which is consistent with the findings of this study. Thus, beyond the computational accuracy, the classification model accurately distinguished SCC from IPF.

The top 20 genes identified based on their high feature importance were linked to carcinogenesis rather than fibrosis. Through the analysis of SCC samples, somatic mutations regulating these genes were detected, with a focus on TFs. ChIP-Seq data from ENCODE revealed a connection between these 20 genes and four TFs: *BHLHE40*, *MYC*, *STAT1*, and *E2F4*. Somatic mutations in these four TFs were identified using the TCGA-LUSC dataset despite the low mutation frequency, suggesting that they are not dataset-specific mutations. Somatic mutations in these four TF genes may influence the expression of downstream genes associated with SCC pathology, including *VCX2*, *TMPRSS11B*, *PRUNE2*, *PRG4*, *PZP*, *SCARA5*, *DES*, *HPSE2*, *HOXD11*, *S100A7A*, and *PLA2G2A*. In addition, the gene expression levels of *PRUNE2*, *PRG4*, *PZP*, *SCARA5*, *DES*, *HPSE2*, *S100A7A*, and *PLA2G2A* were related to pathological tumor size and/or stage (Supplementary Fig. 8). In certain cases where a usual interstitial pneumonia pattern is suspected on CT, IPF can be diagnosed without a surgical lung biopsy. However, patients cannot be diagnosed with IPF definitively solely based on clinical findings or CT imaging; additional examinations including surgical lung biopsy may be considered (Lynch et al., 2018). Such mutations and altered gene expression, if present in the biopsy specimen, could be considered potential carcinogenic risk factors. Therefore, these mutations and gene expression levels may serve as biomarker candidates for diagnosing SCC complications in patients with IPF, and for predicting clinical outcomes in patients with both IPF and SCC. With more clinical data available in the future, our results can be directly compared with them.

ML models for classifying patients with SCC and healthy controls, as well as models for classifying SCC subtypes, have been reported with AUCs of 0.965 and 0.819, respectively (Duan et al., 2025; Joon et al., 2023). However, the genes with high feature importance identified by these models are completely different from those identified by the ML models in this study. Therefore, the ML models in this study may specifically capture the genes that differ between IPF and SCC.

In this study, only the RNA-Seq data were used. However, if other omics data could be collected from the lung tissues of patients with IPF and SCC, the differences between the two diseases could be further clarified. Particularly, epigenetic differences between these two diseases are known (Antoniou et al., 2015), making them important omics layers. Future multilayer analysis may enable a clear understanding of the pathophysiological differences between these two diseases by integrating multi-omics data.

5. Conclusions

In this study, we developed ML models to distinguish SCC from IPF, achieving high accuracy even with omics data where $n << p$. The model identified 20 DEGs between IPF and SCC samples, indicating their potential roles in SCC pathogenesis. We also detected somatic mutations in four TFs regulating 11 of the 20 identified genes. These findings enhance our understanding of the molecular mechanisms underlying SCC complications in patients with IPF, and offer new perspectives for preventing SCC complications in these patients.

Author contributions

Y.N. and K.M. conceived and designed the experiments. Y.N. performed the experiments and analyzed the data. Y.N. contributed to the writing of the draft manuscript. K.M. reviewed and edited the draft manuscript. All authors discussed the data and manuscript. All authors have read and approved the final manuscript.

CRediT authorship contribution statement

Yosui Nojima: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kenji Mizuguchi:** Writing – review & editing, Supervision, Resources, Funding acquisition

Funding

This study was supported by the Japan Society for the Promotion of Science KAKENHI grant number JP20K15422, the Uehara Memorial Foundation to Y.N., and the Ministry of Health, Labour and Welfare Program grant number JPMH19AC5001 to K.M.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the members of the Artificial Intelligence Center for Health and Biomedical Research at the National Institutes of Biomedical Innovation, Health and Nutrition for their valuable support and discussions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2025.108560.

References

Ali, M.S., Singh, J., Alam, M.T., Chopra, A., Arava, S., Bhalla, A.S., Mittal, S., Mohan, A., Mitra, D.K., Hadda, V., 2022. Non-coding RNA in idiopathic interstitial pneumonia and Covid-19 pulmonary fibrosis. Mol. Biol. Rep. 49, 11535–11546. https://doi.org/10.1007/s11033-022-07820-4.

Antoniou, K.M., Tomassetti, S., Tsitoura, E., Vancheri, C., 2015. Idiopathic pulmonary fibrosis and lung cancer: a clinical and pathogenesis update. Curr. Opin. Pulm. Med. 21, 626. https://doi.org/10.1097/MCP.0000000000000217.

Ballester, B., Milara, J., Cortijo, J., 2019. Idiopathic pulmonary fibrosis and lung cancer: mechanisms and molecular targets. Int. J. Mol. Sci. 20, 593. https://doi.org/10.3390/ijms20030593.

Bauer, Y., Tedrow, J., de Bernard, S., Birker-Robaczewska, M., Gibson, K.F., Guardela, B. J., Hess, P., Klenk, A., Lindell, K.O., Poirey, S., Renault, B., Rey, M., Weber, E., Nayler, O., Kaminski, N., 2015. A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. Am. J. Respir. Cell Mol. Biol. 52, 217–231. https://doi.org/10.1165/rcmb.2013-0310OC.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bruin, E.C., McGranahan, N., Mitter, R., Salm, M., Wedge, D.C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A.J., Grönroos, E., Muhammad, M.A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rassl, D.M., Rintoul, R.C., Janes, S.M., Lee, S.-M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T.T., Lee, C.C., Tom, W., Teefe, E., Chen, S.-C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P., Swanton, C., 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science 346, 251–256. https://doi.org/10.1126/science.1253462.

Bzdok, D., 2017. Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. 11.

Bzdok, D., Krzywinski, M., Altman, N., 2017. Machine learning: a primer. Nat. Methods 14, 1119–1120. https://doi.org/10.1038/nmeth.4526.

Bzdok, D., Altman, N., Krzywinski, M., 2018. Statistics versus machine learning. Nat. Methods 15, 233–234. https://doi.org/10.1038/nmeth.4642.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, J., Wang, X., Ma, A., Wang, Q.-E., Liu, B., Li, L., Xu, D., Ma, Q., 2022. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. Nat. Commun. 13, 6494. https://doi.org/10.1038/s41467-022-34277-7.

Chen, K., Zheng, T., Chen, C., Liu, L., Guo, Z., Peng, Y., Zhang, X., Yang, Z., 2023. Pregnancy zone protein serves as a prognostic marker and favors immune infiltration in lung adenocarcinoma. Biomedicines 11, 1978. https://doi.org/10.3390/biomedicines11071978.

Consortium, T.E.P., 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLOS Biol. 9, e1001046. https://doi.org/10.1371/journal.pbio.1001046.

Dituri, F., Scialpi, R., Schmidt, T.A., Frusciante, M., Mancarella, S., Lupo, L.G., Villa, E., Giannelli, G., 2020. Proteoglycan-4 is correlated with longer survival in HCC patients and enhances sorafenib and regorafenib effectiveness via CD44 in vitro. Cell Death Dis. 11, 1–14. https://doi.org/10.1038/s41419-020-03180-8.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Duan, G., Huo, Q., Ni, W., Ding, F., Ye, Y., Tang, T., Dai, H., 2025. Integrative machine learning model for subtype identification and prognostic prediction in lung squamous cell carcinoma. Discov. Onc 16, 886. https://doi.org/10.1007/s12672-025-02560-w.

Fallahian, F., Moosavi, S.A.J., Mahjoubi, F., Shabani, S., Babaheidarian, P., Majidzadeh, T., 2018. Evaluation of desmin, α-SMA and hTERT expression in pulmonary fibrosis and lung cancer. J. Clin. Intensive Care Med. 3 001–009. https://doi.org/10.29328/journal.jcicm.1001011.

Gakii, C., Mukami, V., Too, B., 2023. Feature selection for classification using WGCNA and Spread Sub-Sample for an imbalanced rheumatoid arthritis RNASEQ data. Inform. Med. Unlocked 43, 101402. https://doi.org/10.1016/j.imu.2023.101402.

Goldberg, Y., 2017. Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02165-7.

Hadjicharalambous, M.R., Lindsay, M.A., 2020. Idiopathic pulmonary fibrosis: pathogenesis and the emerging role of long non-coding RNAs. Int. J. Mol. Sci. 21, 524. https://doi.org/10.3390/ijms21020524.

Hata, A., Nakajima, T., Matsusaka, K., Fukuyo, M., Nakayama, M., Morimoto, J., Ito, Y., Yamamoto, T., Sakairi, Y., Rahmutulla, B., Ota, S., Wada, H., Suzuki, H., Iwata, T., Matsubara, H., Ohara, O., Yoshino, I., Kaneda, A., 2021. Genetic alterations in squamous cell lung cancer associated with idiopathic pulmonary fibrosis. Int. J. Cancer 148, 3008–3018. https://doi.org/10.1002/ijc.33499.

Jaiswal, A., Rehman, R., Dutta, J., Singh, S., Ray, A., Shridhar, M., Jaisankar, J., Bhatt, M., Khandelwal, D., Sahoo, B., Ram, A., Mabalirajan, U., 2023. Cellular distribution of secreted phospholipase A2 in lungs of IPF patients and its inhibition in bleomycin-induced pulmonary fibrosis in mice. Cells 12, 1044. https://doi.org/10.3390/cells12071044.

Joon, H.K., Thalor, A., Gupta, D., 2023. Machine learning analysis of lung squamous cell carcinoma gene expression datasets reveals novel prognostic signatures. Comput. Biol. Med. 165, 107430. https://doi.org/10.1016/j.compbiomed.2023.107430.

Kayal, Y., Barash, U., Naroditsky, I., Ilan, N., Vlodavsky, I., 2023. Heparanase 2 (Hpa2)- a new player essential for pancreatic acinar cell differentiation. Cell Death Dis. 14, 1–14. https://doi.org/10.1038/s41419-023-05990-y.

Kim, A., Lim, S.M., Kim, J.-H., Seo, J.-S., 2021. Integrative genomic and transcriptomic analyses of tumor suppressor genes and their role on tumor microenvironment and immunity in lung squamous cell carcinoma. Front. Immunol. 12.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The human transcription factors. Cell 172, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174. https://doi.org/10.2307/2529310.

Li, T., Huang, S., Yan, W., Zhang, Y., Guo, Q., 2022. FOXF2 regulates PRUNE2 transcription in the pathogenesis of colorectal cancer, 15330338221118717 Technol. Cancer Res Treat. 21. https://doi.org/10.1177/15330338221118717.

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

Liñares Blanco, J., Gestal, M., Dorado, J., Fernandez-Lozano, C., 2019. Differential gene expression analysis of RNA-seq data using machine learning for cancer research. In: Tsihrintzis, G.A., Virvou, M., Sakkopoulos, E., Jain, L.C. (Eds.), Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems. Springer International Publishing, Cham, pp. 27–65. https://doi.org/10.1007/978-3-030-15628-2_3.

Luzina, I.G., Salcedo, M.V., Rojas-Peña, M.L., Wyman, A.E., Galvin, J.R., Sachdeva, A., Clerman, A., Kim, J., Franks, T.J., Britt, E.J., Hasday, J.D., Pham, S.M., Burke, A.P., Todd, N.W., Atamas, S.P., 2018. Transcriptomic evidence of immune activation in macroscopically normal-appearing and scarred lung tissues in idiopathic pulmonary fibrosis. Cell. Immunol. 325, 1–13. https://doi.org/10.1016/j.cellimm.2018.01.002.

Lynch, D.A., Sverzellati, N., Travis, W.D., Brown, K.K., Colby, T.V., Galvin, J.R., Goldin, J.G., Hansell, D.M., Inoue, Y., Johkoh, T., Nicholson, A.G., Knight, S.L., Raoof, S., Richeldi, L., Ryerson, C.J., Ryu, J.H., Wells, A.U., 2018. Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. Lancet Respir. Med. 6, 138–153. https://doi.org/10.1016/S2213-2600(17)30433-2.

Mahin, K.F., Robiuddin, Md, Islam, M., Ashraf, S., Yeasmin, F., Shatabda, S., 2022. PanClassif: improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. Genomics 114, 110264. https://doi.org/10.1016/j.ygeno.2022.01.001.

Matsushita, H., Tanaka, S., Saiki, Y., Hara, M., Nakata, K., Tanimura, S., Banba, J., 1995. lung cancer associated with usual interstitial pneumonia. Pathol. Int. 45, 925–932. https://doi.org/10.1111/j.1440-1827.1995.tb03417.x.

Mezheyeuski, A., Bergsland, C.H., Backman, M., Djureinovic, D., Sjöblom, T., Bruun, J., Micke, P., 2018. Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. J. Pathol. 244, 421–431. https://doi.org/10.1002/path.5026.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., Ping, P., 2019. Machine learning and integrative analysis of biomedical big data. Genes 10, 87. https://doi.org/10.3390/genes10020087.

Nance, T., Smith, K.S., Anaya, V., Richardson, R., Ho, L., Pala, M., Mostafavi, S., Battle, A., Feghali-Bostwick, C., Rosen, G., Montgomery, S.B., 2014. Transcriptome analysis reveals differential splicing events in IPF lung tissue. PLOS ONE 9, e92111. https://doi.org/10.1371/journal.pone.0092111.

Ono, H., Arai, Y., Furukawa, E., Narushima, D., Matsuura, T., Nakamura, H., Shiokawa, D., Nagai, M., Imai, T., Mimori, K., Okamoto, K., Hippo, Y., Shibata, T., Kato, M., 2021. Single-cell DNA and RNA sequencing reveals the dynamics of intra-tumor heterogeneity in a colorectal cancer model. BMC Biol. 19, 207. https://doi.org/10.1186/s12915-021-01147-5.

Ozawa, Y., Suda, T., Naito, T., Enomoto, N., Hashimoto, D., Fujisawa, T., Nakamura, Y., Inui, N., Nakamura, H., Chida, K., 2009. Cumulative incidence of and predictive factors for lung cancer in IPF. Respirology 14, 723–728. https://doi.org/10.1111/j.1440-1843.2009.01547.x.

Peikert, T., Specks, U., Farver, C., Erzurum, S.C., Comhair, S.A.A., 2006. Melanoma antigen A4 Is expressed in non–small cell lung cancers and promotes apoptosis. Cancer Res. 66, 4693–4700. https://doi.org/10.1158/0008-5472.CAN-05-3327.

Peng, Q., Liu, Y., Kong, X., Xian, J., Ye, L., Yang, L., Guo, S., Zhang, Y., Zhou, L., Xiang, T., 2021. The novel methylation biomarker SCARA5 sensitizes cancer cells to DNA damage chemotherapy drugs in NSCLC. Front. Oncol. 11.

Raghu, G., Rochwerg, B., Zhang, Y., Garcia, C.A.C., Azuma, A., Behr, J., Brozek, J.L., Collard, H.R., Cunningham, W., Homma, S., Johkoh, T., Martinez, F.J., Myers, J., Protzko, S.L., Richeldi, L., Rind, D., Selman, M., Theodore, A., Wells, A.U., Hoogsteden, H., Schünemann, H.J., 2015. An official ATS/ERS/JRS/ALAT clinical practice guideline: treatment of idiopathic pulmonary fibrosis. an update of the 2011 clinical practice guideline. Am. J. Respir. Crit. Care Med 192, e3–e19. https://doi.org/10.1164/rccm.201506-1063ST.

Saha, S.K., Choi, H.Y., Yang, G.-M., Biswas, P.K., Kim, K., Kang, G.-H., Gil, M., Cho, S.-G., 2020. GPR50 promotes hepatocellular carcinoma progression via the notch signaling pathway through direct interaction with ADAM17. Mol. Ther. Oncolytics 17, 332–349. https://doi.org/10.1016/j.omto.2020.04.002.

Salameh, A., Lee, A.K., Cardó-Vila, M., Nunes, D.N., Efstathiou, E., Staquicini, F.I., Dobroff, A.S., Marchiò, S., Navone, N.M., Hosoya, H., Lauer, R.C., Wen, S., Salmeron, C.C., Hoang, A., Newsham, I., Lima, L.A., Carraro, D.M., Oliviero, S., Kolonin, M.G., Sidman, R.L., Do, K.-A., Troncoso, P., Logothetis, C.J., Brentani, R.R., Calin, G.A., Cavenee, W.K., Dias-Neto, E., Pasqualini, R., Arap, W., 2015. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA

PCA3. Proc. Natl. Acad. Sci. 112, 8403–8408. https://doi.org/10.1073/pnas.1507882112.
Sato, T., Teramukai, S., Kondo, H., Watanabe, A., Ebina, M., Kishi, K., Fujii, Y., Mitsudomi, T., Yoshimura, M., Maniwa, T., Suzuki, K., Kataoka, K., Sugiyama, Y., Kondo, T., Date, H., 2014. Impact and predictors of acute exacerbation of interstitial lung diseases after pulmonary resection for lung cancer. J. Thorac. Cardiovasc. Surg. 147, 1604–1611.e3. https://doi.org/10.1016/j.jtcvs.2013.09.050.
Sato, T., Watanabe, A., Kondo, H., Kanzaki, M., Okubo, K., Yokoi, K., Matsumoto, K., Marutsuka, T., Shinohara, H., Teramukai, S., Kishi, K., Ebina, M., Sugiyama, Y., Meinoshin, O., Date, H., 2015. Long-term results and predictors of survival after surgical resection of patients with lung cancer and interstitial lung diseases. J. Thorac. Cardiovasc. Surg. 149, 64–70.e2. https://doi.org/10.1016/j.jtcvs.2014.08.086.
Schafer, M.J., White, T.A., Iijima, K., Haak, A.J., Ligresti, G., Atkinson, E.J., Oberg, A.L., Birch, J., Salmonowicz, H., Zhu, Y., Mazula, D.L., Brooks, R.W., Fuhrmann-Stroissnigg, H., Pirtskhalava, T., Prakash, Y.S., Tchkonia, T., Robbins, P.D., Aubry, M.C., Passos, J.F., Kirkland, J.L., Tschumperlin, D.J., Kita, H., LeBrasseur, N. K., 2017. Cellular senescence mediates fibrotic pulmonary disease. Nat. Commun. 8, 14532. https://doi.org/10.1038/ncomms14532.
Schmitt, A.M., Chang, H.Y., 2016. Long Noncoding RNAs in Cancer Pathways. Cancer Cell 29, 452–463. https://doi.org/10.1016/j.ccell.2016.03.010.
Seo, J.-S., Lee, J.W., Kim, A., Shin, J.-Y., Jung, Y.J., Lee, S.B., Kim, Y.H., Park, S., Lee, H. J., Park, I.-K., Kang, C.-H., Yun, J.-Y., Kim, J., Kim, Y.T., 2018. Whole Exome and Transcriptome Analyses Integrated with Microenvironmental Immune Signatures of Lung Squamous Cell Carcinoma. Cancer Immunol. Res 6, 848–859. https://doi.org/10.1158/2326-6066.cir-17-0453.
Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M., 2003. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. JNCI Journal National Cancer Institute 95, 14–18. https://doi.org/10.1093/jnci/95.1.14.
Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 100, 9440–9445. https://doi.org/10.1073/pnas.1530509100.
Su, L., Zhang, G., Kong, X., 2020. Prognostic Significance of Pregnancy Zone Protein and Its Correlation with Immune Infiltrates in Hepatocellular Carcinoma. CMAR 12, 9883–9891. https://doi.org/10.2147/CMAR.S269215.
Taguchi, A., Taylor, A.D., Rodriguez, J., Çeliktaş, M., Liu, H., Ma, X., Zhang, Q., Wong, C.-H., Chin, A., Girard, L., Behrens, C., Lam, W.L., Lam, S., Minna, J.D., Wistuba, I.I., Gazdar, A.F., Hanash, S.M., 2014. A Search for Novel Cancer/Testis Antigens in Lung Cancer Identifies VCX/Y Genes, Expanding the Repertoire of Potential Immunotherapeutic Targets. Cancer Res. 74, 4694–4705. https://doi.org/10.1158/0008-5472.CAN-13-3725.
Tan, X., Liu, Z., Wang, Y., Wu, Z., Zou, Y., Luo, S., Tang, Y., Chen, D., Yuan, G., Yao, K., 2022. miR-138-5p-mediated HOXD11 promotes cell invasion and metastasis by activating the FN1/MMP2/MMP9 pathway and predicts poor prognosis in penile squamous cell carcinoma. Cell Death Dis. 13, 1–14. https://doi.org/10.1038/s41419-022-05261-2.
Teschendorff, A.E., 2019. Avoiding common pitfalls in machine learning omic data science. Nat. Mater. 18, 422–427. https://doi.org/10.1038/s41563-018-0241-z.
Tian, Q., Liu, X., Li, A., Wu, H., Xie, Y., Zhang, H., Wu, F., Chen, Y., Bai, C., Zhang, X., 2023. LINC01936 inhibits the proliferation and metastasis of lung squamous cell carcinoma probably by EMT signaling and immune infiltration. PeerJ 11, e16447. https://doi.org/10.7717/peerj.16447.
Tomassetti, S., Gurioli, C., Ryu, J.H., Decker, P.A., Ravaglia, C., Tantalocco, P., Buccioli, M., Piciucchi, S., Sverzellati, N., Dubini, A., Gavelli, G., Chilosi, M., Poletti, V., 2015. The Impact of Lung Cancer on Survival of Idiopathic Pulmonary Fibrosis. Chest 147, 157–164. https://doi.org/10.1378/chest.14-0359.
Turner-Warwick, M., Lebowitz, M., Burrows, B., Johnson, A., 1980. Cryptogenic fibrosing alveolitis and lung cancer. Thorax 35, 496–499. https://doi.org/10.1136/thx.35.7.496.
Updegraff, B.L., Zhou, X., Guo, Y., Padanad, M.S., Chen, P.-H., Yang, C., Sudderth, J., Rodriguez-Tirado, C., Girard, L., Minna, J.D., Mishra, P., DeBerardinis, R.J., O'Donnell, K.A., 2018. Transmembrane Protease TMPRSS11B Promotes Lung Cancer Growth by Enhancing Lactate Export and Glycolytic Metabolism. Cell Rep. 25, 2223–2233.e6. https://doi.org/10.1016/j.celrep.2018.10.100.
Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W., 2013. Cancer Genome Landscapes. Science 339, 1546–1558. https://doi.org/10.1126/science.1235122.
Vukmirovic, M., Herazo-Maya, J.D., Blackmon, J., Skodric-Trifunovic, V., Jovanovic, D., Pavlovic, S., Stojsic, J., Zeljkovic, V., Yan, X., Homer, R., Stefanovic, B., Kaminski, N., 2017. Identification and validation of differentially expressed transcripts by RNA-sequencing of formalin-fixed, paraffin-embedded (FFPE) lung tissue from patients with Idiopathic Pulmonary Fibrosis. BMC Pulm. Med. 17, 15. https://doi.org/10.1186/s12890-016-0356-4.
Wang, J., Li, Z., Ge, Q., Wu, W., Zhu, Q., Luo, J., Chen, L., 2015. Characterization of microRNA transcriptome in tumor, adjacent, and normal tissues of lung squamous cell carcinoma. J. Thorac. Cardiovasc. Surg. 149, 1404–1414.e4. https://doi.org/10.1016/j.jtcvs.2015.02.012.
Wang, R., Li, Y., Hu, E., Kong, F., Wang, J., Liu, J., Shao, Q., Hao, Y., He, D., Xiao, X., 2017. S100A7 promotes lung adenocarcinoma to squamous carcinoma transdifferentiation, and its expression is differentially regulated by the Hippo-YAP pathway in lung cancer cells. Oncotarget 8, 24804–24814. https://doi.org/10.18632/oncotarget.15063.
Wang, Y., Xiao, H., Zhao, F., Li, H., Gao, R., Yan, B., Ren, J., Yang, J., 2020. Decrypting the crosstalk of noncoding RNAs in the progression of IPF. Mol. Biol. Rep. 47, 3169–3179. https://doi.org/10.1007/s11033-020-05368-9.
Wu, Z., Wang, Y.-M., Dai, Y., Chen, L.-A., 2020. POLE2 serves as a prognostic biomarker and is associated with immune infiltration in squamous cell lung cancer. Med Sci. Monit. 26, e921430-1–e921430-11. https://doi.org/10.12659/MSM.921430.
Yin, Q., Strong, M.J., Zhuang, Y., Flemington, E.K., Kaminski, N., de Andrade, J.A., Lasky, J.A., 2020. Assessment of viral RNA in idiopathic pulmonary fibrosis using RNA-seq. BMC Pulm. Med. 20, 81. https://doi.org/10.1186/s12890-020-1114-1.
Yoon, J.H., Nouraie, M., Chen, X., Zou, R.H., Sellares, J., Veraldi, K.L., Chiarchiaro, J., Lindell, K., Wilson, D.O., Kaminski, N., Burns, T., Trejo Bittar, H., Yousem, S., Gibson, K., Kass, D.J., 2018. Characteristics of lung cancer among patients with idiopathic pulmonary fibrosis and interstitial lung disease – analysis of institutional and population data. Respir. Res. 19, 195. https://doi.org/10.1186/s12931-018-0899-4.
Zaman, T., Lee, J.S., 2018. Risk factors for the development of idiopathic pulmonary fibrosis: a review. Curr. Pulmonol. Rep. 7, 118–125. https://doi.org/10.1007/s13665-018-0210-7.
Zhang, F., Chen, X., Wei, K., Liu, D., Xu, X., Zhang, X., Shi, H., 2017. Identification of key transcription factors associated with lung squamous cell carcinoma. Med Sci. Monit. 23, 172–206. https://doi.org/10.12659/MSM.898297.
Zhang, H., Zhao, Q., Chen, Y., Wang, Y., Gao, S., Mao, Y., Li, M., Peng, A., He, D., Xiao, X., 2008. Selective expression of S100A7 in lung squamous cell carcinomas and large cell carcinomas but not in adenocarcinomas and small cell carcinomas. Thorax 63, 352–359. https://doi.org/10.1136/thx.2007.087015.
Zhang, L., Ren, H., Wu, Y., Xue, L., Bai, Y., Wei, D., Wu, Q., 2024. PRG4 represses the genesis and metastasis of osteosarcoma by inhibiting PDL1 expression. Tissue Cell 88, 102409. https://doi.org/10.1016/j.tice.2024.102409.
Zhang, Y., Parmigiani, G., Johnson, W.E., 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinforma. 2, lqaa078. https://doi.org/10.1093/nargab/lqaa078.