# LUAD And LUSC Cancer Classification, Biomarker Identification, and Pathway Analysis Using Overlapping Feature Selection Methods

**Joseph M. Dhahbi** ( dhahbij@cusm.org )
California University of Science and Medicine

**joe w. Chen**
California University of Science and Medicine

**Research Article**

**Keywords:** Lung cancer, feature selection, biomarker, pathway analysis, cancer classification, Overlapping

**Posted Date:** January 29th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-151050/v1

# Abstract

Lung cancer is one of the deadliest cancers in the world. Two of the most common subtypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), have drastically different biological signatures, yet they are often treated similarly and classified together as non-small cell lung cancer (NSCLC). LUAD and LUSC biomarkers are scarce and their distinct biological mechanisms have yet to be elucidated. Many studies have attempted to improve traditional machine learning algorithms or develop novel algorithms to identify biomarkers, but few have used overlapping machine learning or feature selection methods for cancer classification, biomarker identification, or pathway analysis. This study proposes selecting overlapping features as a way to differentiate between cancer subtypes, especially between LUAD and LUSC. Overall, this method achieved classification results comparable to, if not better than, the traditional algorithms. It also identified multiple known biomarkers, and five potentially novel biomarkers with high discriminating values between the two subtypes. Many of the biomarkers also exhibit significant prognostic potential, particularly in LUAD. Our study also unraveled distinct biological pathways between LUAD and LUSC.


# Introduction

Lung cancer is the most commonly diagnosed malignant tumor and is a leading cause of cancer-associated mortality. It is the second highest cause of new cancer cases in both genders in the United States and is the second leading cause of cancer deaths in females globally [1, 2]. The most common subtypes of lung cancers are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), classified together as non-small cell lung cancer (NSCLC) [3, 4]. However, recent studies have suggested that LUAD and LUSC should be classified and treated as different cancers [5].

Identifying the mechanisms underlying LUAD and LUSC is needed to develop useful biomarkers for better diagnosis, and to design therapeutic interventions. Multiple gene expression and immunohistochemistry studies have identified biological pathways and biomarkers that differentiate between LUAD and LUSC [6–8]. Other studies classified cancers using both novel and traditional machine learning or feature selection methods [9–12]. However, few have investigated cancers by applying multiple feature selection methods and selecting the overlapping features. This approach is promising because different feature selection methods select features using different criteria; nevertheless, each method has its strengths and weaknesses. Focusing on overlapping features, will optimize the strength of each method and minimize the probability of false positive features.

In this study, we downloaded LUAD and LUSC RNA-Seq datasets from The Cancer Genome Atlas (TCGA) [13] and analyzed them with five feature selection methods with ranking abilities: Differential Gene Expression Analysis (DGE), Principal Component Analysis (PCA), Least absolute shrinkage and selection operator (Lasso), minimal-Redundancy-Maximal Relevance (mRMR), and Extreme Gradient boosting (XGboost). We overlapped the outcome of the five methods and identified 131 genes which were used for LUAD and LUSC classification and pathway analysis. The classification results were validated with

random forest, a well-known and effective machine learning method. Among the 131 genes, 17 genes were proposed to be biomarkers and their diagnostic potential was assessed using Area Under Curve (AUC) value in Receiving Operating Characteristics (ROC) curve analysis. The prognostic values of the 17 genes were also assessed. The pathway analysis results were used to elucidate the potential clinical differences between LUAD and LUSC.

## Results

# Study Design

We obtained LUAD and LUSC RNA-Seq data from TCGA [13] and selected discriminatory genes by overlapping DGE, PCA, mRMR, XGboost, and lasso. The genes that were overlapped by two or more algorithms were validated and used for LUAD and LUSC classification as well as pathway analysis. The genes that were overlapped by three or more algorithms were selected as biomarker candidates, and their diagnostic values were assessed using ROC analysis and AUC value. The prognostic values of the biomarker candidates were also assessed using Kaplan Meier Plotter [14].

# Selection of genes

Top 500 genes from DGE (Supplementary Table S1) were selected as top features based on their lowest p-values. Similarly, top 500 genes from the first principal component in PCA and the top 500 genes from mRMR (Supplementary Table S1) were selected based on the ranking of the algorithm. Also, 148 genes in Xgboost (Supplementary Table S1) and 68 genes in lasso (Supplementary Table S1) with threshold of 0.5 were identified and selected. Since each of these methods has its own selection criteria, the overlapping genes must satisfy multiple selection criteria, making them significant candidate biomarkers that differentiate LUAD and LUSC. Therefore, the five independent sets of top genes were compared with a Venn diagram to identify the overlapping genes detected by multiple algorithms. Venn diagram (Fig. 1) comparison detected 131 genes (Supplementary Table S2) overlapped by two or more algorithms and 17 genes (Table 1) overlapped by three or more algorithms.

Table 1
17 Biomarker candidate genes that were selected by three or more algorithms

| Genes | Upregulated or Downregulated | Significantly expressed in LUSC or LUAD | Number of algorithms that selected the gene |
|---|---|---|---|
| KRT17 | Upregulated | LUSC | 4 |
| KRT14 (EBS4) | Upregulated | LUSC | 3 |
| KRT6A (K6C) | Upregulated | LUSC | 3 |
| KRT5 | Upregulated | LUSC | 3 |
| S100A2 | Upregulated | LUSC | 3 |
| TUBA1C | Upregulated | LUSC | 3 |
| CELSR2 | Upregulated | LUSC | 3 |
| TRIM29 | Upregulated | LUSC | 3 |
| REPS1 | Upregulated | LUSC | 3 |
| PERP | Upregulated | LUSC | 3 |
| NECTIN1 (PVRL1) | Upregulated | LUSC | 3 |
| GPC1 | Upregulated | LUSC | 3 |
| MUC1 | Downregulated | LUAD | 3 |
| ELFN2 | Downregulated | LUAD | 3 |
| ARHGEF38 (FLJ20184) | Downregulated | LUAD | 3 |
| ARHGAP12 | Downregulated | LUAD | 3 |
| QSOX1 | Downregulated | LUAD | 3 |

# Validation of Selected Genes

To evaluate how effective the selected genes are in classifying LUAD and LUSC, we used random forest to validate the top 500 genes selected from PCA, mRMR, and DGE, as well as the 148 genes from xgboost and 68 genes from lasso (Supplementary Table S1). All of the validation results for each feature selection method returned high classification accuracies of over 90% (Table 2). To compare to the previous feature selection methods, the overlapping 131 genes were validated the same way as the other algorithms. The binary classification statistics (Table 2) were calculated using LUAD as ‘positive’ and LUSC as ‘negative’. The overlapping 131 genes showed comparable, if not better, results to the other algorithms (Table 2). Heatmaps for the top 131 and the top 17 genes were also generated (Fig. 2a and 2b). Both heatmaps, in particular the heatmap with 17 genes, displayed clear borders separating LUAD from LUSC.

Table 2
LUAD and LUSC Classification Statistics

| Feature Selection Method | Accuracy | Specificity | Sensitivity | Precision | F-measure | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|---|---|
| DGE (Top 500) | 0.932476 | 0.901235 | 0.966443 | 0.9 | 0.932039 | (0.9035, 0.9614) |
| PCA (Top 500) | 0.942122 | 0.901235 | 0.986577 | 0.90184 | 0.942308 | (0.9132, 0.9678) |
| mRMR (Top 500) | 0.916399 | 0.888889 | 0.946309 | 0.886792 | 0.915584 | (0.8842, 0.9453) |
| Lasso (68 Genes) | 0.938907 | 0.907407 | 0.973154 | 0.90625 | 0.938511 | (0.9100, 0.9646) |
| Xgboost (148 Genes) | 0.935691 | 0.901235 | 0.973154 | 0.900621 | 0.935484 | (0.9068, 0.9614) |
| Overlapping 131 Genes | 0.938907 | 0.895062 | 0.986577 | 0.896341 | 0.939297 | (0.9100, 0.9646) |

# Identification of the 17 Potential Biomarkers and their ROC Analysis

Among the 17 potential biomarkers (Table 1) that were overlapped by three methods or more, one gene (Keratin 17, KRT17) was chosen by four methods (PCA, DGE, lasso, xgboost) and 16 genes (Quiescin Sulfhydryl Oxidase 1, QSOX1; Rho GTPase Activating Protein 12, ARHGAP12; Rho Guanine Nucleotide Exchange Factor 38, ARHGEF38; Extracellular Leucine Rich Repeat And Fibronectin Type III Domain Containing 2, ELFN2; Mucin 1, cell surface associated, MUC1;, Glypican 1 GPC1; Nectin Cell Adhesion Molecule 1, NECTIN1; P53 Apoptosis Effector Related To PMP22, PERP; RALBP1 Associated Eps Domain Containing 1, REPS1; Tripartite Motif Containing 29, TRIM29; Cadherin EGF LAG seven-pass G-type receptor 2, CELSR2; Tubulin Alpha 1c, TUBA1C; S100 Calcium Binding Protein A2, S100A2; Keratin 5 KRT5; Keratin 14, KRT14; and Keratin 6A, KRT6A) were chosen by three methods. In particular, QSOX1, ARHGAP12, ARHGEF38, ELFN2, and MUC1 were more significantly expressed by LUAD, and the rest were more significantly expressed by LUSC.

ROC curve analysis was applied to the 17 biomarker candidates. Figure 3a and 3b show the upregulated genes, and Fig. 3c shows the downregulated genes. Most of the genes had areas under the curve (AUC) of over 0.9, with NECTIN1 (0.9514), PERP (0.9529), KRT5 (0.9731), KRT6A (0.9532), and ARHGEF38 (0.9574) having AUC of over 0.95. Among the upregulated genes, KRT5 has the highest AUC of 0.9731, thereby displaying the most significant diagnostic potential in classifying LUAD and LUSC, consistent with the study reported by Jain Xiao et al in which KRT5 also had the highest diagnostic potential [6]. Among the downregulated genes, however, ARHGEF38 has the highest AUC of 0.9574, which is on par

with the upregulated genes and significantly higher than all the downregulated genes reported in Xiao's study, indicating that ARHGEF38 can potentially be a novel biomarker in determining the two cancers.

# Kaplan Meier Plotter Analysis of the 17 Potential Biomarkers

Of the 17 potential biomarkers (Table 1), only CELSR2 shows a significant prognostic value in LUSC, with its higher expression corresponding to a more favorable prognosis in LUSC (Table 3). In contrast, many genes show significant prognostic potential in LUAD. High expressions of KRT17, KRT6A, S100A2, TRIM29, REPS1, and GPC1 correspond to an unfavorable prognosis in LUAD, while high expressions of PERP, ELFN2, ARHGAP12, and QSOX1 correspond to a favorable prognosis in LUAD (Table 3).

Table 3
Kaplan Meier Prognostic Values of the 17 Biomarker

| | LUAD | | LUSC | |
|---|---|---|---|---|
| | HR (95% CIs) | P-value | HR (95% CIs) | P-value |
| KRT17 | **1.28 (1.01–1.61)** | **0.037** | 1.11 (0.88–1.4) | 0.39 |
| KRT14 (EBS4) | 1.19 (0.94–1.5) | 0.14 | 1.2 (0.95–1.52) | 0.13 |
| KRT6A (K6C) | **1.67 (1.32–2.12)** | **1.6e-05** | 0.99 (0.78–1.25) | 0.92 |
| KRT5 | 1.14 (0.9–1.43) | 0.28 | 1 (0.79–1.27) | 1 |
| S100A2 | **1.73 (1.36–2.19)** | **4.3e-06** | 1.07 (0.85–1.36) | 0.55 |
| TUBA1C | 1.1 (0.87–1.39) | 0.43 | 1.2 (0.94–1.52) | 0.14 |
| CELSR2 | 0.92 (0.73–1.16) | 0.47 | **0.79 (0.62–1)** | **0.049** |
| TRIM29 | **1.31 (1.04–1.66)** | **0.022** | 0.93 (0.74–1.18) | 0.57 |
| REPS1 | **1.38 (1.08–1.76)** | **0.0093** | 0.9 (0.66–1.23) | 0.51 |
| PERP | **0.67 (0.52–0.85)** | **0.0012** | 0.85 (0.62–1.16) | 0.3 |
| NECTIN1 (PVRL1) | 1.19 (0.94–1.5) | 0.14 | 0.94 (0.74–1.2) | 0.63 |
| GPC1 | **1.36 (1.08–1.72)** | **0.0091** | 0.98 (0.77–1.23) | 0.83 |
| MUC1 | 1.02 (0.81–1.29) | 0.84 | 1.02 (0.8–1.29) | 0.88 |
| ELFN2 | **0.72 (0.56–0.92)** | **0.0076** | 1.07 (0.78–1.47) | 0.67 |
| ARHGEF38 (FLJ20184) | 0.97 (0.77–1.23) | 0.83 | 1.16 (0.91–1.47) | 0.22 |
| ARHGAP12 | **0.61 (0.48–0.77)** | **2.3e-05** | 1.17 (0.93–1.49) | 0.18 |
| QSOX1 | **0.76 (0.6–0.96)** | **0.021** | 0.95 (0.75–1.2) | 0.66 |

# GO Term Enrichment Analysis

To further understand the biological differences between LUAD and LUSC, we performed pathway analysis by splitting the identified 131 genes into two groups: 57 genes downregulated and 74 upregulated in LUSC compared to LUAD. Functional pathway annotation of these two groups of genes was performed using The Database for Annotation, Visualization and Integrated Discovery (DAVID) [15] analysis tool with Gene Ontology (GO) biological pathway enrichments. GO terms with P-value < 0.05 were obtained (Supplementary Tables S3 and S4). The top 10 most significantly upregulated and downregulated GO terms ranked by p-value are shown in Table 4. In addition, DAVID has the functionality to group similar GO terms into clusters of the same biological pathway. To elucidate the potential biological differences between LUAD and LUSC, the top five most significantly upregulated and downregulated clusters ranked by enrichment scores were determined (Table 5, and Supplementary Tables S5 and S6).

Table 4
Top 10 Upregulated and Downregulated GO Biological Pathways.

| **Top 10 Upregulated Pathways** | | | **Top 10 Downregulated Pathways** | | |
|---|---|---|---|---|---|
| **GO Term** | **Pathway** | **P-value** | **GO Term** | **Pathway** | **P-value** |
| GO:0009888 | Tissue development | 4.45E-07 | GO:0002576 | Platelet degranulation | 2.86E-04 |
| GO:0045104 | Intermediate filament cytoskeleton organization | 8.82E-07 | GO:1901575 | Organic Substance Catabolic Process | 8.18E-03 |
| GO:0045103 | Intermediate filament-based process | 9.95E-07 | GO:0009057 | Macromolecule Catabolic Process | 8.29E-03 |
| GO:0007155 | Cell adhesion | 4.25E-06 | GO:0045055 | Regulated Exocytosis | 1.05E-02 |
| GO:0022610 | Biological adhesion | 4.49E-06 | GO:0009056 | Catabolic process | 1.32E-02 |
| GO:0008544 | Epidermis development | 4.64E-06 | GO:00034613 | Cellular Protein Localization | 1.80E-02 |
| GO:0098609 | Cell-cell adhesion | 5.07E-06 | GO:0070727 | Cellular macromolecule localization | 1.89E-02 |
| GO:0034330 | Cell junction organization | 9.93E-06 | GO:0043129 | Surfactant Homeostasis | 2.36E-02 |
| GO:2001233 | Regulation of apoptotic signaling pathway | 3.06E-05 | GO:0016553 | Base conversion or substitution editing | 2.65E-02 |
| GO:0061436 | Establishment of skin barrier | 5.65E-05 | GO:0048875 | Chemical homeostasis within a tissue | 2.94E-02 |

Table 5
Top 5 Clusters of Upregulated and Downregulated Biological pathways

| Top 5 Clusters of Upregulated Biological Pathways | | Top 5 Clusters of Downregulated Biological Pathways | |
|---|---|---|---|
| **Cluster** | **Enrichment Score** | **Cluster** | **Enrichment Score** |
| Cell Adhesion | 4.05 | Platelet degranulation and exocytosis | 1.34 |
| Intermediate filament organization | 3.87 | Tyrosine Kinase Pathways | 0.74 |
| Cell junction organization | 3.42 | Homeostasis | 0.69 |
| Cell component organization | 3.28 | Protein Translation and Localization | 0.68 |
| Hemidesmosome Assembly | 2.67 | Circulatory System Regulation | 0.63 |

In the upregulated group, most pathways are concentrated in cell adhesion, intermediate filament organization, and cell junction assembly. In the downregulated group, the most significant cluster is platelet degranulation and cell exocytosis, as well as other pathways such as tyrosine kinase signaling pathway, homeostasis, protein translation and circulatory system. These results suggest that LUSC tends to express more genes related to cell adhesion and cytoskeleton organization, and LUAD tends to express more genes involved in platelet degranulation and exocytosis, along with other signaling pathways.

# Reactome Pathway Analysis

Reactome pathways [16] were also generated for both the upregulated and the downregulated groups. The most significantly upregulated pathway is the cornification, or the keratinization pathway (Fig. 4, Supplementary Table S7), along with other similar pathways related to cell adhesion, which is consistent with GO term analysis. TP53 regulation pathway, which is often implicated in cancer, is among the top enriched pathways as well (Supplementary Table S7). For the downregulated group, the most significant pathway is peptide elongation synthesis (Fig. 5, Supplementary Table S8), which GO term analysis also reveals to be significant.

# KEGG Pathway Analysis

Only the p53 signaling pathway is significant in the upregulated group (Table 6) in Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] pathway analysis; this is consistent with Reactome analysis which ranks TP53 regulation as the second most upregulated pathway after keratinization and other cell junction related pathways. Only the lysosome and ribosome pathways are significant in the downregulated group (Table 6). The lysosomal pathway is coherent with platelet degranulation and exocytosis, as reported in GO term analysis. Even though the ribosomal pathway has a p-value slightly

greater than 0.05, it is most likely significant as it is also shown to be significantly enriched in both GO and Reactome term analyses (Supplementary Tables S3 and S8).

Table 6
KEGG Upregulated and Downregulated Pathways

| KEGG Upregulated Pathways | | | KEGG Downregulated Pathways | | |
|---|---|---|---|---|---|
| **KEGG Term** | **Pathway** | **P-value** | **KEGG Term** | **Pathway** | **P-value** |
| Hsa04115 | P53 signaling pathway | 0.0476 | Hsa04142 | Lysosome | 0.00727 |
| NA | NA | NA | Hsa03010 | Ribosome | 0.0749 |

# Discussion

Lung cancer remains the second most common cancer in the world [18]. Among the lung cancers, the two most common subtypes LUAD and LUSC are often categorized together as Non-small cell lung cancer. However, increasing evidence suggests that LUAD and LUSC should be considered as different diseases due to their vastly different biological and clinical signature [5]. Identifying biomarkers and unraveling the biological differences between the two can therefore provide a future direction in reaching better diagnosis and treatment for each condition.

Previous studies have utilized traditional feature selection and machine learning methods for cancer diagnosis, detection, and classification [10, 11, 19], but few have extended them to study potential biomarkers and biological pathways to discriminate between LUAD and LUSC. To improve cancer classification accuracy, novel machine learning, and feature selection methods have been developed [12, 20–22]. However, few studies have used overlapping features from different methods for classification, pathway analysis, and biomarker discovery, especially for LUAD and LUSC.

Here we took advantage of the ranking capabilities and the strengths of PCA, mRMR, XGboost, DGE, and Lasso to select 131 overlapping genes for classification and pathway analysis and identify 17 overlapping genes as potential biomarkers. Overall, the overlapping 131 genes showed several high-ranking metrics with lasso and PCA methods. Though the best method may vary depending on the metric, the classification result of using the overlapping 131 genes was by many metrics comparable if not better than the other methods that use more genes. The 131 overlapped genes achieved the highest sensitivity with PCA, the second highest accuracy with lasso, and the second highest F-measure overall, indicating that overlapping feature selection methods can be used to perform cancer classification.

Moreover, this method may prove to be valuable in biomarker discovery. In agreement with our result, previous studies have reported levels of several genes to be greatly elevated in LUSC compared to LUAD; these genes include KRT6 [6, 8, 23, 24], KRT5 [6, 8, 25], KRT14 [8, 23, 24], KRT17 [8, 23], PERP [8, 23], TRIM29 [8, 23], GPC1 [8], CELSR2 [8], S100A2 [8], and TUBA1C [26]. Also, consistent with our result, levels of QSOX1 [23] and MUC1 [8] were reported to be lower in LUSC than in LUAD. Many current biomarkers

such as Tumor Protein P63 (TP63), Napsin A Aspartic Peptidase (NAPSA), Melanophilin (MLPH), Desmocollin 3 (DSC3), and others are also part of the top 131 genes selected by our method [23, 27–30]. To our knowledge, ARHGAP12, ARHGEF38, ELFN2, NECTIN1, and REPS1 are among the top 17 genes in this study to be identified as biomarkers for the first times. Moreover, it is important to note that ARHGEF38 and NECTIN1 have two of the highest diagnostic values among selected genes based on ROC curve analysis, with ARHGEF38 having the highest AUC value among upregulated genes and the second highest AUC value overall. Furthermore, many of the 17 genes show significant prognostic importance, particularly in LUAD (Table 3).

NECTIN1 is a cell adhesion protein that plays a key role in herpes simplex virus type 1 (HSV-1) viral entry and has been shown to be sensitive to herpes oncolytic therapy in squamous cell carcinomas [31, 32]. ELFN2 (extracellular leucine-rich repeat and fibronectin type III domain-containing 2) is also known as protein phosphatase 1 regulatory subunit 29 and belongs to the leucine-rich repeat family. Studies show that ELN2 is prevalent in tumors of the brain and granular cells [33]. REPS1 is a gene that codes for RALBP1 associated Eps domain containing protein 1 and is associated with the endocytosis pathway [34]. ARHGEF38 and ARHGAP12 are both part of the Rho family GTPase regulators. Rho GTPases are essential to cell cytoskeletal structure, motility, and morphogenesis, and they have been implicated in many cancer metastases [35, 36]. ARHGEF38, in particular, has been associated with aggressive prostate cancer [37]. ARHGAP12, though not shown to exhibit invasion potential, has also been implicated in cell proliferation [38].The other upregulated genes ELFN2, QSOX1, and MUC1 have been shown to directly promote metastasis in various cancers [39–43], including lung cancer. Intriguingly, all 5 biomarker candidates (ARHGAP12, ARHGEF38, ELFN2, QSOX1, MUC1) that are upregulated in LUAD are involved in cancer proliferation and metastasis; the most enriched pathway in LUAD, which is platelet degranulation, is associated with metastasis as well [44]. Furthermore, the loss of certain genes upregulated in LUSC such as TRIM29 and KRT6A is associated with more cellular invasion [45, 46]. REPS1 and KRT6A genes which were upregulated in LUSC, were also shown to contribute to metastasis in cancer cells lines [47, 48]. Clinical differences between LUAD and LUSC are well known. In particular, LUAD has a higher metastatic rate than LUSC [49]. Studying these potential biomarkers may provide insight into tumor progression, metastatic, and therapeutic differences between LUAD and LUSC. Overall, the mechanisms by which many of these genes may regulate NSCLC development and metastasis remain unknown; therefore, studies to elucidate the exact mechanisms are warranted.

Different tumor subtypes arise from different types of cells located within each specific region, and consequently, the tumor transcriptome and morphology are thought to reflect this idea. In support of this view, our study, along with previous studies [6, 8, 24, 50], found that pathways specific to squamous cell tumors, including cell adhesion and keratinization, were associated with LUSC (Tables 4 and 5, Fig. 4), and pathways related exocytosis and surfactant homeostasis were associated with LUAD (Table 4).

Aside from cell adhesion or cytoskeleton organization, LUSC demonstrates higher regulation of p53 signaling in both KEGG and Reactome analyses. It is known that TP53 mutation is more common in LUSC than in LUAD [51–53], and that such mutation may predominantly be a non-truncated mutation in

LUSC leading to higher expression levels of genes involved in the p53 regulation pathway [54]. Moreover, P53 mutations often lose their tumor suppression function while gaining oncogenic abilities, leading to increased cell growth and proliferation compared to LUAD [55].

The most prominent pathway associated with LUAD, compared to LUSC, is platelet degranulation and exocytosis (Table 4, Table 5). Interestingly, lung cancer is the most common malignancy to coexist with venous thromboembolism, especially pulmonary embolism [56]. LUAD, in particular, has been shown to be an independent risk factor for pulmonary embolism even among lung cancers [57, 58]. One of the top downregulated clusters also show circulatory system regulation (Supplementary Table S6). Because platelet granulation directly causes thrombus formation, the differential enrichment of platelet granulation pathway can therefore help explain a more active and a more common hypercoagulation and thrombotic process in LUAD compared to LUSC [59]. In addition, platelets have been implicated in both the innate and the adaptive immune systems; platelet degranulation can modulate innate immunity via the release of cytokine, and platelet-leukocyte interactions can lead to leukocyte recruitment and activation in cancer [60]. In fact, CD63, one of the genes in the platelet degranulation pathway (Supplementary Tables S3 and S6), is directly involved in leukocyte recruitment through endothelial P-selectin [61]. LUSC has been associated with a relatively more suppressed immune response, implying a more active immune response in LUAD, which supports our result [55, 62]. The other top enriched LUAD pathways include tyrosine kinase signaling pathways and protein translation, which are known pathogenic pathways in cancers [63–66].

There are several limitations of this study. One of them is that this study does not take into account the RNA expression fold changes, which some groups have used to rank differentially expressed genes [67, 68]. Also, although this study aims to minimize the discovery of false positive biomarkers by overlapping different feature selection methods, the proposed biomarker candidates in this study still lack experimental verification. Nevertheless, these results may shed light into the biological differences between LUAD and LUSC, as well as aid the discovery of better diagnosis and treatment for each [63, 69].

In conclusion, we designed and implemented a workflow of overlapping five different feature selection methods to perform cancer classification, identify novel biomarkers, and study biological differences in NSCLC. We identified ARHGAP12, ARHGEF38, ELFN2, NECTIN1, and REPS1 as novel biomarkers, along with 12 other biomarker candidates. We also provided insight into potential explanations for different clinical findings and biological characteristics between LUSC and LUAD through pathway analysis. Further validation studies of these biomarkers and biological mechanisms are therefore warranted.

# Method

# RNA-Seq data processing

The LUAD and LUSC HTSeq read counts data were downloaded from TCGA[13] using TCGAbiolinks from R [70, 71]. As of June 2020, there were 533 LUAD and 502 LUSC samples. The samples were normalized

using TMM method and standardized using the CPM (read counts per million) function in R. Genes < 1 CPM in over 600 samples were considered noise and discarded to obtain 14,010 genes. The filtered genes were analyzed with different gene selection methods to further narrow down potential gene candidates for biomarkers and pathway analyses.

# Gene Selection and Cancer Classification

Gene selection analysis was performed using five different selection methods to generate five independent sets of top genes. The 5 independent sets were compared, and the resulting overlapped genes were used for cancer classification, biomarker identification, and pathway analysis. The selection methods used were DGE, PCA, xgboost, lasso, and mRMR. DGE between LUAD and LUSC was performed using the edgeR package [72]. Top 500 differentially expressed genes (Supplementary Table 1) were selected as top features based on their lowest p-values; validation of these genes was performed using random forest with the ranger package [73]. The top 500 genes from the first principle component in PCA and the top 500 genes from mRMR [74] were selected and validated the same way as the differentially expressed genes. Genes selected from Xgboost [75] and lasso [76] (Supplementary Table 1) were validated in a similar manner. For each validation, the data were randomly split into a training set and a testing set in a 7:3 ratio, where the training set was used to construct the model and the testing set was used to evaluate the model's performance. To compare each selection method more effectively, we split the training sets and testing sets the same way for all validations. We applied 5-fold cross validation to decide the optimal parameters for each training model and estimated its accuracy by applying the best determined parameters to the test set.

For classification and pathway analysis, we selected genes that were detected by at least two methods, and they were validated using ranger [73]. We also used bootstrapping [77] with 10,000 replicates to calculate the confidence interval for the accuracy of each method, including the proposed method of classification. The genes that were detected by at least 3 methods were considered candidate biomarkers. Their diagnostic potential was determined assessed using receiving operating characteristics (ROC) curve analysis.

# Prognostic Value Analysis Using Kaplan-Meier Plotter

Prognostic values of the identified biomarkers in LUAD and LUSC were evaluated using Kaplan-Meier Plotter. Kaplan-Meier Plotter is an online database that contains comprehensive clinical and microarray data for various cancers, including lung cancer [14].

# Pathway Analysis of Selected Genes

To further investigate and understand the biological difference between LUAD and LUSC, we performed pathway enrichment analysis using KEGG [17], Gene Ontology (GO), and Reactome [16]. Modified Fisher's exact tests were performed using DAVID v6.8 [15]. Pathways with false discovery rate (FDR) < 5% or p-value less than 0.05 were considered significant. These databases were all accessed in November 2020.

# Abbreviations

AUC: Area under Curve

DAVID: The Database for Annotation, Visualization, and Integrated Discovery

DGE: Differential Gene Expression

FPR: False Positive Rate

GO: Gene Ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

Lasso: Least absolute shrinkage and selection operator

LUAD: Lung Adenocarcinoma

LUSC: Lung Squamous Cell Carcinoma

mRMR: minimum Redundancy Maximum Relevance

NSCLC: Non-small cell lung cancer

PCA: Principal Component Analysis

ROC: Receiving Operating Characteristics

TCGA: The Cancer Genome Atlas

TPR: True Positive Rate

Xgboost: Extreme Gradient Boosting

QSOX1: Quiescin Sulfhydryl Oxidase 1

ARHGAP12: Rho GTPase Activating Protein 12

ARHGEF38: Rho Guanine Nucleotide Exchange Factor 38

ELFN2: Extracellular Leucine Rich Repeat And Fibronectin Type III Domain Containing 2

MUC1: Mucin 1, cell surface associated

GPC1: Glypican 1 GPC1

NECTIN1: Nectin Cell Adhesion Molecule 1

PERP: P53 Apoptosis Effector Related To PMP22

REPS1: RALBP1 Associated Eps Domain Containing 1

TRIM29: Tripartite Motif Containing 29

CELSR2: Cadherin EGF LAG seven-pass G-type receptor 2

TUBA1C: Tubulin Alpha 1c

S100A2: S100 Calcium Binding Protein A2

KRT5: Keratin 5

KRT14: Keratin 14

KRT6A: Keratin 6A

TP63: Tumor Protein P63

NAPSA: Napsin A Aspartic Peptidase

MLPH: Melanophilin

DSC3: Desmocollin 3

## Declarations

# Authors' contributions

JC proposed the method of overlapping feature selection methods to investigate LUAD and LUSC. JC obtained, analyzed, and interpreted the data. JC wrote the manuscript and generated the figures and tables. JD supervised the study and prepared the figures. JD made substantial suggestions and revisions of the manuscript. All authors read and approved the final manuscript.

# Competing interests

The authors declare no competing interests.

# Data availability

All data generated and/or analyzed during the current study are included in this published article (and its supplementary information files). The custom code used for data analysis can be accessed at https://github.com/chenjoe569/NSCLC-Research.

# References


1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA Cancer J Clin, 2020. **70**(1): p. 7-30.
2. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
3. Herbst, R.S., J.V. Heymach, and S.M. Lippman, *Lung cancer.* N Engl J Med, 2008. **359**(13): p. 1367-80.
4. Chen, Z., et al., *Non-small-cell lung cancers: a heterogeneous set of diseases.* Nat Rev Cancer, 2014. **14**(8): p. 535-46.
5. Relli, V., et al., *Abandoning the Notion of Non-Small Cell Lung Cancer.* Trends Mol Med, 2019. **25**(7): p. 585-594.
6. Xiao, J., et al., *Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma.* Oncotarget, 2017. **8**(42): p. 71759-71771.
7. Lu, C., et al., *Identification of differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by gene expression profiling.* Mol Med Rep, 2016. **14**(2): p. 1483-90.
8. Zhan, C., et al., *Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma.* J Thorac Dis, 2015. **7**(8): p. 1398-405.
9. Tian, S., *Identification of Subtype-Specific Prognostic Genes for Early-Stage Lung Adenocarcinoma and Squamous Cell Carcinoma Patients Using an Embedded Feature Selection Algorithm.* PLoS One, 2015. **10**(7): p. e0134630.
10. Zhengyan Huang, L.C., Chi Wang, *Classifying Lung Adenocarcinoma and Squamous Cell Carcinoma using RNA-Seq Data.* Cancer Studies And Molecular Medicine Open Journal, 2017. **3**(2).
11. Cai, Z., et al., *Classification of lung cancer using ensemble-based feature selection and machine learning methods.* Mol Biosyst, 2015. **11**(3): p. 791-800.
12. Liu, X.Y., et al., *Novel Regularization Method for Biomarker Selection and Cancer Classification.* IEEE/ACM Trans Comput Biol Bioinform, 2020. **17**(4): p. 1329-1340.
13. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
14. Gyorffy, B., et al., *Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer.* PLoS One, 2013. **8**(12): p. e82241. http://kmplot.com/analysis/. Accessed 15 November 2020.
15. Huang, D.W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biol, 2007. **8**(9): p. R183. https://david.ncifcrf.gov/tools.jsp. Accessed 15 November 2020.
16. Croft, D., et al., *The Reactome pathway knowledgebase.* Nucleic Acids Res, 2014. **42**(Database issue): p. D472-7. http://www.reactome.org. Accessed 15 November 2020.
17. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

18. de Groot, P.M., et al., *The epidemiology of lung cancer.* Transl Lung Cancer Res, 2018. **7**(3): p. 220-233.
19. Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* J Bioinform Comput Biol, 2005. **3**(2): p. 185-205.
20. Danaee, P., R. Ghaeini, and D.A. Hendrix, *A Deep Learning Approach for Cancer Detection and Relevant Gene Identification.* Pac Symp Biocomput, 2017. **22**: p. 219-229.
21. Jiang, L., et al., *Bayesian Hyper-LASSO Classification for Feature Selection with Application to Endometrial Cancer RNA-seq Data.* Sci Rep, 2020. **10**(1): p. 9747.
22. Huang, H.H., X.Y. Liu, and Y. Liang, *Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2 +2 Regularization.* PLoS One, 2016. **11**(5): p. e0149675.
23. Relli, V., et al., *Distinct lung cancer subtypes associate to distinct drivers of tumor progression.* Oncotarget, 2018. **9**(85): p. 35528-35540.
24. Chang, H.H., J.M. Dreyfuss, and M.F. Ramoni, *A transcriptional network signature characterizes lung cancer subtypes.* Cancer, 2011. **117**(2): p. 353-60.
25. Miettinen, M. and M. Sarlomo-Rikala, *Expression of calretinin, thrombomodulin, keratin 5, and mesothelin in lung carcinomas of different types: an immunohistochemical analysis of 596 tumors in comparison with epithelioid mesotheliomas of the pleura.* Am J Surg Pathol, 2003. **27**(2): p. 150-8.
26. Liu, S., et al., *Transcription Factors Contribute to Differential Expression in Cellular Pathways in Lung Adenocarcinoma and Lung Squamous Cell Carcinoma.* Interdiscip Sci, 2018. **10**(4): p. 836-847.
27. Travis, W.D., et al., *Pathologic diagnosis of advanced lung cancer based on small biopsies and cytology: a paradigm shift.* J Thorac Oncol, 2010. **5**(4): p. 411-4.
28. Khayyata, S., et al., *Value of P63 and CK5/6 in distinguishing squamous cell carcinoma from adenocarcinoma in lung fine-needle aspiration specimens.* Diagn Cytopathol, 2009. **37**(3): p. 178-83.
29. Ao, M.H., et al., *The utility of a novel triple marker (combination of TTF1, napsin A, and p40) in the subclassification of non-small cell lung cancer.* Hum Pathol, 2014. **45**(5): p. 926-34.
30. Travis, W.D., et al., *International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma.* J Thorac Oncol, 2011. **6**(2): p. 244-85.
31. Yu, Z., et al., *Nectin-1 expression by squamous cell carcinoma is a predictor of herpes oncolytic sensitivity.* Mol Ther, 2007. **15**(1): p. 103-13.
32. Rikitake, Y., K. Mandai, and Y. Takai, *The role of nectins in different types of cell-cell adhesion.* J Cell Sci, 2012. **125**(Pt 16): p. 3713-22.
33. Liu, C., et al., *LINC00470 Coordinates the Epigenetic Regulation of ELFN2 to Distract GBM Cell Autophagy.* Mol Ther, 2018. **26**(9): p. 2267-2281.
34. Xu, J., et al., *Cloning, expression and characterization of a novel human REPS1 gene.* Biochim Biophys Acta, 2001. **1522**(2): p. 118-21.

35. Cook, D.R., K.L. Rossman, and C.J. Der, *Rho guanine nucleotide exchange factors: regulators of Rho GTPase activity in development and disease.* Oncogene, 2014. **33**(31): p. 4021-35.
36. Porter, A.P., A. Papaioannou, and A. Malliri, *Deregulation of Rho GTPases in cancer.* Small GTPases, 2016. **7**(3): p. 123-38.
37. Liu, K., et al., *ARHGEF38 as a novel biomarker to predict aggressive prostate cancer.* Genes Dis, 2020. **7**(2): p. 217-224.
38. Gentile, A., et al., *Met-driven invasive growth involves transcriptional regulation of Arhgap12.* Oncogene, 2008. **27**(42): p. 5590-8.
39. Zhang, Y.Q., et al., *Overexpression of CST4 promotes gastric cancer aggressiveness by activating the ELFN2 signaling pathway.* Am J Cancer Res, 2017. **7**(11): p. 2290-2304.
40. Knutsvik, G., et al., *QSOX1 expression is associated with aggressive tumor features and reduced survival in breast carcinomas.* Mod Pathol, 2016. **29**(12): p. 1485-1491.
41. Xu, T., et al., *MUC1 downregulation inhibits non-small cell lung cancer progression in human cell lines.* Exp Ther Med, 2017. **14**(5): p. 4443-4447.
42. Kohlgraf, K.G., et al., *Contribution of the MUC1 tandem repeat and cytoplasmic tail to invasive and metastatic properties of a pancreatic cancer cell line.* Cancer Res, 2003. **63**(16): p. 5011-20.
43. Hollingsworth, M.A. and B.J. Swanson, *Mucins in cancer: protection and control of the cell surface.* Nat Rev Cancer, 2004. **4**(1): p. 45-60.
44. Schlesinger, M., *Role of platelets and platelet receptors in cancer metastasis.* J Hematol Oncol, 2018. **11**(1): p. 125.
45. Yanagi, T., et al., *Loss of TRIM29 Alters Keratin Distribution to Promote Cell Invasion in Squamous Cell Carcinoma.* Cancer Res, 2018. **78**(24): p. 6795-6806.
46. Chen, C. and H. Shan, *Keratin 6A gene silencing suppresses cell invasion and metastasis of nasopharyngeal carcinoma via the betacatenin cascade.* Mol Med Rep, 2019. **19**(5): p. 3477-3484.
47. Wu, Z., et al., *RalBP1 is necessary for metastasis of human cancer cell lines.* Neoplasia, 2010. **12**(12): p. 1003-12.
48. Yang, B., et al., *KRT6A Promotes EMT and Cancer Stem Cell Transformation in Lung Adenocarcinoma.* Technol Cancer Res Treat, 2020. **19**: p. 1533033820921248.
49. Milovanovic, I.S., M. Stjepanovic, and D. Mitrovic, *Distribution patterns of the metastases of the lung carcinoma in relation to histological type of the primary tumor: An autopsy study.* Ann Thorac Med, 2017. **12**(3): p. 191-198.
50. Su, R., et al., *Identification of expression signatures for non-small-cell lung carcinoma subtype classification.* Bioinformatics, 2020. **36**(2): p. 339-346.
51. Herbst, R.S., D. Morgensztern, and C. Boshoff, *The biology and management of non-small cell lung cancer.* Nature, 2018. **553**(7689): p. 446-454.
52. Petitjean, A., et al., *TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes.* Oncogene, 2007. **26**(15): p. 2157-65.

53. Labbe, C., et al., *Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC).* Lung Cancer, 2017. **111**: p. 23-29.
54. Wang, X. and Q. Sun, *TP53 mutations, expression and interaction networks in human cancers.* Oncotarget, 2017. **8**(1): p. 624-643.
55. Chen, M., et al., *Differentiated regulation of immune-response related genes between LUAD and LUSC subtypes of lung cancers.* Oncotarget, 2017. **8**(1): p. 133-144.
56. Lee, J.E., et al., *Clinical characteristics of pulmonary embolism with underlying malignancy.* Korean J Intern Med, 2010. **25**(1): p. 66-70.
57. Chew, H.K., et al., *The incidence of venous thromboembolism among patients with primary lung cancer.* J Thromb Haemost, 2008. **6**(4): p. 601-8.
58. Zhang, Y., et al., *Prevalence and associations of VTE in patients with newly diagnosed lung cancer.* Chest, 2014. **146**(3): p. 650-658.
59. Papageorgiou, C., et al., *Lobectomy and postoperative thromboprophylaxis with enoxaparin improve blood hypercoagulability in patients with localized primary lung adenocarcinoma.* Thromb Res, 2013. **132**(5): p. 584-91.
60. Stoiber, D. and A. Assinger, *Platelet-Leukocyte Interplay in Cancer Development and Progression.* Cells, 2020. **9**(4).
61. Doyle, E.L., et al., *CD63 is an essential cofactor to leukocyte recruitment by endothelial P-selectin.* Blood, 2011. **118**(15): p. 4265-73.
62. Lucchetta, M., et al., *Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response*. BMC Cancer, 2019. **19**(1): p. 824.
63. Xie, Z. and D. Liu, *[Pathogenesis of Molecular Signaling Pathways Changes in Smoking-induced Lung Cancer.].* Zhongguo Fei Ai Za Zhi, 2009. **12**(11): p. 1202-5.
64. Bos, J.L., *ras oncogenes in human cancer: a review.* Cancer Res, 1989. **49**(17): p. 4682-9.
65. Fernandez-Medarde, A. and E. Santos, *Ras in cancer and developmental diseases.* Genes Cancer, 2011. **2**(3): p. 344-58.
66. Shenoy, N., et al., *Alterations in the ribosomal machinery in cancer and hematologic disorders.* J Hematol Oncol, 2012. **5**: p. 32.
67. Farztdinov, V. and F. McDyer, *Distributional fold change test - a statistical approach for detecting differential expression in microarray experiments.* Algorithms Mol Biol, 2012. **7**(1): p. 29.
68. Dembele, D. and P. Kastner, *Fold change rank ordering statistics: a new method for detecting differentially expressed genes.* BMC Bioinformatics, 2014. **15**: p. 14.
69. Li, Y., et al., *Lung Cancer and Pulmonary Embolism: What Is the Relationship? A Review.* J Cancer, 2018. **9**(17): p. 3046-3057.
70. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.* Nucleic Acids Res, 2016. **44**(8): p. e71.

71. Silva, T.C., et al., *TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages.* F1000Res, 2016. **5**: p. 1542.
72. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
73. Wright MN, Z.A., *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R* Journal of Statistical Software, 2017. **77**(1): p. 1-17.
74. De Jay, N., et al., *mRMRe: an R package for parallelized mRMR ensemble feature selection.* Bioinformatics, 2013. **29**(18): p. 2365-8.
75. Tianqi Chen, T.H., Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyan Xie, Min Lin, Yifeng Geng, and Yutian Li, *xgboost: Extreme Gradient Boosting*. 2020.
76. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.
77. Canty A, R.B., *boot: Bootstrap R (S-plus) Functions*. 2020.

# Figures

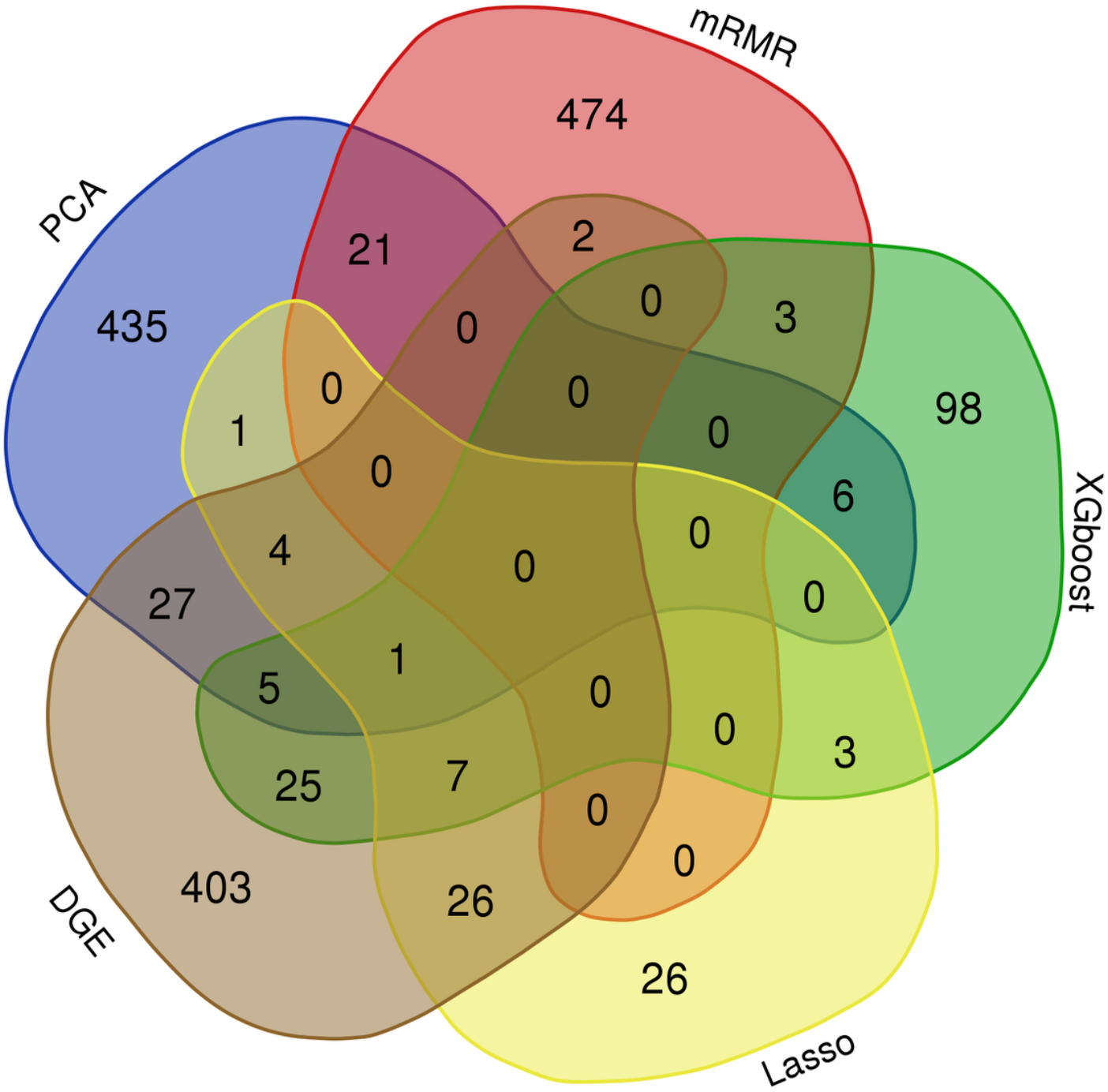
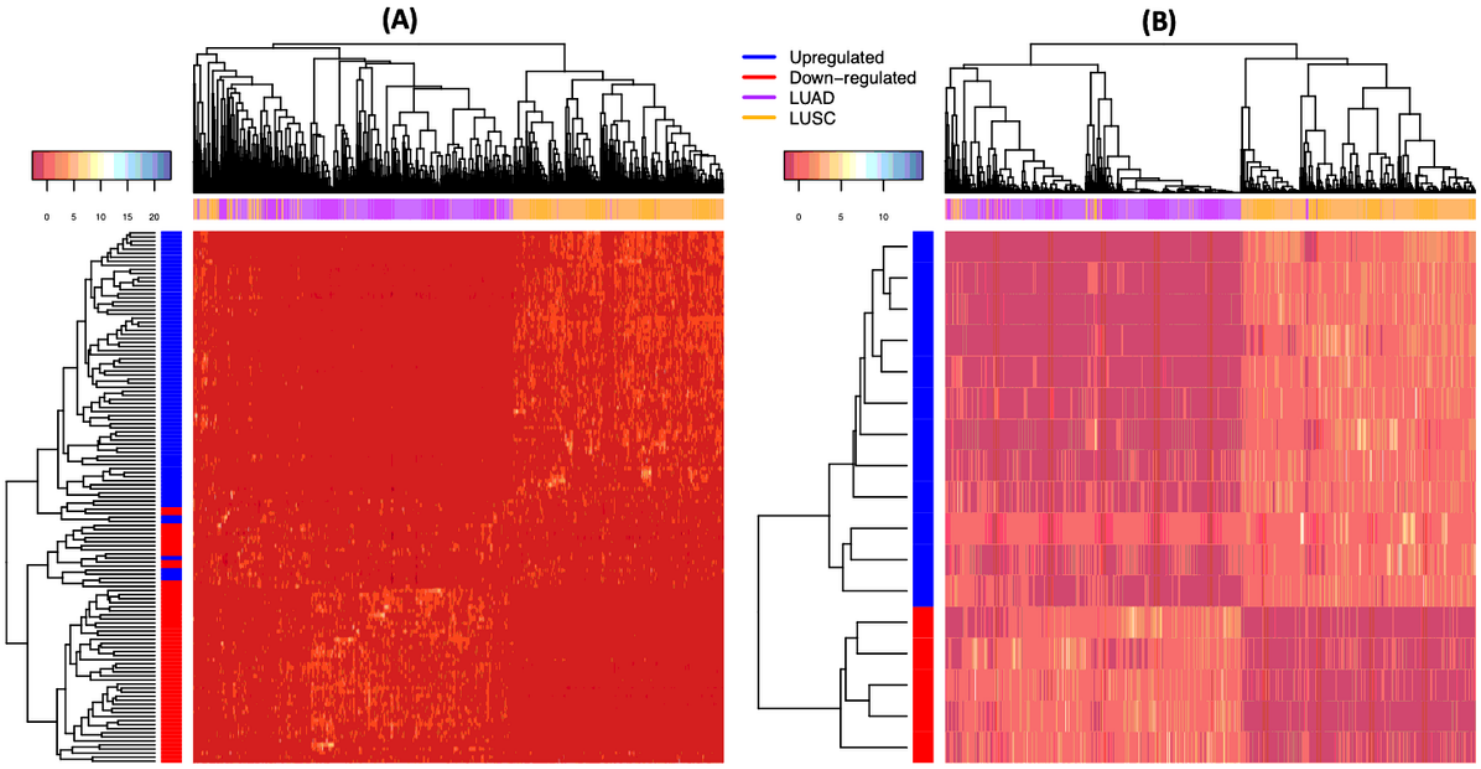


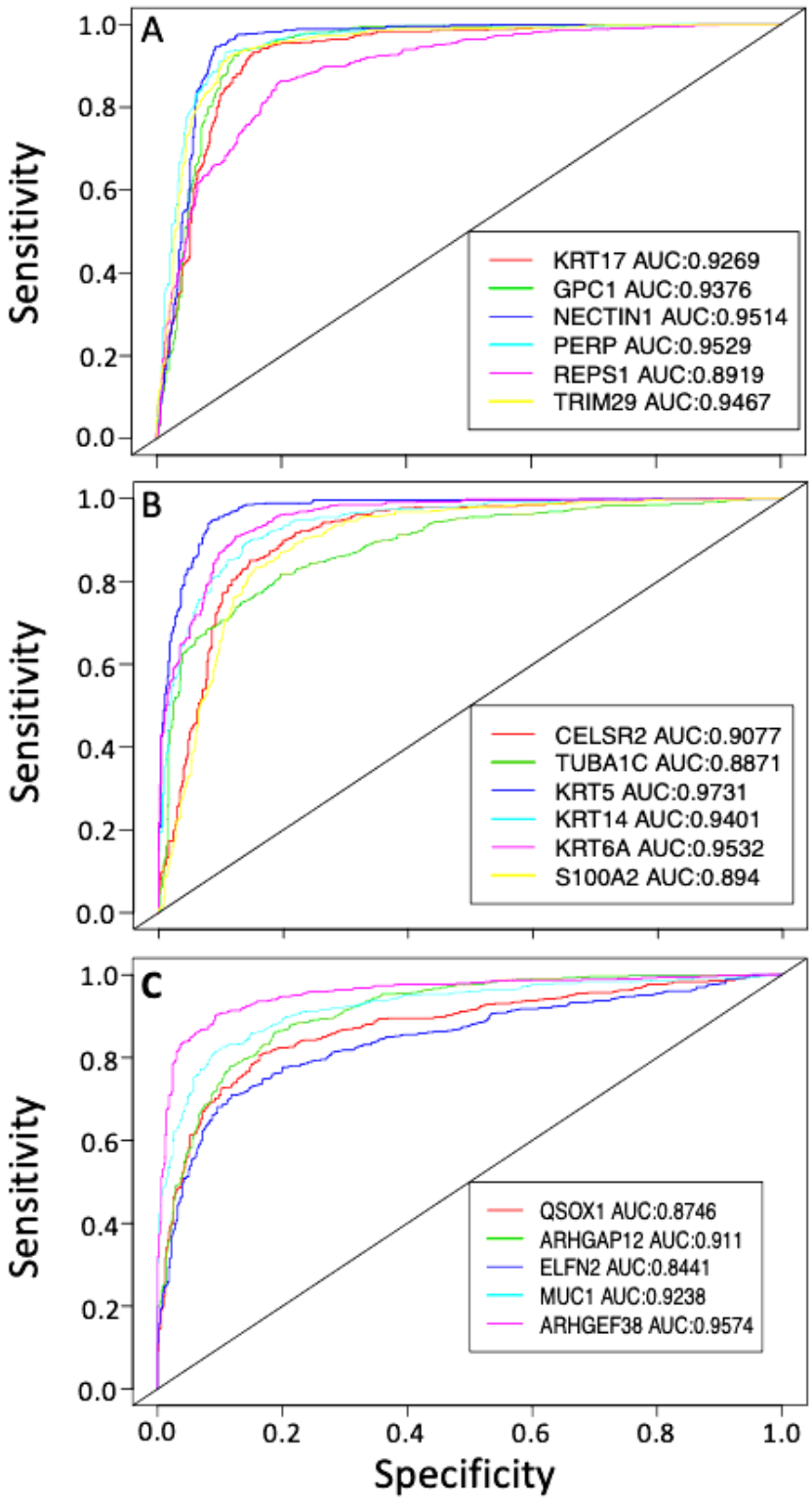

**Figure 1**

Venn diagram shows overlapping genes selected by each algorithm. Venn diagram of selected genes from PCA, mRMR, DGE, Lasso, and XGboost



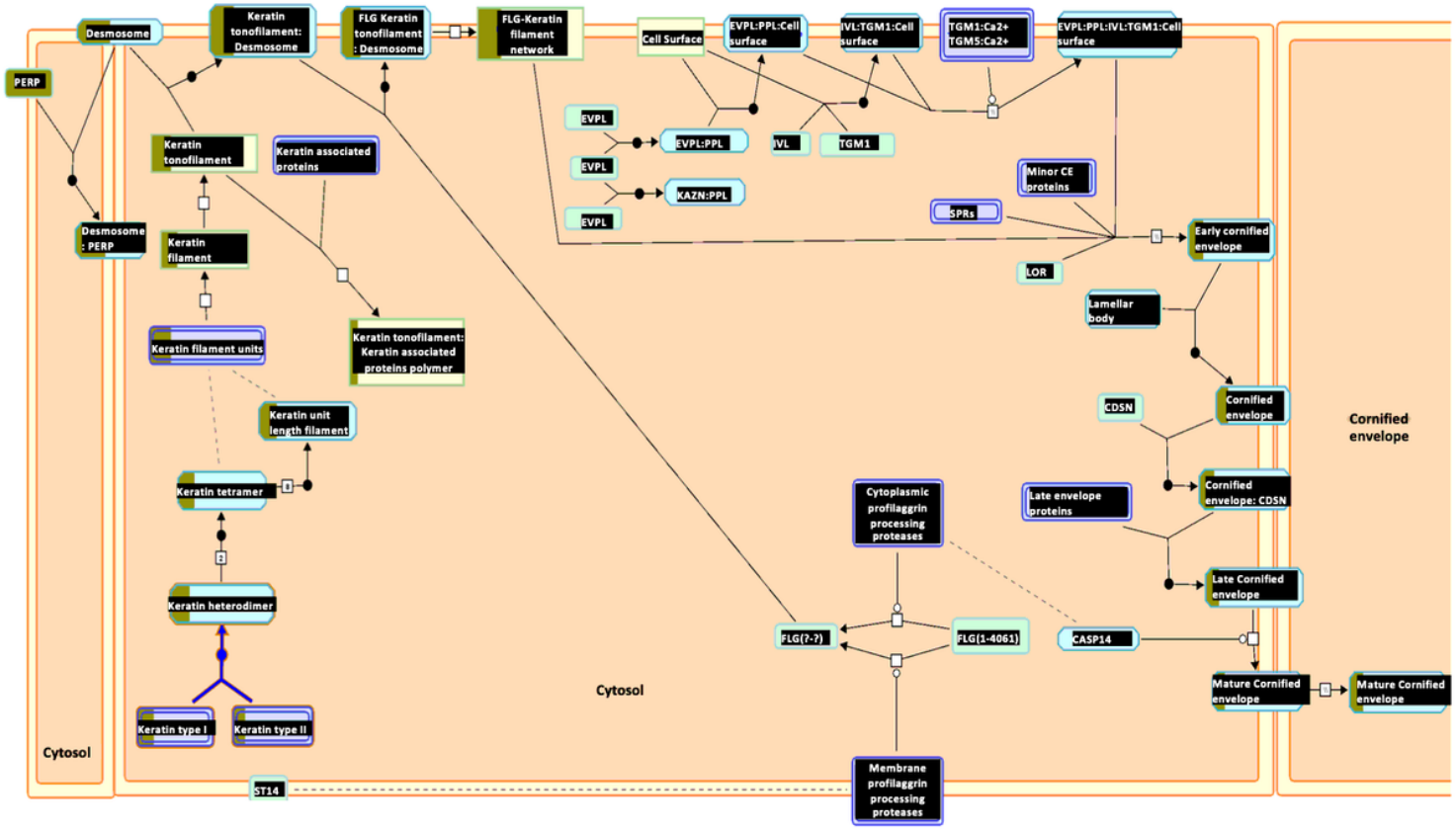

## Figure 2

Heatmap shows the 131 selected genes (Panel a) for pathway analysis and the 17 selected genes (Panel b) as biomarker candidates. The x-axis represents the genes, and the y-axis represents the samples




## Figure 3

ROC and AUC analysis demonstrate discriminating potential for Upregulated (Panel a and b) and Downregulated (Panel c) Genes. X-axis is sensitivity, or true positive rate (TPR). The Y-axis is 1-Specificity, or false positive rate (FPR). Higher AUC indicates higher discriminating potential for the gene.




## Figure 4

Keratinization Pathway is upregulated in LUSC. The Keratinization pathway is the most upregulated pathway according to Reactome analysis with p-value 3.33E -15 and FDR 1.95E-12. The boxes partially highlighted in brown indicate the number of genes identified in the analysis that are associated with each box.

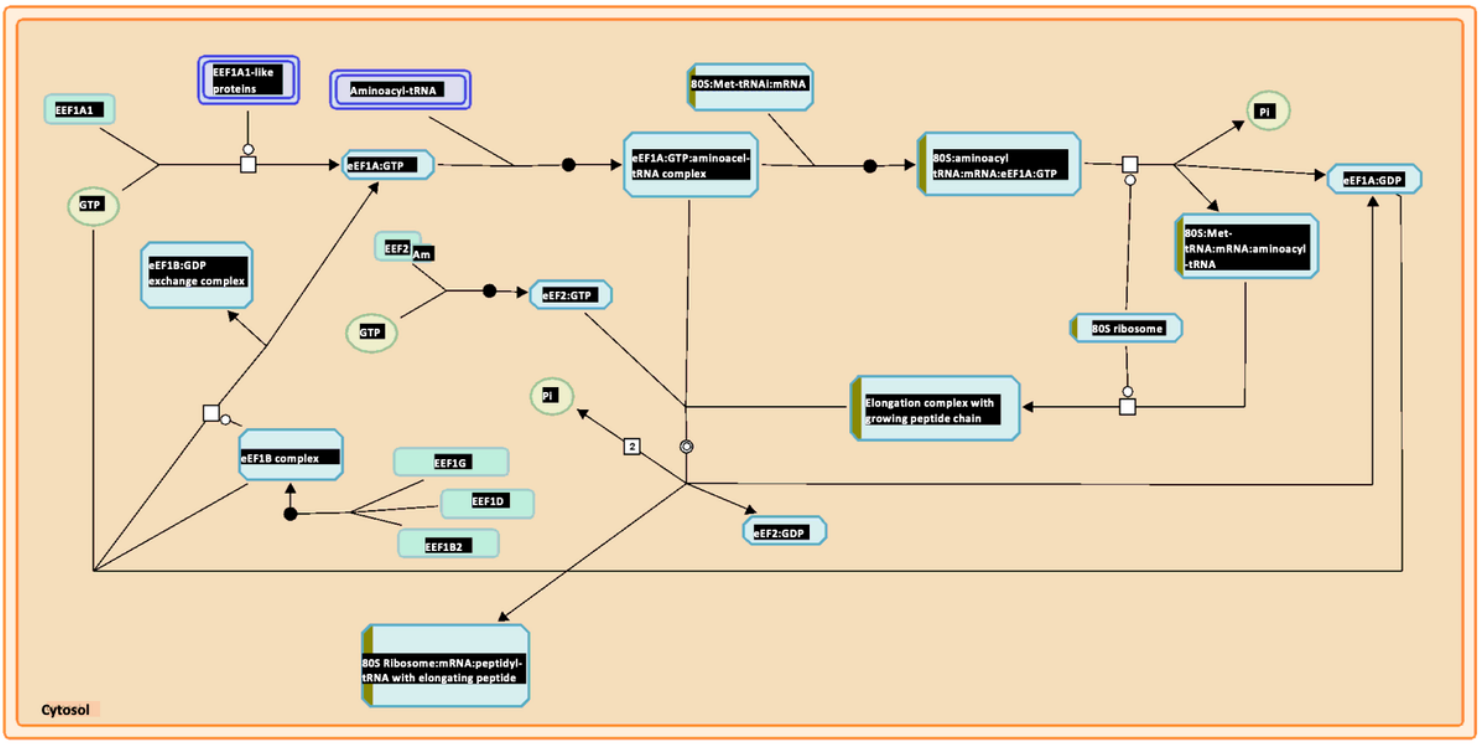


**Figure 5**

Peptide Elongation Pathway is downregulated in LUSC when compared to LUAD. The peptide elongation pathway is the most down-regulated pathway according to Reactome analysis with p-value 9.72E -6 and FDR 0.00157. The boxes partially highlighted in brown indicate the number of genes identified in the analysis that are associated with each box.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx