Heliyon 8 (2022) e11966

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Identification and validation of the mitochondrial function related hub genes by unsupervised machine learning and multi-omics analyses in lung adenocarcinoma

Xing Jin [a,1], Huan Zhang [a,1], Qihai Sui [a,1], Ming Li [a], Jiaqi Liang [a], Zhengyang Hu [a], Ye Cheng [b], Yuansheng Zheng [a], Zhencong Chen [a], Miao Lin [a,**], Hao Wang [a,***], Cheng Zhan [a,*]

[a] Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, No. 180, Fenglin Road, Shanghai, 200032, China
[b] Institutes of Biomedical Sciences and Children's Hospital, Fudan University, Shanghai 201102, China

ARTICLE INFO

Keywords:
LUAD
Mitochondrial function
Clinical outcomes
DARS2
COX5B

ABSTRACT

Background: The mitochondrion and its associated genes were heavily implicated in developing and therapy tumors as the primary cellular organelle in charge of metabolic reprogramming and ferroptosis. Our work focuses on discovering new potential targets while analyzing the multi-omics data of mitochondria-related genes in lung adenocarcinoma (LUAD).

Methods: The Cancer Genome Atlas (TCGA) database provided multi-omics data for LUAD patients. Based on the expression profile of the genes associated with mitochondria, the patients were grouped by the unsupervised clustering method. R was used to explore the differential expressed protein-code gene, miRNA, and lncRNA, as well as their enriched functions and ceRNA networks. Additionally, the discrepancy between immune infiltration and genetic variation was comprehensively characterized. Our clinical samples and in vitro experiments investigated the hub gene determined by LASSO and batch analysis.

Results: Two clusters are distinguished using unsupervised consensus clustering based on mitochondrial heterogeneity. The integrated analysis emphasized that patients in cluster B had a worse prognosis, higher mutation frequencies, and less immune cell infiltration. The hub genes DARS2 and COX5B are identified by further analysis using LASSO penalization. In vitro experiments indicated that DARS2 and COX5B knockdown inhibited tumor cell proliferation. The specimen of our hospital cohort conducted the immunohistochemistry analysis and validated that DARS2 and COX5B's expression was significantly higher in the tumor than in adjacent normal tissue and correlated to LUAD patients' prognosis.

Conclusion: Our observations implied that LUAD patients' tumors had distinct mitochondrial function heterogeneity with different clinical and molecular characteristics. DARS2 and COX5B might be critical genes involved in mitochondrial alterations and potential therapeutic targets.

1. Introduction

Lung cancer accounts for the highest rate of tumor-correlated mortality worldwide [1, 2], especially lung adenocarcinoma (LUAD), the most common subtype of lung cancer. Despite the advance in oncology, the five-year survival rate of patients with LUAD has not improved significantly over the past decades [2, 3]. Hence, continued efforts to identify novel potential targets and therapeutical strategies in LUAD are urgently needed.

Since Warburg described the metabolic feature of tumor cells as "aerobic glycolysis" in 1956 [4], Hanahan and Weinberg extended the concept of "aerobic glycolysis" and proposed a crucial cancer hallmark, namely deregulating cellular metabolism [5, 6]. As cellular ATP is mainly produced by mitochondria metabolism, The cellular metabolic

[*] Corresponding author.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: lin.miao@zs-hospital.sh.cn (M. Lin), wang.hao@zs-hospital.sh.cn (H. Wang), czhan10@fudan.edu.cn (C. Zhan).

[1] These authors contributed equally: Xing Jin, Huan Zhang, and Qihai Sui.

https://doi.org/10.1016/j.heliyon.2022.e11966
Received 15 September 2022; Received in revised form 7 November 2022; Accepted 22 November 2022

reprogramming procedure of tumor cells requires mitochondria biology [7]. Meanwhile, mitochondria are cellular suicidal weapons storage rooms linked to another hallmark of cancer: evasion of cell death. Both apoptosis and ferroptosis have been widely reported to be mediated by the mitochondria through the diverse biological process like the production of ROS, lipid peroxidation, and intrinsic mitochondrial apoptotic signaling pathways [8, 9]. Mitochondria, the pivotal organelle in cellular metabolism and cell death, was a promising target for developing anticancer therapy [10, 11].

In recent years, machine learning has been widely used in biomedicine [12, 13, 14]. Though the application in the clinical setting remains limited, the unsupervised algorithm in multi-omics data for identifying novel molecular targets holds excellent promise. In previous research, Wang demonstrated that clustering analysis helped to dissect cancer heterogeneity [15]. Woolley used the unsupervised cluster analysis to identify the subgroups of patients with heart failure [16]. By bioinformatics analysis, Shi et al. identified hypoxia-derived gene signatures to predict clinical outcomes in Stage one LUAD. However, Unsupervised

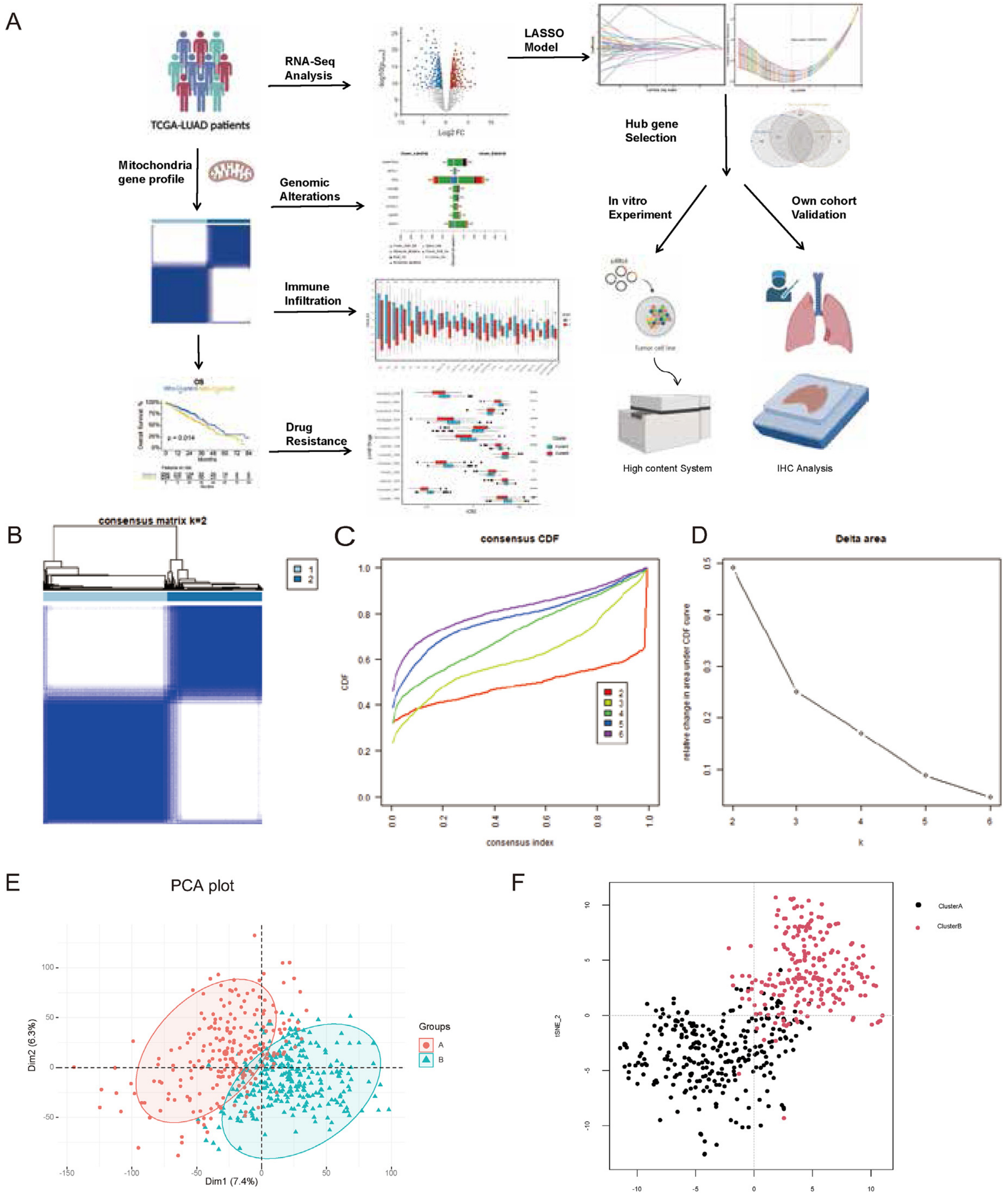

Figure 1. The flow chart of our research and the clustering procedure. (A) The flowchart shows the main flow of the fundamental research. (B) The consensus matrix is represented as heat maps for k = 2. (C) The Cumulative distribution function (CDF) curve is used to select the optimal Cluster. (D) The delta curve diagram indicates the relative change in area under the CDF curve. (E) The principal component analysis (PCA) and T-distributed stochastic neighbor embedding (t-SNE) (F) plot of patients in clusters A and B by transcriptomes.

clustering has not been investigated for mining mitochondria-associated genes in LUAD to discover novel targets and biomarkers.

In this work, we first used the unsupervised machine learning clustering approach to cluster genes related to mitochondrial function and then investigated the multi-omics discrepancies across the clusters to dissect the mitochondrial function molecular pattern in LUAD. Our findings should contribute to a better understanding of the molecular mechanisms behind mitochondrial changes and suggest potential mitochondrial target genes.

2. Methods and materials

2.1. Multi-omics data collection and mitochondrial function related genes query

The RNA sequencing (RNA-seq) expression, copy number variation (CNV), and somatic mutation data of LUAD samples, and the correspondent clinical features of The Cancer Genome Atlas (TCGA) were collected from the UCSC Xena (https://gdc.xenahubs.net). The CNV data was integrated into the MAF file using the R package maftools [17]. Lastly, all the multi-omics analyses were based on the TCGA patients' data to acquire accurate and consistent results unless expressly stated. 1136 Mitochondrial function-related genes are obtained from the Human MitoCarta 3.0 database [18] and 1626 genes from the Integrated Mitochondrial Protein Index database, containing 1184 known and 442 predicted from experimental data using the SVM algorithm [19]. The detailed information and classification of Mitochondrial function-related genes are listed in Supplementary Table S1.

2.2. The unsupervised consensus clustering methods

The Unsupervised consensus clustering was conducted using ConsensusClusterPlus R package [20]. Using the k-means method, setting the euclidean distance as the measurement of distance, and the maximum number of evaluated clusters was k = 6, we performed the machine learning cluster. Empirical cumulative distribution function (CDF) plots show consensus distributions for each k. Moreover, the delta area plot showed a relative increase in cluster stability. The optimal K is obtained based on the minimum PAC algorithm.

2.3. Differential analysis, enrichment function, ceRNA construction, and genomics analysis

The differential analysis between the two Mito-function clusters was conducted using the R package limma [21]. Moreover, the DE (Differentially expressed) mRNAs, DEmiRNAs, and DElncRNAs were identified by the threshold (FDR<0.05 and Log (Foldchange) > 1). The ClusterProfiler package was used for GO (gene ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), and GSEA (Gene Set Enrichment Analysis). The protein-protein interaction network was established using STRING, and the network's visualization, shrinkage, and hub gene selection were made in Cytoscape. The ceRNA interactions were identified in the miRWalk [22], Targetscan [23], and lncSEA [24] under the threshold $p < 0.05$. The maftools R package17 analyzed the distribution of mutated genes and the summary of frequently mutated genes. Tumor mutational burden (TMB) was calculated as the number of somatic base substitutions or indels per megabase (Mb) of the coding region target territory of the test (currently, 1.11 Mb).

2.4. Immune infiltration estimation and drug resistance prediction

Patients' immune infiltration and cell composition estimation was based on multiple approaches. The R package 'GSVA' [25] was performed based on the bulk RNA-seq data and the 24 gene sets on Bindea's publications [26]. Cibersort, Xcell, and MCPCounter deconvolution methods for cell fraction estimate were also conducted to improve the accuracy of tumor microenvironment estimate. The Genomics of Drug Sensitivity in Cancer (GDSC) database was used to predict the half-maximum inhibitory concentration (IC50) in 198 kinds of drug response. The R package" pRRophetic" was used to predict the drug response [27, 28].

Table 1. The baseline demographic data of cluster A and B.

Characteristics	Mito-Cluster A	Mito-Cluster B	*P* value
	N = 279	N = 218	
Age	67.0 [59.8; 74.0]	65.0 [58.0; 71.0]	**0.002**
Age_group:			**0.015**
<60	69 (26.4%)	67 (33.3%)	
60–70	87 (33.3%)	79 (39.3%)	
>70	105 (40.2%)	55 (27.4%)	
Sex:			**0.003**
Female	168 (60.2%)	101 (46.3%)	
Male	111 (39.8%)	117 (53.7%)	
Race:			**0.079**
White	226 (81.0%)	158 (72.5%)	
Black	24 (8.60%)	27 (12.4%)	
Other	29 (10.4%)	33 (15.1%)	
Smoke:			**<0.001**
No	141 (51.8%)	58 (27.5%)	
Yes	131 (48.2%)	153 (72.5%)	
Tstage:			**0.008**
T1	111 (40.1%)	55 (25.3%)	
T2	136 (49.1%)	131 (60.4%)	
T3	21 (7.58%)	22 (10.1%)	
T4	9 (3.25%)	9 (4.15%)	
Nstage:			0.062
N0	191 (70.7%)	130 (60.2%)	
N1	47 (17.4%)	47 (21.8%)	
N2	31 (11.5%)	38 (17.6%)	
N3	1 (0.37%)	1 (0.46%)	
Mstage:			**0.007**
M0	176 (63.8%)	155 (71.4%)	
M1	9 (3.26%)	15 (6.91%)	
MX	91 (33.0%)	47 (21.7%)	
Stage_group:			**0.006**
Early stage	229 (83.3%)	156 (72.6%)	
Later stage	46 (16.7%)	59 (27.4%)	
Radiotherapy:			**0.036**
No	217 (89.3%)	142 (81.6%)	
Yes	26 (10.7%)	32 (18.4%)	
Tumor_site:			0.255
Lower lobe	94 (33.7%)	74 (33.9%)	
Middle lobe	10 (3.58%)	11 (5.05%)	
other site	6 (2.15%)	11 (5.05%)	
Upper lobe	169 (60.6%)	122 (56.0%)	

2.5. LASSO algorithm model construction

LASSO is a widely used regression method appropriate for analyzing data with high dimensions and strong relationships like high throughput data. We used glmnet [7] implementation of LASSO in R language to establish a LASSO model, including a built-in cross-validation function to adjust the L1 regularization parameter lambda for variables selection and identify marker genes.

2.6. Cell culture, transfection, viability, and proliferation assays

A549 and H538 were purchased from the Chinese Academy of Science Cell Bank and cultured in Dulbecco's Modified Eagle's Medium

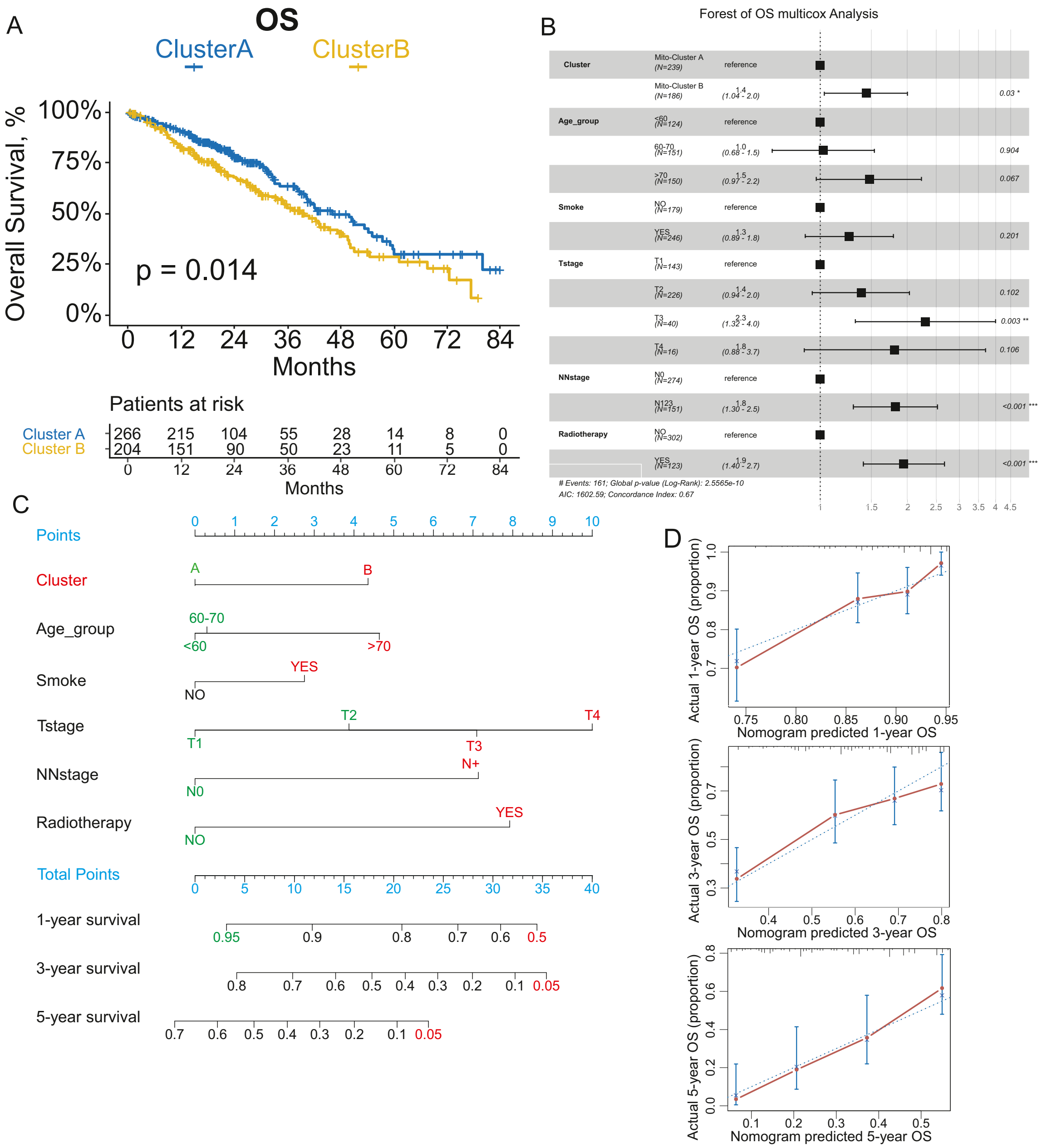

Figure 2. The survival analysis of Cluster A and B (a) The Kaplan Meier survival plot of overall survival (OS); (b) The forest plot of OS multi-variable COX model analysis; (c) The nomogram plot predicting the overall survival probability. (d) The Calibration curves of the nomogram using the bootstrap method in internal validation, Error bars validated standard error of the mean within the current data set.

(DMEM Beyotime, CN) supplemented with 10% Fetal bovine serum (FBS, EveryGreen, Zhejiang, China) and 1% antibiotics at 37 °C in 5% CO_2. Small interfering RNA (siRNA) and the corresponding negative control (siNC) were purchased from Ribobio (Shanghai, China). A549 and H358 cells were transfected with siDARS2, siCOX5B, and siNC using the Lipo8000 Transfection Reagent (Beyotime Biotechnology, China). A549 cells with GFP fluorescence were obtained as reported previously [29]. Cell viability assay was also performed using Cell Counting Kit -8 (CCK-8) assay (Beyotime, CN) according to the previous study. The relative cell viability was calculated as follows: Relative Cell Viability = (Test absorbance/Mean absorbance of control wells). At the logarithmic growth phase, 1500 cells were seeded into 24-well plates (Corning, NY, USA) at 100 μL of cell suspension per well. Cell proliferation was measured according to corresponding fluorescence intensity after incubation for 0, 24, 48, 72, 96, and 120 h at 37 °C. Images were acquired on the Opera Phenix High Content Screening System (PerkinElmer) and analyzed on Harmony High-Content Imaging and Analysis Software.

2.7. RNA isolation and real-time quantitative polymerase chain reaction (RT-qPCR)

Total RNA from cells was extracted using TRIzol reagent (TIANGEN Biotech, Shanghai, China). The cDNA synthesis was performed using the PrimeScript™ RT Master Mix (Yeasen, Shanghai, China). Real-time PCR was conducted with the SYBR-Green kit (Yeasen) to detect the expression levels of DARS2 and COX5B. The gene and primer were listed in: DARS2:forward primer sequence, 5′–ATGTGGAGAGTTGCGTTCGTC, reverse primer sequence, 5′-TGTTTTGCCTTCGGTACTGAATC, COX5B: forward primer sequence, 5′-ATGGCTTCAAGGTTACTTCGC, reverse primer sequence, 5′-CCCTTTGGGGCCAGTACATT. GAPDH was used as a background control gene. The primer sequences for GAPDH were as follows: forward primer sequence, 5′-TCTGCTCCTCCTGTTCGACAGT-3′, reverse primer sequence, 5′-ACCAAATCCGTTGACTCCGAC-3'.

2.8. Patients and LUAD specimens

Tissue specimens, including tumor and adjacent noncancerous tissue, were obtained from 112 patients with LUAD who underwent surgical resection at the Department of Thoracic Surgery, Zhongshan Hospital, Fudan University from February 2012 to November 2016 (30–75 years old). The diagnosis of lung adenocarcinoma was confirmed in each case by histopathological analysis.

2.9. Immunohistochemical staining analysis

One hundred and twelve tissue specimens were performed immunohistochemical staining. Primary antibodies used in IHC, including DARS2 (ab154606, 1:200 for IHC), and COX5B (ab110263, 1:250 for IHC), all antibodies were purchased from Abcam, Cambridge, UK. The procedure was constructed as previously reported [30]. For the quantification of IHC images, the ImageJ IHC Toolbox plugin was used in ImageJ software (NIH).

2.10. Statistical analysis

Statistical analysis was conducted using R Statistical Software. (Version 4.0.4; R Foundation for Statistical Computing). When appropriate, categorical variables were compared by the chi-square test and Fisher's exact test. The normality of data was determined by the Shapiro-Wilk W-tests and determined to use the unpaired t-test or the two-tailed nonparametric Mann-Whitney test when comparing two groups. Spearman's nonparametric analysis determined correlations between all profiles. The Log-rank survival analysis and univariate and multivariate Cox proportional hazards regression by the stepwise method were performed under the R language. The nomogram construction, validation, and calibration were performed and plotted using "rms" and "Hmisc" R packages. For all tests, a p-value $\leq$ 0.05 was considered significant. The exact values are listed. Furthermore, One asterisk represents $p < 0.05$, two asterisks represent $p < 0.01$, three asterisks represent $p < 0.001$, and ns represent not significantly.

3. Results

3.1. The unsupervised machine learning clustering based on mitochondria-related genes

The research flow diagram is shown in Figure 1A. LUAD patients with transcriptome data were collected from the TCGA database. Genes with mitochondrial-associated functions, total 1626, discarding the genes detected no significant difference in the paired analysis of tumor and normal tissues (Table S2), cleaning the genes with ubiquitous low expression levels and excluding the duplicated genes from the Mitocarta 3.0 and MitoMap database.

We carried out the unsupervised consensus clustering using the log (FPKM+1) of the mitochondrial gene expression profile, as shown in Figure 1B. The most appropriate k value was two after checking out the heatmaps of the consensus matrices and the CDF plot (Figure 1C, D). The durability of our clusters was indecently confirmed by the consistent outcome of different clustering algorithm approaches (Figure 1E, F). Based on the aforementioned results, we divided the TCGA-cohort patients into two clusters linked to mitochondrial function and gave them names Cluster A and Cluster B.

3.2. Clinical and survival differences between mitochondrial-function-related clusters

Table 1 shows that Cluster A's patients were more likely to be older women (Age, $p = 0.002$; Sex $p = 0.003$) and with a lower pathological stage (Stage group $p = 0.006$), highlighting Cluster A's favorable prognostic clinical characteristics. To investigate the clinical differences between Cluster A and B, we enrolled all 497 LUAD patients in the Kaplan Meier survival analysis. The overall survival (OS) results revealed that patients in Cluster A exhibited a better survival prognosis (Figure 2A). The median time to the survival of Cluster A was 46.0 (95% CI: 40.3–57.5) months, whereas patients in Cluster B had a considerably shorter median survival time (39.8 [33.3, 48.5] months).

Cluster B was found to be an independent prognostic factor, as shown in Table 2 (Univariate cox: HR 1.59 [1.16, 2.17], $p = 0.004$ and

Table 2. Univarible and multivarible resluts of overall survival identifying the prognostic value of cluster.

	Univarible analysis		Multivarible analysis	
Characteristics	HR (95% CI)	*P*-Value	HR (95% CI)	*P*-Value
Group Cluster A	Ref		Ref	
Cluster B	1.59 (1.16–2.17)	**0.004**	1.44 [1.04, 1.99]	**0.03**
Age_group <60	1.19 (0.87–1.62)	0.269		
60-70	0.95 (0.64–1.41)	0.798	1.03 [0.68, 1.54]	0.904
>70	1.23 (0.84–1.8)	0.297	1.47 [0.97, 2.22]	0.067
Sex female	Ref			
Male	1.18 (0.87–1.61)	0.285		
Smoke No	Ref			
Yes	1.18 (0.86–1.62)	0.301	1.26 [0.89, 1.79]	0.201
T stage T1	Ref		Ref	
T2	1.59 (1.09–2.32)	**0.015**	1.38 [0.94, 2.03]	0.102
T3	3.05 (1.78–5.22)	**<0.001**	2.30 [1.32, 4.00]	**0.003**
T4	3.17 (1.62–6.22)	**0.001**	1.80 [0.88, 3.69]	0.106
N stage N0	Ref		Ref	
N+	2.23 (1.64–3.04)	**<0.001**	1.81 [1.30, 2.53]	**<0.001**
M stage M0	Ref			
M1	1.91 (1.05–3.48)	**0.034**		
MX	0.75 (0.51–1.11)	0.147		
Stage I	Ref			
II	2.23 (1.52–3.27)	**<0.001**		
III	3.03 (2.04–4.51)	**<0.001**		
IV	3.32 (1.77–6.23)	**<0.001**		
Stage_group early stage				
Later stage				
Tumor_site:Lower lobe	Ref			
Middle lobe	1.37 (0.54–3.43)	0.508		
Upper lobe	1.36 (0.62–2.97)	0.447		
other site	0.9 (0.65–1.25)	0.515		
Radiotherapy:No/ unknown	Ref		Ref	
YES	2.26 (1.65–3.08)	**<0.001**	1.93 [1.40, 2.67]	**<0.001**

Significant *p*-values given in bold. 95% CI, 95% confidence interval; HR, hazard ratio.

Multivariate cox: HR 1.44 [1.04, 1.99], $p = 0.030$) (Figure 2B). A nomogram incorporating the predictors, including group, pathologic T stage, pathologic N+ stage, and radiotherapy, was then constructed with a 0.673 C-index (Figure 2C). According to Figure 2D's calibration curves, the internal validation of the nomogram model performed well in terms of reproducibility and predictability.

3.3. The enrichment analysis verified the correlation between mitochondrial function and clusters

The transcriptome profiling difference between the two Clusters was further investigated using the multi-omics data and bioinformatics analysis. Due to the diversity of expression patterns, we conducted the

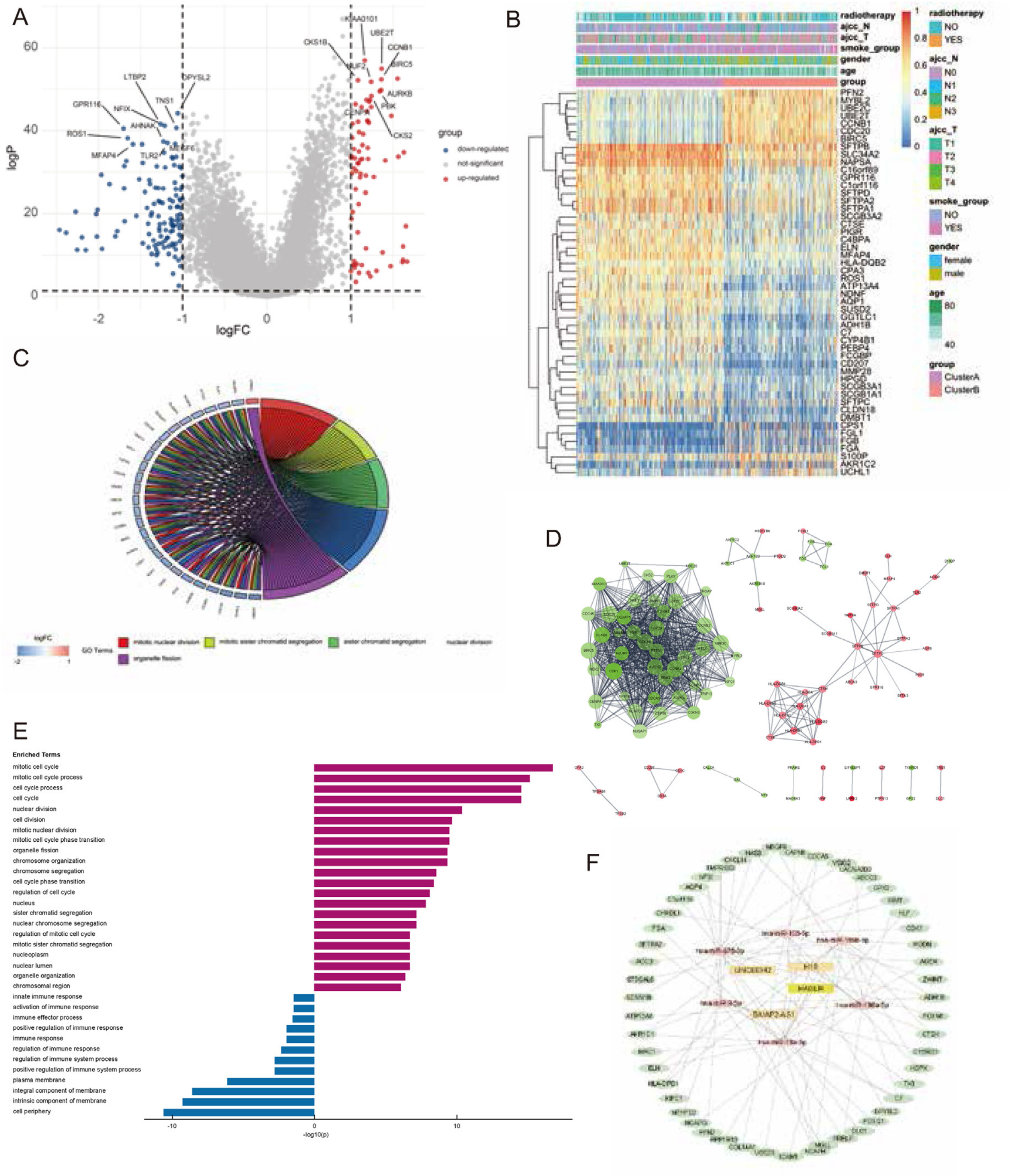

Figure 3. The differential and enrichment analysis of Cluster A and B (a) The volcano plot of differential expressed protein-coding genes, red represents upregulated genes, gray represents no significant difference, and blue represents downregulated genes. (b) The heatmap plot displaying the differential expressed gene of Cluster A and B, the top bars were colored and indicated the clinical features of the patients; (c) The Chord plot displaying the relationship between the top 5 GO enrichment analysis terms and the DE genes. (d) The PPI network of the DE genes, green and red, relatively indicates the downregulated and upregulated genes. (e) The GO terms are enriched in down-regulated and upregulated genes (purple: upregulated DE genes in Cluster B, blue: down-regulated DE genes). (f) The ceRNA network of the DElncRNA, DEmiRNA, and DEgenes.

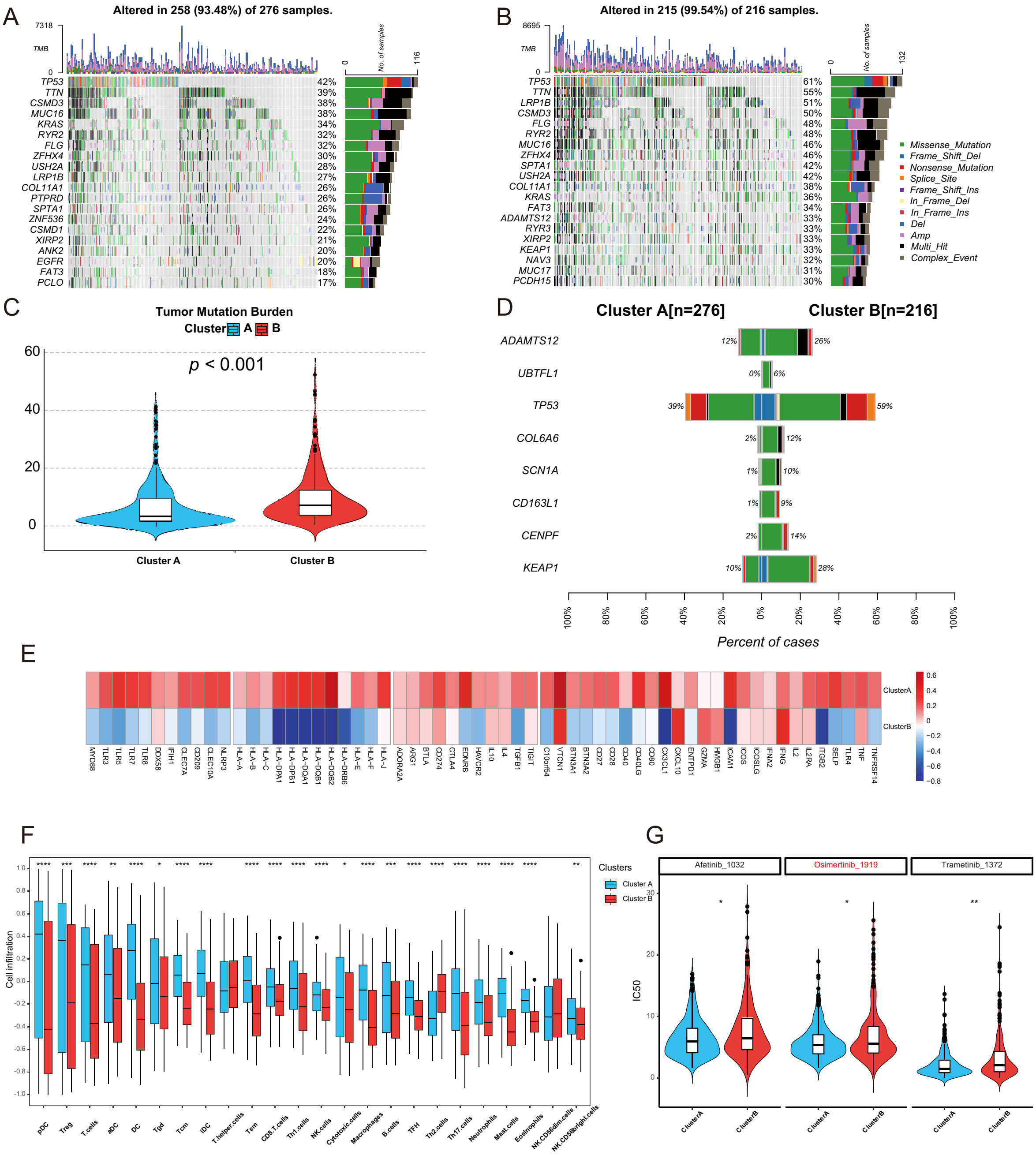

Figure 4. The genomic analysis of Cluster A and B The waterfall (oncoplot) displays the somatic and methylated landscape of Cluster A (a) and Cluster B(b). (c) The Box plot shows that tumor mutation burden (TMB) is higher in Cluster B vs. Cluster A; Boxes show the median ±25–75th percentiles, and whiskers show a 1.5 × interquartile range below above the 25th and 75th percentiles, respectively. *P* values are derived via a two-sided T-test (d). The co-bar plot showed the differentially mutated genes in Cluster A and B; the fisher test was performed. (e) The heat plot displayed the relative expression level of immune-related genes from the ImmPort database. (f) Comparison of the immune cell differences between the two Clusters. (g) The comparison of the predicted IC50 value of three targeted drugs between Cluster A and B.

differential analysis separately in a different type of transcriptome. Of all 19971 protein-coding genes, 123 downregulated and 73 upregulated were identified in Cluster B (threshold: LogFC > 1, adjust *p* value < 0.05). Similarly, we obtained 99 differential expressed miRNAs of 2248 miRNA isoforms, and 82 differentials expressed lncRNAs(Figure 3A, Supplementary Figure S2A, B, C, D). As illustrated in Figure 3B, the heatmap display of the expression profile could distinguish the two clusters and show the fold change difference.

Further enrichment analysis on the DEGs was interesting. GO and KEGG analysis showed that most of the enriched terms or pathways were related to the mitochondria's function, like cell cycle checkpoint, Oxidative phosphorylation, ROS formation, etc. (Figure 3C, Supplementary Figure S3 A, B, C, D)

Additionally, we discovered the protein-protein interaction (PPI) network of DEGs using the STRING online database and displayed it using Cytoscape. Cytoscape calculated the DEGs' degree levels. Figure 3D also shows the PPI network, from which the DEGs were found to have interacted with each other and could be divided into two main modules. The upregulated modular genes in Cluster B were enriched in the mitochondrial function terms, while the signaling and immune function-related terms were significantly enriched among the downregulated module genes, as shown in Figure 3E. Similar results were further

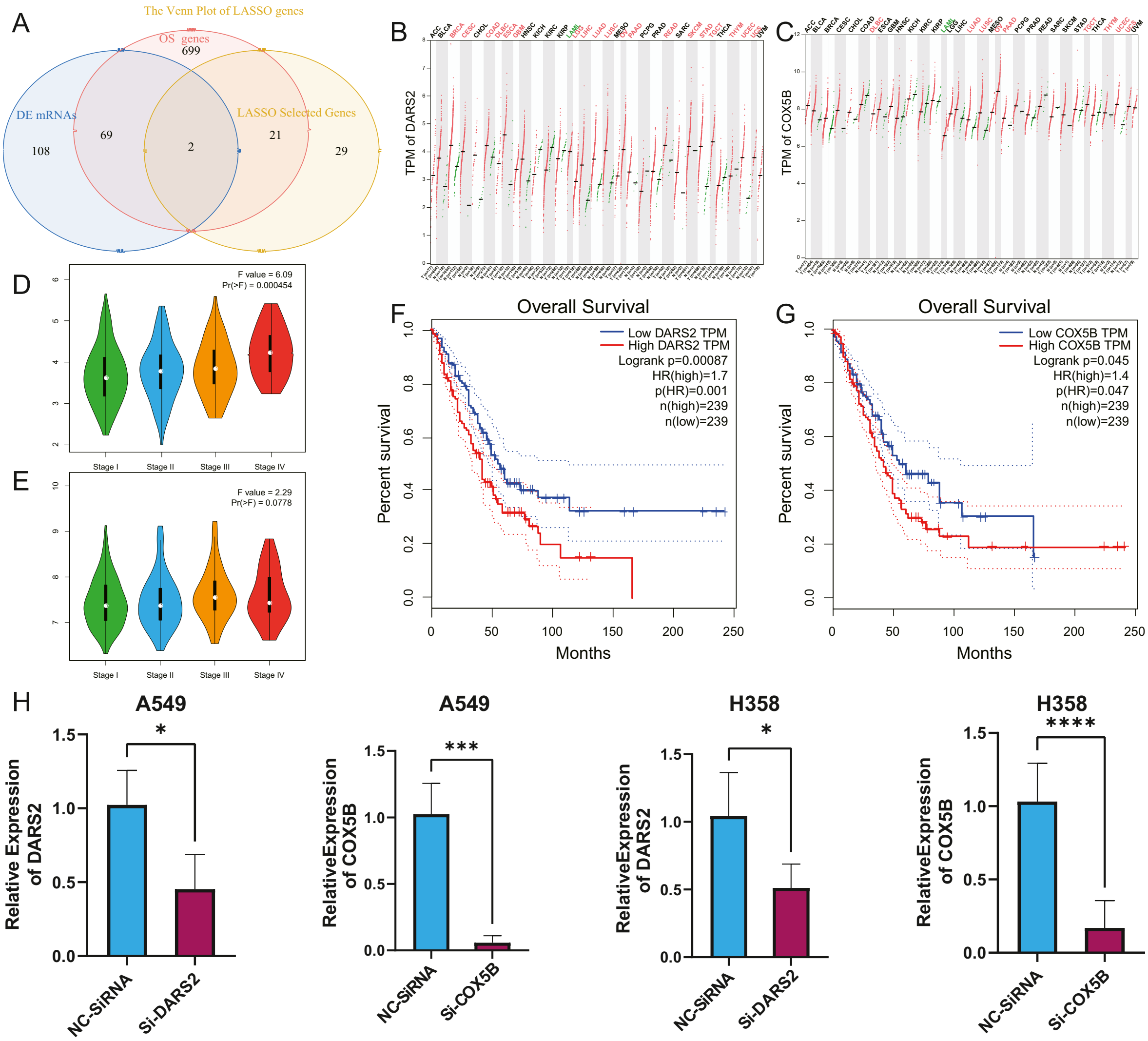

Figure 5. The hub gene identification and prognostic validation. The Venn plot(a) showed the identification of the two hub genes. The pan-cancer dot plot identified the DARS2(b) and COX5B(c) expression profiles across all tumor samples and normal tissues. The violin plot showed the correlation between clinical stage and the expression level of DARS2(d) and COX5B(e). KM overall survival analysis of TCGA-LUAD tumor samples with high or low DARS2(f) and COX5B(g) expression. The effect of siRNA to knockdown DARS2 and COX5B mRNA levels in A549 and H358 cell lines was measured by RT–qPCR(h).

endorsed by an unbiased Gene Set Enrichment Analysis (GSEA) (Supplementary Figure S3E, F, G, H).

The specific mitochondrial function-related ceRNA network shown in Figure 3F was constructed by 195 pairs of lncRNA-miRNA-gene, involving four lncRNAs, six miRNAs, and 156 genes bioinformatic analysis. Among the network, the downregulated lncRNA BAIAP2-AS1, HAGLR, and LINC00342 and the upregulated lncRNA H19 were identified as critical regulatory lncRNAs in the ceRNA network.

3.4. Cluster B harbored a heavier tumor mutation load, poorer immune cell infiltration, and higher targeted drug tolerance

After detecting the RNA-seq alterations and the enriched functions in the above section, the investigation of the two clusters in the genomic layer was conducted for further comparison. Our study examined single-nucleotide polymorphism (SNP), insertion and deletion (INDEL), and copy number mutation changes in genomics (CNV). We integrated the genomic CNV data into the MAF file using the maftools package in the data handling procedure.

The frequencies and types of variations observed in the two clusters are shown in Figures 4A and B and Supplementary Figure S4A and B. Clusters A and B shared a similar distribution of top mutant genes and variant classification. However, cluster B had significantly more variations per sample than cluster A. (Median: cluster B 240 vs. Cluster A 109.5). Notably, as shown in Figure 4D and Supplementary Figure S4, the bar plots illustrated that Cluster B exhibited greater frequencies of somatic mutations in the genes KEAP1 (10% *vs.* 20%, $p < 0.001$), TP53 (39% *vs.* 59%, $p < 0.001$), and ADAMTS12 (12% *vs.* 26%, $p < 0.001$).

The following pathway forest plot verified that the higher level of mutation variations gives rise to the activation and interaction of various oncopathways in ClusterB (for example, Hippo, A *vs.* B, Odds Ratio: 0.373, $P < 0.001$). Consistent with the mutation analysis above, cluster B holds an elevated TMB (Figure 4C). The above results suggested that patients in Cluster B typically bear more genomics abnormity.

The tumor immune microenvironment, including the immune cells and other mediated immune molecules, is an increasingly important area in the integrated analysis of tumor features. Our further GSVA algorithm analysis based on the 24 immune cells revealed that Cluster A showed a relatively higher immune cell composition (Figure 4F & Supplementary Figure S5D). Moreover, the investigation of other immune-related gene expressions (Innate immune, antigen presentation, immune stimulator and immune inhibitor) exhibits a similar pattern to cell composition (Figure 4E & Supplementary Figure S5A and B). The boxplot in Supplementary Figure S5C provides a broad overview of the checkpoint gene landscape. Immune checkpoint gene expression was lower in Cluster B patients compared to Cluster A patients. Additionally, we identified the expression pattern of other targeted therapeutic genes undergoing clinical trials. Surprisingly, Cluster A showed much higher expression of practically all medication target genes than Cluster B.

The higher immune infiltration activated more potential for immunotherapy and targeted treatments. We extensively examined the IC50 values of numerous medications and substances in patients having transcriptome data using the GDSC database. Patients in Cluster B demonstrated a decreased IC50 value for targeted therapy medications such as Afatinib, Osimertinib, and Trametinib, which is consistent with the earlier findings. The lower sensitivity to targeted drugs also inspired us to find novel therapeutic target genes related to mitochondrial function.

3.5. A combination of LASSO and batch survival analysis identified two hub genes

Further analysis to seek hub genes was followed, and the mass univariate cox regression was performed to identify OS-related DEmRNAs. Meanwhile, the LASSO regression analysis (Supplementary Figure S6A) narrowed the range of hub DEmRNAs selection. The overlap of two genes, DARS2 (Aspartyl-TRNA Synthetase 2, Mitochondrial) and COX5B (Cytochrome C Oxidase Subunit 5B), was shown to be associated with

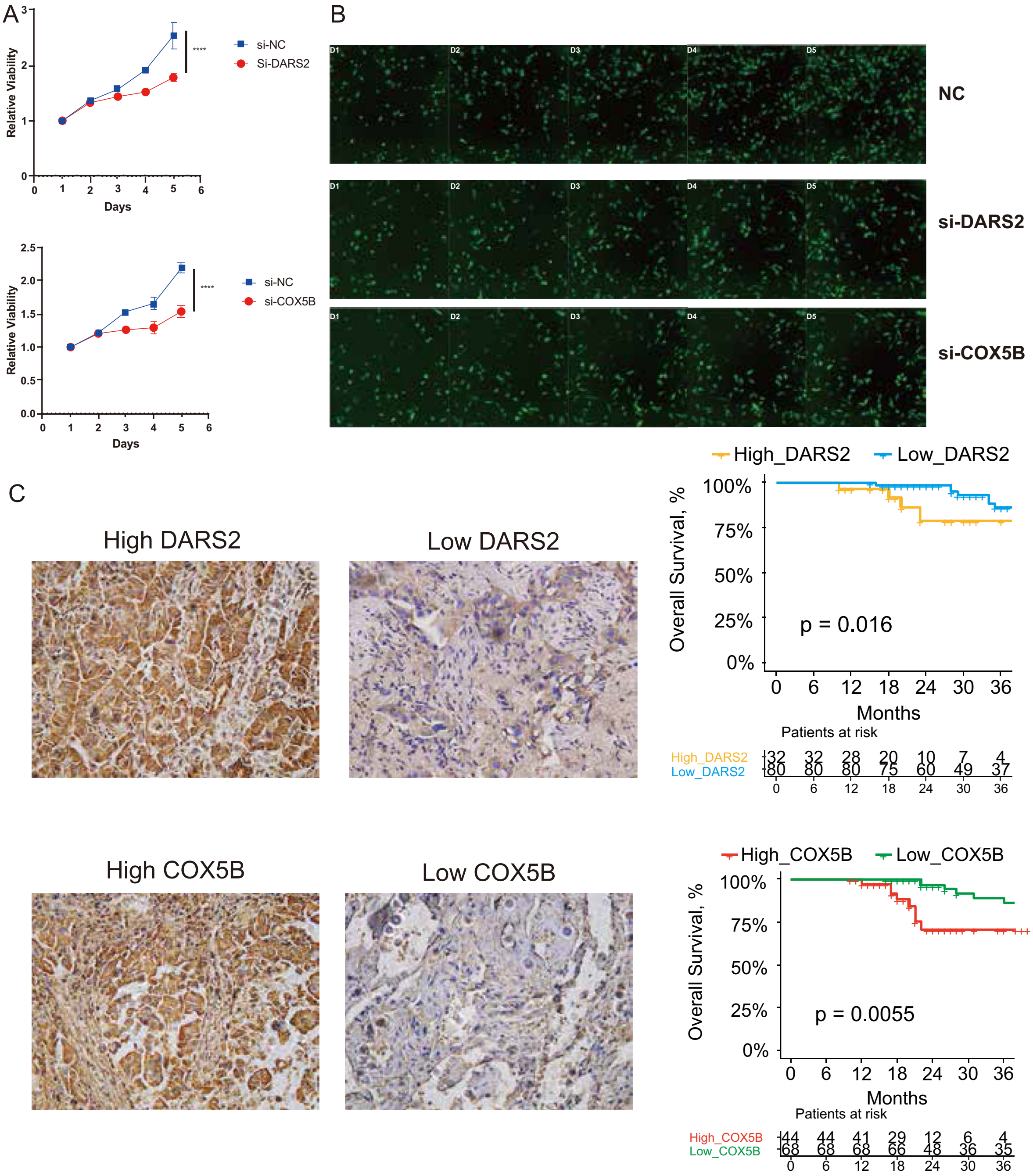

Figure 6. Further verification of the potential therapeutic. Line chart(a) and the high-content imaging(b) illustrate decreased cell viability in A549 after transfection with si-DARS2 and si-COX5B. (c) The representative IHC staining image of low and high expression levels of DARS2 and COX5B in LUAD tumor samples. (d) The Unsupervised univariate analysis using Kaplan-Meier survival analysis of patients with lung adenocarcinoma in our hospital showed that patients with high expression of DARS2 or COX5B had shorter OS and were associated with poor prognosis.

DEGs, overall survival outcomes, and the LASSO analysis (Figure 5A). DARS2 and COX5B expression was independently second confirmed in the PANCANCER data (Figure 5B and C). With little expression in healthy tissues, DARS2 and COX5B were expressed at higher levels in several tumor types, such as colon, urinary, cervical, and ovarian cancer. Moreover, their increased gene expression was correlated to the worse survival outcome of LUAD patients (Figure 5F and G) and closely linked to the LUAD tumor staging (Figure 5D and E).

3.6. Knockdown of DARS2 and COX5B significantly attenuated cells' proliferation

To confirm the bioinformatics findings, we next conducted a series of cell-based experiments to investigate the function of these two genes in lung adenocarcinoma cell lines A549 and H358. First, using RT-qPCR analysis, we validated the knockdown of DARS2 and COX5B (Figure 5H). Then, the CCK-8 cell viability studies showed that DARS2 and COX5B knockdown reduced tumor cell proliferation (Figure 6A). Furthermore, the LUAD cells with GFP reporter protein were transfected with DARS2 and COX5B siRNAs and observed in the high-content system for five consecutive days (Figure 6B). These findings clearly showed that suppressing DARS2 and COX5B reduced cell growth.

3.7. DARS2 and COX5B's expression was significantly correlated with lung adenocarcinoma patients' survival in IHC staining

The next section of our study explored the clinical significance of these two genes in Real-World Data. Again, we proved that DARS2 and COX5B are risk factors of prognosis for the LUAD patients in multiple curated GEO public datasets (Supplementary Figure S6B and C). Next, we collected the paraffin-embedded sections from our cohort, including 112 lung adenocarcinoma patients who underwent surgery. The IHC results confirmed that DARS2 and COX5B expression was elevated in LUAD tumor tissues compared with that in paired normal tissues. Figure 6C displays statistics for scores as well as typical photos of IHC staining. We further confirmed that patients with high expression of DARS2 or COX5B had significantly worse prognoses based on the KM survival analysis.

4. Discussion

This study first collected and integrated the mitochondrial function-related genes with different databases and then divided the patients into clusters A and B with distinct clinical and genetic characteristics using the unsupervised machine learning method. Patients in Cluster B displayed significant enrichments in cell proliferation, DNA replication, and essential mitochondrial functions like oxidative phosphorylation, according to the DE and GSEA analysis, whereas Cluster A's profiles were enriched in a variety of immune response pathways. Our research also demonstrated and validated that patients in Cluster A had good survival.

We created the mitochondrial function-related ceRNA network through databases search and literature analysis based on the differentially expressed lncRNAs, miRNAs, and mRNAs. HAGLR, as a prognostic biomarker, was identified in various tumor types, including lung cancer [31]. An earlier investigation revealed that HAGLR promoted cell migration and invasion by targeting miRNAs like miR-147a and miR-133b and promoting lipogenesis in NSCLC [32]. Interestingly, Our network analysis proposed a new miRNA target of HAGLR, miR-18a-5p, which was reported to play a dual role in oncogenic processes [33]. The discovery of the ceRNA networks in our findings offers a fresh perspective on understanding the mitochondrial function in LUAD.

Furthermore, our study revealed that the mitochondrial function plays pivotal roles in the tumor immune milieu, including innate immunity, checkpoint therapy, and immune cell infiltration. Mitochondria are essential players in innate immunity pathways, both in signaling and response, according to West and Banoth's review [34, 35]. Additionally, Klein et al. discovered that by reducing ROS production, limiting mitochondrial dysfunction could improve tumor killer T cell survival [36]. Our studies also indicated that the EGFR-TKI medication Osimertinib resistance might be a result of mitochondrial malfunction from the comparison of Cluster A and B, Recent research revealed that chemoresistance in malignant tumors was linked to mitochondrial dysfunction [37]. Zhou et al. proved that the acquired drug-resistant A549 cells continue to have mitochondrial abnormalities [38]. For individuals resistant to treatment, identifying candidate genes associated with mitochondrial function can aid in developing more potent therapy plans.

Subsequently, we identified the two hub genes DARS2 and COX5B through continuous screening using the LASSO algorithm, validated their clinical value as a potential biomarker in LUAD patients, and explored their protumorigenic tumor phenotype. DARS2 is a mitochondrial enzyme aminoacylated aspartyl-tRNA. Previous work indicated its deficiency mainly caused leukoencephalopathy and cardiomyopathy [39, 40, 41, 42], but recent studies have identified that DARS2 has prognostic significance in urinary system cancers [43, 44], hematological malignancies [45], and lung cancer [46]. Jiang et al. demonstrated that DARS2 modulates the proliferation, invasion, and apoptosis of LUAD cells [47], which aligns with our results. COX5B is the terminal enzyme of the mitochondrial respiratory chain and is associated with tumor growth and migration in several cancers [48, 49, 50, 51]. The loss of COX5B inhibits proliferation and promotes senescence via mitochondrial dysfunction in breast cancer, as indicated by Gao et al. [52]. For the first time, our work methodically uncovered the clinical significance and pro-tumor character of DARX2 and COX5B in LUAD.

Our research has several restrictions. More comprehensive tests were required better to understand the underlying mechanisms of these two hub genes. Still, to some extent, we first demonstrated these two proteins' positive effects on LUAD tumor cell proliferation. Besides, our cohort's patients' prognostic data was unavailable due to the short follow-up time. Additional open-source GEO data with RNA-seq data was added to confirm the clinical value of our study.

5. Conclusion

In conclusion, the purpose of the current study was to reveal the correlation between the mitochondrial function pattern and multi-omics characteristics of lung adenocarcinoma and discover the potential targets. The patients clustered into two patterns based on the transcriptome of mitochondrial function-related genes showed different clinical outcomes, mitochondrial activities, genomic features, and immune infiltrations. Next, we conducted the LASSO-COX algorithm and PPI network to anchor the two candidate genes: DARS2 and COX5B. Furthermore, experimental and clinical specimen results ascertained that these two genes could promote tumor growth phenotype. The study has enhanced our understanding of the relationship between mitochondria activity and metabolism reprogramming, genomic abnormality, and immune dysfunction in LUAD, suggesting the crosstalk of mitochondrial activity and tumor progression. DARS2 and COX5B had been initially identified and validated as potential therapeutic candidate targets in LUAD.

Declarations

Author contribution statement

Cheng Zhan, Miao Lin, Hao Wang: Conceived and designed the experiments.

Xing Jin, Qihai Sui: Performed the experiments; Wrote the paper.

Ming Li; Jiaqi Liang: Analyzed and interpreted the data.

Zhencong Chen, Zhengyang Hu, Ye Cheng, Yuansheng Zheng: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by the Natural Science Foundation of Shanghai [No. 22ZR1411900].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e11966.

Acknowledgements

We thank all the patients and investigators who participated in the study. We thank TCGA and GEO database for supplying the platforms and the generous grantors for sharing their precious data sets.

References

[1] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin (2021).

[2] R. Zheng, et al., Cancer incidence and mortality in China, 2016, J. Nat. Cancer Center 2 (2022) 1–9.

[3] G. de Castro, et al., Five-year outcomes with pembrolizumab versus chemotherapy as first-line therapy in patients with non-small-cell lung cancer and programmed death ligand-1 tumor proportion score $\geq$ 1% in the KEYNOTE-042 study, J. Clin. Oncol. (2022), JCO2102885.

[4] O. Warburg, On respiratory impairment in cancer cells, Science 124 (1956) 269–270.

[5] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (2011) 646–674.

[6] D. Hanahan, Hallmarks of cancer: new dimensions, Cancer Discov. 12 (2022) 31–46.

[7] L. Guo, Mitochondria and the permeability transition pore in cancer metabolic reprogramming, Biochem. Pharmacol. 188 (2021), 114537.

[8] H. Yu, P. Guo, X. Xie, Y. Wang, G. Chen, Ferroptosis, a new form of cell death, and its relationships with tumourous diseases, J. Cell Mol. Med. 21 (2017) 648–657.

[9] A.M. Battaglia, et al., Ferroptosis and cancer: mitochondria meet the 'iron maiden' cell death, Cells 9 (2020) E1505.

[10] C.-L. Kuo, et al., Mitochondrial oxidative stress in the tumor microenvironment and cancer immunoescape: foe or friend? J. Biomed. Sci. 29 (2022) 74.

[11] P. Ghosh, C. Vidal, S. Dey, L. Zhang, Mitochondria targeting as an effective strategy for cancer therapy, Int. J. Mol. Sci. 21 (2020) E3363.

[12] A.L. Beam, I.S. Kohane, Big data and machine learning in health care, JAMA 319 (2018) 1317–1318.

[13] K.A. Tran, et al., Deep learning in cancer diagnosis, prognosis and treatment selection, Genome Med. 13 (2021) 152.

[14] A.J. Combes, et al., Discovering dominant tumor immune archetypes in a pan-cancer census, Cell 185 (2022) 184–203.

[15] X. Wang, F. Markowetz, F. De Sousa E Melo, J.P. Medema, L. Vermeulen, Dissecting cancer heterogeneity–an unsupervised classification approach, Int. J. Biochem. Cell Biol. 45 (2013) 2574–2579.

[16] R.J. Woolley, et al., Machine learning based on biomarker profiles identifies distinct subgroups of heart failure with preserved ejection fraction, Eur. J. Heart Fail. 23 (2021) 983–991.

[17] A. Mayakonda, D.-C. Lin, Y. Assenov, C. Plass, H.P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer, Genome Res. 28 (2018) 1747–1756.

[18] S. Rath, et al., MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations, Nucleic Acids Res. 49 (2021) D1541–D1547.

[19] A.C. Smith, A.J. Robinson, MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases, Nucleic Acids Res. 47 (2019) D1225–D1228.

[20] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (2010) 1572–1573.

[21] M.E. Ritchie, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47.

[22] H. Dweep, N. Gretz, C. Sticht, miRWalk database for miRNA-target interactions, Methods Mol. Biol. 1182 (2014) 289–305.

[23] V. Agarwal, G.W. Bell, J.-W. Nam, D.P. Bartel, Predicting effective microRNA target sites in mammalian mRNAs, Elife 4 (2015).

[24] J. Chen, et al., LncSEA: a platform for long non-coding RNA related sets and enrichment analysis, Nucleic Acids Res. 49 (2021) D969–D980.

[25] S. Hänzelmann, R. Castelo, J.G.S.V.A. Guinney, Gene set variation analysis for microarray and RNA-Seq data, BMC Bioinf. 14 (2013) 7.

[26] G. Bindea, et al., Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer, Immunity 39 (2013) 782–795.

[27] D. Maeser, R.F. Gruener, R.S. Huang, oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data, Briefings Bioinf. (2021), bbab260.

[28] P. Geeleher, N. Cox, R.S. Huang, pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels, PLoS One 9 (2014), e107468.

[29] G. Bi, et al., Knockdown of GTF2E2 inhibits the growth and progression of lung adenocarcinoma via RPS4X in vitro and in vivo, Cancer Cell Int. 21 (2021) 181.

[30] C. Zhan, et al., Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma, J. Thorac. Dis. 7 (2015) 8.

[31] C. Yang, et al., Long noncoding RNA HAGLR acts as a microRNA-143-5p sponge to regulate epithelial-mesenchymal transition and metastatic potential in esophageal cancer by regulating LAMP3, Faseb. J. 33 (2019) 10490–10504.

[32] C. Lu, J. Ma, D. Cai, Increased HAGLR expression promotes non-small cell lung cancer proliferation and invasion via enhanced de novo lipogenesis, Tumour Biol 39 (2017), 1010428317697574.

[33] T. Kolenda, et al., Good or not good: role of miR-18a in cancer biology, Rep. Practical Oncol. Radiother. 25 (2020) 808–819.

[34] A.P. West, Mitochondrial dysfunction as a trigger of innate immune responses and inflammation, Toxicology 391 (2017) 54–63.

[35] B. Banoth, S.L. Cassel, Mitochondria in innate immune signaling, Transl. Res. 202 (2018) 52–68.

[36] K. Klein, et al., Role of mitochondria in cancer immune evasion and potential therapeutic approaches, Front. Immunol. 11 (2020), 573326.

[37] N. Guaragnella, S. Giannattasio, L. Moro, Mitochondrial dysfunction in cancer chemoresistance, Biochem. Pharmacol. 92 (2014) 62–72.

[38] X. Zhou, et al., Dichloroacetate restores drug sensitivity in paclitaxel-resistant cells by inducing citric acid accumulation, Mol. Cancer 14 (2015) 63.

[39] G.C. Scheper, et al., Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation, Nat. Genet. 39 (2007) 534–539.

[40] S.A. Dogan, et al., Tissue-specific loss of DARS2 activates stress responses independently of respiratory chain deficiency in the heart, Cell Metabol. 19 (2014) 458–469.

[41] W.B.V. Pinto, P.V.S. de Souza, DARS2 gene clinical spectrum: new ideas regarding an underdiagnosed leukoencephalopathy, Brain 137 (2014) e289.

[42] M.W. Friederich, et al., Pathogenic variants in glutamyl-tRNAGln amidotransferase subunits cause a lethal mitochondrial cardiomyopathy disorder, Nat. Commun. 9 (2018) 4065.

[43] F. Chen, Q. Wang, Y. Zhou, The construction and validation of an RNA binding protein-related prognostic model for bladder cancer, BMC Cancer 21 (2021) 244.

[44] Y. Wu, et al., Identification of the functions and prognostic values of RNA binding proteins in bladder cancer, Front. Genet. 12 (2021), 574196.

[45] Y. Zhang, et al., Identification of biomarkers for acute leukemia via machine learning-based stemness index, Gene 804 (2021), 145903.

[46] L. Yang, et al., Development and validation of a prediction model for lung adenocarcinoma based on RNA-binding protein, Ann. Transl. Med. 9 (2021) 474.

[47] Y. Jiang, et al., High expression of DARS2 indicates poor prognosis in lung adenocarcinoma, J. Clin. Lab. Anal. 36 (2022), e24691.

[48] J. Stein, J. Tenbrock, G. Kristiansen, S.C. Müller, J. Ellinger, Systematic expression analysis of the mitochondrial respiratory chain protein subunits identifies *COX5B* as a prognostic marker in clear cell renal cell carcinoma, Int. J. Urol. 26 (2019) 910–916.

[49] T. Hu, J. Xi, Identification of COX5B as a novel biomarker in high-grade glioma patients, OTT 10 (2017) 5463–5470.

[50] S.-P. Gao, et al., High expression of COX5B is associated with poor prognosis in breast cancer, Future Oncol. 13 (2017) 1711–1719.

[51] Y.-D. Chu, et al., COX5B-Mediated bioenergetic alteration regulates tumor growth and migration by modulating AMPK-UHMK1-ERK cascade in hepatoma, Cancers 12 (2020) 1646.

[52] S.-P. Gao, et al., Loss of COX5B inhibits proliferation and promotes senescence via mitochondrial dysfunction in breast cancer, Oncotarget 6 (2015) 43363–43374.