Received: 9 February 2021 | Revised: 21 September 2021 | Accepted: 14 October 2021

DOI: 10.1002/cam4.4386

RESEARCH ARTICLE

Cancer Medicine Open Access WILEY

Nonnegative matrix factorization-based bioinformatics analysis reveals that TPX2 and SELENBP1 are two predictors of the inner sub-consensuses of lung adenocarcinoma

Haiwei Wang[1,2] | **Xinrui Wang**[1,2] | **Liangpu Xu**[1,2] | **Hua Cao**[1,2] | **Ji Zhang**[3]

[1]Fujian Key Laboratory for Prenatal Diagnosis and Birth Defect, Fujian Maternity and Child Health Hospital, Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian, China

[2]Key Laboratory of Technical Evaluation of Fertility Regulation for Non-human Primate, National Health and Family Planning Commission, Fuzhou, Fujian, China

[3]State Key Laboratory for Medical Genomics, Shanghai Institute of Hematology, Rui-Jin Hospital Affiliated to School of Medicine, Shanghai Jiao Tong University, Shanghai, China

Correspondence

Haiwei Wang, Fujian Key Laboratory for Prenatal Diagnosis and Birth Defect, Fujian Maternity and Child Health Hospital, Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian, China.
Email: hwwang@sibs.ac.cn

Hua Cao, Key Laboratory of Technical Evaluation of Fertility Regulation for Non-human Primate, National Health and Family Planning Commission, Fuzhou, Fujian, China.
Email: caohua69@fjmu.edu.cn

Ji Zhang, State Key Laboratory for Medical Genomics, Shanghai Institute of Hematology, Rui-Jin Hospital Affiliated to School of Medicine, Shanghai Jiao Tong University, Shanghai, China.
Email: Zj11222@rjh.com.cn

Abstract

Background: Lung adenocarcinoma (LUAD) is a heterogeneous disease. However the inner sub-groups of LUAD have not been fully studied. Markers predicted the sub-groups and prognosis of LUAD are badly needed.

Aims: To identify biomarkers associated with the sub-groups and prognosis of LUAD.

Materials and Methods: Using nonnegative matrix factorization (NMF) clustering, LUAD patients from The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) datasets and LUAD cell lines from Genomics of Drug Sensitivity in Cancer (GDSC) dataset were divided into different sub-consensuses based on the gene expression profiling. The overall survival of LUAD patients in each sub-consensus was determined by Kaplan-Meier survival analysis. The common genes which were differentially expressed in each sub-consensus of LUAD patients and LUAD cell lines were identified using TBtools. The predictive accuracy of TPX2 and SELENBP1 for theinner sub-consensuses of LUAD was determined by Receiver operator characteristic (ROC) analysis. The Kaplan-Meier survival analysis was also used to test the prognostic significance of TPX2 and SELENBP1 in LUAD patients.

Results: Using nonnegative matrix factorization clustering, LUAD patients in The Cancer Genome Atlas (TCGA), GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets were divided into three sub-consensuses. Sub-consensus3 LUAD patients were with low overall survival and were with high TP53 mutations. Similarly, LUAD cell lines were also divided into three sub-consensuses by NMF method, and sub-consensus2 cell lines were resistant to EGFR inhibitors. Identification of the common genes which were differentially expressed in different sub-consensuses of LUAD patients and LUAD cell lines revealed that TPX2 was highly expressed in sub-consensus3 LUAD patients and sub-consensus2

Haiwei Wang and Xinrui Wang contributed equally to this work.

Funding information
The present study was supported by grants from the Fujian Maternity and Child Health Hospital (grant nos. YCXB 18-10 and YCXM 19-04). This study was also supported by Fujian Natural Science Foundation (grant nos. 2020J01337).

LUAD cell lines. On the contrary, SELENBP1 was highly expressed in sub-consensus1 LUAD patients and sub-consensus1 LUAD cell lines. The expression levels of TPX2 and SELENBP1 could distinguish sub-consensus3 LUAD patients or sub-consensus2 LUAD cell lines from other sub-consensuses of LUAD patients or cell lines. Moreover, compared with normal lung tissues, TPX2 was highly expressed, while, SELENBP1 was lowly expressed in LUAD tissues. Furthermore, the higher expression levels of TPX2 were associated with the lower relapse-free survival and the lower overall survival of LUAD patients. While, the higher expression levels of SELENBP1 were associated with the higher relapse-free survival and higher overall survival. At last, we showed that TP53 mutant LUAD patients were with higher TPX2 and lower SELENBP1 expressions.

Discussion: Both iCluster and NMF method are proved to be robust LUAD classification systems. However, the LUAD patients in different iclusters had no significant clinical overall survival, while, sub-consensus3 LUAD patients from NMF classification were with lower overall survival than other sub-consensuses.

Conclusions: By integrated analysis of 1765 LUAD patients and 64 LUAD cell lines, we showed that NMF was a robust inner sub-consensuses classification method of LUAD. TPX2 and SELENBP1 were differentially expressed in different LUAD sub- consensuses, and predicted the inner sub-consensuses of LUAD with high accuracy. TPX2 was an unfavorable prognostic biomarker of LUAD which was up-regulated in LUAD tissues and associated with the low overall survival of LUAD. SELENBP1 was a favorable prognostic biomarker of LUAD which was down-regulated in LUAD tissues and associated with the prolonged overall survival of LUAD.

KEYWORDS
nonnegative matrix factorization, SELENBP1, sub-consensus of lung adenocarcinoma, TPX2

1 | BACKGROUND

Lung adenocarcinoma (LUAD) is one of common and lethal type of non-small cell lung cancer (NSCLC).[1,2] The incidence and mortality of LUAD are increasing every year.[3,4] Genetic alterations of LUAD are extensively studied. Somatically mutant tumor suppressor gene TP53 and activated oncogenes EGFR and KRAS are commonly detected in LUAD.[5] Although molecular-targeted therapies, like EGFR inhibition therapy[6,7] and checkpoint blockade immune therapy[8,9] have achieved some improvements in clinical outcomes, the 5-year survival rate of LUAD remains very low. LUAD is a heterogeneous disease. Previously, based on the genetic alterations,[10,11] the mRNA,[12–14] microRNA[15–17] or long non-coding RNA expression signature,[18–20] and the immune cells infiltration signature,[21,22] LUAD could be further divided into different sub-groups. In the year of 2014, the Cancer Genome Atlas (TCGA) research groups divided LUAD into six clusters using iCluster analysis by integrating the gene expression, DNA methylation, and genetic alterations of LUAD.[5] Each cluster of LUAD showed different molecular features. For example, TP53 mutations were enriched in clusters 1–3 and SETD2 mutations were enriched in cluster 4. This study provided deep understanding of the molecular heterogeneity of LUAD. However, the clinical overall survival of those six clusters was not significantly different. So, new classification methods are needed to further reveal the inner sub-consensuses of LUAD. And more prognostic makers are needed to predict the therapeutic responses and clinical outcomes of LUAD.

Nonnegative matrix factorization (NMF) is an unsupervised sub-consensus clustering system.[23] By calculating the approximation of the factors, NMF could reveal the basic patterns of multidimensional data.[24,25] NMF has been used successfully in many fields, including cancers.

Colon cancer patients could be divided into goblet-like, enterocyte, stem-like, inflammatory, and transit-amplifying five sub-consensuses based on NMF classification.[26] And this sub-consensus classification system could predict the clinical outcomes and therapeutic responses of colon cancer patients. We also used NMF method to identify the sub-consensus of colon cancer cell lines and found a sub-consensus of colon cancer cells was sensitive to BRAF inhibitors and PI3K-mTOR inhibitors.[27] NMF was also used for liver cancer classification[28–30] and pancreatic cancer classification.[31,32] Lung squamous cell carcinoma (LUSC) is another type of NSCLC.[33] Using NMF method, LUSC could be divided into LUSC-A and LUSC-B two sub-consensuses with different biological characteristics and clinical outcomes.[34] All those results highlighted that NMF was a robust cancer classification system. However, the robustness of NMF classification system in LUAD was not analyzed.

To the best of our knowledge, this is the first integrated bioinformatics study of large cohorts of LUAD patients and LUAD cell lines to reveal the inner heterogeneity of LUAD by NMF method. Our results provide deep understanding of the inner heterogeneity of LUAD. Our data also suggests the potential prognostic biomarkers and therapeutic targets of LUAD.

2 | MATERIALS AND METHODS

2.1 | Data collection and processing

TCGA LUAD gene expression dataset, DNA mutation dataset, and LUAD clinical dataset were downloaded from the TCGA hub (https://tcga.xenahubs.net).[5]

The gene expression matrix of LUAD patients along with the clinical survival information was downloaded from the Gene Expression Omnibus (GEO) website (www.ncbi.nlm.nih.gov/geo), including GSE30219,[35] GSE42127,[36] GSE50081,[37] GSE68465,[38] and GSE72094[39] datasets. The gene expression series matrix of normal lung tissues and LUAD tissues was downloaded from GSE7670,[40] GSE10072,[41] GSE18842,[42] GSE27262,[43] and GSE32863[44] datasets. Raw CEL data and clinical data of LUAD patients in MSKCC dataset 1 and MSKCC dataset 2 were available athttp://cbio.mskcc.org/Public/lung_array_data/.[45] All the GEO datasets were annotated using R software (version 3.5.0). The expression values were averaged by "plyr" package (version 1.8.5).

Gene expression matrix and drug sensitivity of LUAD cell lines were downloaded from Genomics of Drug Sensitivity in Cancer (GDSC) project (https://www.cancerrxgene.org/).[46]

2.2 | NMF classification of LUAD patients and LUAD cell lines

LUAD patients and LUAD cell lines were divided into two sub-consensuses, three sub-consensuses, or four sub-consensuses by "NMF" package in R software based on the globe gene expression levels (version 0.23.0l). The number of sub-consensuses was determined by the number of ranks.

2.3 | Survival analysis

The overall survival of LUAD patients in each sub-consensus was determined by the Kaplan–Meier survival analysis which was performed using "survival" package (version 3.1-8) in R statistics software. The prognostic values of TPX2 and SELENBP1 on the relapse-free survival or overall survival were also determined using "survival" package. *P* values were calculated by the log-rank test.

2.4 | Heatmap presentation

The heatmaps were generated using "pheatmap" package (version 1.0.12) in R statistics software. The "average" method determined the clustering scale and "correlation" method determined the clustering distance.

2.5 | Venn diagram

The Venn diagram was generated using Wonderful Venn in TBtools software (version x32_1_064).[47]

2.6 | ROC analysis

The ROC curves were plotted by "pROC" package (version 1.16.2) in R statistics software. The area under the ROC curve (AUC) was also calculated by "pROC" package.

2.7 | Biological process enrichment analysis

The enriched biological process was identified using The Database for Annotation, Visualization, and Integrated Discovery (DAVID) website (version 6.8; https://david.ncifcrf.gov).[48,49] The enriched biological processes with p values < 0.05 were considered to be statistically significant.

2.8 | Statistical analysis

The box plots were generated from GraphPad Prism software (version 5.0). Statistical analysis was performed using the two-tailed paired Student's *t*-test. *P* value <0.05 was chosen to be significantly different.

3 | RESULTS

3.1 | Identification of the molecular sub-consensuses of LUAD by NMF method using TCGA dataset

We designed a working process to study the molecular sub-consensuses of LUAD based on NMF classification (Figure 1). First, 515 LUAD patients were collected from TCGA dataset.[5] Based on the gene transcriptional profiling, those patients were divided into two sub-consensuses using NMF method. In total, 249 LUAD patients were in sub-consensus 1 and 266 LAUD patients were in sub-consensus 2. Consensus map showed the high correlation of LUAD patients in each sub-consensus (Figure 2A). Moreover, we found significantly different overall survival in sub-consensus 1 and sub-consensus 2 LUAD patients. LUAD patients in sub-consensus 2 had lower overall survival (Figure 2B). Furthermore, LUAD patients in sub-consensus 2 were with higher TP53 mutations (Figure 2C). However, the KRAS, EGFR, or BRAF mutations in sub-consensus 1 and sub-consensus 2 were not significantly different (Figure 2C).

One advantage of NMF classification is that the number of sub-consensuses can be easily determined by the number of ranks.[34] We further divided the 515 LUAD patients into three or four sub-consensuses using NMF method (Figure 2A). Overall survival was significantly different in different sub-consensuses of LUAD patients (Figure 2B). LUAD patients in sub-consensus 3 had the lowest overall survival. TP53, KRAS, and EGFR mutations, but not BRAF mutations, were significantly higher in sub-consensus 3 of LUAD patients (Figure 2D). Also, in the four sub-consensuses classification, compared with other sub-consensuses, LUAD patients in sub-consensus 4 had the lowest overall survival (Figure 2B). Those results suggested that, using NMF method, LUAD patients from TCGA dataset could be divided into different sub-consensuses with different molecular characteristics and clinical overall survival.

3.2 | Validation of the sub-consensuses classification of LUAD using five independent GEO datasets

Using the expression datasets deposited in GEO website, we further validated the sub-consensuses NMF

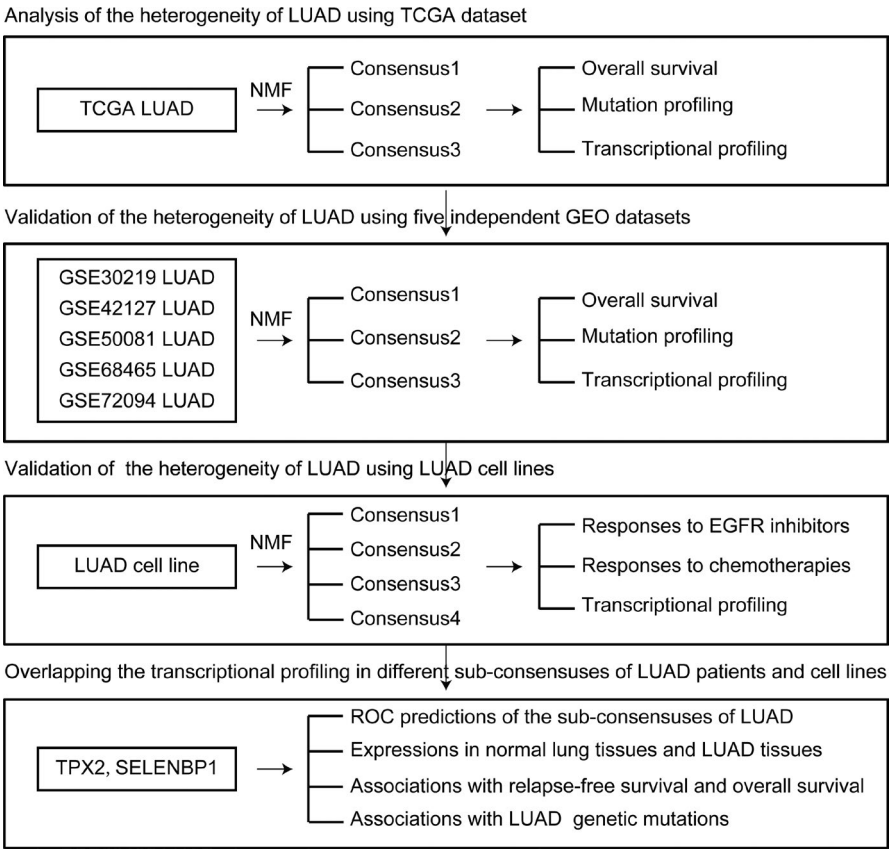


FIGURE 1 A working process to study the molecular sub-consensuses of LUAD based on NMF classification

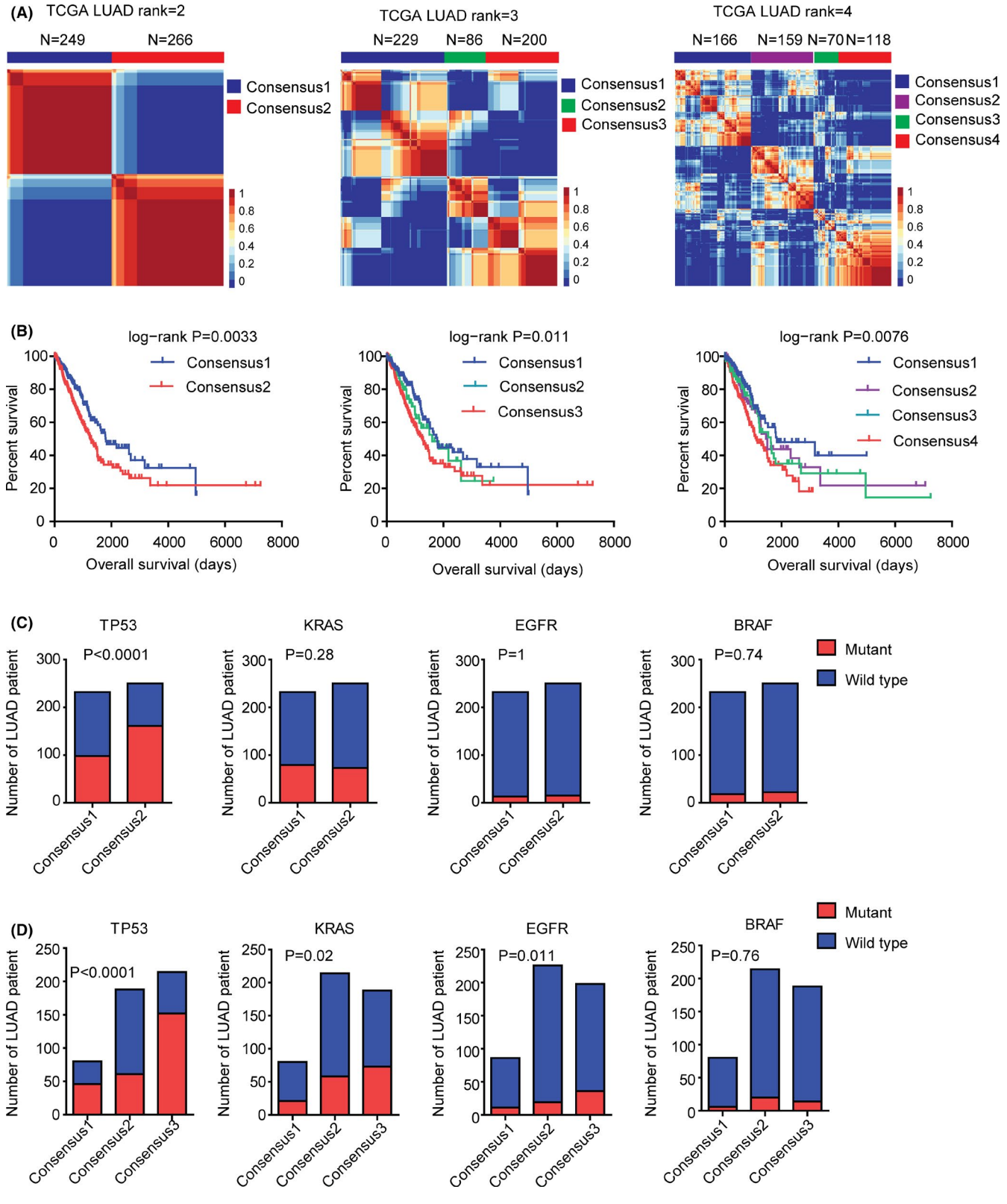
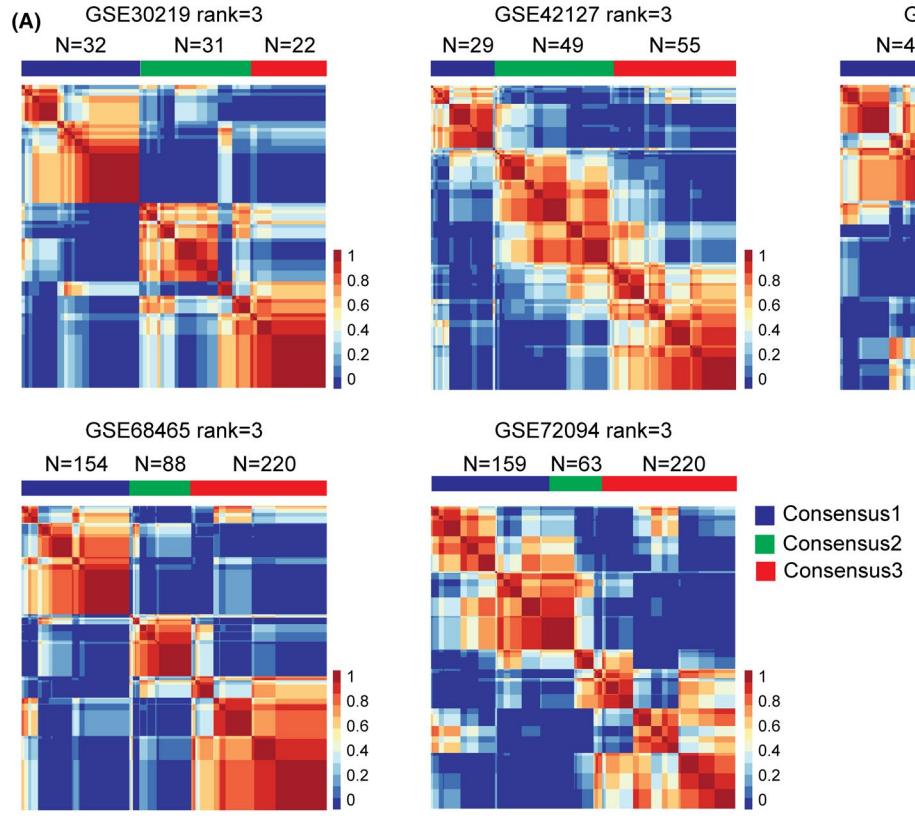


FIGURE 2 Identification of the molecular sub-consensuses of LUAD by NMF method using TCGA dataset. (A) Primary LUAD patients from TCGA dataset were divided into two, three, or four sub-consensuses based on the gene transcriptional profiling using NMF method. Consensus maps showed the correlation profiling of LUAD derived from two sub-consensuses, three sub-consensuses, or four sub-consensuses. (B) The Kaplan–Meier survival analysis was used to determine the overall survival of LUAD patients in each sub-consensus derived from two, three, or four sub-consensuses classification. The overall survival *p* values were calculated by the log-rank test. (C) Contingency graphs showed the number of LUAD patients with TP53, KRAS, EGFR, or BRAF mutations in each sub-consensus derived from the two sub-consensuses classification. *p* values were determined using the Chi-squared test. (D) Contingency graphs showed the number of LUAD patients with TP53, KRAS, EGFR, or BRAF mutations in each sub-consensus derived from the three sub-consensuses classification


(A)
GSE30219 rank=3
N=32 N=31 N=22
GSE42127 rank=3
N=29 N=49 N=55
GSE50081 rank=3
N=46 N=42 N=39
GSE68465 rank=3
N=154 N=88 N=220
GSE72094 rank=3
N=159 N=63 N=220
1
0.8
0.6
0.4
0.2
0
Consensus1
Consensus2
Consensus3

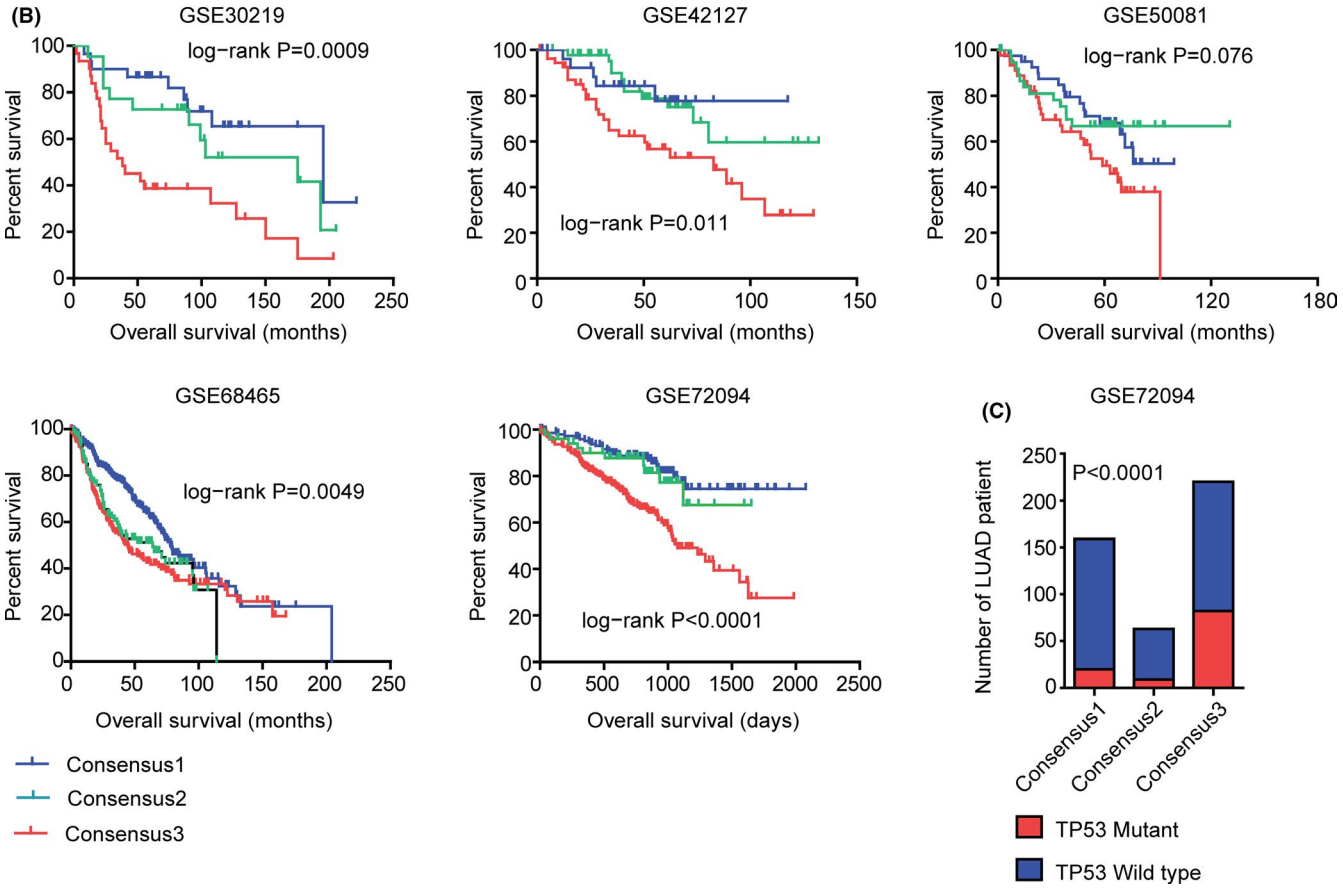

(B)
GSE30219
log−rank P=0.0009
GSE42127
log−rank P=0.011
GSE50081
log−rank P=0.076
GSE68465
log−rank P=0.0049
GSE72094
log−rank P<0.0001
Percent survival
Overall survival (months)
Overall survival (days)
Consensus1
Consensus2
Consensus3
(C)
GSE72094
P<0.0001
Number of LUAD patient
TP53 Mutant
TP53 Wild type

FIGURE 3 Validation of the sub-consensuses classification of LUAD using five independent GEO datasets. (A) Primary LUAD patients from GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets were divided into three sub-consensuses based on the gene expression profiling. (B) The Kaplan–Meier survival analysis was used to determine the different overall survival of sub-consensus 1, sub-consensus 2, and sub-consensus 3 LUAD patients derived from GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. The overall survival *p* values were determined by the log-rank test. (C) Contingency graphs showed the number of LUAD patients with TP53 mutations in each sub-consensus derived from the three sub-consensuses classification in GSE72094 dataset

classification of LUAD. Collectively, 85 LUAD patients from GSE30219,[35] 133 LUAD patients from GSE42127,[36] 127 LUAD patients from GSE50081,[37] 462 LUAD patients from GSE68465,[38] and 442 LUAD patients from GSE72094 datasets[39] were used for further studies. Similarly, using NMF classification, LUAD patients in each GEO dataset were divided into two sub-consensuses (Figure S1A). Then, the overall survival of each sub-consensus of LUAD patients was determined. LUAD patient in sub-consensus 2 had more unfavorable prognosis than LUAD patients in sub-consensus 1 in GSE30219, GSE50081, GSE68465, and GSE72094 datasets (Figure S1B).

Using same strategy, LUAD patients derived from GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets were divided into three sub-consensuses, as illustrated in the consensus maps (Figure 3A). Similar to the results derived from TCGA dataset, the Kaplan–Meier survival analysis showed that LUAD patient in sub-consensus 3 had the worst overall survival than LUAD patients in other sub-consensuses in GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets (Figure 3B). Furthermore, LUAD patients in sub-consensus 3 were with higher TP53 mutations in GSE72094 dataset (Figure 3C).

So, by integrated analysis of total 1765 LUAD patients from TCGA and GEO datasets, we concluded that three sub-consensuses of NMF method was a robust classification method of LUAD.

3.3 | Validation of the sub-consensuses classification of LUAD using LUAD cell lines

Primary LUAD tissues include LUAD tumor cells, tumor stroma cells, and infiltrated immune cells. The tissue complexity may influence the classification of LUAD. On the contrary, LUAD cell lines are relatively homogenous, and the expression profiling of cell lines may represent the intrinsic sub-consensus of LUAD. So, we used LUAD cell lines to validate the three sub-consensuses classification of primary LUAD patients.

Gene expression data and drug responses of 64 LUAD cell lines were downloaded from Genomics of Drug Sensitivity in Cancer (GDSC) project.[46] First, the 64 LUAD cell lines were divided into two (Figure S2A), three, or four sub-consensuses based on the gene expression profiling using NMF method (Figure 4A). Then, we tested the cell responses to EGFR inhibitors and chemotherapeutic drugs in different LUAD sub-consensuses. When the LUAD cell lines were classified into two sub-consensuses, cells in sub-consensus 1 were more sensitive to EGFR inhibitors afatinib and gefitinib (Figure S2B). However, there was no significant difference in the cetuximab and pelitinib sensitivity between sub-consensus 1 and sub-consensus 2 LUAD cells (Figure S2B).

We then divided the LUAD cell lines into three or four sub-consensuses. However, the number of cell lines of sub-consensus 1 in the three sub-consensuses classification and sub-consensus 3 in the four sub-consensuses classification was very low (Figure 4A). Therefore, only the sub-consensus 1, 2, and 4 in the four sub-consensuses classification were further studied. LUAD cells in sub-consensus 2 were more resistant to EGFR inhibitors afatinib, cetuximab, gefitinib, and pelitinib treatment compared with LUAD cells in sub-consensus 1 and sub-consensus 4 (Figure 4B). However, there was no significant difference in the 5-fluorouracil, irinotecan, docetaxel, and etoposide sensitivity in different LUAD sub-consensuses (Figure 4C).

3.4 | Transcriptional characteristics of the sub-consensus 3 LUAD

In all the TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets, the sub-consensus 3 LUAD patients had unfavorable prognosis, while, the sub-consensus 1 LUAD patients had favorable prognosis. Based on the absolute fold changes >2 and *p* values <0.001 criterion, the differentially expressed genes between sub-consensus 1 and sub-consensus 3 were determined. As demonstrated in the clustering heatmaps, 1903 genes were differentially expressed in TCGA dataset, 490 genes were differentially expressed in GSE30219 dataset, 1466 genes were differentially expressed in GSE42127 dataset, 528 genes were differentially expressed in GSE50081 dataset, 153 genes were differentially expressed in GSE68465 dataset, and 727 genes were differentially expressed in GSE72094 dataset, respectively (Figure 5A).

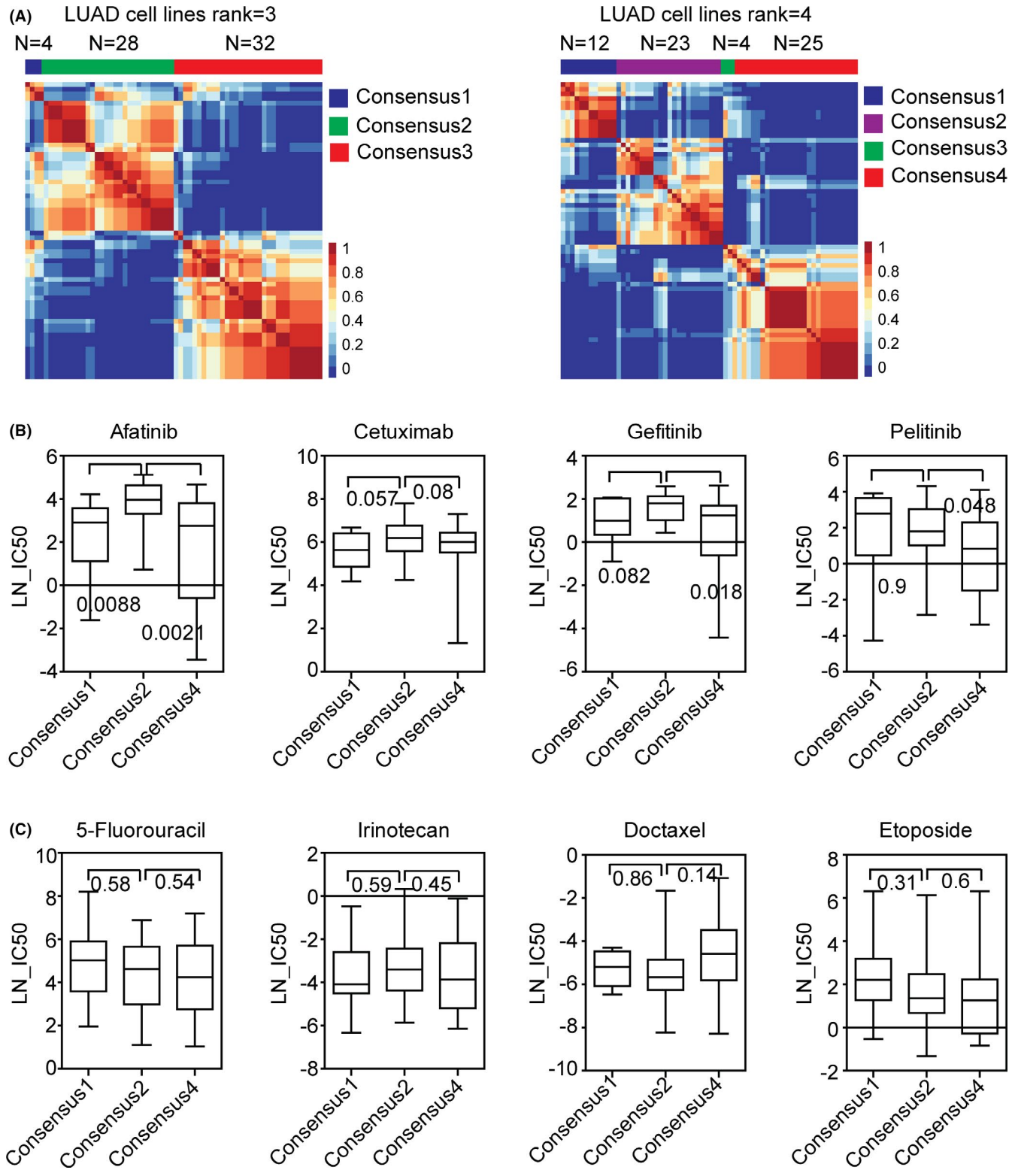


FIGURE 4 Validation of the sub-consensuses classification of LUAD using LUAD cell lines. (A) LUAD cell lines were divided into three or four sub-consensuses based on the gene expression profiling using NMF. Consensus maps showed the correlation profiling of LUAD cell lines from three sub-consensuses or four sub-consensuses. (B) Box plots showed the LN-IC50 of EGFR inhibitors afatinib, cetuximab, gefitinib, and pelitinib in each LUAD sub-consensus of cell lines derived from the four sub-consensuses classification. *P* values were generated by two-tailed paired Student's *t*-test. (C) Box plots showed the LN-IC50 of chemotherapeutic drugs 5-fluorouracil, irinotecan, docetaxel, and etoposide in each LUAD sub-consensus of cell lines derived from the four sub-consensuses classification

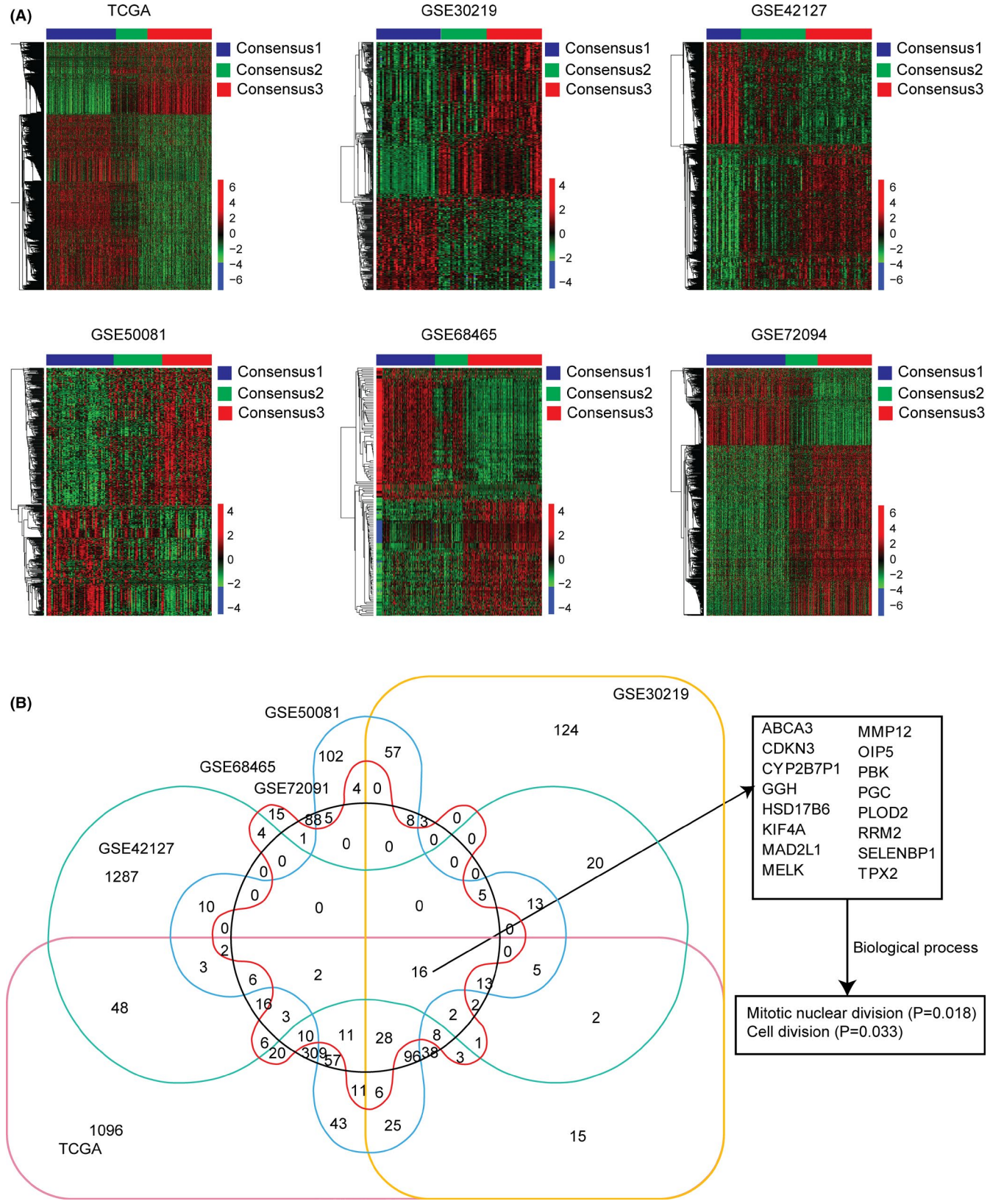


FIGURE 5 Transcriptional characteristics of the sub-consensus 3 LUAD. (A) Clustering heatmaps showed the differentially expressed genes between sub-consensus 1 and sub-consensus 3 LUAD patients in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. Upregulated (red) and downregulated (blue) genes in sub-consensus 3 LUAD patients are shown. (B) Venn diagram showed the common genes which were differentially expressed between sub-consensus 1 and sub-consensus 3 LUAD patients in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. The enriched biological processes of the 16 commonly and differentially expressed genes were determined

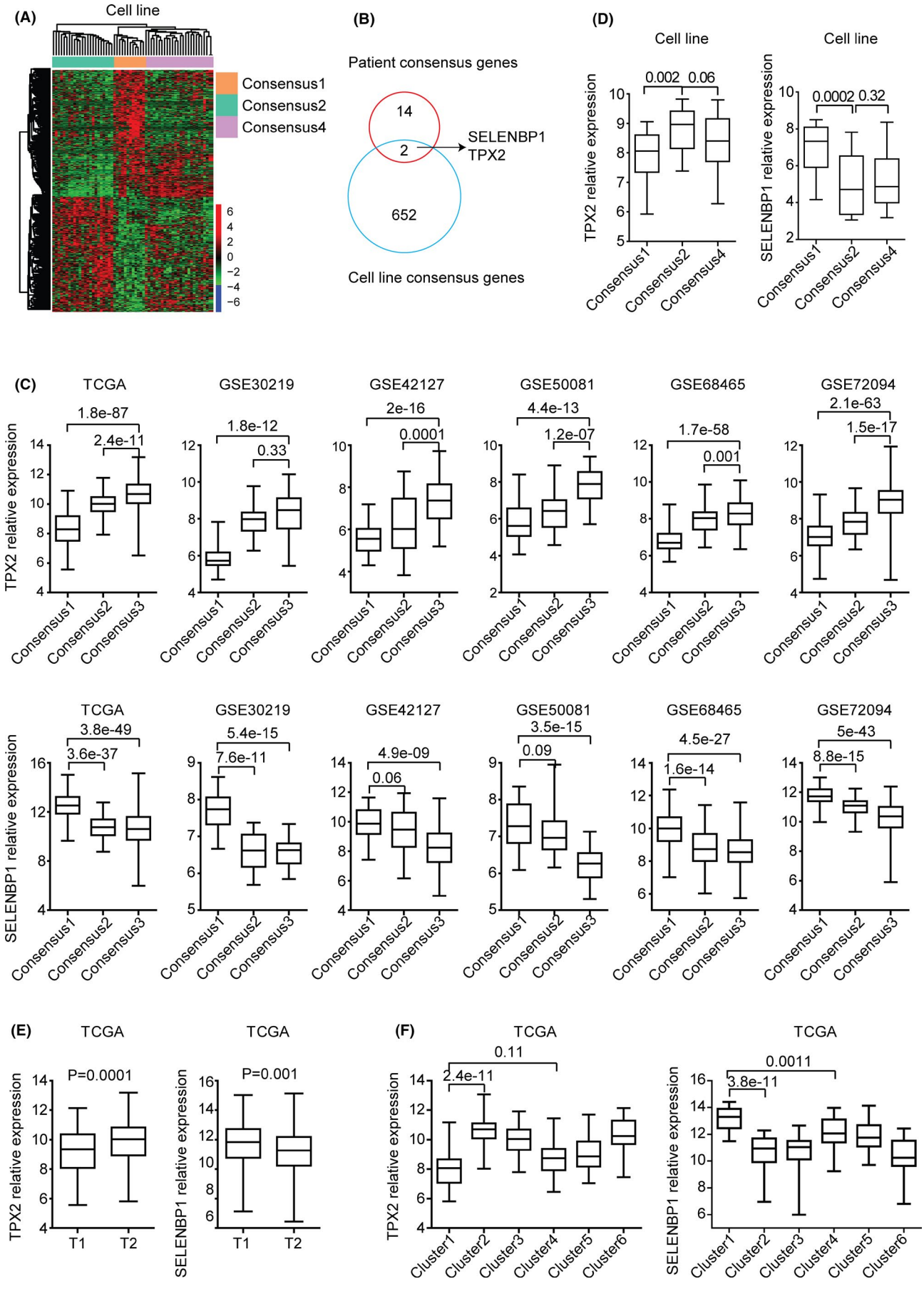

(A)
Cell line
Consensus1
Consensus2
Consensus4
(B)
Patient consensus genes
14
2
652
SELENBP1
TPX2
Cell line consensus genes
(D)
Cell line
TPX2 relative expression
0.002 0.06
Consensus1 Consensus2 Consensus4
Cell line
SELENBP1 relative expression
0.0002 0.32
Consensus1 Consensus2 Consensus4
(C)
TCGA
GSE30219
GSE42127
GSE50081
GSE68465
GSE72094
TPX2 relative expression
1.8e-87
2.4e-11
1.8e-12
0.33
2e-16
0.0001
4.4e-13
1.2e-07
1.7e-58
0.001
2.1e-63
1.5e-17
SELENBP1 relative expression
3.8e-49
3.6e-37
5.4e-15
7.6e-11
4.9e-09
0.06
3.5e-15
0.09
4.5e-27
1.6e-14
5e-43
8.8e-15
Consensus1 Consensus2 Consensus3
(E)
TCGA
TCGA
TPX2 relative expression
P=0.0001
T1 T2
SELENBP1 relative expression
P=0.001
T1 T2
(F)
TCGA
TCGA
TPX2 relative expression
0.11
2.4e-11
SELENBP1 relative expression
0.0011
3.8e-11
Cluster1 Cluster2 Cluster3 Cluster4 Cluster5 Cluster6

FIGURE 6 TPX2 and SELENBP1 are differentially expressed in different LUAD sub-consensuses. (A) Un-supervised clustering heatmap demonstrated the differentially expressed genes in sub-consensus 2 LUAD cell lines. (B) Venn diagram depicted the common genes associated with the different sub-consensuses of LUAD patients and cell lines. (C) Box plots showed the TPX2 and SELENBP2 expression levels in each sub-consensus of LUAD patients derived from TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. *p* values were generated using the two-tailed paired Student's *t*-test. (D) Box plots showed the TPX2 and SELENBP2 expression levels in each sub-consensus of LUAD cell lines. (E) Box plots showed the TPX2 and SELENBP2 expression levels in T1 and T2 stage of TCGA LUAD patients. (F) Box plots showed the expression levels of TPX2 and SELENBP2 in each cluster based on the TCGA LUAD iCluster classification

Next, using overlapping analysis, we tried to identify the common genes which were differentially expressed in sub-consensus 3 LUAD patients. As shown in the Venn diagram, 16 genes were differentially expressed between sub-consensus 1 and sub-consensus 3 LUAD patients in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets (Figure 5B). The expression levels of those 16 genes were further demonstrated in heatmaps (Figure S3). In TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets, MMP12, PLOD2, GGH, OIP5, PBK, RRM2, MELK, KIF4A, TPX2, CDKN3, and MAD2L1 were all upregulated in sub-consensus 3 LUAD patients, while, HSD17B6, CYP2B7P1, SELENBP1, ABCA3, and PGC were all downregulated in sub-consensus 3 LUAD patients. Biological process enrichment analysis showed that those 16 genes were associated with mitotic nuclear division and cell division processes (Figure 5B).

3.5 | TPX2 and SELENBP1 are differentially expressed in different LUAD sub-consensuses

Genes differentially expressed in sub-consensus 2 LUAD cell lines were further identified. Based on the *p* values <0.001 criterion, 654 genes were differentially expressed in sub-consensus 2 LUAD cell lines, compared with sub-consensus 1 LUAD cell lines (Figure 6A). Further overlapping with the genes associated with different sub-consensuses of LUAD patients, we found that TPX2 and SELENBP1 were two potential predictors of the inner sub-consensuses of primary LUAD patients and LUAD cell lines (Figure 6B).

TPX2 is a microtubule-associated gene and is critical to the spindle formation during cell cycle progress.[50,51] TPX2 was highly expressed in sub-consensus 3 of LUAD patients in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets (Figure 6C). On the contrary, SELENBP1 was highly expressed in sub-consensus 1 of LUAD patients (Figure 6C). Moreover, in the EGFR inhibitors resistant sub-consensus 2 LUAD cell lines, TPX2 was highly expressed and SELENBP1 was lowly expressed (Figure 6D). Furthermore, compared with T1 stage, the expression levels of TPX2 were relatively higher in T2 stage. While, the expression levels of SELENBP1 were relatively lower in T2 stage of LUAD patients (Figure 6E). Previously, TCGA LUAD research groups used iCluster analysis and divided the TCGA LUAD patients into six clusters.[5] TPX2 was highly expressed and SELENBP1 was lowly expressed in cluster 1 LUAD patients (Figure 5F).

3.6 | TPX2 and SELENBP1 are two predictors of the sub-consensuses of LUAD

Furthermore, we attempted to determine the predictive values of TPX2 and SELENBP in distinguishing the sub-consensuses of LUAD. The ROC analysis in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets indicated that the expression levels of TPX2 could distinguish sub-consensus 3 from sub-consensus 1 LUAD patients with high specificity and sensitivity (Figure 7A). Similar predictive specificity and sensitivity of SELENBP1 in distinguishing sub-consensus 3 from sub-consensus 1 LUAD patients were observed in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets (Figure 7A). Furthermore, the expression levels of TPX2 or SELENBP1 also distinguished sub-consensus 2 from sub-consensus 1 of LUAD cell lines with high accuracy (Figure 7B).

Importantly, the expression levels of TPX2 or SELENBP1 had robust predictive values in the iCluster classification. In TCGA LUAD cohort, ROC curves showed the similar specificity and sensitivity of TPX2 or SELENBP1 to distinguish cluster 1 from cluster 2 TCGA LUAD patients (Figure 7C). All those results highlighted the sub-types predictive values of TPX2 and SELENBP1 in LUAD.

3.7 | TPX2 is upregulated in LUAD while SELENBP1 is downregulated in LUAD

Next, we analyzed the expression levels of TPX2 and SELENBP1 in normal lung tissues and LUAD tissues. Gene expression profiling of 261 normal lung tissues and 271 LUAD tissues was downloaded from TCGA,[5] GSE7670,[40] GSE10072,[41] GSE18842,[42] GSE27262,[43] and GSE32863[44] datasets. In all those six datasets, TPX2

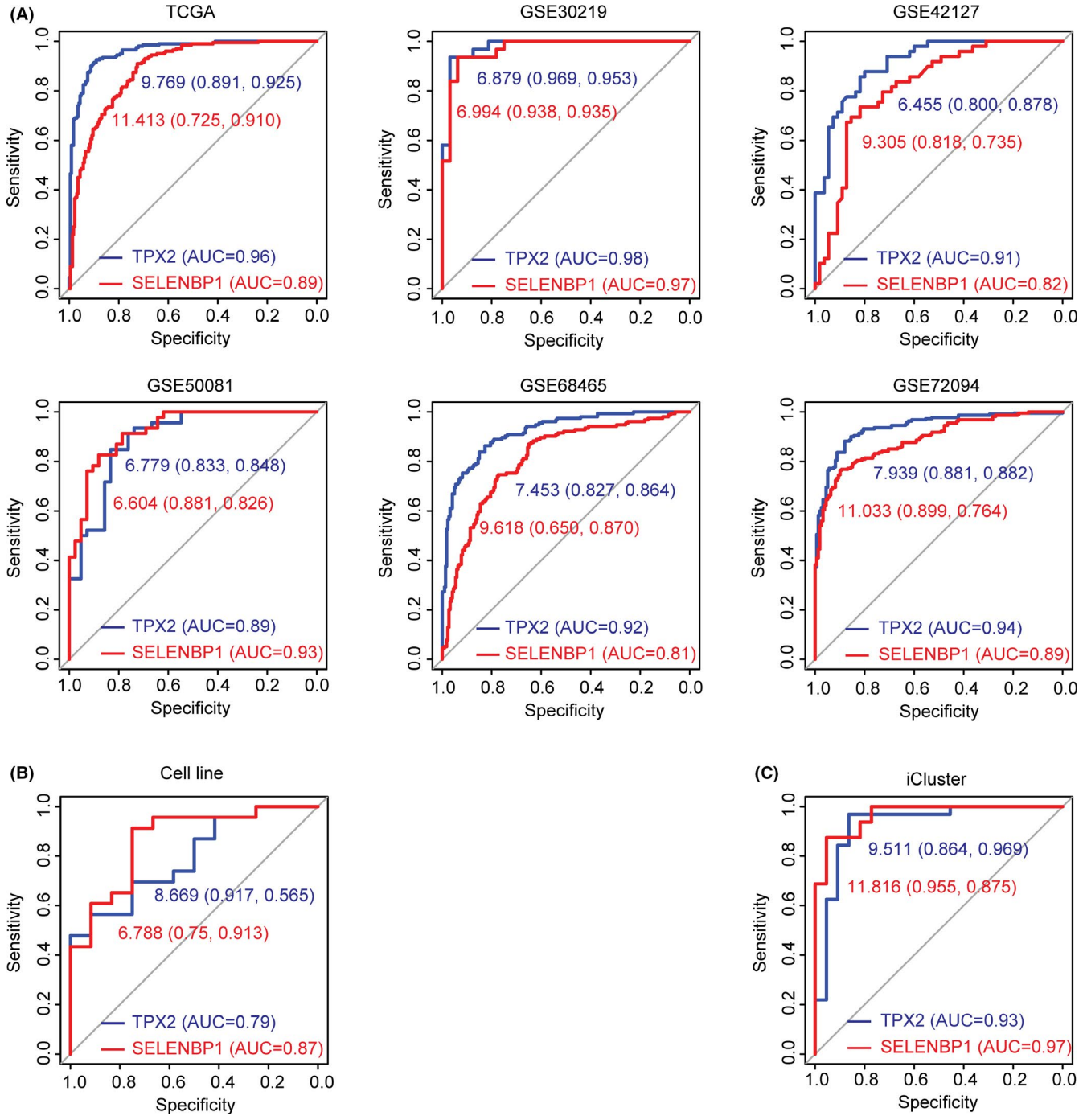


FIGURE 7 TPX2 and SELENBP1 are two predictors of the sub-consensuses of LUAD. (A) ROC curves showed the predictive specificity and sensitivity of TPX2 or SELENBP1 to distinguish sub-consensus 3 from sub-consensus 1 LUAD patients derived from TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. (B) ROC curve showed the predictive accuracy of TPX2 or SELENBP1 to distinguish sub-consensus 2 from sub-consensus 1 LUAD cell lines. (C) ROC curve showed the predictive values of TPX2 or SELENBP1 in distinguishing cluster 1 from cluster 2 LUAD patients based on the TCGA LUAD iCluster classification. AUC, area under the ROC curve

was overexpressed in LUAD tissues, compared with normal lung tissues (Figure 8A). On the contrary, SELENBP1 was significantly downregulated in LUAD tissues in TCGA, GSE7670, GSE10072, GSE18842, and GSE32863 datasets (Figure 8B). Moreover, in GSE30219 dataset, TPX2 was upregulated in LUAD tissues, while SELENBP1 was downregulated in LUAD tissues (Figure 8A,B).

Furthermore, TPX2 and SELENBP1 distinguished LUAD tissues from normal lung tissues with high accuracy. The ROC curve analysis demonstrated significant predictive values of TPX2 in TCGA, GSE10072, and

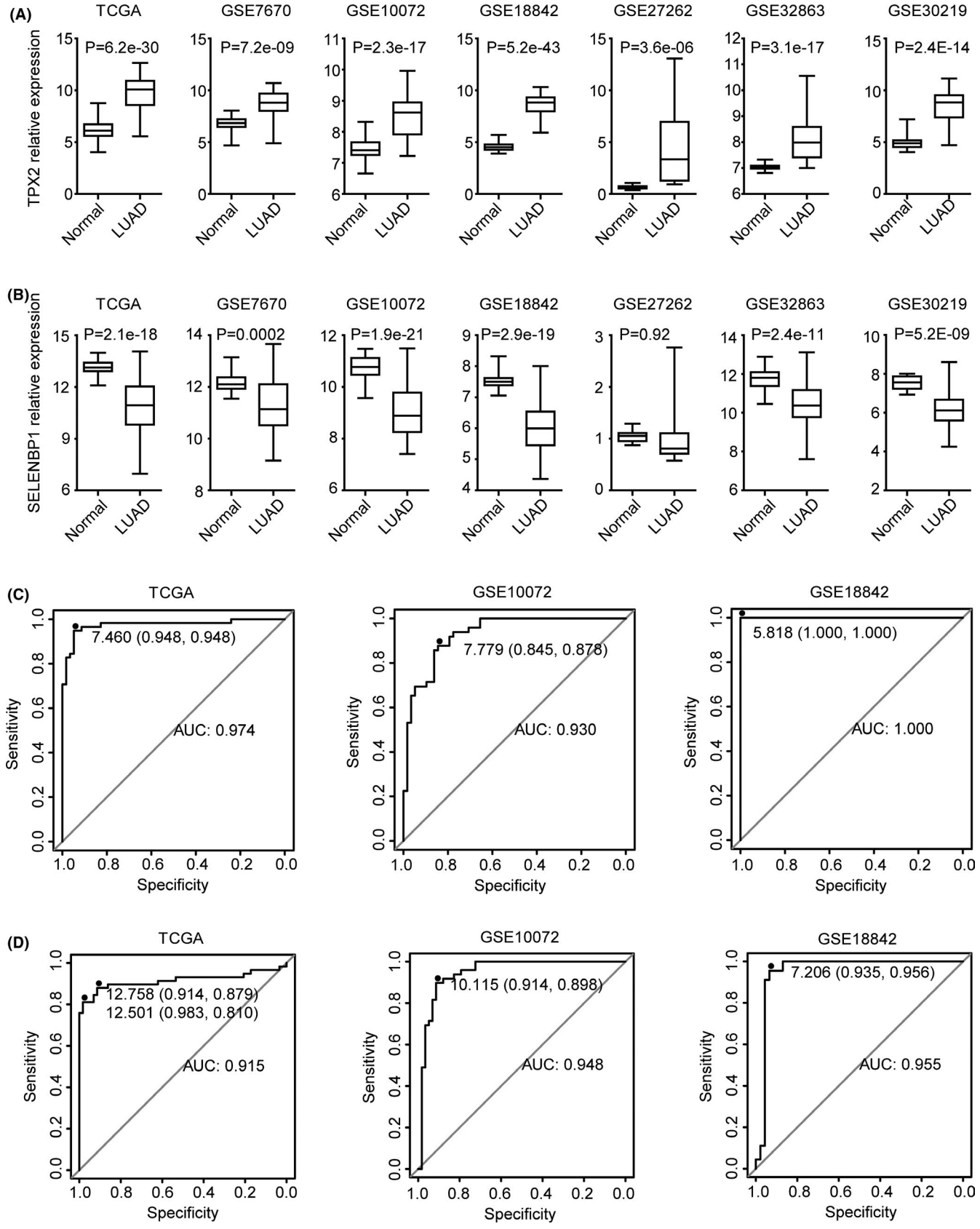


FIGURE 8 TPX2 is upregulated in LUAD while SELENBP1 is downregulated in LUAD. (A–B) Box plots demonstrated the different expression levels of TPX2 (A) and SELENBP1 (B) in normal lung tissues and LUAD tissues in TCGA, GSE7670, GSE10072, GSE18842, GSE27262, GSE32863, and GSE30219 datasets. *p* values were generated using the two-tailed paired Student's *t*-test. (C–D) ROC curves showed the predictive accuracy of TPX2 (C) or SELENBP1 (D) to distinguish LUAD tissues from normal lung tissues in TCGA, GSE10072, and GSE18842 datasets

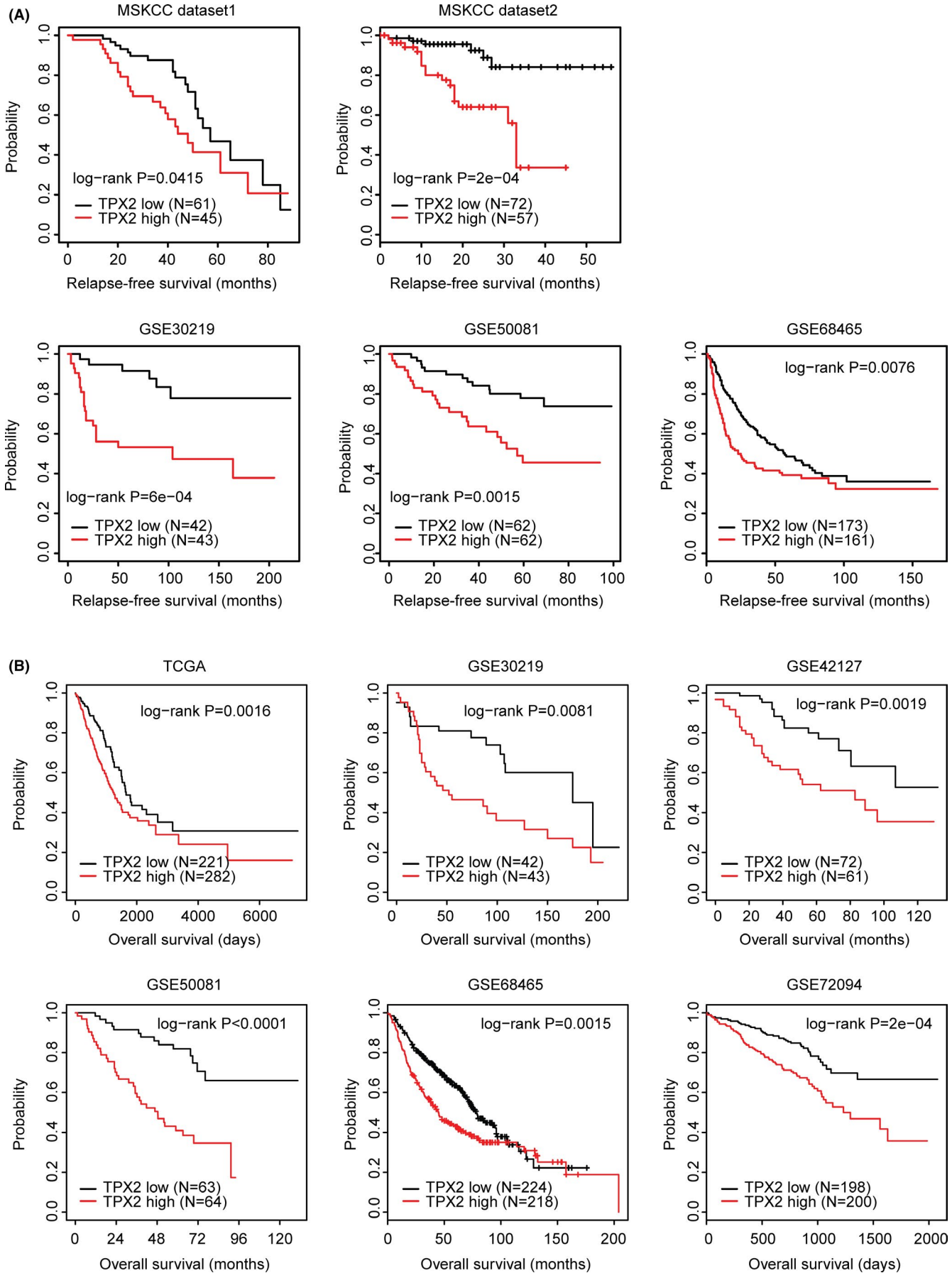

(A)
MSKCC dataset1
log−rank P=0.0415
TPX2 low (N=61)
TPX2 high (N=45)
Probability
Relapse-free survival (months)
MSKCC dataset2
log−rank P=2e−04
TPX2 low (N=72)
TPX2 high (N=57)
Probability
Relapse-free survival (months)
GSE30219
log−rank P=6e−04
TPX2 low (N=42)
TPX2 high (N=43)
Probability
Relapse-free survival (months)
GSE50081
log−rank P=0.0015
TPX2 low (N=62)
TPX2 high (N=62)
Probability
Relapse-free survival (months)
GSE68465
log−rank P=0.0076
TPX2 low (N=173)
TPX2 high (N=161)
Probability
Relapse-free survival (months)
(B)
TCGA
log−rank P=0.0016
TPX2 low (N=221)
TPX2 high (N=282)
Probability
Overall survival (days)
GSE30219
log−rank P=0.0081
TPX2 low (N=42)
TPX2 high (N=43)
Probability
Overall survival (months)
GSE42127
log−rank P=0.0019
TPX2 low (N=72)
TPX2 high (N=61)
Probability
Overall survival (months)
GSE50081
log−rank P<0.0001
TPX2 low (N=63)
TPX2 high (N=64)
Probability
Overall survival (months)
GSE68465
log−rank P=0.0015
TPX2 low (N=224)
TPX2 high (N=218)
Probability
Overall survival (months)
GSE72094
log−rank P=2e−04
TPX2 low (N=198)
TPX2 high (N=200)
Probability
Overall survival (days)

FIGURE 9 Higher expression levels of TPX2 are associated with the lower relapse-free survival and the lower overall survival of LUAD. (A) The Kaplan–Meier Plotters demonstrated the associations between TPX2 and the LUAD relapse-free survival in MSKCC dataset 1, MSKCC dataset 2, GSE30219, GSE50081, and GSE68465 datasets. The *p* values showed the different relapse-free survival between TPX2 highly expressed LUAD patients (red) and TPX2 lowly expressed LUAD patients (black). (B) The Kaplan–Meier Plotters demonstrated the different overall survival of TPX2 highly expressed LUAD patients (red) and TPX2 lowly expressed LUAD patients (black) in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets

GSE18842 datasets (Figure 8C). Particularly in GSE18842 dataset, the distinguishing of LUAD tissues from normal lung tissues through TPX2 expression was completely accurate (AUC=1). Moreover, high predictive specificity and sensitivity of SELENBP1 in TCGA, GSE10072, and GSE18842 datasets were also observed (Figure 8D).

3.8 | Higher expression levels of TPX2 are associated with the lower relapse-free survival and the lower overall survival of LUAD

Then, we determined the prognostic effects of TPX2 in LUAD relapse-free survival. First, in two MSKCC datasets,[45] higher expression levels of TPX2 were correlated with lower relapse-free survival in patients with LUAD (Figure 9A). Furthermore, the unfavorable prognosis of TPX2 in LUAD was validated in GSE30219, GSE50081, and GSE68465 datasets. The Kaplan–Meier Plotters demonstrated that TPX2 highly expressed LUAD patients had lower relapse-free survival than TPX2 lowly expressed LUAD patients (Figure 9A).

Previously, we showed that TPX2 was highly expressed in sub-consensus 3 LUAD patients, which had low overall survival in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. Consistent with those observations, in all those six datasets, TPX2 highly expressed LUAD patients resulted lower overall survival than TPX2 lowly expressed LUAD patients (Figure 9B), suggesting the importance of TPX2 as a negative marker in the clinical outcome prediction of LUAD.

3.9 | Lower expression levels of SELENBP1 are associated with the lower relapse-free survival and the lower overall survival of LUAD

Unlike TPX2, SELENBP1 may serve as a positive marker in the clinical outcome prediction of LUAD. First, as we previously showed, compared with the normal lung tissues, SELENBP1 was downregulated in LUAD tissues (Figure 8B). Second, LUAD patients with higher expression levels of SELENBP1 were associated with better relapse-free survival in GSE30219, GSE50081, and GSE68465 datasets (Figure 10A). Less significantly, SELENBP1 highly expressed LUAD patients also had higher relapse-free survival than SELENBP1 lowly expressed LUAD patients in MSKCC1 and MSKCC2 datasets (Figure 10A).

Third, in TCGA, GSE30219, GSE50081, GSE68465, and GSE72094 datasets, significantly different overall survival between SELENBP1 highly expressed LUAD patients and SELENBP1 lowly expressed LUAD patients was demonstrated (Figure 10B). All those results highlighted the prognostic effects of SELENBP1 in patients with LUAD.

3.10 | Expression levels of TPX2 and SELENBP1 are correlated with TP53 mutations

Previously, we showed the different TP53, KRAS, and EGFR mutations in different sub-consensuses of LUAD patients (Figure 2D). So, we detected the expression levels of TPX2 and SELENBP1 in TP53, KRAS, or EGFR-mutant LUAD patients and TP53, KRAS, or EGFR wild-type LUAD patients derived from TCGA dataset. Compared with TP53 wild-type LUAD patients, TPX2 was overexpressed in TP53-mutant LUAD patients (Figure 11A). However, SELENBP1 was downregulated in TP53-mutant LUAD patients (Figure 11A). Moreover, the upregulations of TPX2 and downregulations of SELENBP1 in TP53-mutant LUAD patients were confirmed in GSE72094 dataset (Figure 11B). Furthermore, in both TCGA and GSE72094 datasets, TPX2 was downregulated, while, SELENBP1 was upregulated in EGFR-mutant LUAD patients (Figure 11A,B). However, there was no significant difference in TPX2 and SELENBP1 expression levels between KRAS-mutant and wild-type LUAD patients (Figure 11A,B). Next, we tried to determine if TPX2 or SELENBP1 could also distinguish TP53-mutant from TP53 wild-type LUAD patients. The ROC analysis showed significant AUC values of TPX2 or SELENBP1 in the prediction of TP53 status in TCGA and GSE72094 datasets (Figure 11C).

4 | DISCUSSION

Heterogeneity poses great challenges in cancer prognosis and treatment. With the advancements in gene microarray and RNA-Seq technologies, classification

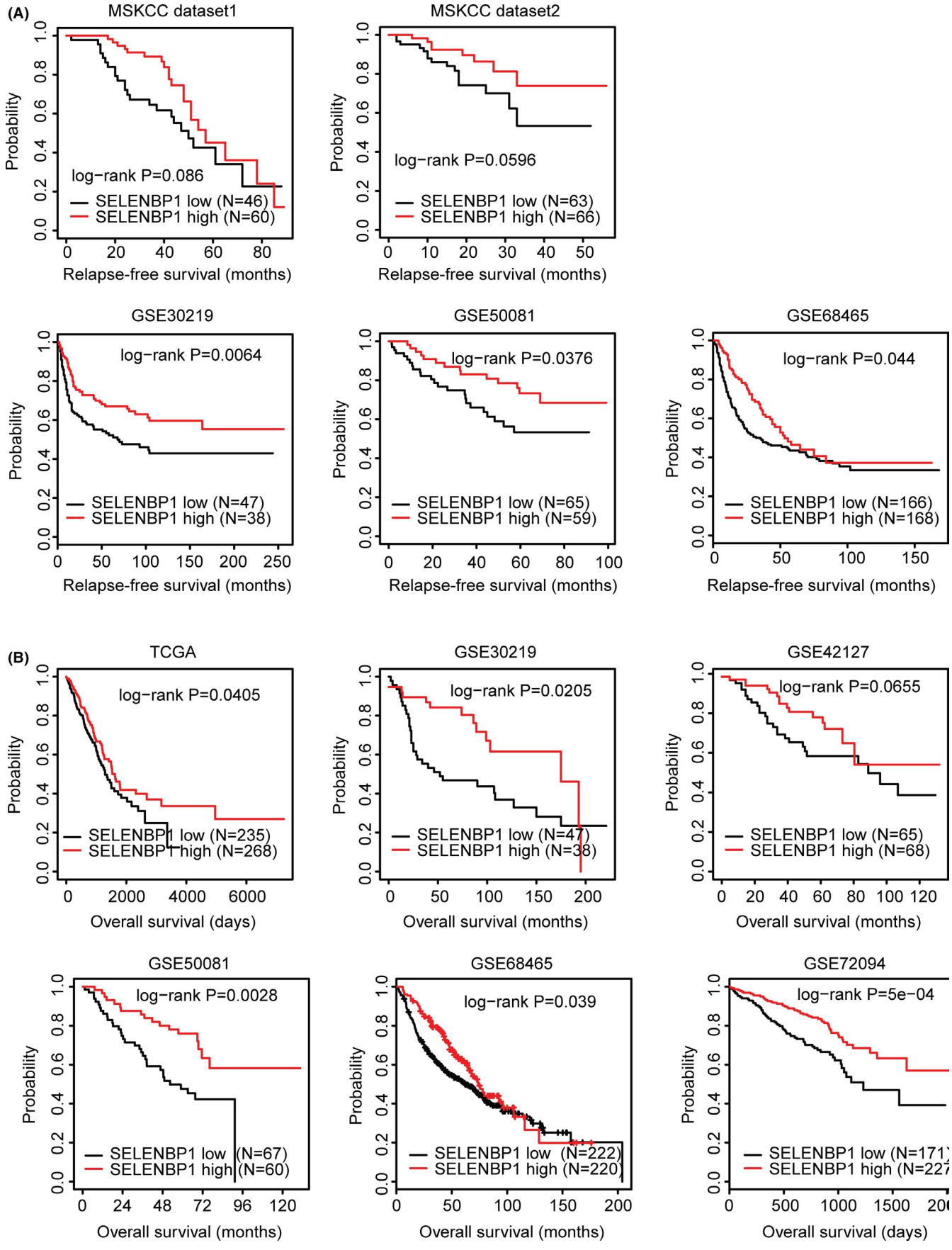

(A)
MSKCC dataset1
log−rank P=0.086
SELENBP1 low (N=46)
SELENBP1 high (N=60)
Probability
Relapse-free survival (months)
MSKCC dataset2
log−rank P=0.0596
SELENBP1 low (N=63)
SELENBP1 high (N=66)
Probability
Relapse-free survival (months)
GSE30219
log−rank P=0.0064
SELENBP1 low (N=47)
SELENBP1 high (N=38)
Probability
Relapse-free survival (months)
GSE50081
log−rank P=0.0376
SELENBP1 low (N=65)
SELENBP1 high (N=59)
Probability
Relapse-free survival (months)
GSE68465
log−rank P=0.044
SELENBP1 low (N=166)
SELENBP1 high (N=168)
Probability
Relapse-free survival (months)
(B)
TCGA
log−rank P=0.0405
SELENBP1 low (N=235)
SELENBP1 high (N=268)
Probability
Overall survival (days)
GSE30219
log−rank P=0.0205
SELENBP1 low (N=47)
SELENBP1 high (N=38)
Probability
Overall survival (months)
GSE42127
log−rank P=0.0655
SELENBP1 low (N=65)
SELENBP1 high (N=68)
Probability
Overall survival (months)
GSE50081
log−rank P=0.0028
SELENBP1 low (N=67)
SELENBP1 high (N=60)
Probability
Overall survival (months)
GSE68465
log−rank P=0.039
SELENBP1 low (N=222)
SELENBP1 high (N=220)
Probability
Overall survival (months)
GSE72094
log−rank P=5e−04
SELENBP1 low (N=171)
SELENBP1 high (N=227)
Probability
Overall survival (days)

FIGURE 10 Lower expression levels of SELENBP1 are associated with the lower relapse-free survival and the lower overall survival of LUAD. (A) The Kaplan–Meier Plotters demonstrated the associations between SELENBP1 and the LUAD relapse-free survival in MSKCC dataset 1, MSKCC dataset 2, GSE30219, GSE50081, and GSE68465 datasets. The *p* values showed the different relapse-free survival between SELENBP1 highly expressed LUAD patients (red) and SELENBP1 lowly expressed LUAD patients (black). (B) The Kaplan–Meier Plotters demonstrated the different overall survival of SELENBP1 highly expressed LUAD patients (red) with SELENBP1 lowly expressed LUAD patients (black) in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets

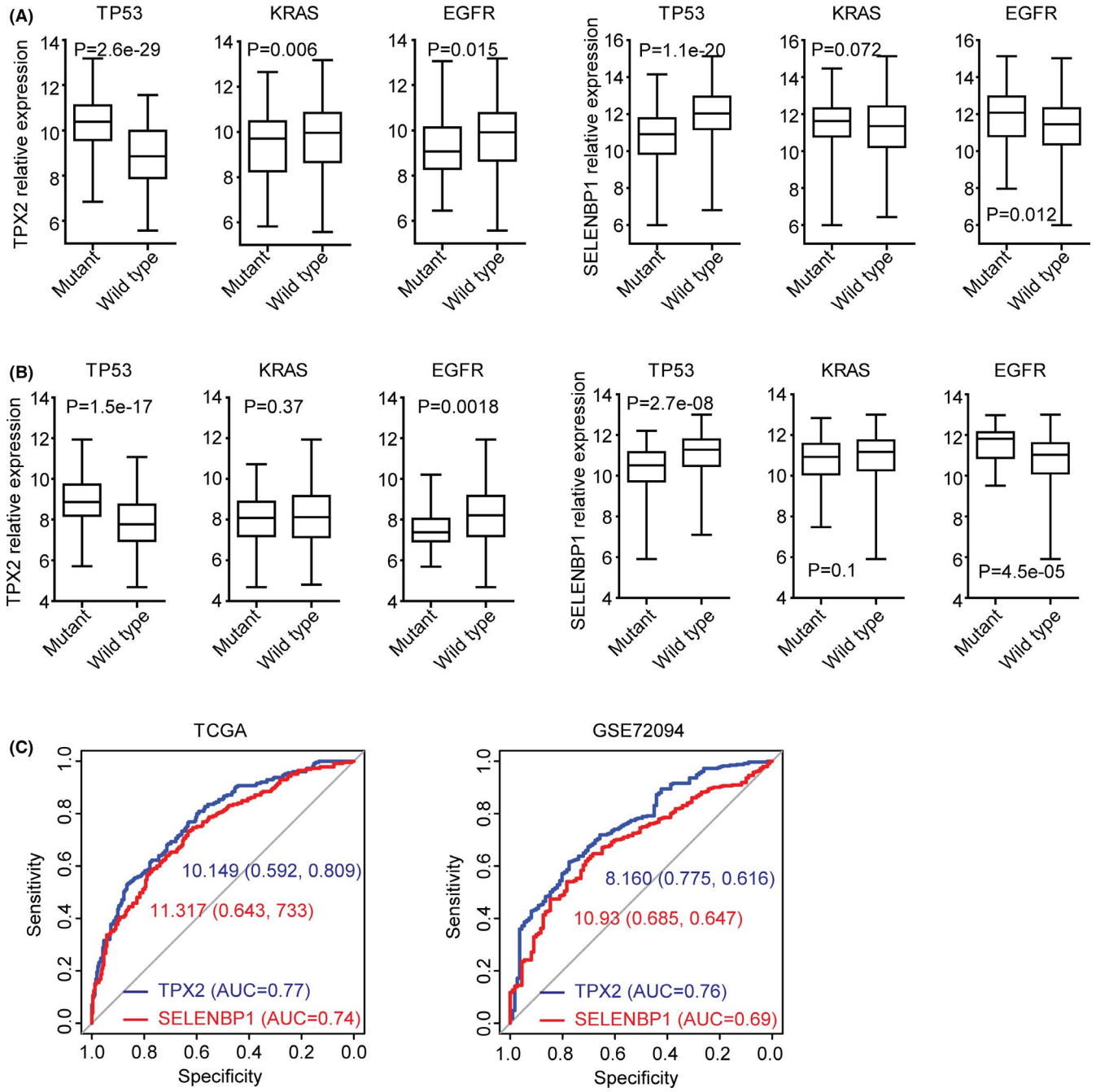


FIGURE 11 The expression levels of TPX2 and SELENBP1 are correlated with TP53 mutations. (A) The expression levels of TPX2 and SELENBP1 in TP53, KRAS, or EGFR-mutant LUAD patients and TP53, KRAS, or EGFR wild-type LUAD patients derived from TCGA dataset. *p* values were generated using the two-tailed paired Student's *t*-test. (B) The expression levels of TPX2 and SELENBP1 in TP53, KRAS, or EGFR-mutant LUAD patients and TP53, KRAS, or EGFR wild-type LUAD patients derived from GSE72094 dataset. (C) ROC curves showed the predictive accuracy of TPX2 or SELENBP1 to distinguish TP53-mutant from TP53 wild-type LUAD in TCGA and GSE72094 datasets

systems based on the transcriptional profiling of cancers are developed. Self-organizing maps (SOMs) clustering,[52,53] integrative iCluster clustering,[5,54,55] network-based consensus molecular classification (CMS),[56,57] and nonnegative matrix factorization (NMF) clustering[26,28,31,34] are all proved to be robust cancer classification systems. However, for the same cohort of tumor patients, distinct clustering methods may result different classification. For example, LUAD patients in TCGA dataset were divided into six clusters by iCluster clustering.[5] In this study, the same LUAD patients in TCGA dataset were classified into three sub-consensuses using NMF classification. Both iCluster and NMF method divided the LUAD patients into different groups with distinct molecular characteristics. However, the LUAD patients in different clusters had no significant clinical overall survival, while, sub-consensus 3 LUAD patients from NMF classification were with lower overall survival than other sub-consensuses.

Different cohorts of patients and gene expression technologies pose another level of complexity and further influence the classification results.[58,59] Through same NMF method, LUAD patients from TCGA, GSE30219, GSE50081, GSE68465, and GSE72094 datasets were divided into two-consensuses with different overall survival. However, in GSE42127 dataset, there was no significantly different overall survival between sub-consensus 1 and sub-consensus 2 LUAD patients. Integrating various independent datasets may increase the statistical robustness. So, in this study, we analyzed 1765 LUAD patients from TCGA and five independent GEO datasets, we found that three sub-consensuses of NMF method was a robust classification method of LUAD. Moreover, TPX2 and SELENBP1 were differentially expressed in the different sub-consensuses of LUAD patients in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. Furthermore, TPX2 and SELENBP1 predicted the inner sub-consensuses of LUAD with high accuracy.

Except transcriptional profiling, the immune cells infiltration signature or tumor immune microenvironment is another important factor influencing the heterogeneity of LUAD[21,22] and making it difficult to interpret the results from different LUAD cohorts. On the contrary, classifications of LUAD cell lines may represent the intrinsic heterogeneity of LUAD tumors. In this study, the NMF method was used in the classification of 64 LUAD cell lines. LUAD cell lines in different sub-consensuses had different responses to EGFR inhibitors. Also, the expression levels of TPX2 and SELENBP1 were different in the different sub-consensuses of LUAD cell lines.

Previous reports showed that, in NSCLC and in some smoking related LUAD patients, TPX2 was associated with poor prognosis.[60-62] In this study, we further showed that, compared with normal lung tissues, TPX2 was overexpressed in LUAD tissues in TCGA, GSE7670, GSE10072, GSE18842, GSE27262, GSE32863, and GSE30219 datasets. And the overexpression of TPX2 was associated with low relapse-free survival and low overall survival in TCGA, GSE30219, GSE42127, GSE50081, GSE68465, and GSE72094 datasets. SELENBP1 was reported as a tumor suppressor gene in many types of tumors.[63–65] The prognosis of SELENBP1 in LUAD was unknown. Our data showed that SELENBP1 was downregulated in LUAD tissues and LAUD patients with higher expression levels of SELENBP1 were associated with better relapse-free survival and overall survival. However, those results were derived from TCGA and GEO datasets, further clinical validations of the prognosis of TPX2 and SELENBP1 were needed. Moreover, the detailed mechanisms of TPX2 and SELENBP1 in LUAD development should be further studied.

5 | CONCLUSIONS

By integrated analysis of 1765 LUAD patients and 64 LUAD cell lines, we showed that NMF was a robust inner sub-consensuses classification method of LUAD. Sub-consensus 3 LUAD patients were with low overall survival and were with high TP53 and KRAS mutations. And sub-consensus 2 LUAD cell lines were resistant to EGFR inhibitors. TPX2 and SELENBP1 were differentially expressed in different LUAD sub-consensuses, and predicted the inner sub-consensuses of LUAD with high accuracy. TPX2 was an unfavorable prognostic biomarker of LUAD which was upregulated in LUAD tissues and associated with the low overall survival of LUAD. SELENBP1 was a favorable prognostic biomarker of LUAD which was downregulated in LUAD tissues and associated with the prolonged overall survival of LUAD.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during the current study are available in TCGA (tcga.xenahubs.net) and GEO (www.ncbi.nlm.nih.gov/geo) repositories.

ORCID

Haiwei Wang https://orcid.org/0000-0002-9675-4039

REFERENCES

1. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. *Lancet Oncol*. 2011;12(2):175-180.
2. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. 2015;10(9):1243-1260.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin*. 2018;68(1):7-30.
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
5. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550.
6. Miller VA, Hirsh V, Cadranel J, et al. Afatinib versus placebo for patients with advanced, metastatic non-small-cell lung cancer after failure of erlotinib, gefitinib, or both, and one or two lines of chemotherapy (LUX-Lung 1): a phase 2b/3 randomised trial. *Lancet Oncol*. 2012;13(5):528-538.
7. Moll HP, Pranz K, Musteanu M, et al. Afatinib restrains K-RAS-driven lung tumorigenesis. *Sci Transl Med*. 2018;10(446).
8. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med*. 2016;375(19):1823-1833.
9. Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med*. 2012;366(26):2443-2454.
10. Wu K, Zhang X, Li F, et al. Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. *Nat Commun*. 2015;6:10131.
11. Jiang J, Gu Y, Liu J, et al. Coexistence of p16/CDKN2A homozygous deletions and activating EGFR mutations in lung adenocarcinoma patients signifies a poor response to EGFR-TKIs. *Lung Cancer*. 2016;102:101-107.
12. Clemenceau A, Gaudreault N, Henry C, et al. Tumor-based gene expression biomarkers to predict survival following curative intent resection for stage I lung adenocarcinoma. *PLoS One*. 2018;13(11):e0207513.
13. Qi L, Li T, Shi G, et al. An individualized gene expression signature for prediction of lung adenocarcinoma metastases. *Mol Oncol*. 2017;11(11):1630-1645.
14. Wang H, Wang X, Xu L, Zhang J, Cao H. High expression levels of pyrimidine metabolic rate-limiting enzymes are adverse prognostic factors in lung adenocarcinoma: a study based on The Cancer Genome Atlas and Gene Expression Omnibus datasets. *Purinergic Signal*. 2020;16(3):347-366.
15. Li X, Shi Y, Yin Z, Xue X, Zhou B. An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J Transl Med*. 2014;12:159.
16. Siriwardhana C, Khadka VS, Chen JJ, Deng Y. Development of a miRNA-seq based prognostic signature in lung adenocarcinoma. *BMC Cancer*. 2019;19(1):34.
17. Sui J, Yang RS, Xu SY, et al. Comprehensive analysis of aberrantly expressed microRNA profiles reveals potential biomarkers of human lung adenocarcinoma progression. *Oncol Rep*. 2017;38(4):2453-2463.
18. Sui J, Yang S, Liu T, et al. Molecular characterization of lung adenocarcinoma: a potential four-long noncoding RNA prognostic signature. *J Cell Biochem*. 2019;120(1):705-714.
19. Zheng S, Zheng D, Dong C, et al. Development of a novel prognostic signature of long non-coding RNAs in lung adenocarcinoma. *J Cancer Res Clin Oncol*. 2017;143(9):1649-1657.
20. Liao M, Liu Q, Li B, Liao W, Xie W, Zhang Y. A group of long noncoding RNAs identified by data mining can predict the prognosis of lung adenocarcinoma. *Cancer Sci*. 2018;109(12):4033-4044.
21. Song Q, Shang J, Yang Z, et al. Identification of an immune signature predicting prognosis risk of patients in lung adenocarcinoma. *J Transl Med*. 2019;17(1):70.
22. Yue C, Ma H, Zhou Y. Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. *PeerJ*. 2019;7:e8128.
23. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11:367.
24. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multimodal data. *Bioinformatics*. 2016;32(1):1-8.
25. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One*. 2017;12(5):e0176278.
26. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med*. 2013;19(5):619-625.
27. Wang H, Wang X, Xu L, Zhang J, Cao H. A molecular sub-cluster of colon cancer cells with low VDR expression is sensitive to chemotherapy, BRAF inhibitors and PI3K-mTOR inhibitors treatment. *Aging (Albany NY)*. 2019;11(19):8587-8603.
28. Ke K, Chen G, Cai Z, et al. Evaluation and prediction of hepatocellular carcinoma prognosis based on molecular classification. *Cancer Manag Res*. 2018;10:5291-5302.
29. Zhang Q, Yu X, Zheng Q, He Y, Guo W. A molecular subtype model for liver HBV-related hepatocellular carcinoma patients based on immune-related genes. *Front Oncol*. 2020;10:560229.
30. Ma X, Gu J, Wang K, et al. Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. *Sci Rep*. 2019;9(1):7073.
31. Zhao L, Zhao H, Yan H. Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes. *BMC Cancer*. 2018;18(1):603.
32. Mishra NK, Guda C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget*. 2017;8(17):28990-29012.
33. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525.
34. Inamura K, Fujiwara T, Hoshida Y, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*. 2005;24(47):7105-7113.
35. Rousseaux S, Debernardi A, Jacquiau B, et al. Ectopic activation of germline and placental genes identifies aggressive

metastasis-prone lung cancers. *Sci Transl Med*. 2013;5(186):186ra166.
36. Tang H, Xiao G, Behrens C, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin Cancer Res*. 2013;19(6):1577-1586.
37. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol*. 2014;9(1):59-64.
38. Director's Challenge Consortium for the Molecular Classification of Lung A, Shedden K, Taylor JM, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008;14(8):822-827.
39. Schabath MB, Welsh EA, Fulp WJ, et al. Differential association of STK11 and TP53 with KRAS mutation-associated gene expression, proliferation and immune surveillance in lung adenocarcinoma. *Oncogene*. 2016;35(24):3209-3216.
40. Su L-J, Chang C-W, Wu Y-C, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genom*. 2007;8:140.
41. Landi MT, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*. 2008;3(2):e1651.
42. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, et al. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer*. 2011;129(2):355-364.
43. Wei T-Y, Juan C-C, Hisa J-Y, et al. Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade. *Cancer Sci*. 2012;103(9):1640-1650.
44. Selamat SA, Chung BS, Girard L, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res*. 2012;22(7):1197-1211.
45. Nguyen DX, Chiang AC, Zhang XH, et al. WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis. *Cell*. 2009;138(1):51-62.
46. Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. 2016;166(3):740-754.
47. Chen C, Chen H, Zhang Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*. 2020;13(8):1194-1202.
48. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57.
49. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13.
50. Aguirre-Portolés C, Bird AW, Hyman A, Cañamero M, Pérez de Castro I, Malumbres M. Perez de Castro I, Malumbres M: Tpx2 controls spindle integrity, genome stability, and tumor development. *Cancer Res*. 2012;72(6):1518-1528.
51. Zhou F, Wang M, Aibaidula M, et al. TPX2 promotes metastasis and serves as a marker of poor prognosis in non-small cell lung cancer. *Med Sci Monit*. 2020;26:e925147.
52. Borkowska EM, Kruk A, Jedrzejczyk A, et al. Molecular subtyping of bladder cancer using Kohonen self-organizing maps. *Cancer Med*. 2014;3(5):1225-1234.
53. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*. 1999;96(6):2907-2912.
54. Shen R, Mo Q, Schultz N, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*. 2012;7(4):e35236.
55. Xie H, Xu H, Hou Y, et al. Integrative prognostic subtype discovery in high-grade serous ovarian cancer. *J Cell Biochem*. 2019;120(11):18659-18666.
56. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015;21(11):1350-1356.
57. Morris JS, Luthra R, Liu Y, et al. Development and validation of a gene signature classifier for consensus molecular subtyping of colorectal carcinoma in a CLIA-certified setting. *Clin Cancer Res*. 2020.
58. Takeuchi T, Tomida S, Yatabe Y, et al. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol*. 2006;24(11):1679-1688.
59. Wang S, Liu F, Wang Y, et al. Integrated analysis of 34 microarray datasets reveals CBX3 as a diagnostic and prognostic biomarker in glioblastoma. *J Transl Med*. 2019;17(1):179.
60. He R, Zuo S. A Robust 8-Gene prognostic signature for early-stage non-small cell lung cancer. *Front Oncol*. 2019;9:693.
61. Dai B, Ren LQ, Han XY, Liu DJ. Bioinformatics analysis reveals 6 key biomarkers associated with non-small-cell lung cancer. *J Int Med Res*. 2020;48(3):300060519887637.
62. Zhang MY, Liu XX, Li H, Li R, Liu X, Qu YQ. Elevated mRNA Levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis. *Int J Med Sci*. 2018;15(14):1676-1685.
63. Schott M, de Jel MM, Engelmann JC, et al. Selenium-binding protein 1 is down-regulated in malignant melanoma. *Oncotarget*. 2018;9(12):10445-10456.
64. Zeng GQ, Yi H, Zhang PF, et al. The function and significance of SELENBP1 downregulation in human bronchial epithelial carcinogenic process. *PLoS One*. 2013;8(8):e71865.
65. Caswell DR, Chuang CH, Ma RK, Winters IP, Snyder EL, Winslow MM. Tumor suppressor activity of Selenbp1, a direct Nkx2-1 target in lung adenocarcinoma. *Mol Cancer Res*. 2018;16(11):1737-1749.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wang H, Wang X, Xu L, Cao H, Zhang J. Nonnegative matrix factorization-based bioinformatics analysis reveals that TPX2 and SELENBP1 are two predictors of the inner sub-consensuses of lung adenocarcinoma. *Cancer Med*. 2021;10:9058–9077. doi:10.1002/cam4.4386