Integration of genomic data analysis for demonstrating potential targets in the subgroup populations of squamous cell lung cancer patients

Yongcui Wang[1,2], **Weiling Zhao**[1], **Xiaobo Zhou**[1,3]

[1]Center for Bioinformatics & Systems Biology, Department of Radiology, Wake Forest School of Medicine, Winston Salem, NC, USA

[2]Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, China

[3]School of Life Science and Biotechnology, Xi'an Jiaotong University, Xi'an, China

Correspondence to: *Xiaobo Zhou,* ***email:*** *xizhou@wakehealth.edu*

Keywords: *squamous lung cancer, integrated genomic analysis, mutation, methylation, patients' survival*

Received: *October 27, 2015* ***Accepted:*** *May 14, 2016* ***Published:*** *June 15, 2016*

ABSTRACT

Squamous cell carcinoma (SCC) is the second most frequent histologic subtype of non-small cell lung cancer (NSCLC), causing approximately 400,000 deaths per year worldwide. Although targeted therapies have improved outcomes in patients with adenocarcinoma, the most common subtype of NSCLC, the genomic alterations in SCC have not been comprehensively characterized and no therapeutic agents have been approved specifically for the patients with SCC. Therefore, development of novel therapeutic approaches is urgently needed. Here, we developed an integrative approach, called DLSA, to integrate genomic, epigenomic and transcriptomic data. DLSA stratified SCC patients into distinct survival subgroups and identified the potential molecular drivers in individual survival subtypes. Three subgroups of SCC patients with diverse molecular and clinical characteristics were unveiled through DLSA. Combined analysis of clinical and molecular data on those subgroups suggested that the molecular features in the stratified subgroups are not only consistent with the previous findings, but also provide a guide to targeted agents that worth to be evaluated in clinical trials for SCC patients with poor survival. In conclusion, DLSA offers the possibility for faster, safer, and cheaper the development of novel anti-cancer therapeutics in the early-stage clinical trails.

INTRODUCTION

Lung cancer is the leading cause of cancer death, resulting approximate 159,000 deaths in the United States in 2014 [1]. Adenocarcinomas and squamous cell carcinoma (SCC) are the most two frequent histologic subtype of non-small cell lung cancer (NSCLC). Although targeted therapeutics with EGFR tyrosine kinase and ALK inhibitors have been afforded benefits to the patients with lung adenocarcinomas, they are not effective for those with lung SCC. Therefore, development of novel therapeutic approaches for the treatment of lung SCC is urgently needed.

Large-scale cancer genome databases, including The Cancer Genome Atlas (TCGA), Cancer Genome Project (CGP), and Cancer Cell line Encyclopedia (CCLE), provide a great opportunity for uncovering the landscape of genetic alterations that underlie cancer patients' survival in a comprehensive genetic background. Through investigating the molecular features and survival outcomes of intrinsic subtypes, several novel targets for histologic diagnosis and therapy have been identified for lung adenocarcinoma and other cancers [2-7]. Currently, a clinically important challenge for SCC is to discover the novel survival subtypes and their molecular drivers through an integrative framework.

The standard approach for integration of multiple cancer genomic datasets is conducting cluster analyses on individual data platforms and then integrating these diverse platforms-specific cluster assignments into subtypes. Each subtype shares features across multiple datasets [8-10]. Shen et al. developed a joint latent variable model, called as iCluster, to incorporate diverse data types simultaneously, including binary (somatic mutation), categorical (copy number gain, normal, loss), and continuous (gene expression), and generate a single

integrated cluster assignment [11, 12]. iCluster captures the major biological variation observed across cancer genomes [13] and has been widely used to classify tumor subtypes [8, 9, 12, 14, 15]. However, the subtypes of tumors identified by iCluster are not associated with patients' survival. Therefore, approaches integrating tumor subtypes and survival are needed for clinical application.

In our study, we developed a novel integrative model to incorporate multiple genomic data sources and identify novel survival subtypes simultaneously. Each type of genomic data was considered as one representation of patients, thus integrated diverse genomic datasets for patient stratification was converted as clustering patients with various representations. Recently, deep learning framework (DLF) has been developed to learn objects with multiple levels of representations by transforming inputs through multiple non-linear processing layers [16-18]. It has been widely applied in video and audio classification [19-22]. Inspired by this, we proposed an integrative framework, called DLSA (**D**eep **L**earning for **S**urvival **A**nalysis) to stratify SCC patients into distinct survival subgroups and identify the potential molecular drivers in individual survival subtypes based on the concepts of DLF. Three subgroups of SCC patients were identified through DLSA. The clinical data and molecular data analyses on those subtypes suggest that our DLSA-based subtypes are agreed well with the previous findings in SCC. Furthermore, DLSA identified potential therapeutic candidates for further evaluation and validation.

RESULTS

Here, we developed a novel computational method to discover the subtypes of SCC through incorporating genomic, epigenomic and transcriptomic data, called DLSA. The schematic illustration of DLSA is shown in the Figure 1. The first step was to learn the survival signatures by associating a deep learning network with patients' survival. This effort leads to incorporation of multiple platform cancer data sources simultaneously in one model. Then the survival subgroups of patients were identified based on the discovered survival-specific signatures. Finally, the clinical and genomic features for each subtype were analyzed to identify potential targets for SCC treatment.

Survival-associated subtypes of lung SCC

Through learning patients' survival by DLSA in an integrative framework, the survival-specific signatures were identified. Patients with SCC were stratified into three subgroups based on the profiles of those signatures

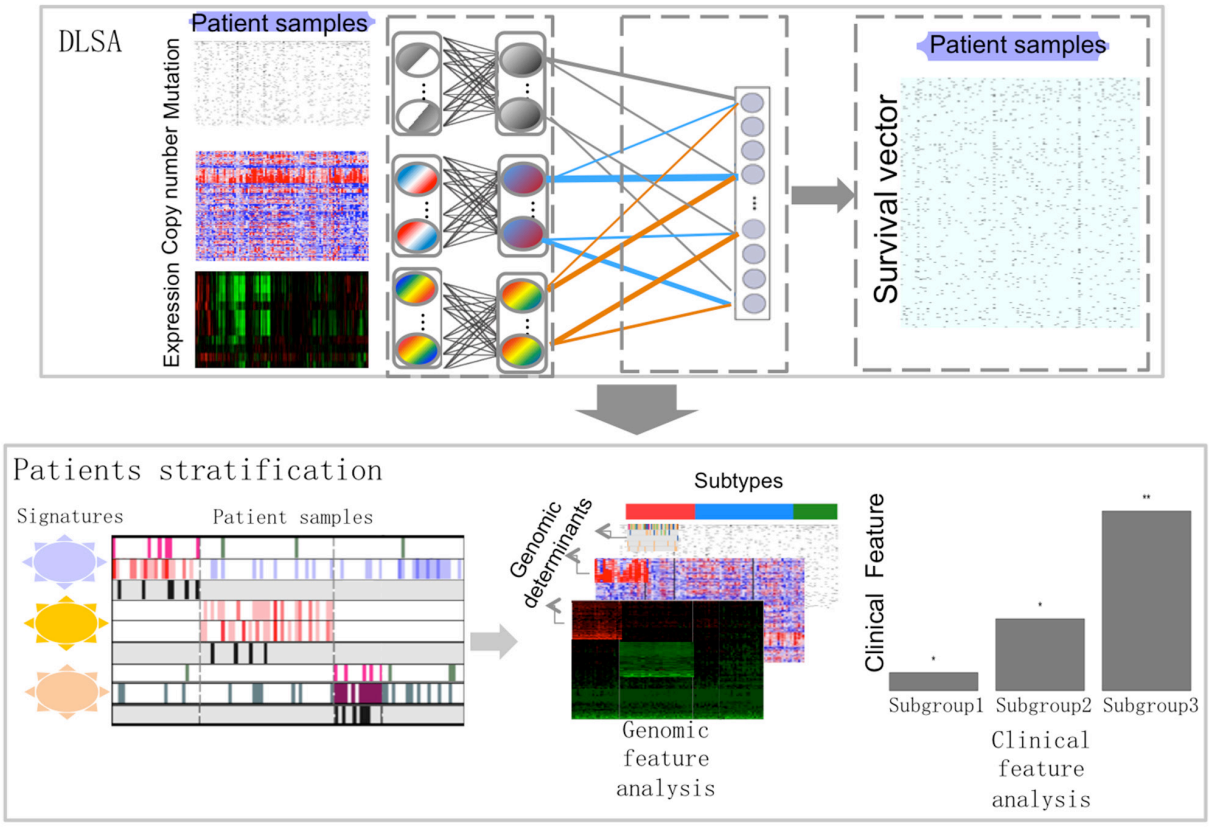


Figure 1: The schematic illustration of DLSA. Patients with lung SCC was classified into distinct survival subgroups based on the profile of survival-specific signatures, which were learned by associating a deep learning network with patients' survival. The potential molecular targets were identified by analyzing the genomic and clinical features in each subgroup.

(Supplementary Figure S2). Specifically, the subgroup 1 (Clus 1) patients were dominated with deletions of AKT1 and BRAF, hyper-methylation of NFE2L2, over-expression of FOXP1, and PTEN mutation. Subgroup 2 (Clus 2) patients had increased copy numbers of SOX2 and PIK3CA, and mutation of PIK3CA, MET, BRAF and RB1. Increased amplification of WHSClL1 and FGFR1, and mutation of EGFR were seen in the subgroup 3 (Clus 3) patients (Supplementary Figure S2). The molecular (upper panel of the Figure 2) and clinical data (bottom panel of the Figure 2) were presented with the three subgroups. The different molecular patterns across three subgroups were shown in Figure 2. For example, the gene expression levels were higher in the Clus 1; Gain of copy number and variation of methylation were seen in the Clus 2; The Clus 3 appeared to fall between Clus 1 and Clus 2, including medium mutation rate and gene expression levels.

Clinical characteristics of SCC subgroups

We exanimated the association of three subgroups with clinical factors, including patients' survival, tumor pathological stages, tobacco-consuming history, tumor location, patients' gender, and age at initial diagnosis. The number of patients, survival-specific signatures and median survival time for SCC subgroups were shown in the Figure 3A. The median survival days were 699, 397.5 and 496.5 in Clus 1, Clus 2 and Clus 3, respectively. We further analyzed the patients' survival in three subgroups using Kaplan-Meier survival analysis via 'survfit' function in R 'survival' package (Figure 3B). Figure 3B revealed

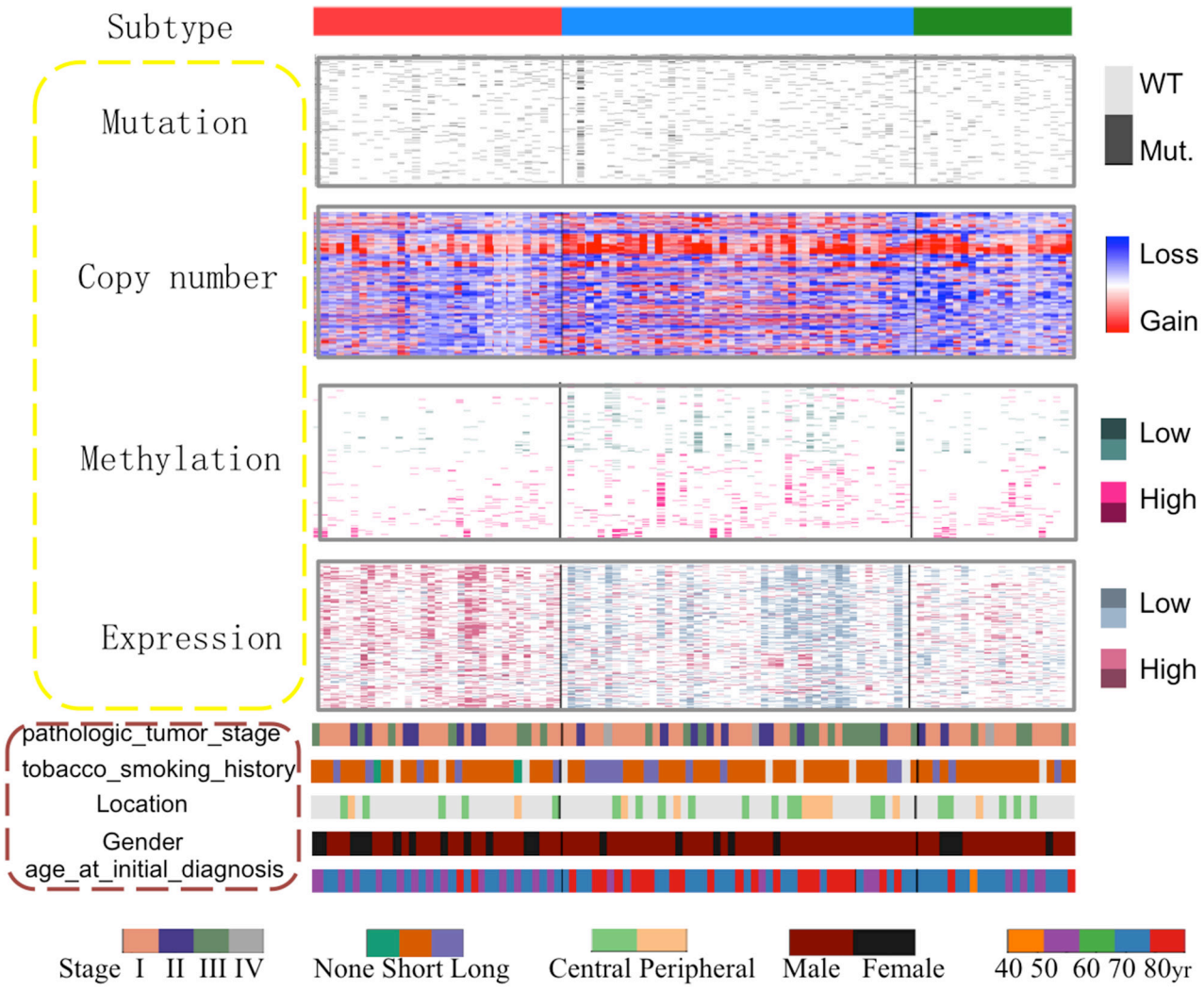


Figure 2: DLSA survival-associated subtypes. Four platform data were integrated with DLSA and three subtypes identified in the lung SCC patients as shown in red (Clus 1), blue (Clus 2) and green (Clus 3). Molecular (top) and clinical (bottom) data were grouped into the three subtypes. Tumor pathological stages (I, II, III, IV), tobacco-consuming history (none: lifelong Non-smoker; short: smoking less than 15 years; and long: smoking more than 15 years), tumor location (central or peripheral of lung), gender (male or female), and age at initial diagnosis (age range of 40-50, 50-60, 60-70 or 70-80 years old) are shown with different colors.

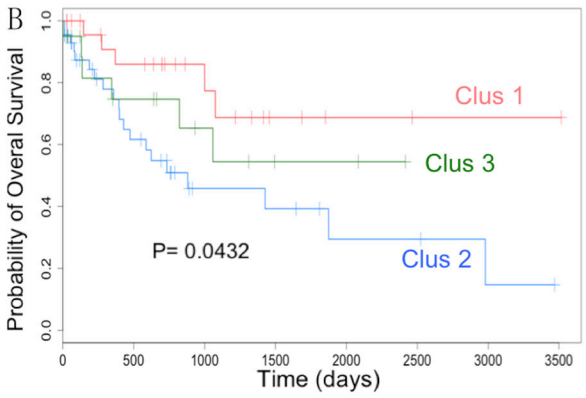
significant differences in the probability of overall survival among patients in the three subgroups. Clus 1 had better survival outcomes when compared with the other two groups. The worst survival outcomes were seen in the Clus 2. In terms of tumor stages, patients with stage I tumors were 58%, 51% and 62% in the Clus 1, 2, and 3, respectively. More than 21% patients in the Clus 1 carried with stage 1 tumors, when compared with those in the Clus 2. The percentage of patients with stage III tumors in the Clus 2 patients (28%) was higher than did in other two groups (20% for the Clus 1 and 19% for the Clus 3). The percentage of patients with central SCC tumors was 83%, 59% and 72% in the Clus 1, Clus 2, and Clus 3, respectively (Figure 3D). Apparently, Clus 1 group had ~ 24% more patients with central SCC tumors, compared with those in the Clus 2. In contrast, the patients with peripheral SCC tumors was 17%, 41% and 28% in the Clus 1, Clus 2 and Clus 3, respectively, implicating that the peripheral location of SCC tumors was possibly associated with the worse survival outcome of Clus 2.

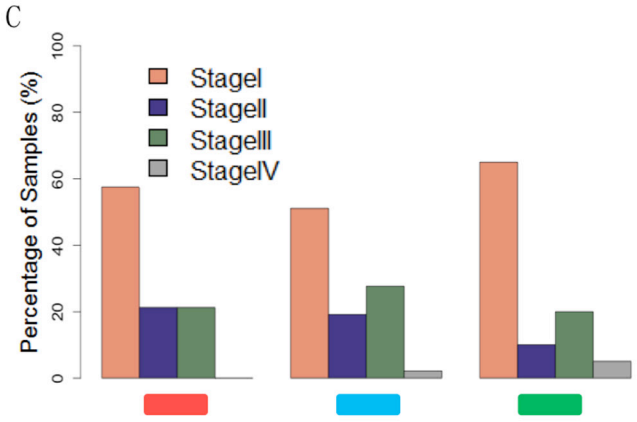
We assessed the association of patients' survival with tobacco consuming history by Kaplan-Meier survival analysis. The majority of SCC patients had tobacco smoking history and only a small portion of them was non-smokers. The *p*-value depicted the significance of Kaplan-Meier survival curves for patients with different smoking history groups (non-smoker, smoking for less than 15 years, and smoking for more than 15 years) was 0.03 (Supplementary Figure S3), indicating that patients' tobacco consuming history had a strong association with their survival.

We analyzed the gender distribution in the three subgroups. As shown in the Figure 4A, the majority of patients in the three groups were male and accounted for 70%, 80% and 76% of the total number of patients in the Clus 1, Clus 2 and Clus 3, respectively. To assess the association between smoking history and DLSA survival subtypes, we drew the barplot to show the percentage of patients with diverse smoking history in individual DLSA survival subtypes in the Figure 4B. The patients with more than 15 years smoking history in Clus 1, Clus 2, and Clus 3 were 15%, 40% and 22%, respectively. A high percentage of patients with a long smoking history in the Clus 2 suggested a possible association between tobacco consuming history and worse survival outcomes. We also analyzed the distribution of initial age at diagnosis in the three subgroups. In general, the average initial age of diagnosis in all of three groups were aged over 65 (Figure 4C). The average age at diagnosis in the Clus 2 was higher than those in the Clus 1 and Clus 3. In summary,

A

Subtypes	Signatures	No. of patients	Median survival
Clus 1	FOX1 overexpression, NFE2L2 hyper-methylation, AKT1 and BRAF deletion PTEN mutation	33	699 days
Clus 2	SOX2 and PIK3CA amplification PIK3CA, MET, BRAF and RB1 mutation	47	397.5 days
Clus 3	WHSC1L1 and FGFR1amplification EGFR mutation	21	496.5 days





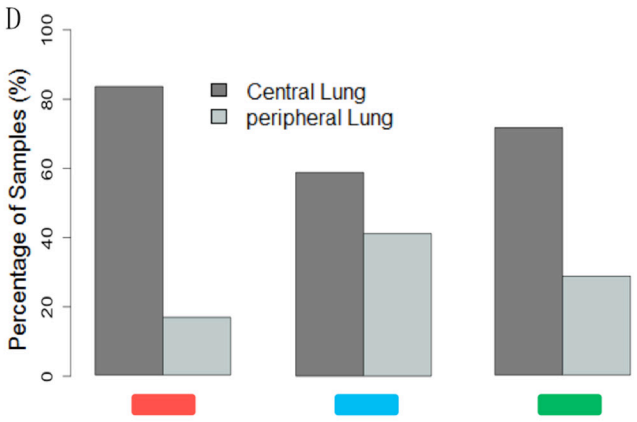


Figure 3: Clinical and pathological characteristics for individual DLSA subtypes. A. The number of patients, key molecular signatures and median survival time in three DLSA subgroups. **B.** Kaplan–Meier survival curves of Clus 1 (red), Clus 2 (blue) and Clus 3 (green). **C.** Distribution of SCC patients with different tumor stages in three subgroups. **D.** Tumor sites of SCC patients in three subgroups. DLSA results in survival subtypes with diverse clinical and pathological characteristics.

our analysis showed that a shorter smoking history and central localization of SCC tumors were related to a better survival outcome (Clus 1); while a long smoking history and peripheral distribution of tumors appeared to associate with a poor survival (Clus 2). Furthermore, our findings about the associations between survival outcomes and clinical factors in SCC patients are consistent with previous reports [24-27]. Moreover, statistical analysis on tobacco consuming history and initial age of diagnosis indicated that the old people with a long smoking history are more vulnerable.

Genomic characteristics for DLSA subtypes of SCC

Somatic genomic alterations

Significantly mutated genes were identified using MutSig algorithm [28]. There were 10 genes with a false discovery rate (FDR) Q value less than 0.1, including TP53, CDKN2A, PTEN, PIK3CA, KEAP1, MLL2, HLA-A, NFE2L2, NOTCH1, and RB1. The mutation rates of genes characterized previously in SCC (AKT1, DDR2, EGFR, BRAF, MET) [24] were also listed in the Figure 5A. Most of genes had missense mutation, which may result in changes in a coding sequence. The most commonly mutated gene is TP53 (82%), while only 1% of patients had AKT1 mutation. Although only a small fraction of SCC patients had DDR2 mutation (3%), lung SCC cell lines harboring DDR2 mutations were shown to be sensitive to dasatinib (DDR2 inhibitor), suggesting the clinical relevance of DDR2. Mutations in MLL2 (28%), PIK3CA (16%) and NFE2L2 (15%) were also common. PIK3CA mutations were enriched in patients with a history of tobacco consuming. About 95% patients with PIK3CA mutations have tobacco consuming history and 43% patients with tobacco consuming history for more than 15 years. As we showed earlier, a long smoking history was associated with poor survival in SCC patients. The association of PIK3CA mutation with tobacco consuming history implied that a possible link of PIK3CA mutation with poor survival outcomes. Indeed, increased mutation of PIK3CA was observed in the Clus 2 (62%).

Mutations in oncogenes BRAF and MET, and tumor suppressors KEAP1 and RB1 were also common in the Clus 2 (barplot in the Figure 5A). Mutation in

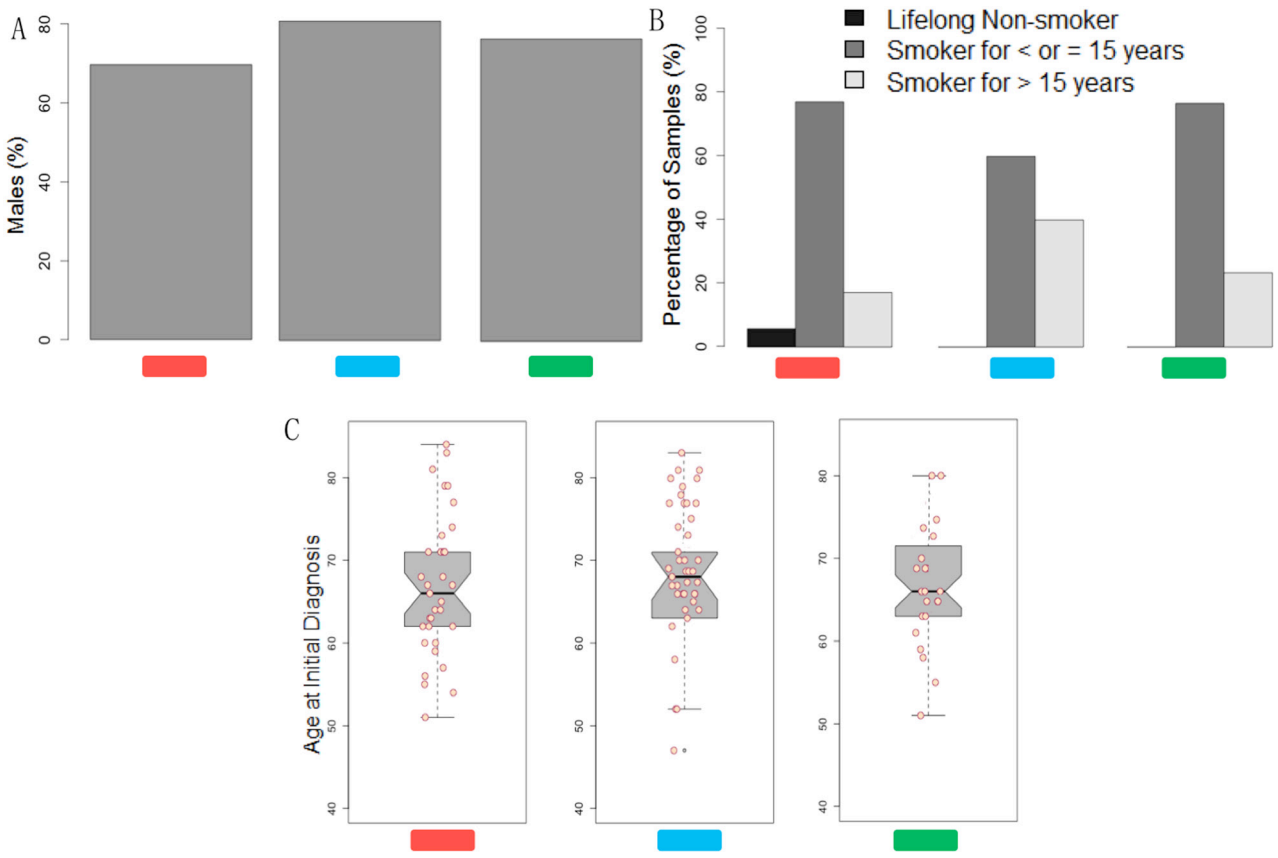


Figure 4: DLSA reveals survival subtypes with diverse clinical features. Panel **A.** shows the percentage of male patients in the Clus1 (red), Clus2 (blue) and Clus3 (green), respectively, indicating that most patients with SCC of the lung were male. Panel **B.** shows the percentage of nonsmoker (black), smoking history ≤ 15 years (grey) and smoking history > 15 years (light bars) in each subgroup, respectively, suggesting that more patients had a longer smoking history fall in Clus 2. Panel **C.** shows the age at initial diagnosis in 3 subgroups, and Clus 2 patients displayed higher diagnosis age.

PTEN was significant in the Clus 1 cohort (50%). The Kaplan-Meier survival curves analysis displayed that there was significant difference in patient's survival between SCC patients with PTEN mutation and wide type (Supplementary Figure S13), indicating the strong associations between PTEN mutation and better survival. EGFR and AKT mutation were often seen from the patients in the Clus 3 (75% patients with EGFR mutation, 100% with AKT mutation). The left panel of Figure 5B shows amplification of oncogenes, including PIK3CA, SOX2, PDGFRA, KIT, EGFR and MET, and deletion of tumor suppressors, including FOXP1, CDKN2A, PTEN, RB1 and NF1. Copy number gain in PIK3CA and SOX2 was seen in the 42% and 55% of Clus 2 patients, respectively, which were not seen in the other two groups. The percentage of Clus 2 patients with a copy number gain in PDGFRA, KIT, EGFR, MET, and BRAF were 12%, 10%, 11%, 13%, and 11%, respectively, which were higher than those in the other two groups. While deletion of tumor suppressors CDKN2A, PTEN were seen from 38% and 12% of Clus 2 patients, respectively, which were higher than those in the other two groups (Figure 5B). BRAF and AKT1 deletion frequencies were 24% and 27% in the Clus 1 patients, respectively (Figure 5B). While, the deletion in BRAF and AKT were not observed in the other two groups. 47% and 57% Clus 3 patients carried copy number gain in WHSC1L1 and FGFR1 (Figure 5B), respectively. In conclusion, the occurrence rate in oncogene amplification and tumor suppressor deletion were higher in the SCC tumors from Clus 2 patients, which appeared to be associated with poor survival. While lack of mutation and copy number alteration in these genes linked with a better survival outcome.

DNA methylation and mRNA expression profiling

As we showed above, tumor oncogenes including PIK3CA, EGFR, BRAF, FGFR1 and MET were found frequently mutated in SCC patients. These genes were also reported as genetic alterations in SCC in previous works [14,24-27]. We then further analyzed the expression and methylation of those genes. They were shown as the heatmap in Supplementary Figure S4. Consistent with the gene mutation profiles, above oncogenes was often overexpressed in the SCC patients. The percentage of patients with

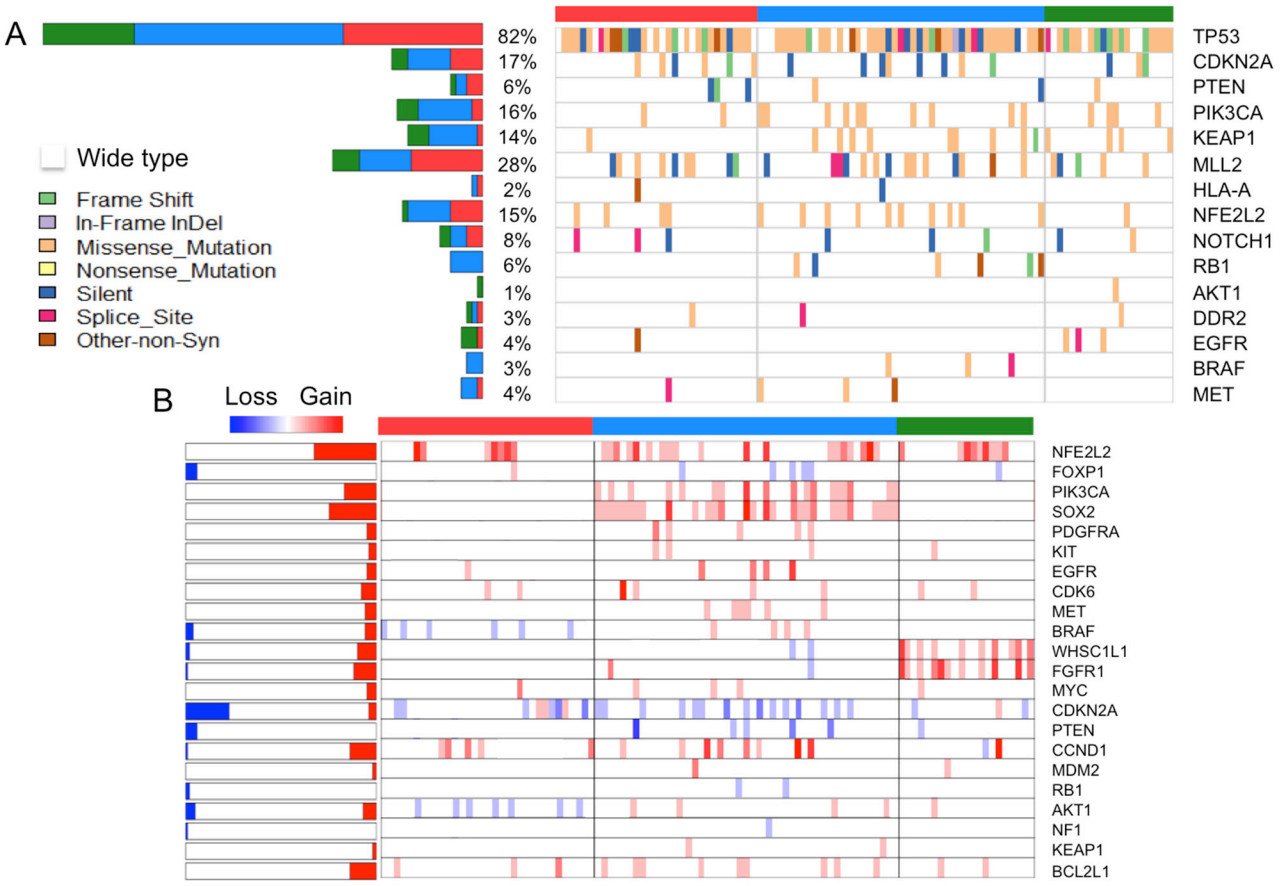


Figure 5: Somatic genomic alterations in individual DLSA subtypes. Panel **A.** shows significantly mutated genes' profile (with q-value less than 0.1 calculated by MutSigCV1.4) in the subgroups. The left part of Panel A shows the percentage of patients with mutation and mutation distribution across three subtypes. Panel **B.** shows the copy number variation for those significantly deletion or amplification genes (reported by TCGA investigators) in SCC patients. The left part of panel B displays the percentage of gain, norm and loss for significantly altered genes across all of the patients.

increased expression of PIK3CA, EGFR, BRAF, FGFR1 and MET were 17%, 51%, 10%, 26%, 14%, respectively. The percentage of Clus 2 patients with overexpressed PIK3CA, BRAF, FGFR, and MET were 52%, 70%, 65%, and 71%, respectively, and much higher than those in the other two groups. About 57% patients in the Clus 3 have increased expression of EGFR. These results together suggest a potential association between overexpression of PIK3CA, BRAF, FGFR1 and MET and the poor survival phenotype of Clus 2. EGFR was reported as a significantly mutated gene in lung adenocarcinoma [14]. The targeted therapy of EGFR inhibitors has been reported benefits to the patients with lung adenocarcinoma [2], but not to the patients with lung SCC. Mutation and overexpression of EGFR were seen in the Clus 3 patients, indicating EGFR-targeted therapy might just benefit a subset of patients with lung SCC.

We then compared the gene expression profiles of tumors from Clus 1 and Clus 2 subgroups. As shown in the Supplementary Figure S4, The expression of NFE2L2 was overexpressed in 21% Clus 2 patients compared with 10% in the Clus 1. NFE2L2 was also hypo-methylated in 9% Clus 2 patients (Supplementary Figure S4). The expression level of tumor suppressor genes CDKN2A, RB1 and FOXP1 were low in 50%, 7% and 8% of the Clus 2 patients, respectively. Correspondingly, these genes in the Clus 2 patients were often highly methylated (21% for CDKN2A, 25% for RB1, and 21% for FOXP1) (Supplementary Figure S4). In contrast, the expression and methylation of above genes were opposite in the Clus 1 patients. For example, PIK3CA and AKT1 in Clus 1 patients were often hypermethylated and underexpressed. While CDKN2A was frequently hypomethyled and mRNA level was high in Clus 1 patients. In addition, reduced expression TTF1 and overexpression of TP63, two well-known diagnosis markers for lung SCC [24,29,30], were seen in the poor surviving patients (Supplementary Figure S4). In conclusion, our results are generally consistent with the previous reports in SCC tumors [24,27], including overexpression of oncogenes and reduced levels of tumor suppressors in the poor survival group. Decreased expression of NFE2L2 and high expression of FOXP1 are correlated with the better survival group of SCC patients.

In addition, considering the fact that the genes showing significant different expressions between the poor and good survival groups may represent the goldmines for the development of therapeutic plan in treatment of SCC patients with poor survival. Thus we performed gene expression analysis on our DLSA Clus1 (good survival) and Clus2 (poor survival) patients through "limma" R package, and identified 81 differentially expressed genes (DEGs) with p-value less than 0.05 and the absolute value of log fold change greater than 1.5 (Table S1). Interestingly, all of these DEGs were down-regulated in the poor survival group, indicating the patients' survival might be improved by up-regulation of those genes.

Independent data test

Clinical and molecular analyses of DLSA subgroups indicated some novel findings about SCC. For example, we found that the NSCLC-associated typical genomic variations were related with poor survival of SCC patients, such as PIK3CA mutation and CDKN2A deletions. Furthermore, we found some variations, which were previously reported as the markers of lung SCC [25, 31], were also probably linked with poor survival, such as the overexpression of p63 and low expression of TTF1. To validate those findings, we introduced an independent dataset, which characterized the genome alterations in 594 clinically annotated lung tumors of SCC [31]. According to the analysis on somatic mutations in this new dataset, we found that PIK3CA mutation was associated with longer smoking history and poor survival outcome, which is consistent with our earlier outcomes (Figure 6). Beside PIK3CA mutation, the correlation of CDKN2A deletion with poor survival was also confirmed in this dataset (Supplementary Figure S10). Furthermore, the association of enhanced expression of p63 and reduced expression of TTF1 with poor survival outcome was also seen in this dataset (Supplementary Figure S11) [31].

In addition, to validate the results based on above DEGs analysis, we introduced another dataset, which characterized the gene expression levels on primary SCCs from 129 patients using Affymetrix U133A gene chips2 [32]. Two clinically relevant subsets of SCC patients were classified based on those expression data through unsupervised hierarchical clustering method, including the cluster 1 with poor survival outcome and the cluster 2 with good survival outcome. There were 121 genes (non-unique) showing significantly different expressions between two groups, and the majority of these genes (118) were down-regulated in the poor survival group [32], which is agreed with our findings. In addition, among four genes further validated by TaqMan quantitative RT-PCR2 in [32], the expression level of NTRK2 was much lower in the poor survival SCC subgroup, consistent with our result (Supplementary Figure S12). All these results confirm our findings in DEGs analysis, and most importantly they indicate a potential therapeutic agent for SCC patients with poor survival: NTRK2 is worthy of future experimental validation.

Through above independent data validation, some testable hypotheses are suggested for SCC. For example, PIK3CA mutation is linked with longer smoking history and poor survival, CDKN2A deletion is related with poor survival, overexpression of p63 and low expression of TTF1 are mainly seen in the poor survival subgroups, and low expression level of NTRK2 is linked with poor survival. All those findings provide great opportunities for biomarker discovery and therapeutic applications in treatment of SCC patients with poor survival.

The potential targets

The somatic alterations identified in this study include key signaling molecules important for initiation or progression of cancer. Most of them were mainly overexpressed in the Clus 2 patients (Supplementary Figure S5A). For example, overexpression of SOX2 was seen in 81% of Clus 2 patients. Increased expression of SOX2 has been proposed to maintain some characteristics of cancer cells [33,34]. Knocking-down or downregulation of SOX2 inhibits cancer cell growth and metastatic potential [35-37]. The expression of SOX2 is critical to maintain the self-renewal in embryonic stem cells and neural progenitor cells [38-40]. Recent studies indicated that SOX2 regulates tumor initiation and cancer stem-cell functions in squamous-cell carcinoma [41]. NFEL2 is a basic leucine zipper (bZIP) protein and activate the antioxidant proteins that protect against oxidative damage. Increased expression of NFE2L2 has been associated with radioresistance of cancer cells [42, 43]. Overexpression in NFE2L2 was seen in 49% of Clus 2 patients. Therefore, inhibition of NFE2L2 may sensitize the SCC cancer to radio- or checmothrapies. FGFR1, PIK3CA and AKT are involved in PI3K-Akt signaling. Increased expression of FGFR1, PIK3CA and AKT were observed in 43%, 53% and 38% of Clus 2 patients (Supplementary Figure S5A), indicating an activation of the PI3/AKT signaling transduction pathway. While the percentage of patients with highly expressed those genes in the Clus 1 or Clus 3 were low. Previous studies suggested that blockade of FGFR1 might be a promising target in the treatment of SCC and multiple FGFR inhibitors were in early clinical development [27]. Activation of the PI3K kinase signaling system results in AKT activation and enabled cancer cells to acquire multiple 'hallmark' characteristics [44]. Thus, the

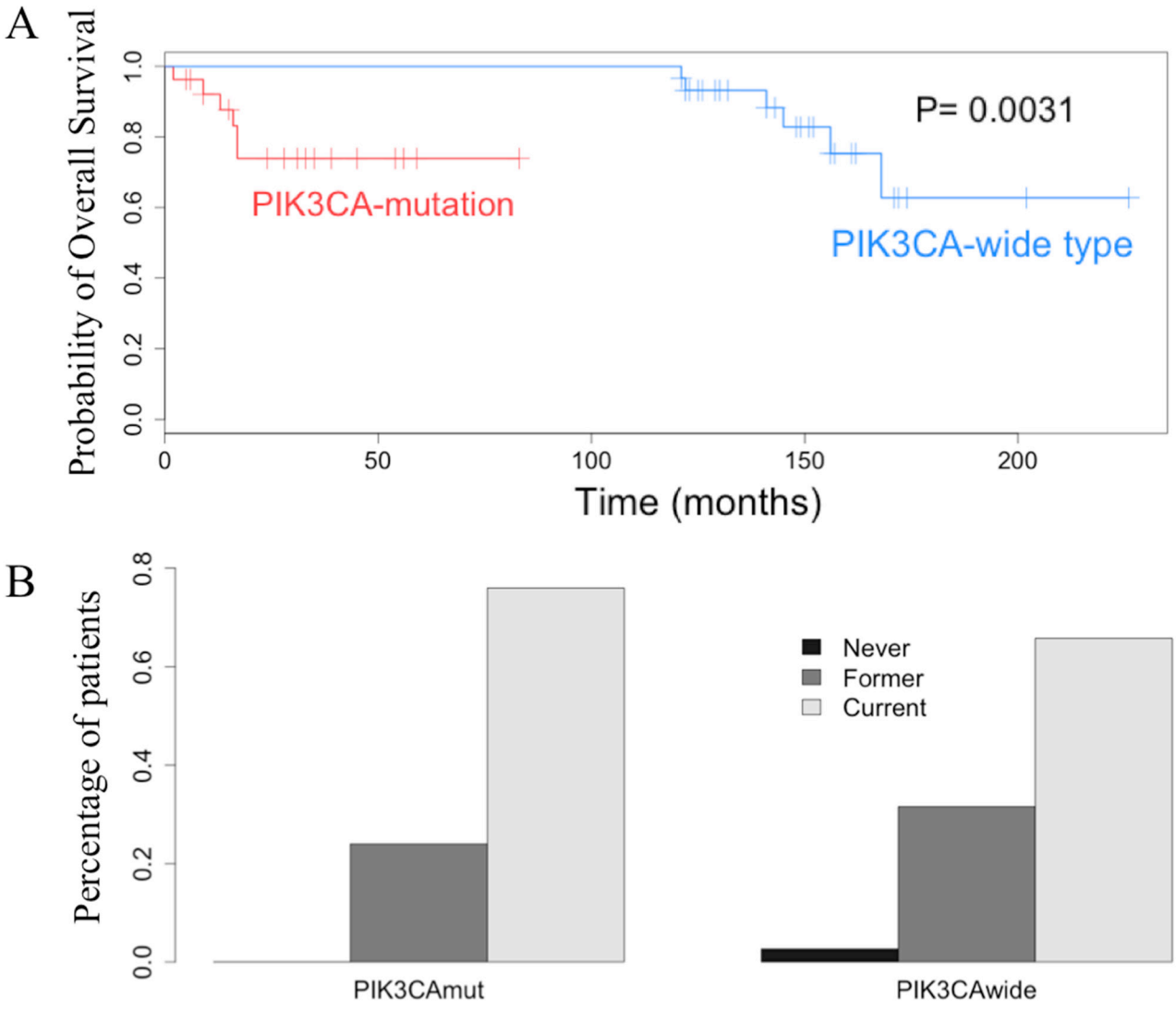


Figure 6: Verification of PIK3CA mutation in an independent dataset. A. Kaplan-Meier survival curves analysis on SCC patients with PIK3CA mutation and wide type. **B.** The smoking history for patients with PIK3CA mutation and wide type. It shows that PIK3CA mutation correlated with poor survival and longer smoking history.

elucidation of genes involved in PI3K-Akt signaling might provide the novel therapeutic strategies in the treatment of SCC.

Besides above significantly mutated and altered genes, we also uncovered another promising therapeutic targets: ERBB2. ERBB2 is a proto-oncogene located on the long arm of chromosome 17 (17q12) and encodes HER2/Neu, a receptor tyrosine kinase of the epidermal growth factor receptor family. Evidence suggests a role for ERBB2 signaling in mediating resistance of lung cancers to EGFR-targeting therapies, suggesting inhibition of ERBB2 could potentially benefit a subset of patients with SCC of lung [45, 46]. In addition, EGFR is in the up-stream of ErbB signaling, of which downstream will participate in several cancers, including NSCL, glioma and endometrial cancer (referred by KEGG human signaling pathway of ErbB). Here, we found that mutation and amplification in ERBB2 were only seen in the Clus 2 (Supplementary Figure S5B). In addition, enhanced expression of ERBB2 was also enriched in this subgroup (85%). These results together suggest that therapy targeting ERBB2 could potentially benefit a subset of SCC patients with poor survival.

Comparison with other integrative frameworks

Here, we integrated diverse genomic data types by DLSA to perform the patient stratification and molecular target discovery. The equation (1) $\sim$ (2) show that DLSA can not only capture the information of each data source by feature extraction from transformation layers, but also model inter-feature relationship by introducing the positive semi-definite matrix Ψ from the fusion layer. Most importantly, DLSA intends to control the diversity of different data types through a regularization constraint, which was lacking in other integrative frameworks, such as iCluster. iCluster was developed for integrative clustering based on the hypothesis that diverse molecular phenotypes can be predicted by a set of orthogonal latent variables that can reveal tumor subgroups of biological and clinical importance [11-13]. However, during modeling, the constraint to control the independence of latent variables was missing. In addition, iCluster reintroduces the original genomic variables via latent variables, which models the variance-covariance structure within data types and the associations among different data type by a joint probability model. While, DLAS reintroduces the original genomic variable via fusion variables, which capture inner- and inter-feature relationship by a multiple sigmoid processing. That is, both iCuster and DLAS can reintroduce the original genomic variables by latent variables, which can characterize and differentiate the original genomic variables. Both of them belong to the framework of feature re-representation. iCluster assumed the genomic variables are correlated with latent variables by a generalized linear system, and learned representing parameters via a joint probability model. While DLSA applied multiple nonlinear processing to model the relationship between genomic variables and latent variables, and learned the representing parameters via a regularized regression model. In conclusion, in addition to the main characteristics of iCluster, DLSA has added independent control in the modeling process.

Besides independent control, the another difference between iCluster and DLSA is that DLSA incorporated patients' survival, while iCluster did not. To display the advantage of adding survival information, we run iCluster on our dataset. The clustering results with three clusters are displayed in the supplementary materials (Supplementary Figure S7, Supplementary Figure S8). Although diverse patterns of copy number, methylation, and gene expression profiles in different subgroup were reveiled by iCluster (Supplementary Figure S7B-S7D), the iCluster subgroups presented inconsistent clinical characteristics. For example, they do not have significant diverse survival outcomes (Supplementary Figure S8A); subgroup two shows the old initial diagnosis age and most comparable central tumor and peripheral tumor (the factors associating with poor survival) (Supplementary Figure S8B, S8D), but have the least male patients and patients with more than 15 years smoking history (the factors associating with poor survival) (Supplementary Figure S8E, S8F). These results show the advantage of our DLSA by incorporating survival data. That is, the subgroups of patients not only display different genomic characteristics, but also have diverse clinical properties, especially show the significant difference of survival outcomes, which provide the insight into the understanding the molecular mechanism of patients' survival and figure out the way to prolong the life span at molecular level.

DISCUSSION

In this study, we attempted to develop a statistically powerful integrative approach: DLSA to incorporate multiple cancer genomic data types for patient stratification. The main contributions include i] Using an integrative algorithm to incorporate heterogeneous genomic data sources, and ii] association of the genomic features with patients' survival to identify novel survival subtypes and potential target agents. Through DLSA, we defined three subgroups of SCC patients with diverse clinical and molecular characteristics: Clus 1: better survival subgroup, Clus 2: poor survival subgroup, and Clus 3: medium survival. The key features for each SCC patients subtypes can be seen in Figure 7. In conclusion, SCC patients in better survival subgroup (Clus 1) had a shorter smoking history, higher PTEN mutation frequency, low activity of oncogene and high activity in tumor suppressors; SCC patients in the poor survival subgroup (Clus 2) had a longer smoking history and higher spreading tumor distribution, frequent oncogene mutation and amplification, and inactivated tumor suppressors. More SCC patients in the Clus 3 carried stage 1 tumors, and display EGFR mutation and amplification

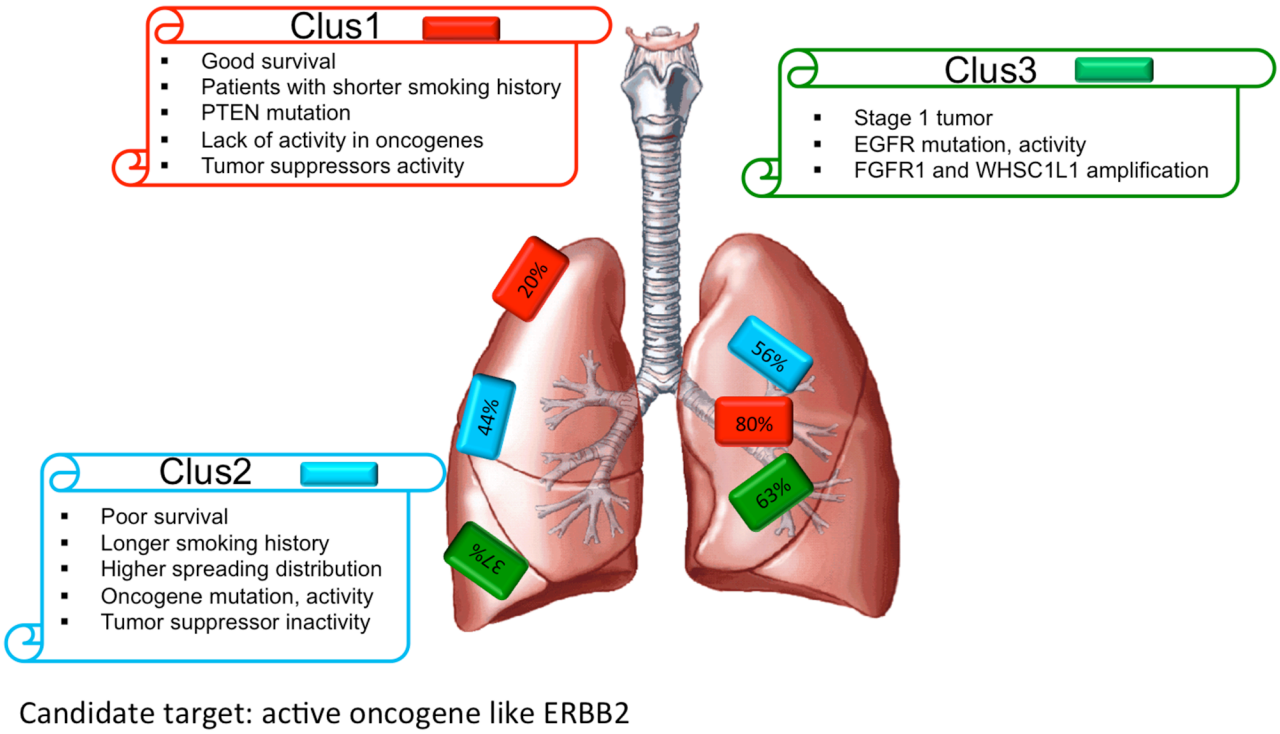


Figure 7: Key features of SCC lung cancer subtypes. This schematic lists some of the important features associated with each of the three survival subgroups of SCC patients. Distribution of survival subtypes in tumors obtained from distinct regions of the lung is represented by inset charts.

of FGFR1 and WHSC1L1. Those discoveries are not only consistent with the previous findings in SCC of lung, but also provide a guide to targeted agents that worth to be evaluated in clinical trials.

Based on the molecular and clinical analyses on SCC subtypes and subsequent validation, the identified gene group could be used as diagnostic markers of SCC. Some of them have the potential to be used as therapeutic targets for SCC treatment. For example, PIK3CA is mutated and overexpressed in the group with poor survival. Other studies indicate mutant PIK3CA stimulates the AKT pathway and promotes cell growth in cancers [24-26]. Therefore, PIK3CA is a potent target for the treatment of SCC. Overexpression of ERBB2 was associated with poor survival outcome of SCC patients and could be a potential target for therapy. Considering the fact that ERBB2 signaling plays an important role in mediating resistance of lung cancers to EGFR-targeting therapies, the future work could also test the role of ERBB2 in mediating resistance of patients to EGFR-targeting therapies.

MATERIALS AND METHODS

Materials

The data sources that we employed here for survival signature-guided patient stratification were collected from Lung squamous cell carcinoma of TCGA database, including DNA sequencing-based somatic mutation, array-based somatic copy number alterations, array-based DNA methylation profiling and messenger RNA expression from 178, 405, 161 and 155 patients, respectively. A total of 101 patients had all of four data types available. The clinical data used for further analysis included patients' survival, tobacco smoking history, gender and initial age of diagnosis, and tumor pathologic stages and location, which were collected from the TXT file of ' nationwidechildrens.org_clinical_patient_lusc' in TCGA data portal. We applied the methods used by TCGA working groups [8, 9, 14, 15] for data processing in our study. Briefly, the mutation MAF file was used for somatic mutation data analysis. A gene-by-sample matrix of binary values (1-mutated, 0-wildtype) was generated for integrative clustering. The top 1000 significantly mutated genes ranked by the MutSig [28] analysis were included for clustering. Array-based level two TSV files were used for analysis of somatic copy number variations. We applied standard deviation to exclude genes with little variance across the samples, and the 2,540 genes were remained for further analysis. The array-based DNA methylation level three TXT files were employed for methylation analysis and 4,000 genes corresponding to the top 4,000 of most variable CpG sites were selected based on median absolute deviation of β-value across the samples. For mRNA data, lowly expressed genes were excluded based on median-normalized counts and 414 highly expressed genes were selected for clustering.

Methods

Learning survival signatures through associating a deep learning network with patients' survival

To capture both inner- and inter- feature relationship, a deep learning network with transformation layer and fusion layer for feature extraction and inter-feature relationship learning was introduced here. Suppose we have a total of N patients with a total of M data types (x_n^m, $m=1,\cdots,M; n=1,\cdots,N$) and L layers in a deep learning network (Figure 1). Denoted a_l^m and a_{l-1}^m as the input and output of the lth layer for m th data type, $l=1,\cdots,E$ (E : the number of transformation layers), $m=1,\cdots,M$, and a_l as the note in l th fusion layer, $l=E+1,\ldots,L-1$. Let W_l^m and b_l^m as the weight matrix and bias vector of the transition function from $l-1$ th to l th transformation layer, respectively, and W_l and b_l as the weight matrix and bias vector of the transition function from $l-1$ th to l th fusion layer, respectively. Given patient with m th data type x^m, the transition function from the $l-1$ th to the l th layer was calculated as:

$$a_l^m=\begin{cases}\sigma\left(W_{l-1}^m a_{l-1}^m+b_{l-1}^m\right), & 1<l\le E,\\ x^m, & l=1\end{cases}$$

$$a_l=\begin{cases}\sigma\sum_{m=1}^{M} W_E^m a_E^m+b_E, & l=E+1\\ \sigma(W_l a_l+b_l), & E+1<l\le L-1\end{cases} \quad (1)$$

where $\sigma(\cdot)$ is sigmoid function, defined as $\sigma(\mu)=\frac{1}{1+\exp(-\mu)}$.

The optimal weights for each layer can be obtained by the following optimization problem:

$$\min_{W,\Psi} \sum_{i=1}^{N}\ell(\hat{y}_i,y_i)+\frac{\lambda_1}{2}\left(\sum_{l=1}^{E}\sum_{m=1}^{M}\|W_l^m\|_F^2+\sum_{l=E+1}^{L-1}\|W_l\|_F^2\right)+\frac{\lambda_2}{2}tr\left(W_E\Psi^{-1}W_E^T\right) \quad (2)$$

$$\text{s.t. } \Psi\succeq 0,\ tr(\Psi)=1,$$

where ℓ is loss function measuring the discrepancy between the survival of patient y and the estimated survival $\hat{y}$ (the output of the network a_L), λ_1 and λ_2 are regularization parameters, positive semi-definite matrix $\Psi\in R^{M\times M}$ models the inter-feature relationship, and $tr(\Psi)=1$ is applied to restrict complexity of the model, as suggested in [23], $\|W\|_F$ is Frobenius norm of matrix W. The first two regularization terms in the optimal function are used for sparsity control and last one for managing the diversity among different features.

Previous works assigned patients into different subgroups by integrating multi-level genomic data, and did not link with patients' survival [8-10]. Here, for survival learning, we associated the fusion variables a_{L-1} with patients' survival to generate survival subtypes of SCC. Patients' survival includes two elements, events and survival times. To obtain a survival output containing both two elements, we introduced a binary vector here, named survival vector, to represent the survival status at a given time point. Suppose we have a total of P time points, for patient x with survival time of s, the corresponding survival vector is $y(s)=(y_1(s),\cdots,y_P(s))$, where $y_j(s)=0$, if $s>t_j$, otherwise $y_j(s)=1$. For example, for patient with survival time $s=3.2$ month, and a total of 60 time points, then the survial vector for this patient is $y=(0,0,0,1,1,\cdots,1)$ (containing 3 zero, and $P-3$ one, where $P=60$).

Once the survival vector has been defined, the loss function could be easily defined. For this purpose, we calculated the probability of observing survival vectors as follows:

$$Prob_\beta\left(y(s)=(y_1(s),\cdots y_P(s))/a_{L-1}\right)=\frac{\prod_{i=1}^{P}\left(e^{\int_0^{t_i}a_{L-1}^T\beta(t_i)du}-1\right)^{y_i(s)}}{\sum_{k=0}^{P}\prod_{i=k+1}^{P}\left(e^{\int_0^{t_i}a_{L-1}^T\beta(t_i)du}-1\right)}, \quad (3)$$

where a_{L-1} is the fusion variables in the last fusion layer, $\beta(t)$ is time-varying parameters in Aalen linear hazard function on given feature θ: $h(t)=\theta^T\beta(t)$. The above probability was obtained based on survival function: $Prob_\beta(T>t_i/\theta)=e^{-\int_0^{t_i}\theta^T\beta(t_i)du}$. We introduced Aalen linear hazard model here because there is no baseline hazard function in Aalen model. The log likelihood of a set of N patients with survival time $s_1,\cdots,s_N$ and fusion variables $a_{L-1}^1,\cdots,a_{L-1}^N$ was formulated as follows:

$$\sum_{j=1}^{N}\left[\sum_{i=1}^{P}y_i(s_j)\log\left(e^{\int_0^{t_i}(a_{L-1}^j)^T\beta_i du}-1\right)-log\left(\sum_{k=0}^{P}\prod_{i=k+1}^{P}\left(e^{\int_0^{t_i}(a_{L-1}^j)^T\beta_i du}-1\right)\right)\right] \quad (4)$$

where $\beta(t_i)=\beta_i, i=1,\cdots P.$ Let $\tau_{\mathcal{B}}(a_{L-1}^j,i)=e^{\int_0^{t_i}(a_{L-1}^j)^T\beta_i du}=e^{t_i(a_{L-1}^j)^T\beta_i}$, then the final optimization problem for survival estimation via deep learning network was formulated as follows:

$$\min_{W,\mathcal{B},\Psi}\ \mathcal{L}+C_1\sum_{i=1}^{P}\|\beta_i\|^2+C_2\sum_{i=1}^{P-1}\|\beta_{i+1}-\beta_i\|^2+\lambda_1\left(\sum_{l=1}^{E}\sum_{i=1}^{M}\|W_l^m\|_F^2+\sum_{l=E+1}^{L-1}\|W_l\|_F^2\right)+\lambda_2 tr\left(W_E\Psi^{-1}W_E^T\right) \quad (5)$$

$$\text{s.t. } \Psi\succeq 0,\ tr(\Psi)=1,$$

where $\mathcal{L}=-\sum_{j=1}^{N}\left[\sum_{i=1}^{P}y_i(s_j)\log\left(\tau_{\mathcal{B}}(a_{L-1}^j,i)-1\right)-log\left(\sum_{k=0}^{P}\prod_{i=k+1}^{P}\left(\tau_{\mathcal{B}}(a_{L-1}^j,i)-1\right)\right)\right]$,

$C_1, C_2, \lambda_1, \lambda_2$ are pre-defined regularization constants, $\|W\|_e$ is Frobenius norm of matrix W. The first regulation term $\left|\left|\beta_i\right|\right|^2$ was used to prevent the norm of the parameter vector from overfitting. The second regulation term $\left\|\beta_{i+1} - \beta_i^2\right\|$ ensured that parameters vary smoothly across consecutive time points, which is especially important for controlling capacity of a model when time points became dense. We applied an alternative optimization method to iteratively minimize the optimization problem (5), the details were described in supplementary materials.

Learning novel survival subtypes of SCC through unveiled survival-specific signatures

The optimal weight matrix W and hazard parameter β indicate the most contributed genomic features during survival learning. These important features were defined as the survival signatures and would guide us for tumor stratification. The flowchart for survival signatures generation after optimization problem (5) solved was displayed in the Supplementary Figure S1. Suppose that the deep learning network only contains four layers, including input, transform, fusion and output, the most contributed fusion notes were found based on the value of β. The weight matrices for top fusion notes provide the evidence for determining the most contributed data types during learning. The genomic variables that could be used as candidate signatures were traced based on the weight matrix of transformation layer. The candidate signatures were generated by looking for the most frequent genomic variables based on different top fusion notes. Then the final survival-specific signatures were uncovered by comparing with the known genetic alterations in SCC [24-27]. Based on the profiles of survival-specific signatures across the SCC patients, the novel survival subgroups of patients were unveiled, and the genomic feature analysis for individual subtype was performed to identify the promising targets for further diagnosis and treatment.

ACKNOWLEDGMENTS

Wang, Zhao and Zhou are partially funded by NIH 1U01CA166886 (Zhou), NIH 1R01LM010185 (Zhou), and NIH 1U01HL111560 (Zhou). This work is also partially funded by NSFC #61373105.

REFERENCES

1. http://www.cancer.gov/cancertopics/types/lung.
2. Ladanyi M, Pao W. Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. Modern pathology, 2008; 21: S16-S22.
3. Greulich H. The Genomics of Lung Adenocarcinoma Opportunities for Targeted Therapies. Genes & cancer, 2010; 1: 1200-1210. doi: 10.1177/1947601911407324.
4. Khvalevsky EZ, Gabai R, Rachmut IH, Horwitza E, Brunschwiga Z, Orbacha A, Shemia A, Golanb T, Dombc AJ., Yavinc E, Giladid H, Rivkind L, Simerzind A, et.al. Mutant KRAS is a druggable target for pancreatic cancer. Proceedings of the National Academy of Sciences, 2013; 110: 20723-20728.
5. Hussain S A, Palmer D H, Spooner D, Rea DW. Molecularly targeted therapeutics for breast cancer. BioDrugs, 2007; 21: 215-224.
6. Roukos DH. Targeting gastric cancer with trastuzumab: new clinical practice and innovative developments to overcome resistance. Annals of surgical oncology, 2010; 17: 14-17.
7. Reardon DA, Turner S, Peters KB, Desjardins A, Gururangan S, Sampson JH, McLendon RE, Herndon JE, Jones LW, Kirkpatrick JP, Friedman AH, Vredenburgh JJ, Bigner DD, Friedman HS. A review of VEGF/VEGFR-targeted therapeutics for recurrent glioblastoma. Journal of the National Comprehensive Cancer Network, 2011; 9: 414-427.
8. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. Nature, 2013; 497: 67-73.
9. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014; 513: 202-209.
10. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014; 158: 929-944.
11. Shen, R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25: 2906-2912.
12. Shen R, Olshen AB, Huse J, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25: 2906-2912.
13. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences. 2013; 110: 4245-4250.
14. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489: 519-525.
15. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511: 543-550.
16. Chin C, Brown DE. Learning in science: A comparison of deep and surface approaches. Journal of research in science teaching. 2000; 37: 109-138.

17. Bengio Y. Learning deep architectures for AI. Foundations and trends® in Machine Learning. 2009; 2: 1-127.
18. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2013; 35: 1798-1828.
19. Schmidhuber J. Deep Learning in Neural Networks: An Overview. arXiv preprint arXiv:1404.7828, 2014.
20. Wu Z, Jiang YG, Wang J, Pu J, Xue X. Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification. Proceedings of the ACM International Conference on Multimedia. ACM, 2014.
21. Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 2009.
22. Le QV, Zou WY, Yeung SY. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.
23. Zhang Y., Yeung DY. A convex formulation for learning task relationships in multi-task learning. In UAI, 2010.
24. Sos ML, Thomas RK. Genetic insight and therapeutic targets in squamous-cell lung cancer. Oncogene. 2012; 31: 4811-4814.
25. Perez-Moreno P, Brambilla E, Thomas R, Soria JC. Squamous cell carcinoma of the lung: molecular subtypes and therapeutic opportunities. Clinical Cancer Research. 2012; 18: 2443-2451.
26. Cumberbatch M, Tang X, Beran G, Eckersley S, Wang X, Ellston RP, Dearden S, Cosulich S, Smith PD, Behrens C, Kim ES, Su X, Fan S, et al. Identification of a subset of human non-small cell lung cancer patients with high PI3Kβ and low PTEN expression, more prevalent in squamous cell carcinoma. Clin Cancer Res., 2014; 20: 595-603.
27. Rooney M, Devarakonda S, Govindan R. Genomics of squamous cell lung cancer. The oncologist. 2013;18: 707-716.
28. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499: 214-218.
29. Terry J, Leung S, Laskin J, Leslie KO, Gown AM, Ionescu DN. Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples. The American journal of surgical pathology. 2010; 34: 1805-1811.
30. Bishop JA, Teruya-Feldstein J, Westra WH, Pelosi G, Travis WD, Rekhtman N. p40 (DNp63) is superior to p63 for the diagnosis of pulmonary squamous cell carcinoma. Mod Pathol. 2011; 25: 405-15.
31. Project, The Clinical Lung Cancer Genome, and Network Genomic Medicine NGM. A genomics-based classification of human lung tumors. Science translational medicine. 2013; 5: 209ra153.
32. Raponi M., Zhang Y., Yu J. et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer research. 2006; 66: 7466-7472.
33. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005; 122:947-956.
34. Hussenet T, du Manoir S. SOX2 in squamous cell carcinoma: Amplifying a pleiotropic oncogene along carcinogenesis. Cell Cycle. 2010; 9:1480-1486.
35. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L, Tamayo P, Gat-Viks I, Ramos AH, Woo MS, Weir BA, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nature Gen., 2009; 41:1238-1242.
36. Xiang R, Liao D, Cheng T, Zhou H, Shi Q, Chuang TS, Markowitz D, Reisfeld RA, Luo Y. Downregulation of transcription factor SOX2 in cancer stem cells suppresses growth and metastasis of lung cancer. Br J Cancer., 2011; 104:1410-1417.
37. Fang WT, Fan CC, Li SM, Jang TH, Lin HP, Shih NY, Chen CH, Wang TY, Huang SF, Lee AY, Liu YL, Tsai FY, Huang CT, et al. Downregulation of a putative tumor suppressor BMP4 by SOX2 promotes growth of lung squamous cell carcinoma. International Journal of Cancer., 2014; 135: 809-819.
38. Rao RR, Calhoun JD, Qin X, Rekaya R, Clark JK, Stice SL. Comparative transcriptional profiling of two human embryonic stem cell lines. Biotechnology and bioengineering. 2004; 88: 273-286.
39. Wang J, Rao S, Chu J, Shen X, Levasseur DN, Theunissen TW, Orkin SH. A protein interaction network for pluripotency of embryonic stem cells. Nature. 2006; 444: 364-368.
40. Adachi K, Suemori H, Yasuda SY, Nakatsuji N, Kawase E. Role of SOX2 in maintaining pluripotency of human embryonic stem cells. Genes Cells, 2010; 15: 455-470.
41. Boumahdi S, Driessens G, Lapouge G, Rorive S, Nassar D, Le Mercier M, Delatte B, Caauwe A, Lenglez S, Nkusi E, Brohée S, Salmon I, Dubois C, et al. SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. Nature. 2014; 511: 246-250.
42. Kobayashi A, Kang MI, Okawa H, Ohtsuji M, Zenke Y, Chiba T, Igarashi K, Yamamoto M. Oxidative stress sensor Keap1 functions as an adaptor for Cul3-based E3 ligase to

regulate proteasomal degradation of Nrf2. Mol Cell Biol., 2004; 24:7130 -7139.

43. Solis LM, Behrens C, Dong W, Suraokar M, Ozburn NC, Moran CA, Corvalan AH, Biswal S, Swisher SG, Bekele BN, Minna JD, Stewart DJ, Wistuba II. Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. Clin Cancer Res., 2010; 16:3743-3753.

44. Vivanco I, Sawyers CL. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. Nat Rev Cancer., 2002; 2:489-501.

45. Yonesaka K, Zejnullahu K, Okamoto I, Satoh T, Cappuzzo F, Souglakos J, Ercan D, Rogers A, Roncalli M, Takeda M, Fujisaka Y, Philips J, Shimizu T, et al. Activation of ERBB2 signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. Science Transl Med., 2011; 3:99ra86.

46. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov., 2012; 2:401-404.