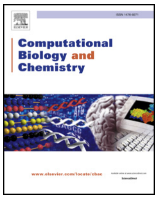
Computational Biology and Chemistry 87 (2020) 107277

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



Research Article

Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication

Tzong-Yi Lee[a,b,c], Kai-Yao Huang[a,b,c], Cheng-Hsiang Chuang[d], Cheng-Yang Lee[e], Tzu-Hao Chang[e,f,*]

[a] Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China
[b] School of Life and Health Science, The Chinese University of Hong Kong, Shenzhen, China
[c] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[d] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan
[e] Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei City, Taiwan
[f] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei City, Taiwan

ARTICLE INFO

Keywords:
Lung adenocarcinoma
Multi-omics
Autoencoder
Deep learning
Machine learning
Survival analysis
Prognosis prediction

ABSTRACT

Lung cancer is the most occurring cancer type, and its mortality rate is also the highest, among them lung adenocarcinoma (LUAD) accounts for about 40 % of lung cancer. There is an urgent need to develop a prognosis prediction model for lung adenocarcinoma. Previous LUAD prognosis studies only took single-omics data, such as mRNA or miRNA, into consideration. To this end, we proposed a deep learning-based autoencoding approach for combination of four-omics data, mRNA, miRNA, DNA methylation and copy number variations, to construct an autoencoder model, which learned representative features to differentiate the two optimal patient subgroups with a significant difference in survival (P = 4.08e-09) and good consistency index (C-index = 0.65). The multi-omics model was validated though four independent datasets, i.e. GSE81089 for mRNA (n = 198, P = 0.0083), GSE63805 for miRNA (n = 32, P = 0.018), GSE63384 for DNA methylation (n = 35, P = 0.009), and TCGA independent samples for copy number variations (n = 94, P = 0.0052). Finally, a functional analysis was performed on two survival subgroups to discover genes involved in biological processes and pathways. This is the first study incorporating deep autoencoding and four-omics data to construct a robust survival prediction model, and results show the approach is useful at predicting LUAD prognostication.

1. Introduction

Lung cancer is the most occurring cancer and the leading cause of cancer-related death (Woods et al., 2011; Wu et al., 2014; Siegel et al., 2019). Lung cancer patient's five-year survival rate after diagnosis is lower than most other cancer (Wu et al., 2014; Nanavaty et al., 2014). Non-small cell lung cancer (NSCLC) occupies 85 % of lung cancers, and lung adenocarcinoma (LUAD) is the most occurring subtype of NSCLC (Travis, 2011). Prognosis is a very important first step in making the most informed medical decisions for patients with non-small cell lung cancer. However, the heterogeneity of cancer leads to a very large difference in patient prognosis (Kratz et al., 2012). Therefore, a more critical risk stratification tool could help allocate health care resources better (Onn and Dickey, 2006).

Omics usually refers to a systematic study of the collection of various research objects (generally biomolecules) in biology, such as genomics, epigenomics or transcriptomics. Messenger RNA (mRNA) is a large part of RNA molecules. It is the transcription of DNA, with the corresponding genetic information to provide the information needed for the next translation into protein. MicroRNA, which is a non-coding RNA molecule, is composed by approximately 21–23 nucleotides. The main function of microRNA is to regulate posttranscriptional expression of mRNA by binding its target. DNA methylation is a process of chemical modification of DNA that changes DNA segment activity without altering the sequence. DNA methylation adds methylation to DNA molecules, which typically acts to repress gene transcription. Copy number variation (CNV) is caused by rearrangement of the genome, generally refers to the increase or decrease of the copy number of large fragments of the genome above 1 kb in length. Numerous studies have proved that copy number variations are associated with prognosis in

* Corresponding author at: Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei City, Taiwan.
E-mail addresses: leetzongyi@cuhk.edu.cn (T.-Y. Lee), kaiyao.tw@gmail.com (K.-Y. Huang), a379152468@gmail.com (C.-H. Chuang), nathanlee@tmu.edu.tw (C.-Y. Lee), kevinchang@tmu.edu.tw (T.-H. Chang).

https://doi.org/10.1016/j.compbiolchem.2020.107277
Received 31 March 2020; Accepted 30 April 2020
Available online 12 May 2020

cancer (Kumaran et al., 2017; Hieronymus et al., 2018; Zhang et al., 2018a; Smith and Sheltzer, 2018). In terms of genomic, transcriptomics, epigenomics has been explored to reveal the omics heterogeneity for histologically tumor (Lawrence et al., 2013; Alexandrov et al., 2013; Witte et al., 2014; Yoshihara et al., 2013; Jacobsen et al., 2013; Gentles et al., 2015). The complexity of the tumor genome is a big challenge for the assessment of cancer prognostic.

The Cancer Genome Atlas (TCGA) provides several types of data, with more production of omics data, multi-omics integration is much needed. While multi-omics data integration method has been widely used in subtype identification (Mankoo et al., 2011; Kim et al., 2017; Huang et al., 2017), and recent studies showed that deep learning approach can effectively use non-linear patterns to transform multi-omics data to provide a better prognosis model (Chaudhary et al., 2018; Tan et al., 2015). For instance, Chaudhary and *et al.* proposed a deep learning-based mode on hepatocellular carcinoma (HCC) with autoencoder to combine three-omics data, mRNA, miRNA and methylation (Chaudhary et al., 2018), and transform into new features that can represent the original input data. Two survival subtypes of HCC with significant survival differences were identified by using K-means clustering algorithms.

However, for LUAD, most previous studies focus on exploring the molecular subgroups or facilitating the prognosis analysis based on single-omics data. Therefore, to use of more comprehensive omics information, here we applied four-omics data, RNA-Seq, miRNA-Seq, DNA methylation and copy number variations, to construct an accurate survival risk stratification model for LUAD patients based on deep learning autoencoding approach. Additionally, four independent datasets were collected to validate the robustness of the model. Functional enrichment analysis was performed to discover the biological mechanisms associated to LUAD prognosis.

2. Materials and methods

The workflow of the study was given in Fig. 1, and can be divided into two parts, one part is inferring survival groups and the other one is robustness validation. In training part, first we stacked the four unit-norm scaled matrices as input of autoencoder, and generating new features from autoencoder. Then, we conducted a univariate Cox proportional hazards (Cox-PH) model on each of the new features and selected the features that significantly (log-rank P-value < 0.05) associated with overall survival (OS). Subsequently, a covariate Cox-PH was performed with penalization through the Lasso regression to select features by minimizing the log partial likelihood (Tibshirani, 1997). After all, we used these selected features to cluster samples using K-means algorithm, and construct a random forest model with cross-validation for model evaluation. In validation part, we built supervised classification models to predict 4-omics independent datasets. ANOVA (F-value) and the DESeq2 of R package (Love et al., 2014) were applied to selected the top features between survival subgroups, and these features were used for construction of random forest models for each type of omics data and model validation by using independent datasets.

2.1. Dataset and data preprocessing

Four types of omics data were used in this study, including mRNA, miRNA, DNA methylation and copy number variations. TCGA 4-omics data was used as training data and four independent datasets (GSE81089, GSE63805, GSE63384, and TCGA independent CNV) were used for validation. The clinical information of five datasets is listed in Additional file 1: Table S1.

Training dataset: TCGA Lung Adenocarcinoma (LUAD) data was obtained from the Genomics Data Commons web portal, consisting 399 LUAD samples with level 3 Illumina HiSeq RNA sequencing (RNA-Seq) gene expression data, Illumina HiSeq miRNASeq miRNA sequencing (miRNA-Seq) data, HumanMethylation450 DNA methylation data, Affymetrix SNP 6.0 array Copy number variations data and the clinical information data. Both of the RNA-Seq gene expression data and the miRNA-Seq data have been converted into FPKM and RPM, respectively. For methylation data, probe sites in CpG island within 1500 base pairs (bp) upstream of the transcriptional start site (TSS) were remained. The beta values were transformed into M values (Ching et al., 2015, 2014) using the lumi package in R, and mapped them to gene level by averaging their methylation values. For Copy number variations data, Gistic2.0 was applied to generate each genes copy number value by default extreme method (Mermel et al., 2011). The extreme method chooses whichever of min or max is furthest from diploid. In dealing with missing value, two steps were performed on mRNA, miRNA and CNV data (Mermel et al., 2011). First, features which has more than 20 % NA or zero value were removed. Second, samples which has more than 20 % NA or zero value were removed. For methylation data, features with any NA or zero value were removed.

Independent datasets: For the independent data, three datasets were collected from the Gene Expression Omnibus (GEO) database and one CNV independent dataset was collected from TCGA LUAD independent samples which possessed CNV information but lack of mRNA, miRNA or methylation information. In GSE81089 (RNA-Seq, Illumina HiSeq 2500), fresh frozen non-small cell lung cancer (NSCLC) tumor tissue with RNA-Seq and survival data from 198 samples and surgically treated at the Uppsala University Hospital, Uppsala, Sweden in

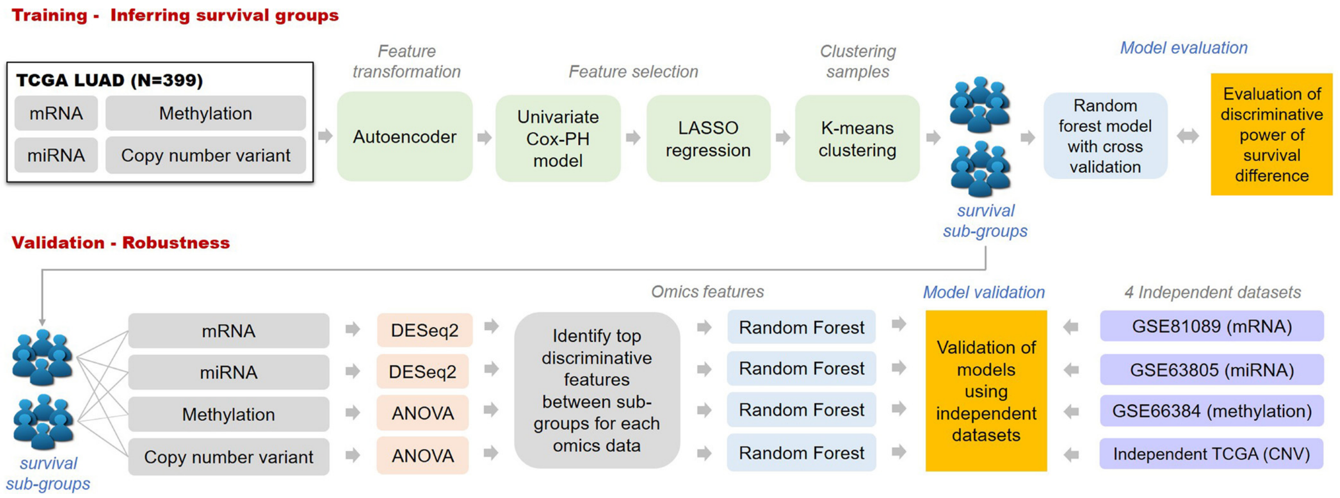


Fig. 1. The workflow of the study.

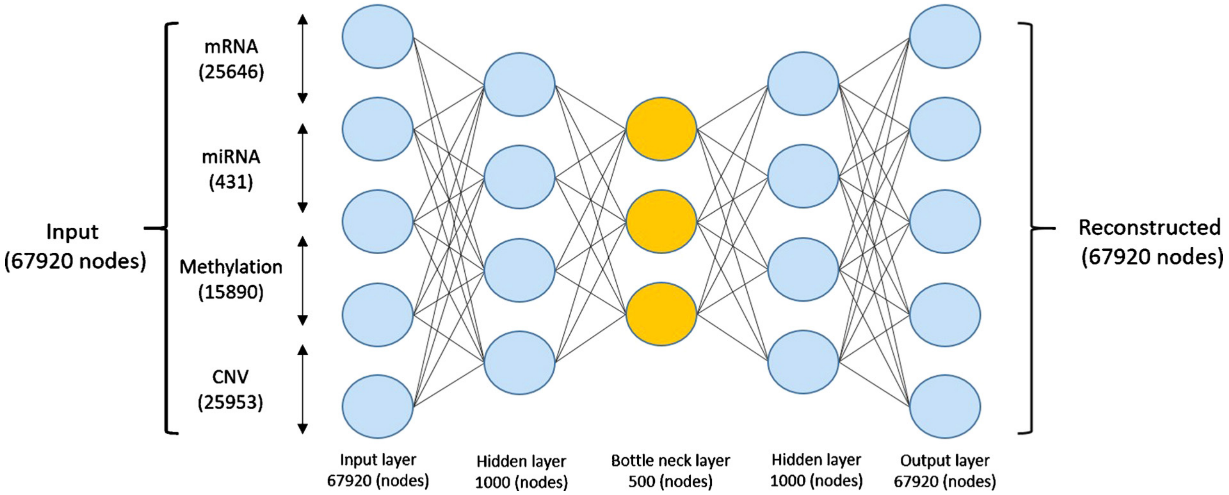


Fig. 2. Autoencoder architecture for integration of 4-omics data of LUAD.

2006 – 2010. RNA-Seq gene expression data was quantified into FPKM. In GSE63805 (miRNA, NanoString nCounter Human miRNA assay), 32 LUAD samples tissue with miRNA expression and clinical data were downloaded. In GSE63384 (DNA methylation, Illumina HumanMethylation27 BeadChip), 35 LUAD samples tissue with DNA methylation and clinical data were used. In TCGA independent CNV dataset (copy number variations, Affymetrix SNP 6.0 array), 94 LUAD samples with copy number variations data and clinical data were used.

Feature transformation – Autoencoder: Four omics preprocessed TCGA LUAD data for a total of 399 patients were used as input to the antoencoder. We used four-omics matrices that are unit-norm scaled by sample and stacked them to a new matrix (Liu et al., 2016). An autoencoder is one of unsupervised artificial neural network use to learn efficient data by minimizing the error between the reconstructed output data and the original input data. The most important purpose of the autoencoder is to learn a vital feature of the input while reserving the necessary information. There are five layers in our model, one is input layer, three hidden layers (1000, 500, 1000) of the middle one is bottleneck layer, last is output layer. For the input layer $x = \{x_1, x_2, \cdots, x_m\}$ of the dimension m, y= $\{y_1, y_2, \cdots, y_n\}$ of the number of node is n in the hidden layer. We used relu as activation function between each layer, except for output layer, we used tanh, for a layer i,and could be expressed as:

$$y_i = \text{relu}(W_i \bullet x + b_i) = f_i(x)$$

So for the output $\hat{x}$ could be given by:

$$\hat{x} = \tanh(f_3(f_2(f_1(x))) + b) = F(x)$$

The error between the input x and the output $\hat{x}$ was measured by the function binary_crossentropy. That is:

$$\text{loss}(x, \hat{x}) = -\frac{1}{n}\sum_{i=1}^{n}[x_i \log \hat{x}_i + (1 - x_i)\log(1 - \hat{x}_i)]$$

Dealing with overfitting, L1 regularization β_1, and L2 regularization β_2 penalty on the weight vector W_i and the layer without the output layer respectively, L1 and L2 regularization were set to 0.001 and 0.0001. That is:

$$\text{L}(x, \hat{x}) = \text{loss}(x, \hat{x}) + \sum_{i=1}^{k}(\beta_1 \|W_i\|_1 + \beta_2 \|F_{k-1}(x)\|_2^2)$$

An autoencoder was constructed using Python with Keras library (https://github.com/fchollet/keras). The bottleneck layer of the autoencoder was extracted to represent new features from the original 4-omics data. Adam with learning rate 0.001 as the optimizers was used to find the minimum error of the model with 3 epochs and 0.5 dropout.

2.2. Analysis of survival associated features and survival subgroups

Univariate Cox regression analysis: After the original number of features were reduced by the autoencoder to 500 new transformed features get from the bottleneck layer, a univariate Cox-PH model on each feature was produced by the autoencoder, and top features with significantly association with survival (P-value < 0.05) were selected using the R *survival* package (Chen et al., 2014).

Lasso regression: Lasso regression was used to perform penalization, and the model variables may shrink all the way to zero and result in variable selection process. Lasso regression performs L1 regularization, which involves penalizing the absolute size of the regression coefficients. Some coefficients can become zero and eliminated from the model. A 10-fold cross validation was performed to estimate the best lambda using the R *glmnet* package (Friedman et al., 2010).

K-means clustering algorithms: To identify survival subgroups, here we used the reduced features which associated with survival to make clustering using K-means algorithms. The optimal number of clusters was determined by Silhouette index (Rousseeuw, 1987) and elbow methods (Zhang et al., 2018b).

2.3. Prediction model construction

To elucidate the robustness of survival subgroups, differentially expressed omics features were identified for construction of supervised classification models using random forest algorithm (Breiman, 2001a), and validated using independent datasets.

Differentially expressed genes: To identify differentially expressed gene between survival subgroups for each omics data, we used the R *DESeq2* package to find the differentially expressed mRNA and miRNA (absolute fold change > 1.5 and P-value < 0.05). For methylation and CNV data, ANOVA test was applied for identify differentially expressed genes based on M value and segmentation mean value, respectively.

Random forest: Random Forest is an ensemble learning method for classification (Breiman, 2001b). It constructs multiple decision trees and output the mode of the classification of the individual trees. Each individual decision tree in the random forest spits out a class prediction and the class with the most votes prediction was chosen as a result. Random forest can deal the overfitting for the decision tree.

Feature selection and scaling: In order to applying data to build random forest models, the top features that are the most different between the subgroups were collected. For mRNA and miRNA data, top 100 and 50 features that have the greatest difference by log2-fold-change using the R *Deseq2* package (Love et al., 2014) were collected, respectively. For methylation data, top 50 features based on ANOVA F-

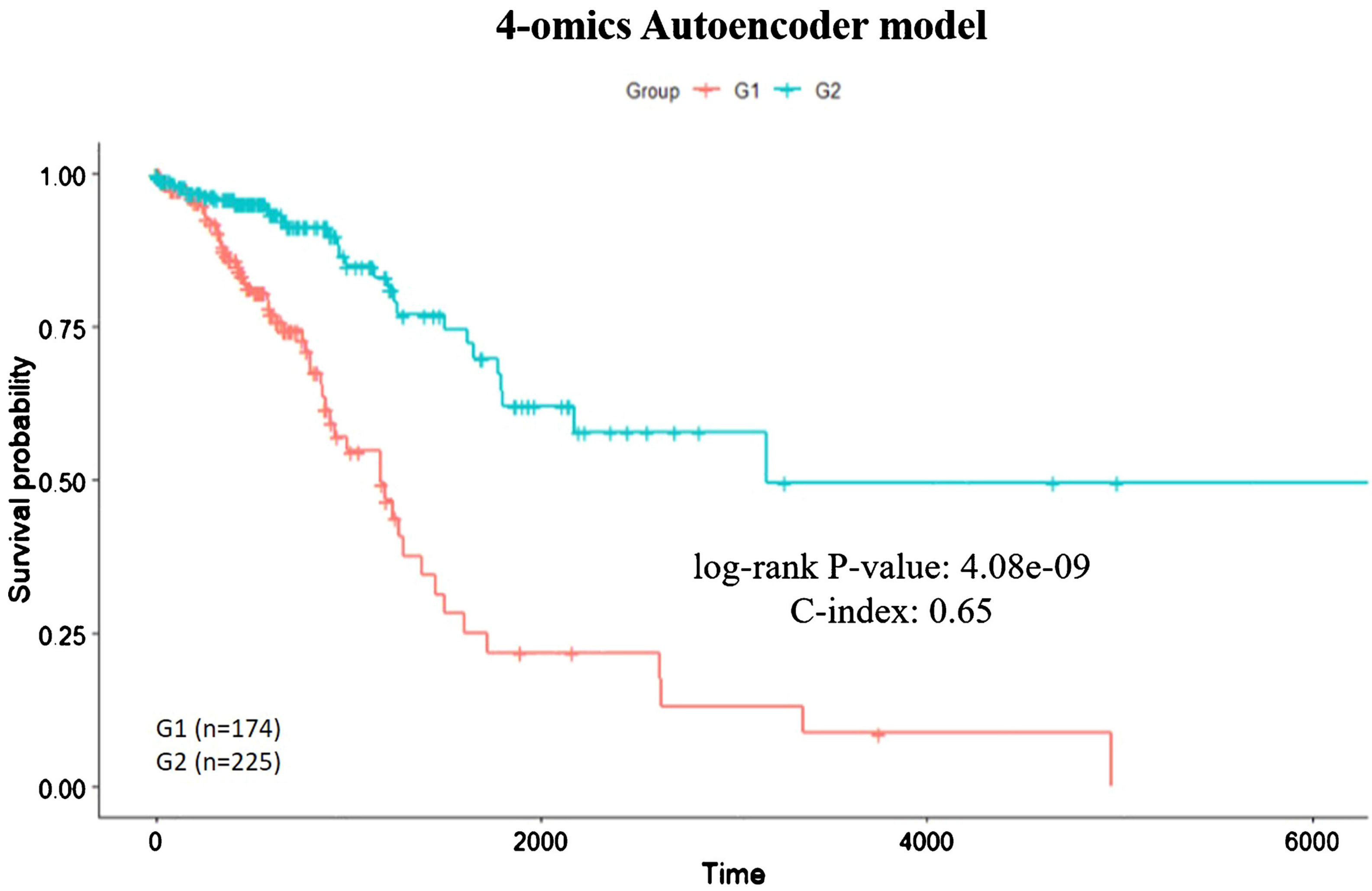


Fig. 3. Kaplan-Meier Curves between two survival subgroups of 4-omics autoencoder model.

Table 1
Evaluation of prediction model of survival subgroups on training dataset and testing dataset.

	Log-rank P value[a]	C-index[a]
Training dataset	1.3e-05 ± 6.2e-06	0.65 ± 4.5e-04
Testing dataset	9.3e-03 ± 0.005	0.64 ± 0.02

[a] Mean ± Standard deviation.

values were remained. For CNV data, more features (top 400) were collected based on ANOVA F-values, because the genes in a close region may possess similar copy number values.

Two steps scaling on both training dataset and independent datasets were performed, and the normalization procedures showed overall decent better performance for independent datasets in previous study (Chaudhary et al., 2018). We first applied Median absolute deviation (mad) scaling on both training dataset and independent datasets, this approach was used to scale samples from RNA-seq data previously (Zhang et al., 2014). Next, for mRNA, StandardScaler normalization was applied on training dataset and independent datasets using the means and the standard deviations of the training dataset (Angermueller et al., 2016). For miRNA data, DNA methylation data and CNV data, the l2-norm normalization was used for both training dataset and independent datasets (Chaudhary et al., 2018). The *scikit − learn* class were used to implement processing on these dataset (Pedregosa et al., 2011).

2.4. Performance evaluation

Two values, C-index, Log-rank P value, which could closely reflect the accuracy of the survival predictions in our survival subgroups, were applied for performance evaluation.

Concordance index: The concordance index (C-index) can be viewed as the score of all individual pairs that correctly predict the survival time (Raykar et al., 2007), and based on the Harrell C statistic (Harrell et al., 1996). The calculation method of C-index is to randomly combine all the samples in the studied data into two pairs, if one with a longer survival time has a predicted survival time longer than the other, or a predicted survival probability is high. Then the prediction result is consistent with the actual result. We construct Cox-PH model and calculate the C index using the R *survival* package (Pedregosa et al., 2011).

Log-rank P value of Cox-PH regression: The Log-rank test is the nonparametric hypothesis test to compare that two or more survival curves. The Kaplan-Meier (KM) survival curves of the two risk groups were plotted and calculated the log-rank P values of the differences in survival between them using the R *survminer* package (Chang et al., 2011).

Additionally, three approaches, Cox-PH model, PCA model, and iClusterPlus model, were constructed for performance comparison with deep learning-based model. In the first approach, the top features were collected in all 4-omics data according to the Log-rank P value computed by univariate Cox-PH model, and K-means was used to cluster the samples. In the second approach, the principal component analysis (PCA) were performed, due to the maximum number limitation of the principal component, 100 components were applied. The subsets (Chen et al., 2002) of PCA features associated with survival selected by univariate Cox-PH models were identified, and clustered samples using the same procedure in the first approach. For the third approach, we first selected all features in 4-omics data significantly associated with survival using univariate Cox-PH models, and then clustered samples through the integrative clustering plus (iClusterPlus) analysis (Shen et al., 2009).

2.5. Functional analysis

pregulated and downregulated expression genes were used for the Ingenuity Pathway Analysis (IPA) which is All-in-one, web-based analysis software and database. IPA identify local networks that are particularly enrich in the input gene set by using the computational algorithms. The p value that measures overlap of observed and predicted gene sets determined by Fisher's exact test, and the Z-score analysis method (Kramer et al., 2014) assessing the match of the activation states of predicted functions, pathways, or transcriptional regulators (> :0 increased, < 0: decreased).

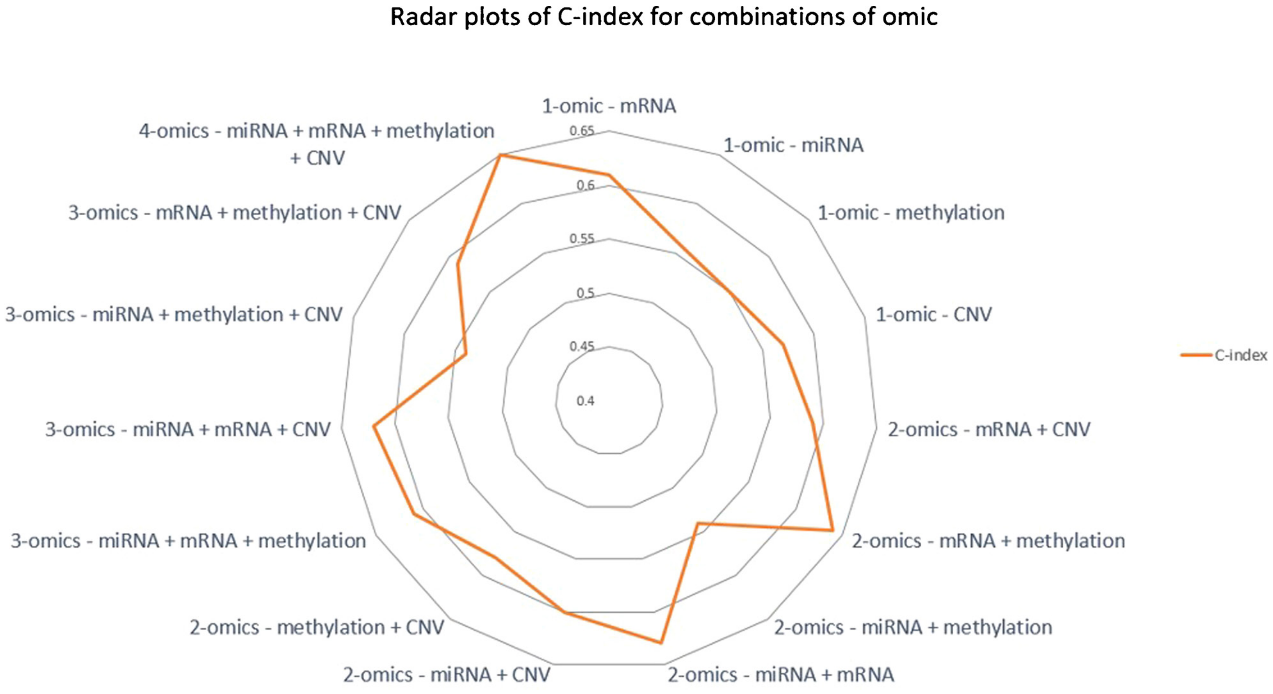


Fig. 4. Rader plot of C-index for the models with different omics combinations.

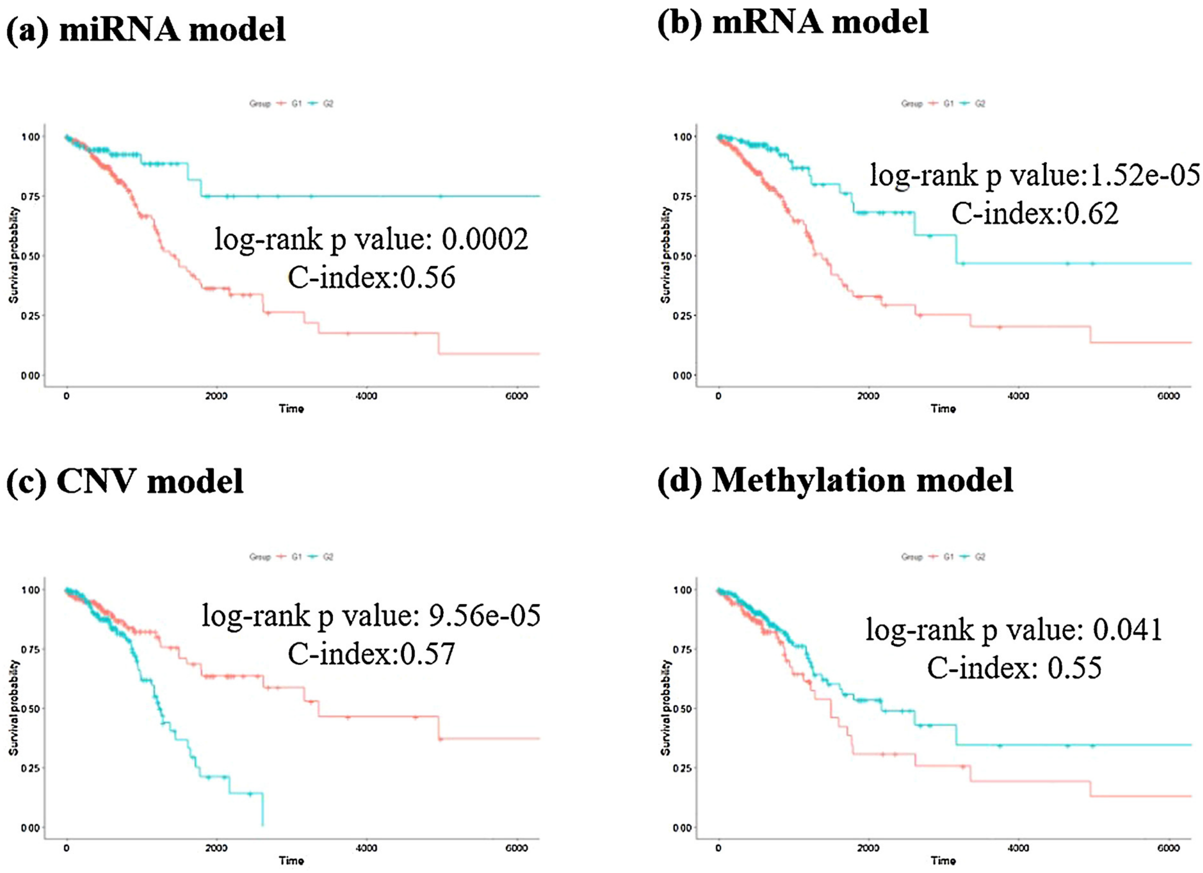


Fig. 5. Kaplan-Meier curves of the single-omics autoencoder-based models.

3. Results

3.1. Two differential survival subgroups are identified in TCGA 4-omics data

Three hundred and ninety-nine LUAD samples were obtained from TCGA that both have RNA-Seq, miRNA-Seq, DNA methylation and CNV data. For these 399 samples, we performed preprocessing on the data as mentioned in the materials and methods, and got 25,646 genes from RNA-Seq, 431 miRNAs from miRNA-Seq, 15,890 features from DNA methylation, and 25,953 features from CNV data for the input. Fig. 2 showed the architecture of the autoencoder we used, and represented the deep learning framework stacked these four omics type features together. The value of 500 nodes from the bottleneck layer was used as new features for Cox-PH model. We then built univariate Cox-PH regression on each of the 500 features to identify significantly (Wald test P-value < 0.05) associated with overall survival and retained 95 features. Next a multivariate Cox-PH model was conducted and performed penalization through Lasso regression, and 22 features were obtained and used to cluster the samples (Additional file 2: Fig. S1). For the optimal number of clustering, we used the silhouette index and the elbow method to determine. As shown in Additional file 3: Fig. S2, K = 2 achieved the highest clustering score. We then assessed the prognostic difference between these two subgroups by the survival analysis using the full TCGA LUAD data, and Fig. 3 showed the difference between two subgroups were extremely significant (log-rank P-value = 4.08e-09). Moreover, our deep learning framework model also generated a good concordance index (C-index) value of 0.65.

Additionally, to check the robustness of the performance on the labels and features generated from autoencoder-based model, 3-folds cross validation with 5 repetitions was performed to get the mean C-index and the Log-rank P value. As shown in Table 1, on average, the survival difference between two survival subgroups showed a good C-index (0.65 ± 4.5e-04) and significant Log-rank P value (1.3e-05 ± 6.2e-06) on training dataset. Similarly, the 4-omics held-out testing also got a steady C-index (0.64 ± 0.02) and Log- rank P value (9.3e-03 ± 0.005) on testing dataset.

3.2. Performance for Autoencoder-based model with different hyper-parameter

Different hyper-parameter combinations were applied for autoencoding model construction, and the performances were shown in Additional file 4: Table S2. With more than three hidden layers, with a higher number of hidden nodes, or with more epochs dramatic decrease the performances instead of improving. Conversely, with more number of bottleneck layer nodes can usually increase the performance. Fig. 4 presented Rader plots of C-index for the models with different omics combinations, and results showed 4-omics model achieved the highest C-index. The different omics combinations were compared in terms of prognostic performance measured by C-index. Among all combinations, mRNA shows the highest performance. The C-index of most of the combination adding mRNA can be higher than 0.6. Except for two combinations (mRNA + CNV and mRNA + methylation + CNV) with the C-index of 0.59. Incorporating miRNA can increase the C-index, despite the single miRNA with a low prediction power. Interestingly, C-index for combination of mRNA and DNA methylation reached 0.64, but the addition of CNV reduces the prognosis to 0.59. Results showed that the prognostic ability of each omics might be not an additive relationship, but a combined connection.

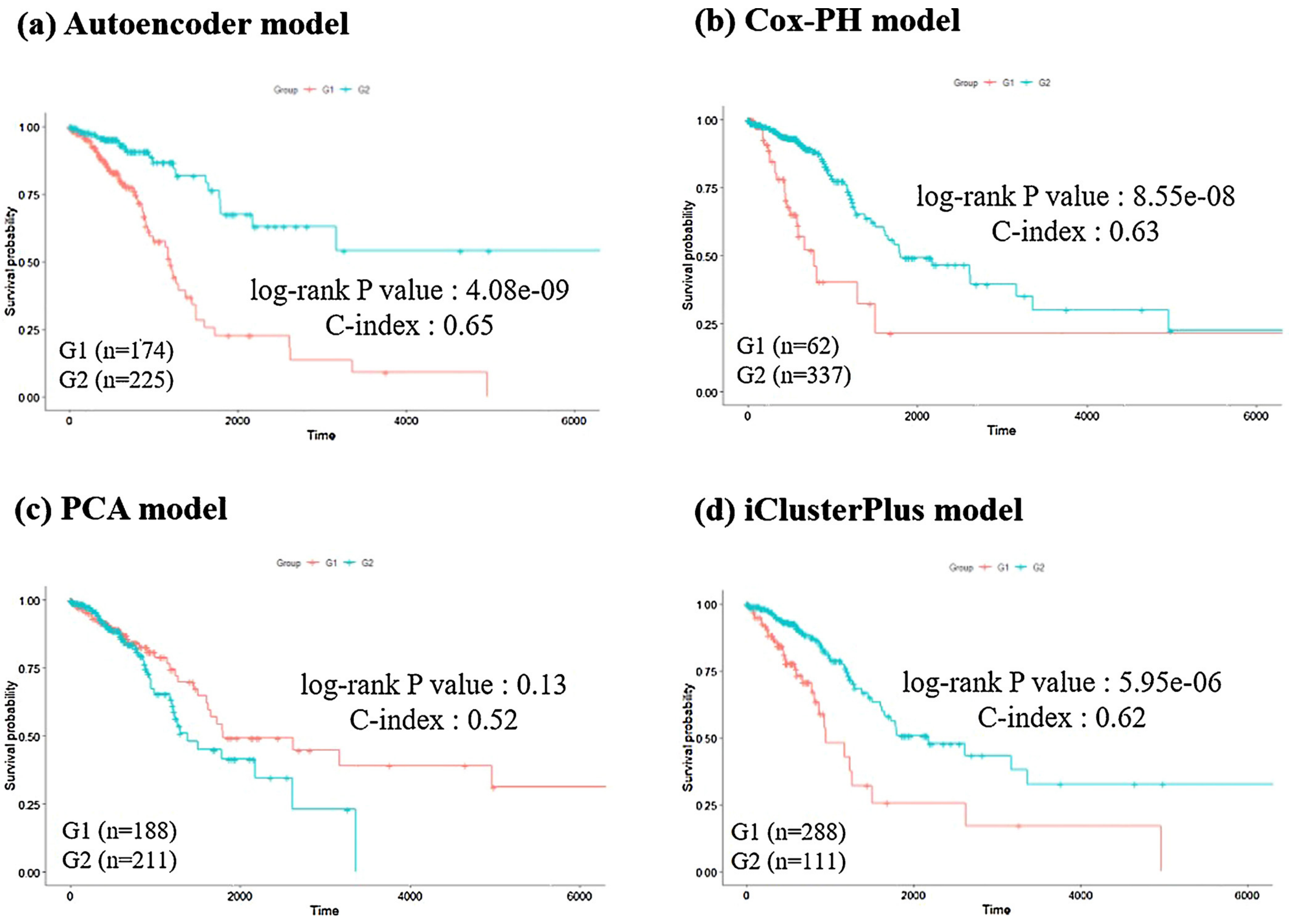


Fig. 6. Kaplan-Meier curves of (a) autoencoder model (b) Cox-PH model (d) PCA model and (D) iClusterPlus model.

3.3. *Autoencoder-based model using multi-omics data performed better than models using single omics data or other approaches*

For comparison of prediction performance between the model using 4-omics data and single omics data, Fig. 5 depicted the Kaplan-Meier survival curves of survival subgroups generated using the autoencoder-based models with mRNA, miRNA, DNA methylation and CNV, respectively. As compared to multi-omics model (Fig. 3), 4-omics model has the better C-index (0.65) and log-rank P-value (4.08e-09), which demonstrated multi-omics model perform better than single-omics model.

For comparison of prediction performance between autoencoder approach and three alternative approaches (Cox-ph, PCA and iClusterPlus), Fig. 6 showed the Kaplan-Meier survival curves of survival subgroups generated using different approaches. In univariate Cox-PH model (Fig. 6b), the top 22 features, same number of features used in autoencoder model, were selected from 4-omics data, and used for survival subgroups clustering. Results showed the model gave a significant P value of 8.55e-08 which is comparable with autoencoder-based model (P value 4.08e-09, Fig. 6a). The principal component analysis (PCA) model (Fig. 6c) was used as the conventional dimension reduction method instead of autoencoder. One-hundred principal components were obtained, and selected by univariate Cox-PH. Only 3 PCA features were found associated with survival, and therefore it performed a non-significant discriminative power of survival subgroups (Log-rank P value = 0.13). The iClusterPlus model (Fig. 6d) clustered the sample into groups based on multi-type genomic data directly. iClusterPlus show a good predicted power with Log-rank P value of 5.95e-06 and C-index value of 0.62 in two survival subgroups, but still less significant as compared to the autoencoder-based model. The results demonstrated that multi-omics integration based on autoencoder is superior to these alternative approaches.

3.4. *Validate survival subgroups in independent datasets*

To test the robustness of the classification at predicting prognosis, four independent datasets from GEO (GSE81089, GSE63805 and GSE63384) and TCGA CNV independent samples were collected for validation. The RNA-Seq data from GSE81089 (n = 198) achieved a good Log-rank P value of 0.0083 between two predicted survival subgroups (Fig. 7a). The miRNA data from GSE63805 (n = 32) and DNA methylation data from GSE63384 (n = 35) achieved log-rank P-value 0.018 and 0.009 (Fig. 7**b** and Fig. 7c), respectively. The CNV dataset (n = 94) achieved a significant difference with Log-rank P value of 0.0052 between two predicted survival subgroups (Fig. 7d). The validation results by independent datasets demonstrate the robustness of our approaches for lung adenocarcinoma prognostication.

3.5. *Adding clinical information improve performance of autoencoder-based model*

Additional file 5: Table S3 showed the associations between survival subgroups and clinical factors (gender, age, stage, and smoking history), and results demonstrated that stage (P = 0.03) were marginally associated with survival subgroups. Additionally, we assessed the performance of autoencoder-based model combined with clinical variables. As shown in Table 2, the prognostic power by using omics data (C-index = 0.65) was higher than using clinical factors, and the

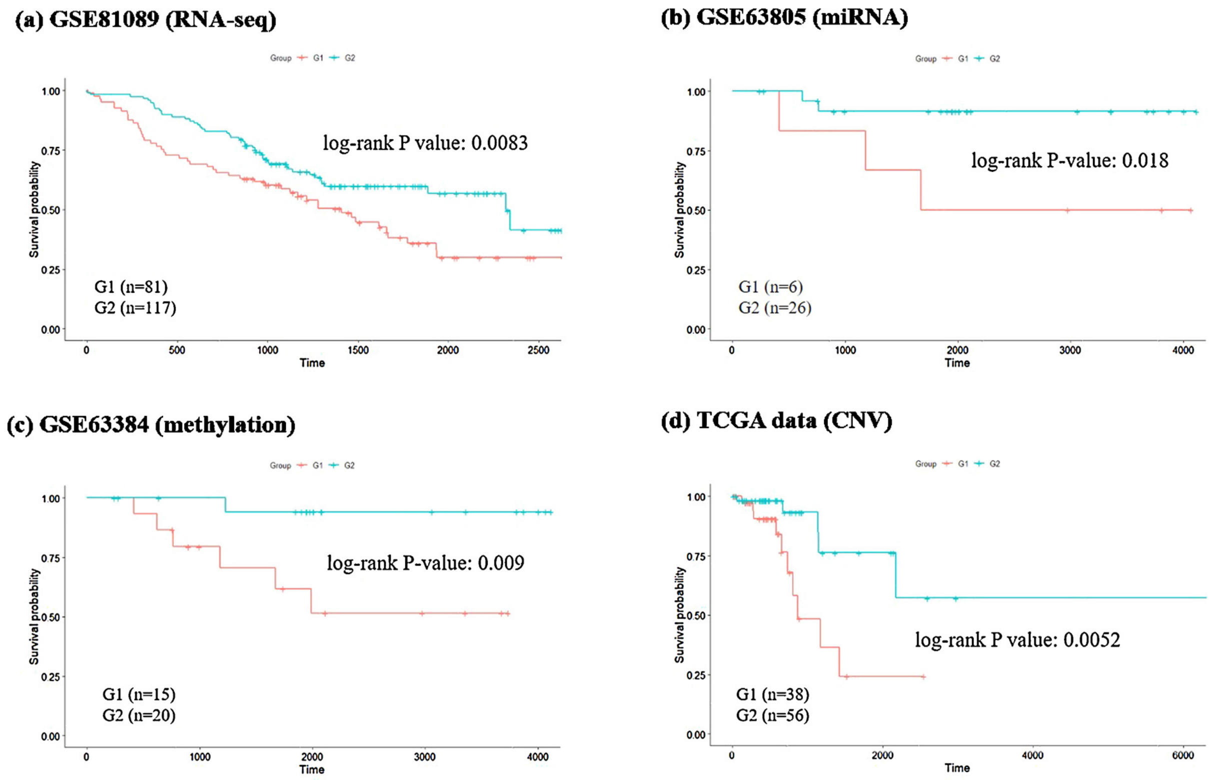


Fig. 7. Kaplan-Meier curves of the models validated using the external independent datasets: (a) GSE81089, (b) GSE63805, (c) GSE63384 and (d) TCGA independent CNV dataset.

Table 2
C-index performance of the models generated from 4-moics data and clinical factors.

	C-index value of Combined features[a]	C-index value of clinical factor
4-omics	0.65	NA
4-omics + Gender	0.68	0.46
4-omics + Age	0.68	0.57
4-omics + Stage	**0.70**	**0.63**
4-omics + Smoking history	0.66	0.53

Bold values represent the highest C-index value of combined features and clinical factors.

[a] Using both 4-omics features and clinical factor.

combination of omics data and stage information achieved a high C-index of 0.7. Different from previous reports in liver cancer (Chaudhary et al., 2018), adding clinical variables improved the C-index in lung adenocarcinoma prognostication. However, combining smoking history was not significantly improving the C-index (0.66). It might be conjectured that the influences of clinical factors on genetic may be variable on different people. Therefore, it is not definite to completely reflect the clinical influence on the genomic features.

3.6. Functional analysis of the survival subgroups in LUAD samples

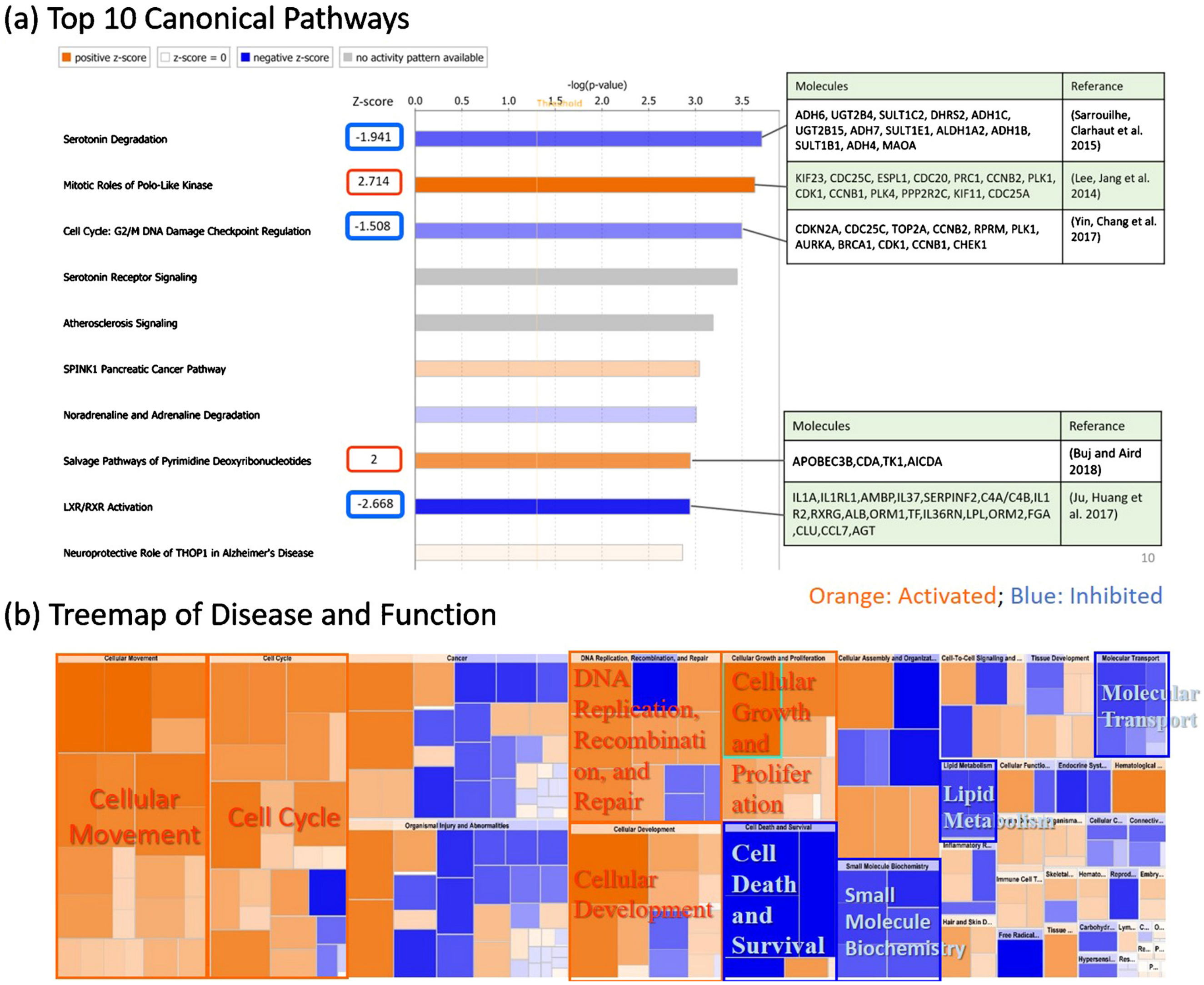
The differential gene expression between the two identified subgroups were determined by the R *DESeq2* package and used for functional analysis. As shown in Additional file 6: Figs. S3, 857 upregulated and 1064 downregulated genes were obtained in high risk survival subgroup (G1), and for miRNA, 20 upregulated and 46 downregulated genes were obtained. In Table 3 showed the top 5 differential genes according to log2 fold change values. At least one study per gene (UPK1B, REG4, PCSK1, PCSK2, KRT6A, hsa-mir-375, hsa-mir-509 − 2, hsa-mir-323b, hsa-mir-509 − 1, hsa-mir-449a) has been shown to be associated with tumor development or poor clinical outcome (Wang et al., 2018; Kaprio et al., 2014; Demidyuk et al., 2013; Holloway et al., 2015; Harris et al., 2012; Chen et al., 2018; Meng et al., 2018; Du et al., 2018; Bou Kheir et al., 2011). In top five methylation and CNV genes selected based on the ANOVA P value, several literatures have pointed that these genes (FER1L4, MFI2, ZBTB18, SSTR2, BPTF, C17orf80) were associated with tumor development or poor prognosis (Xia et al., 2015; Yin et al., 2016; Fedele et al., 2017; Zhou et al., 2009; Zhao et al., 2019; Kim et al., 2018). The detailed top genes of mRNA, miRNA, DNA methylation and CNV data were listed in Additional file 7: Table S4.

Ingenuity pathway analysis (IPA) were applied for functional analysis of the differentially expressed genes. The top 10 enriched categories of canonical pathways were listed in Fig. 8A. the activation z-score analysis method was applied to measure the activation state of the pathways or functions influenced by differentially expressed genes. The z-score makes predictions about a biological function has been significantly activate or inhabit (> 0: increased, < 0: decreased). In practice, the absolute z-score ≥ 2 was considered significant. There are two pathways with significant high z-scores, which are "Mitotic Roles of Polo-Like Kinase" (z-score = 2.714) and "Salvage Pathways of Pyrimidine Deoxyribonucleotides" (z-score = 2). Three pathways with low z-scores, "LXR/RXR Activation" (z-scores = -2.668), "Serotonin Degradation" (z-score = -1.94) and "Cell Cycle: G2/M DNA Damage Checkpoint Regulation" (z-score = -1.508). All of the above pathways have been shown to be associated with tumor progression or poor prognosis (Sarrouilhe et al., 2015; Lee et al., 2014; Yin et al., 2017; Buj and Aird, 2018; Ju et al., 2017). The details of LUAD survival subgroups involved canonical pathways were in Additional file 8: Table S5. Additionally, the treemap of LUAD involved diseases and functions was depicted in Fig. 8B. Results showed that "Cellular Movement", "Cell Cycle", "DNA Replication, Recombination, and Repair", "Cellular Development" and "Cellular Growth and Proliferation" are significantly activated, and numerous biological functions were inhibited, such as

Table 3
Top five differential genes between survival subgroups in each omics data.

Omics Type	Gene	Log2FC[a] or mean diff [b]	P value	Reference
mRNA	UPK1B	4.76	3.82E-53	(Wang et al., 2018)
	REG4	−3.81	2.32E-25	(Kaprio et al., 2014)
	PCSK1	−3.45	9.56E-35	(Demidyuk et al., 2013)
	PCSK2	−3.4	1.85E-28	
	KRT6A	3.38	1.99E-32	(Holloway et al., 2015)
miRNA	hsa-mir-375	−1.38	2.62E-18	(Harris et al., 2012)
	hsa-mir-509 − 2	−1.33	4.21E-08	(Chen et al., 2018)
	hsa-mir-323b	−1.32	1.84E-10	(Meng et al., 2018)
	hsa-mir-509 − 1	−1.31	1.09E-07	(Du et al., 2018)
	hsa-mir-449a	−1.24	7.90E-08	(Bou Kheir et al., 2011)
DNA methylation	FER1L4	0.54	3.57E-12	(Xia et al., 2015)
	MFI2	−0.41	4.88E-10	(Yin et al., 2016)
	ZBTB18	0.19	3.66E-08	(Fedele et al., 2017)
	RP11 − 1E11.1	0.59	4.54E-08	
	RP5 − 998N21.7	0.28	6.49E-07	
CNV	SSTR2	0.17	1.81E-08	(Zhou et al., 2009)
	BPTF	0.18	2.21E-08	(Zhao et al., 2019)
	NOL11	0.17	2.37E-08	
	C17orf80	0.17	2.96E-08	(Kim et al., 2018)
	COG1	0.17	3.09E-08	

[a] Log2FC: Log2FoldChange(G1/G2) for mRNA and miRNA.
[b] Mean diff: mean(G1) – mean(G2) for methylation and CNV.



Molecules	Referance
ADH6, UGT2B4, SULT1C2, DHRS2, ADH1C, UGT2B15, ADH7, SULT1E1, ALDH1A2, ADH1B, SULT1B1, ADH4, MAOA	(Sarrouilhe, Clarhaut et al. 2015)
KIF23, CDC25C, ESPL1, CDC20, PRC1, CCNB2, PLK1, CDK1, CCNB1, PLK4, PPP2R2C, KIF11, CDC25A	(Lee, Jang et al. 2014)
CDKN2A, CDC25C, TOP2A, CCNB2, RPRM, PLK1, AURKA, BRCA1, CDK1, CCNB1, CHEK1	(Yin, Chang et al. 2017)

Molecules	Referance
APOBEC3B,CDA,TK1,AICDA	(Buj and Aird 2018)
IL1A,IL1RL1,AMBP,IL37,SERPINF2,C4A/C4B,IL1R2,RXRG,ALB,ORM1,TF,IL36RN,LPL,ORM2,FGA,CLU,CCL7,AGT	(Ju, Huang et al. 2017)

Fig. 8. LUAD survival subgroups involved (a) Top 10 canonical pathways (b) treemap of disease and biological function.

"Cell Death and Survival", "Small Molecule Biochemistry", "Lipid Metabolism" and "Molecular Transport". The details of diseases and biological functions analysis results were provided in Additional file 9: Table S6.

4. Discussion

Lung adenocarcinoma is the most commonly diagnosed type of lung cancer, although there are many studies for identifying lung adenocarcinoma subtypes, the inclusion of patient survival information in the identification of subtypes of lung adenocarcinoma cancer is rarely reported before. The previous study (Chaudhary et al., 2018) demonstrated deep learning-based model was applicable to the risk stratification of survival of HCC patients by using three omics data, mRNA, miRNA and methylation. They significantly distinguish the survival differences between two groups of HCC patients. In this work, we significantly differentiated the survival differences between two survival groups of lung adenocarcinoma patients (Log-rank P value = 4.08e-09) by using four-omics data, mRNA, miRNA, methylation and CNV, and reached a good performance with C-index of 0.65 between two survival subgroups.

TP53, EGFR, KRAS, ALK is the most frequently mutated genes in LUAD. Here we analyzed their mutation rate in two survival subgroups. As shown in Additional file 10: Table S7, TP53 and ALK is significantly associated with two survival subgroups (P value = 0.006 and P value = 0.015, respectively). Samples in aggressive subtype G1 (high-risk survival) possess higher mutation rate in TP53 (58 %) and ALK (10 %) as compared with samples in G2 (44 % and 4% in TP53 and ALK, respectively). TP53 has been reported to be associated with poor survival in LUAD (Gu et al., 2016; Mogi and Kuwano, 2011; Mitsudomi et al., 1993; Fukuyama et al., 1997).

Several pathways identified though IPA analysis, such as Serotonin has been reported as a mitogenic factor in cells. High doses of serotonin have a role in stimulating tumor growth, whereas low doses of serotonin can inhibit tumor growth. Therefore, the dysregulation of serotonin degradation affects the development of cancer (Sarrouilhe et al., 2015). Polo-like kinase family consists of 5 members (Plk1-Plk5), they participate in the multiple functions of cell division. Recent studies have pointed out that PLKs are related in the development of many tumors and overexpressed in several types of cancer (Lee et al., 2014). The Cell Cycle: G2/M checkpoint regulation was dysregulated. It is the second checkpoint in the cell cycle, which can repair the DNA by preventing damaged DNA cells from entering the M phase. This regulation is vital to prevent cells from undergoing malignant transformation (Yin et al., 2017). Two biosynthetic pathways to produce deoxyribonucleotide triphosphates is de novo and salvage. To make sure rapid replication of genome, cancer cells increase the biosynthesis of dNTP. The upregulation of salvage pathways of pyrimidine deoxyribonucleotides helps tumor growth (Buj and Aird, 2018). As a potential target in cancer therapy, Liver X receptors (LXRs) is involved in many types of cancer, inhibiting tumor growth by regulating different signaling. The loss of gene expression here inhibited pathway of LXR/RXR activation, which may lead to the proliferation of cancer cells (Ju et al., 2017).

5. Conclusion

In this work, we successfully identified two survival subgroups from LUAD with a significant survival differences (Log-rank P value: 4.08e-09) and good model fitness (C-index: 0.65). Furthermore, autoencoder-based model compared with other three approaches demonstrate our approaches was superior than other approaches. Validation through independent dataset showed the robustness of the model. Numerous critical genes, pathway and function were identified related to tumor development or prognosis. This is the first study incorporating deep autoencoding and four-omics data to construct a robust survival prediction model for lung adenocarcinoma. The study significantly contributes to the current understanding of LUAD prognosis, and shows our approach could be useful at predicting LUAD prognostication.

Declaration of Competing Interest

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

Acknowledgement

This work was supported by the Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.compbiolchem.2020.107277.

References

Alexandrov, L.B., et al., 2013. Signatures of mutational processes in human cancer. Nature 500 (7463), 415–421.
Angermueller, C., et al., 2016. Deep learning for computational biology. Mol. Syst. Biol. 12 (7).
Bou Kheir, T., et al., 2011. miR-449 inhibits cell proliferation and is down-regulated in gastric cancer. Mol. Cancer 10 (1), 29.
Breiman, L., 2001a. Random forests %. J Mach. Learn. 45 (1), 5–32.
Breiman, L., 2001b. Mach. Learn. 45 (1), 5–32.
Buj, R., Aird, K.M., 2018. Deoxyribonucleotide triphosphate metabolism in Cancer and metabolic disease. Front. Endocrinol. (Lausanne) 9, 177.
Chang, T.H., et al., 2011. Characterization and prediction of mRNA polyadenylation sites in human genes. Med. Biol. Eng. Comput. 49 (4), 463–472.
Chaudhary, K., et al., 2018. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin. Cancer Res. 24 (6), 1248–1259.
Chen, X., et al., 2002. Gene expression patterns in human liver cancers. Mol. Biol. Cell 13 (6), 1929–1939.
Chen, J.M., et al., 2014. Effects of statins on incident dementia in patients with type 2 DM: a population-based retrospective cohort study in Taiwan. PLoS One 9 (2), e88434.
Chen, Q., et al., 2018. Identification of differentially expressed miRNAs in early-stage cervical cancer with lymph node metastasis across the cancer genome atlas datasets. Cancer Manag. Res. 10, 6489–6504.
Ching, T., et al., 2014. Genome-wide hypermethylation coupled with promoter hypomethylation in the chorioamniotic membranes of early onset pre-eclampsia. Mol. Hum. Reprod. 20 (9), 885–904.
Ching, T., et al., 2015. Genome-scale hypomethylation in the cord blood DNAs associated with early onset preeclampsia. Clin. Epigenetics 7, 21.
Demidyuk, I.V., et al., 2013. Alterations in gene expression of proprotein convertases in human lung cancer have a limited number of scenarios. PLoS One 8 (2), e55752.
Du, P., et al., 2018. MicroRNA-509-3p inhibits cell proliferation and invasion via downregulation of X-linked inhibitor of apoptosis in glioma. Oncol. Lett. 15 (1), 1307–1312.
Fedele, V., et al., 2017. Epigenetic regulation of ZBTB18 promotes glioblastoma progression. Mol. Cancer Res. 15 (8), 998–1011.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.
Fukuyama, Y., et al., 1997. K-ras and p53 mutations are an independent unfavourable prognostic indicator in patients with non-small-cell lung cancer. Br. J. Cancer 75 (8), 1125–1130.
Gentles, A.J., et al., 2015. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat. Med. 21 (8), 938–945.
Gu, J., et al., 2016. TP53 mutation is associated with a poor clinical outcome for non-small cell lung cancer: evidence from a meta-analysis. Mol. Clin. Oncol. 5 (6), 705–713.
Harrell Jr., F.E., Lee, K.L., Mark, D.B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. 15 (4), 361–387.
Harris, T., et al., 2012. Low-level expression of miR-375 correlates with poor outcome and metastasis while altering the invasive properties of head and neck squamous cell carcinomas. Am. J. Pathol. 180 (3), 917–928.
Hieronymus, H., et al., 2018. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife 7.
Holloway, K.R., et al., 2015. Krt6a-positive mammary epithelial progenitors are not at increased vulnerability to tumorigenesis initiated by ErbB2. PLoS One 10 (1), e0117239.
Huang, S., Chaudhary, K., Garmire, L.X., 2017. More is better: recent progress in multi-omics data integration methods. Front. Genet. 8, 84.

Jacobsen, A., et al., 2013. Analysis of microRNA-target interactions across diverse cancer types. Nat. Struct. Mol. Biol. 20, 1325.

Ju, X., et al., 2017. Liver X receptors as potential targets for cancer therapeutics. Oncol. Lett. 14 (6), 7676–7680.

Kaprio, T., et al., 2014. REG4 independently predicts better prognosis in non-mucinous colorectal cancer. PLoS One 9 (10), e109600.

Kim, D., et al., 2017. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. J. Am. Med. Inform. Assoc. 24 (3), 577–587.

Kim, Y., Pierce, C.M., Robinson, L.A., 2018. Impact of viral presence in tumor on gene expression in non-small cell lung cancer. BMC Cancer 18 (1), 843.

Kramer, A., et al., 2014. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics 30 (4), 523–530.

Kratz, J.R., et al., 2012. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. Lancet 379 (9818), 823–832.

Kumaran, M., et al., 2017. Germline copy number variations are associated with breast cancer risk and prognosis. Sci. Rep. 7 (1), 14621.

Lawrence, M.S., et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499 (7457), 214–218.

Lee, S.Y., Jang, C., Lee, K.A., 2014. Polo-like kinases (plks), a key regulator of cell cycle and new potential target for cancer therapy. Dev. Reprod. 18 (1), 65–71.

Liu, F., et al., 2016. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci. Rep. 6, 28517.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15 (12), 550.

Mankoo, P.K., et al., 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One 6 (11), e24709.

Meng, Y., Quan, L., Liu, A., 2018. Identification of key microRNAs associated with diffuse large B-cell lymphoma by analyzing serum microRNA expressions. Gene 642, 205–211.

Mermel, C.H., et al., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12 (4), R41.

Mitsudomi, T., et al., 1993. Mutations of the p53 gene as a predictor of poor prognosis in patients with non-small-cell lung cancer. J. Natl. Cancer Inst. 85 (24), 2018–2023.

Mogi, A., Kuwano, H., 2011. TP53 mutations in nonsmall cell lung cancer. J. Biomed. Biotechnol. 2011, 583929.

Nanavaty, P., Alvarez, M.S., Alberts, W.M., 2014. Lung cancer screening: advantages, controversies, and applications. Cancer Control 21 (1), 9–14.

Onn, A., Dickey, B.F., 2006. A better crystal ball to predict lung-cancer survival? Lancet Oncol. 7 (10), 789–790.

Pedregosa, F., et al., 2011. Scikit-learn: machine learning in python %. J. Mach. Learn. Res. 12, 2825–2830.

Raykar, V.C., et al., 2007. On ranking in survival analysis: bounds on the concordance index. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Curran Associates Inc.: Vancouver, British Columbia, Canada. pp. 1209–1216.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Sarrouilhe, D., et al., 2015. Serotonin and cancer: what is the link? Curr. Mol. Med. 15 (1), 62–77.

Shen, R., Olshen, A.B., Ladanyi, M., 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25 (22), 2906–2912.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics 69 (1), 7–34.

Smith, J.C., Sheltzer, J.M., 2018. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. Elife 7.

Tan, J., et al., 2015. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac. Symp. Biocomput. 132–143.

Tibshirani, R., 1997. The lasso method for variable selection in the Cox model. Stat. Med. 16 (4), 385–395.

Travis, W.D., 2011. Pathology of lung cancer. Clin. Chest Med. 32 (4), 669–692.

Wang, F.H., et al., 2018. UPK1B promotes the invasion and metastasis of bladder cancer via regulating the Wnt/beta-catenin pathway. Eur. Rev. Med. Pharmacol. Sci. 22 (17), 5471–5480.

Witte, T., Plass, C., Gerhauser, C., 2014. Pan-cancer patterns of DNA methylation. Genome Med. 6 (8), 66.

Woods, L.M., et al., 2011. Evidence against the proposition that "UK cancer survival statistics are misleading": simulation study with National Cancer Registry data. BMJ 342, d3399.

Wu, X., et al., 2014. Transgelin overexpression in lung adenocarcinoma is associated with tumor progression. Int. J. Mol. Med. 34 (2), 585–591.

Xia, T., et al., 2015. Long noncoding RNA FER1L4 suppresses cancer cell growth by acting as a competing endogenous RNA and regulating PTEN expression. Sci. Rep. 5, 13445.

Yin, Z., et al., 2016. Overexpression of long non-coding RNA MFI2 promotes cell proliferation and suppresses apoptosis in human osteosarcoma. Oncol. Rep. 36 (4), 2033–2040.

Yin, L., Chang, C., Xu, C., 2017. G2/M checkpoint plays a vital role at the early stage of HCC by analysis of key pathways and genes. Oncotarget 8 (44), 76305–76317.

Yoshihara, K., et al., 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. 4, 2612.

Zhang, B., et al., 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513 (7518), 382–387.

Zhang, L., et al., 2018a. Association analysis of somatic copy number alteration burden with breast cancer survival. Front. Genet. 9 p. 421.

Zhang, L., et al., 2018b. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front. Genet. 9, 477.

Zhao, X., et al., 2019. BPTF promotes hepatocellular carcinoma growth by modulating hTERT signaling and cancer stem cell traits. Redox Biol. 20, 427–441.

Zhou, T., et al., 2009. Overexpression of SSTR2 inhibited the growth of SSTR2-positive tumors via multiple signaling pathways. Acta Oncol. 48 (3), 401–410.