Genes & Diseases (2023) 10, 1190–1193

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

RAPID COMMUNICATION

Development of a molecular feature-based survival prediction model of ovarian cancer using the deep neural network

Ovarian cancer (OC) is one of the most lethal gynecologic cancer worldwide, and survival prediction is meaningful for personalized treatment.[1] The survival outcome of cancer patients mainly depended on the malignancy of the primary tumor which is tightly linked with the expression profile of the molecular features.[2] Therefore, in this study, we developed a molecular feature-based survival prediction model of OC using a deep neural network (DNN).

As described in the workflow diagrams of the whole study (Fig. 1A) and data preprocessing (Fig. S1A), the miRNA/mRNA expression data and the clinical information were obtained in The Cancer Genome Atlas (TCGA). Patient characteristics were summarized in Table S1. Non-primary samples and samples/features with more than 20% missing values were removed. The left missing values were imputed by using K Nearest Neighbor algorithm. This process was performed two times before and after normalization to replace the missing values with the average expression levels of the features in the same subtype.[3] The data were normalized by the quantile normalization algorithm (Table S2), and after normalization, the batch effect was efficiently reduced (Fig. S1B). The data were further scaled by the Z-score algorithm (Table S3). The expression profile of all mRNAs/miRNAs was visualized by heatmap (Fig. S1C), while no obvious difference in survival was obtained across the three groups of clustered patients, indicating that the survival-related features should be extracted.

We next extracted the survival-related features by Cox-PH and Kaplan–Meier algorithm; for Kaplan–Meier analysis, the samples were grouped according to the median or upper and lower quartile, separately. In total, 172 and 155 survival-related oncogenes and tumor suppressors were identified (Fig. S2 and Table S4). By evaluating the discrimination ability for the 3/5-year overall survival rate through the average area under the receiver operating characteristic curve (AUC) value, 41 candidates were selected and ranked (Table S5). The expression profile of the survival-related mRNAs/miRNAs was visualized by heatmap and the patients were clustered into three groups with different survival outcomes (Fig. 1B), indicating that the survival-related mRNA/miRNA features were successfully extracted.

To obtain the optimal combination of mRNA/miRNA features, a comprehensive analysis was performed. The patients were clustered by the K-means clustering program with different combinations of the survival-related features and the optimal combination was selected based on the Silhouette coefficient, Calinski-Harabasz index, and the maximum difference in median survival time. As shown in Figure 1C and Figure S3, clustering the patients into six groups with the top 13 features achieved the best performance. The grouped patients represent three different survival outcomes (high survival group: 3-year survival rate $\approx$ 100% and 5-year survival rate $\approx$ 90%; moderate survival group: 3-year survival rate $\approx$ 90% and 5-year survival rate $\approx$ 50%; poor survival group: 3-year survival rate $\approx$ 40% and 5-year survival rate $\approx$ 18%) (Fig. 1D); the best and the worst subgroups had a median survival time of 4624 and 197 days, respectively (Fig. 1E). Unsupervised principal component analysis of 13 features in the three groups was performed to visualize the profiling differences and the result confirmed the apparent discrimination (Fig. 1F).

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2022.10.011

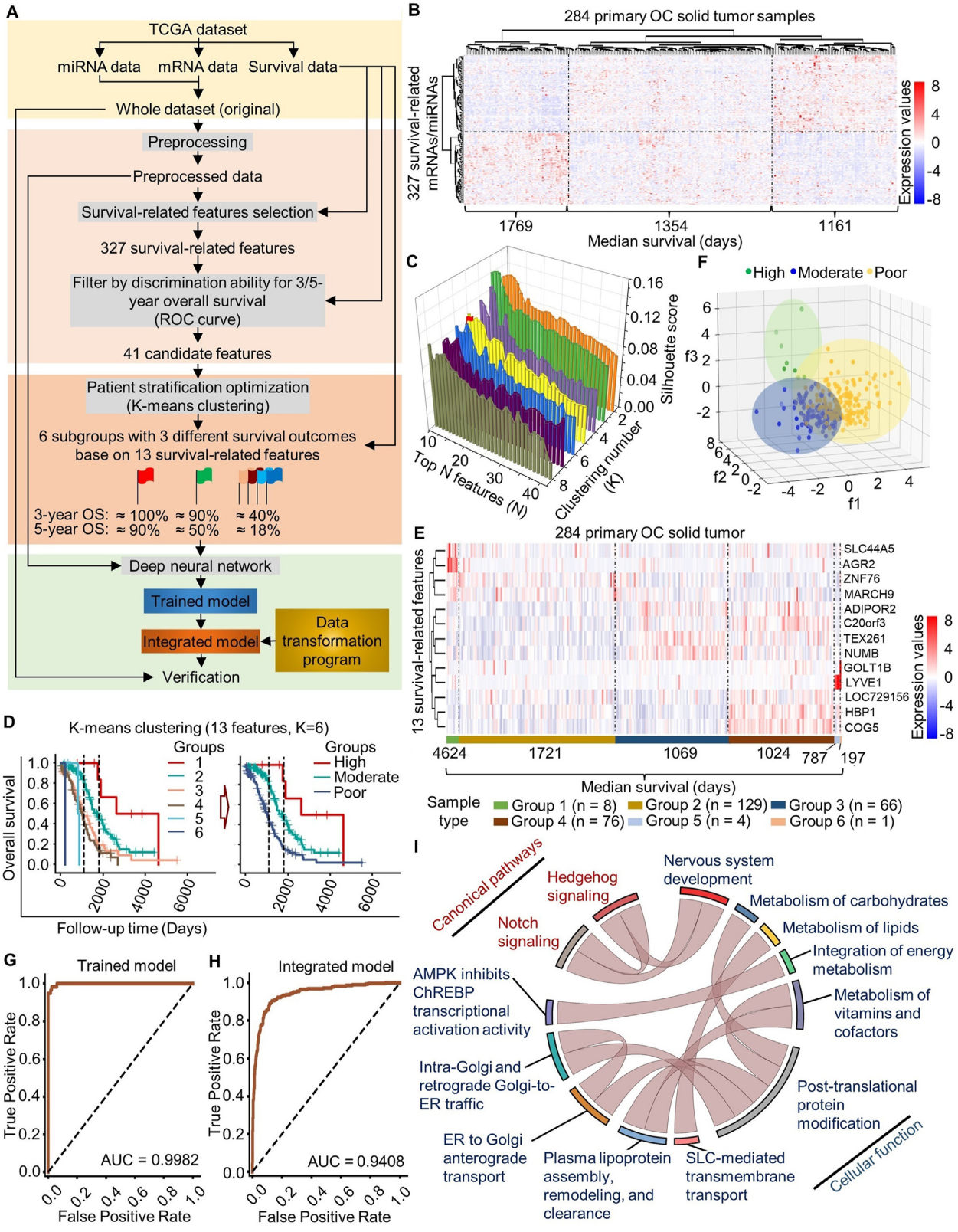

Figure 1 Molecular feature-based survival prediction model of ovarian cancer (OC) using deep neural network. **(A)** Overall workflow. The workflow includes four steps: Step 1, dataset preparation; Step 2, data preprocessing and survival-related feature selection; Step 3, Patient clustering; Step 4, model development and verification. In step 1, the whole dataset compose of miRNA

We next used DNN to develop a survival prediction model. The workflow of model development and verification was described in Figure S4A. In this study, an 8/2 split was used for generating training and testing data, and five-fold cross-validation was used for hyperparameter optimization.[4] The optimal hyperparameters were given in Figure S4B. The model developed through the training sets (named trained model) was evaluated through the testing sets. As shown in Figure 1G, the trained model achieved an average AUC of 0.9982.

For predicting the individual samples tested in reality, a data transformation program base on the parameters used in data preprocessing was developed (the source code was provided on the GitHub website https://github.com/ymy948/OV). As a result, the model (named integrated model) was finally developed by integrating it with the data transformation program. To further confirm the model by mimicking reality, the original TCGA data were used; importantly, the samples were individually subjected to the model. The result showed that the integrated model achieved a high performance (AUC = 0.9408) (Fig. 1H) and the survival differences were reflected (Fig. S4C).

To understand the molecular basis of the discrimination ability of the survival analysis model, the 13 survival-related genes employed by the prediction model were subjected to pathway analysis. These genes were enriched in biological functions, including metabolism of carbohydrates, metabolism of lipids, integration of energy metabolism, metabolism of vitamins and cofactors, post-translational protein modification, SLD-mediated transmembrane transport, plasma lipoprotein assembly, remodeling and clearance, ER to Golgi anterograde transport, intra-Golgi and retrograde Golgi-to-ER traffic, AMPK inhibits chREBP transcriptional activation activity, and canonical pathways, including NOTCH signaling and Hedgehog signaling (Fig. 1I). Deeper analysis indicated that activation of functions/pathways, including retinoid metabolism and transport, GPLD hydrolyses GPI-anchors from proteins, ubiquitination of NOTCH1 by ITCH in the absence of ligand, degradation of GLI1 by proteasome, phosphorylation of ChREBP at serine 568 by AMPK, cis-Golgi t-SNAREs bind YKT6 on tethered vesicle, neddylation, and uptake of hyaluronic acid, and inhibition of choline transports from the extracellular space to the cytosol, are associated with the poor prognosis of OC patients (Fig. S5A). The details of inhibition of NOTCH and Hedgehog signaling in cancer cells within OC patients with poor prognosis were shown in Figure S5B and S5C, respectively. These results also strongly suggested that the stemness of the cancer cells is tightly linked with OC prognosis.[5]

We next wonder if this stratification is correlated with the clinical features. The Chi−square test was used. We found that the clinical features, including patients with tumor-free cancer status, complete remission/response, days to new tumor event after initial treatment <365, and age at initial pathological diagnosis <365, have a higher probability of falling into high or moderate survival groups (Table S6).

In summary, an OC survival prediction DNN model was developed which robustly stratifies patients into three groups with distinct survival outcomes.

Author contributions

Tingyuan Lang: conceptualization, investigation, formal analysis, methodology, writing − original draft, writing − review and editing, supervision. Muyao Yang: methodology, data analysis, programming, plotting. Yunqiu Xia: methodology, data analysis, plotting. Jingshu Liu: methodology, data analysis, plotting. Yunzhe Li: methodology, correlation analysis, plotting. Lingling Yang: methodology, survival analysis, plotting. Chenxi Cui: methodology, pathway analysis, plotting. Yunran Hu: methodology, pathway analysis, plotting. Yang Luo: investigation, formal analysis. Dongling Zou: formal analysis. Lei Zhou: investigation, formal analysis, methodology. Zhou Fu: investigation, formal analysis, methodology. Qi Zhou: supervision.

Conflict of interests

All authors declare no conflict of interests related to the contents of this article.

and mRNA data in The Cancer Genome Atlas was prepared. In step 2, the data were preprocessed followed by survival-related features. In step 3, the patients were optimally grouped into 6 subgroups with three different survival outcomes. In step 4, the trained and integrated models were subsequently developed, followed by verification by individually inputting original data for mimicking reality. **(B)** Heatmap presenting the expression profile of 327 survival-related mRNAs/miRNAs in OC solid tumor samples. The median survival times of the grouped patients were shown. **(C)** The Silhouette coefficient was used to select the optimal parameters for patient clustering. **(D)** The patients were optimally clustered into 6 subgroups with three different survival outcomes (high, moderate, and poor survival) according to 13 survival-related features. **(E)** Heatmap presenting the expression profile of 13 survival-related features used in optimal clustering in OC solid tumor samples. The median survival times of each subgroup were shown. **(F)** Score plot of unsupervised principal component analysis overview of 13 survival-related features among the high, moderate, and poor survival groups. **(G)** Receiver operating characteristic (ROC) curve for trained prediction model to discriminate between high, moderate, and poor survival groups through preprocessed data. **(H)** ROC curve for integrated prediction model to discriminate between high, moderate, and poor survival groups through individually inputted original data. **(I)** Pathway enrichment analysis of 13 survival-related mRNAs used in the prediction model was visualized by a chord diagram.

Funding

This work was supported by Chongqing Science & Technology Bureau (China) (No. CSTB2022NSCQ-MSX1413, cstc2019jscx-msxmX0174 and cstc2021ycjh-bgzxm0134).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2022.10.011.

References

1. Siegel R, Miller K, Fuchs HE, et al. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7—33.
2. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. 2018;173(2):400—416. e11.
3. Ding D, Lang T, Zou D, et al. Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinf*. 2021;22(1):331.
4. Wright LG, Onodera T, Stein MM, et al. Deep physical neural networks trained with backpropagation. *Nature*. 2022; 601(7894):549—555.
5. Raleigh DR, Reiter JF. Misactivation of Hedgehog signaling causes inherited and sporadic cancers. *J Clin Invest*. 2019;129(2): 465—475.

Tingyuan Lang [a,b,1,*], Muyao Yang [c,1], Yunqiu Xia [d,e], Jingshu Liu [f], Yunzhe Li [g], Lingling Yang [g], Chenxi Cui [g], Yunran Hu [h,i], Yang Luo [j], Dongling Zou [a], Lei Zhou [k,l,m], Zhou Fu [n,**], Qi Zhou [a,***]

[a] *Department of Gynecologic Oncology, Chongqing University Cancer Hospital & Chongqing Cancer Institute & Chongqing Cancer Hospital, Chongqing 400030, China*
[b] *Reproductive Medicine Center, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400042, China*
[c] *Bioengineering College of Chongqing University, Chongqing 400044, China*
[d] *Department of Dermatology, Ministry of Education Key Laboratory of Child Development and Disorders, National Clinical Research Center for Child Health and Disorders, China*
[e] *International Science and Technology Cooperation Base of Child Development and Critical Disorders, Chongqing Engineering Research Center of Stem Cell Therapy, Children's Hospital of Chongqing Medical University, Chongqing, China*
[f] *Department of Gynaecology and Obstetrics, Reproductive Medicine Center, The Second Affiliated Hospital, Chongqing Medical University, Chongqing 400030, China*
[g] *School of Medicine, Chongqing University, Chongqing 400044, China*
[h] *Department of Pharmacy, Affiliated Drum Tower Hospital of Nanjing University Medical School, Nanjing 210008, China*
[i] *School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 210009, Jiangsu Province, China*
[j] *Center of Smart Laboratory and Molecular Medicine, School of Medicine, Chongqing University, Chongqing 400044, China*
[k] *Centre for Eye and Vision Research, School of Optometry, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[l] *Centre for Eye and Vision Research, Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[m] *Centre for Eye and Vision Research, Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Hong Kong SAR, China*
[n] *Department of Respiratory, Children's Hospital of Chongqing Medical University & National Clinical Research Center for Child Health and Disorders & Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing 400044, China*

*Corresponding author.

**Corresponding author.

***Corresponding author.
E-mail addresses: michaellang2009@163.com (T. Lang), fuzhou@163.com (Z. Fu), cqzl_zq@163.com (Q. Zhou)

17 June 2022
Available online 26 October 2022

[1] These authors contributed equally to this work.