SUPPLEMENTARY MATERIALS

Data Preprocessing for Machine Learning

Random Forest (RF)[1], an ensemble algorithm of decision trees, was developed to distinguish kidney renal clear cell carcinoma (KIRC) patients and normal patients based on mRNA transcripts (n=24, 4 HOXA gene family members, 9 miRNA, and 11 LncRNA, all derived from elements in the ceRNA regulatory network) and clinical features (n=7, including age, grade, clinic stage, gender, and TNM) from TCGA-KIRC cohort. The open-sourced R package ("randomForest", version = 4.6) was used to perform this progress. We employed random stratified sampling to select 422 patients (70% of patient cohort) as the training set (tumor samples = 370, and normal samples = 52), while the remaining 180 patients were used as the independent testing set (**Figure 4A)**. Moreover, to ensure reproducible results, we obtained an external validation set from the GEO database (**GSE151428**), consisting of 17 normal and 58 KIRC samples. A fixed random number seed was also used to ensure reproducibility of the results. The data in three feature sets was normalized by Log2 transformation.

Feature Selection

Due to irrelevant features were trained and thus may decrease model's performance, we firstly pre-trained all the original features based on the training set. And then based on the ranking of the features importance with average reduced accuracy in model, we filtered out the variables with weaker discriminatory power, including gender and T stage. We finally trained our early screening model using features containing 5 clinically relevant variables and 24 HOXA-ceRNA elements' variables. The importance ranking of the output of the final trained model is shown in **Figure 4B**. Besides, in addition to clinically relevant features, the functions of these genes included in our model are described individually in the discussion section of the text.

Model Training and Model Evaluation

We performed 10-fold cross validation to optimize RF parameters. Two core model parameters were identified as follows: the number of trees was set to 120, and the number of features randomly sampled as candidates was set to 8. To evaluate overall performance of our model, ROC curves with AUROC values for each class and confusion matrices (predicted label as the index of maximum value of the predicted probability vector) were generated for evaluating the performance in the training set, the test set and the validation set, respectively (**Figure 4C,D)**. Of note, even when the validation set is missing clinical features and only genetic features are retained, our model can still achieve good classification results, with a validation set AUC of 0.756. (**Figure 4D**).

Data Visualization

All the figures in our manuscript were plotted using R packages of ggplot2 (version 0.9.0).

Data and source code availability

The data that support the findings of this study, including the training set, the test set, and the validation set have been deposited in the **Supplementary Table S2**. Custom scripts for early screening model in this study were present in **Supplementary Figure 5**.

References

1. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **2003,** *43* (6), 1947-58.