



OPEN

# Identification of 12 immune-related lncRNAs and molecular subtypes for the clear cell renal cell carcinoma based on RNA sequencing data

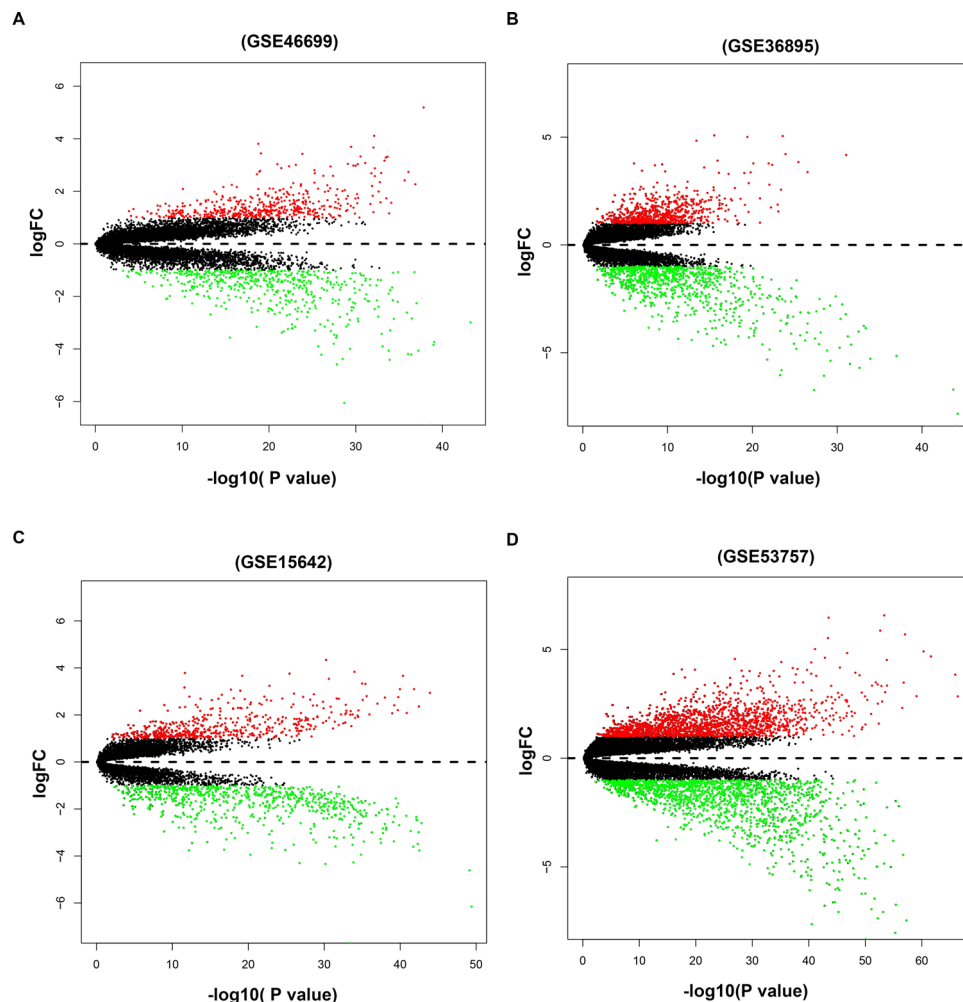
Weimin Zhong<sup>1,5</sup>, Bin Chen<sup>4,5</sup>, Hongbin Zhong<sup>1</sup>, Chaoqun Huang<sup>1</sup>, Jianqiong Lin<sup>1</sup>, Maoshu Zhu<sup>1</sup>, Miaoxuan Chen<sup>1</sup>, Ying Lin<sup>1</sup>, Yao Lin<sup>2✉</sup> & Jiyi Huang<sup>1,3✉</sup>

Clear cell renal cell carcinoma (ccRCC) is the most common type of renal cell carcinoma (RCC). Despite the existing extensive research, the molecular and pathogenic mechanisms of ccRCC are elusive. We aimed to identify the immune-related lncRNA signature and molecular subtypes associated with ccRCC. By integrating 4 microarray datasets from Gene Expression Omnibus database, we identified 49 immune-related genes. The corresponding immune-related lncRNAs were further identified in the TCGA dataset. 12-lncRNAs prognostic and independent signature was identified through survival analysis and survival difference between risk groups was further identified based on the risk score. Besides, we identified 3 molecular subtypes and survival analysis result showed that cluster 2 has a better survival outcome. Further, ssGSEA enrichment analysis for the immune-associated gene sets revealed that cluster 1 corresponded to a high immune infiltration level. While cluster 2 and cluster 3 corresponded to low and medium immune infiltration level, respectively. In addition, we validated the 12-lncRNA prognostic signature and molecular subtypes in an external validation dataset from the ICGC database. In summary, we identified a 12-lncRNA prognostic signature which may provide new insights into the molecular mechanisms of ccRCC and the molecular subtypes provided a theoretical basis for personalized treatment by clinicians.

Renal cell carcinoma (RCC) accounts for over 90% of all kidney cancer cases among human adults<sup>1</sup>. Clear cell renal cell carcinoma (ccRCC) is the most common RCC histological subtype which responsible for 70–80% of RCC cases<sup>2</sup>. Recent studies estimate that 102,000 patients succumb to ccRCC and 202,000 new cases are diagnosed annually throughout the globe<sup>3</sup>. Besides, ccRCC has no clear symptoms in its early stages therefore can only be diagnosed in advanced stages<sup>4</sup>. Despite interventions through chemotherapy and radiotherapy, ccRCC is prone to distant metastasis. These metastases occur in lung, bone, liver, distant lymph nodes, and renal vein<sup>5</sup>. A few effective biomarkers, particularly in both early and advanced stages of ccRCC, have been identified.

The recent version of the human genome annotation transcribes approximately 16,000 long non-coding RNAs (lncRNAs) responsible for 26.7% of the total genes (<https://www.genecodegenes.org/human/>)<sup>6</sup>. lncRNAs are non-coding RNAs composed of about 200 nucleotides in length. They include the antisense lncRNAs, intronic transcripts, large intergenic noncoding RNAs (lincRNAs), promoter-associated lncRNAs, and UTR associated lncRNAs<sup>7</sup>. Previous studies have revealed that lncRNA plays an important role in regulating gene expression, epigenetics, cell differentiation, and ontogeny<sup>8,9</sup>. Additionally, studies have shown that massive aberrantly expressed

<sup>1</sup>The Fifth Hospital of Xiamen, Xiamen 361101, Fujian Province, People's Republic of China. <sup>2</sup>Key Laboratory of Optoelectronic Science and Technology for Medicine of Ministry of Education, College of Life Sciences, Qishan Campus, Fujian Normal University, Fuzhou 350117, Fujian Province, People's Republic of China. <sup>3</sup>Xiang'an Branch, The First Affiliated Hospital of Xiamen University, Xiamen University, Xiamen 361101, Fujian Province, People's Republic of China. <sup>4</sup>The First Affiliated Hospital of Xiamen University, Xiamen University, Xiamen 361003, Fujian Province, People's Republic of China. <sup>5</sup>These authors contributed equally: Weimin Zhong and Bin Chen. ✉email: yaolin@fjnu.edu.cn; hjy0602@163.com



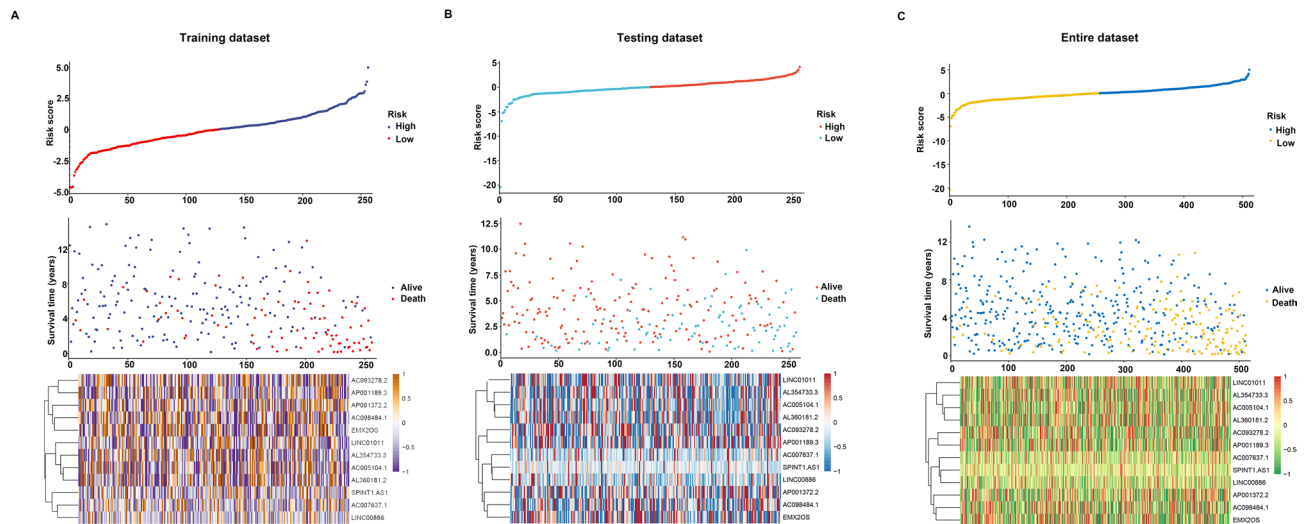
**Figure 1.** Volcano plots for the aberrantly expressed genes between ccRCC tissues and adjacent normal tissues in the four datasets. (A) GSE46699, (B) GSE36895, (C) GSE15642 and (D) GSE53757. The red dots represented up-regulated genes with the cut-off  $\log_{2}FC > 1.0$  and adjusted  $P < 0.05$ . The green dots represented down-regulated genes based on the criterion:  $\log_{2}FC < -1.0$  and adjusted  $P < 0.05$ . Black dots represent the genes have no significance.

lncRNAs are highly linked to tumor diagnosis and metastasis<sup>10</sup>. However, little information has been pronounced on the possible effects of dysregulated lncRNAs on ccRCC. Therefore, more specific and reliable lncRNA-based biomarkers need to be uncovered. This will thereby improve the prognostic accuracy of ccRCC.

This study assessed the immune-related lncRNA by integrating the GEO dataset, TCGA dataset, and ICGA dataset. We identified 12 lncRNA prognostic signature associated with ccRCC survival. Also, 3 molecular subtypes (cluster 1, cluster 2, and cluster 3) based on the immune lncRNA and gene expression profile were identified. The three molecular subtypes were highly correlated with immunity. The identification of immune-related lncRNA and molecular subtypes of ccRCC will advance the treatment of ccRCC patients, and provide a better understanding of the underlying molecular mechanisms of ccRCC.

## Results

**Identifying immune-related DEGs in ccRCC microarray datasets.** Four microarray datasets, including GSE46699, GSE36895, GSE15641, and GSE53757 were retrieved from the GEO database. Differentially expressed analysis was conducted on these datasets. Results showed that the GSE46699 dataset had a total of 1,006 DEGs (484 up-regulated and 522 down-regulated DEGs). In the GSE36895 dataset, 725 up-regulated genes and 870 down-regulated genes were identified. Besides, in the GSE15641 dataset, there were 517 up-regulated genes and 703 down-regulated genes were identified. In addition, 1,689 up-regulated genes and 1585 down-regulated genes were identified in the GSE53757 dataset. The results DEGs from four datasets are shown in Fig. 1. A total of 326 DEGs were identified, including 145 up-regulated and 181 down-regulated DEGs through the integrating of the four datasets using the RRA algorithm (Supplementary Table 1). The top 20 up-regulated DEGs and top 20 down-regulated DEGs are highlighted in Supplementary Fig. 1.



**Figure 2.** Risk plot for the ccRCC patients. (A) Entire dataset (B) Training dataset. (C) Testing dataset. Each panel consists of three rows: top rows showed a risk score distribution for the high risk score group and low risk score group; middle rows represent the ccRCC patients distribution and survival status; the bottom rows showed that the heatmap of 12 prognostic lncRNA expression.

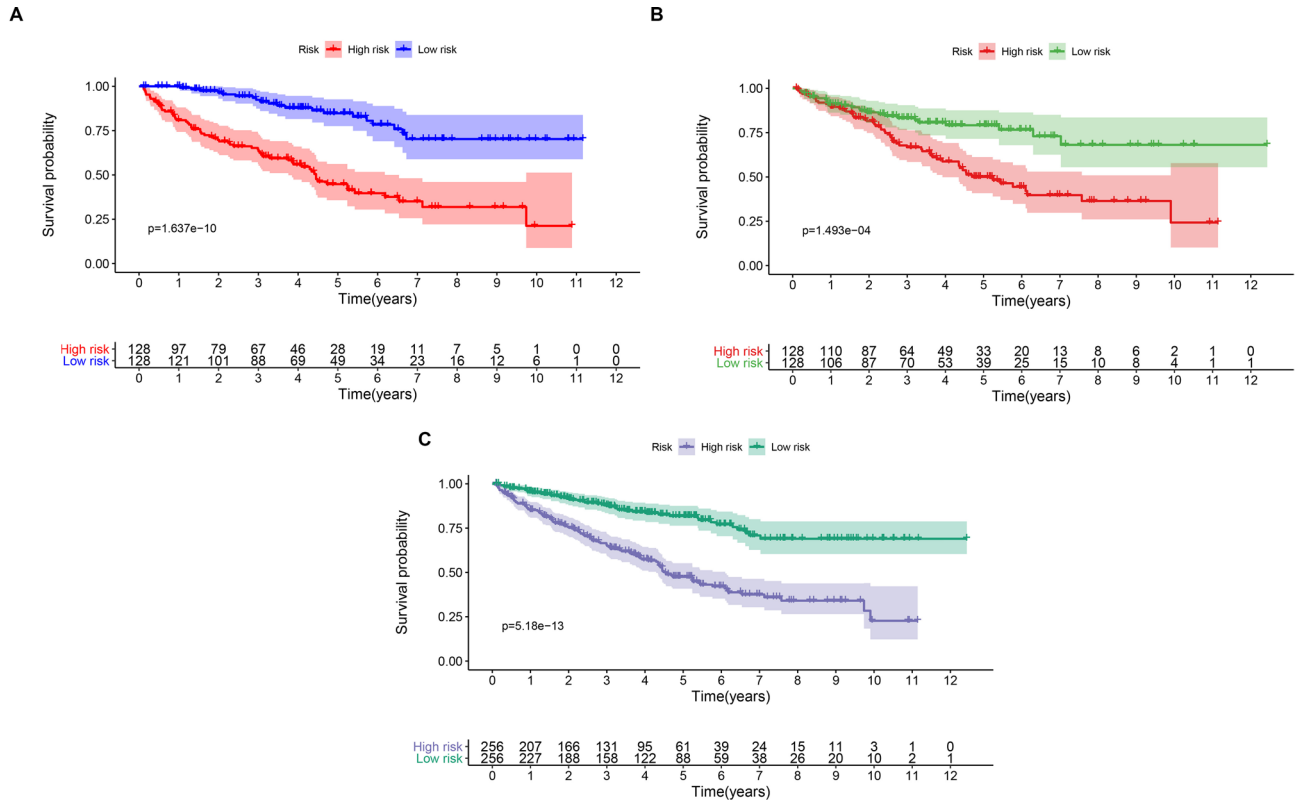
**Identifying immune-associated lncRNAs.** Among the 326 DEGs in the four datasets, 49 immune genes were identified and further extracted from the corresponding gene expression in the TCGA dataset. Accordingly, all information on the expression of lncRNAs was extracted from the TCGA dataset. To obtain high-quality lncRNA, we excluded the lowly expressed lncRNAs with the cut-off criteria that average expression in all samples need more than 0.5. The Pearson coefficient was adopted to estimate the correlation between lncRNA and immune genes. Finally, a total of 140 immune-related lncRNAs were identified based on the criterion:  $|cor|_{\text{pearson}} > 0.5$  and  $p\text{-value} < 0.001$  (Supplementary Table 2).

**Survival prediction for the lncRNAs and GSEA enrichment analysis.** In total, 512 ccRCC patient samples with 140 immune-related lncRNAs were randomly divided into the testing dataset ( $N = 256$ ) and training dataset ( $N = 256$ ) respectively. The correlation of immune-related lncRNA with patients' overall survival time was determined via univariate cox regression analysis in the training dataset. Further, we selected 56 immune-related lncRNAs with  $p\text{-value} < 0.05$  for Lasso-penalized multivariate Cox proportional hazards modeling analysis. After 1,000 iterations, 12 immune-related lncRNA expression signatures were identified. They include: AC005104.1, AC093278.2, AC098484.1, AL360181.2, EMX2OS, LINC01011, SPINT1-AS1, AP001372.2, AC007637.1, AL354733.3, AP001189.3 and LINC00886. Moreover, these selected signatures displayed optimal survival prediction in the training dataset for more than 50 times. Further, a calculation of the 12-lncRNA signature risk score for each patient in the training and testing datasets was conducted using the risk formula:

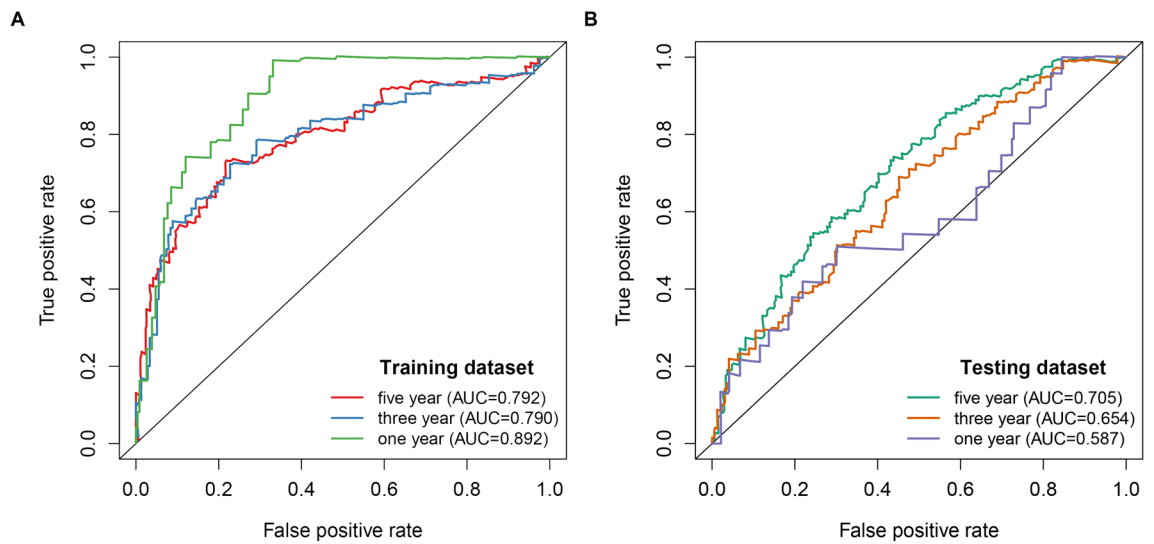
$$\begin{aligned} \text{Risk score} = & 0.062557657 * \text{AC005104.1} + (-0.093367264) * \text{AC093278.2} \\ & + (-0.263904071) * \text{AC098484.1} + 0.048384794 * \text{AL360181.2} + (-0.044967607) \\ & * \text{EMX2OS} + 0.23110957 * \text{LINC01011} + (-0.134140912) * \text{SPINT1-AS1} \\ & + (-0.063359846) * \text{AP001372.2} + (-0.218305812) * \text{AC007637.1} + 0.178920045 \\ & * \text{AL354733.3} + (-0.074645356) * \text{AP001189.3} + (-0.220170832) * \text{LINC00886}. \end{aligned}$$

Thus, the patients were further classified into high-risk ( $N = 128$ ) and low-risk groups ( $N = 128$ ). This was done according to the median risk score in the training, testing, and entire datasets respectively. The patients in the training dataset with a high-risk score corresponding to more deceased cases whereas those with low-risk scores exhibited prolonged survival time (Fig. 2A). Similar results were reported in the testing and entire datasets respectively (Fig. 2B,C). The KM curves and log-rank test identified a significant difference between the high-risk group and low-risk group from all datasets ( $p\text{-value} < 0.001$ ) (Fig. 3). ROC analysis revealed that the 12-lncRNA prognostic model effectively predicted the survival of ccRCC patients in 1-year, 3-year, and 5-year period in the training dataset and testing dataset respectively (Fig. 4A,B). To further evaluate the reliability of the prognostic model, we used an external dataset from the ICGC database for verification. The risk score for each patient was calculated using the KM curve and the results which depicted a significant divergence between risk groups (Supplementary Fig. 2). Also, the ROC curve analysis showed a good performance for the risk model (Supplementary Fig. 3).

In addition, through cox regression analysis the 12-lncRNA signature independently predicted the clinical traits (Fig. 5). To compare the accuracy of the survival prediction between 12-lncRNA signature risk models and clinical traits, we conducted a ROC analysis and observed that our risk model was optimal with an AUC value of up to 0.783 (Supplementary Fig. 4). Moreover, we investigate the association between 12 lncRNAs expression level and clinical trait (Stage and Grade), as highlighted in Supplementary Fig. 5A. The findings revealed that



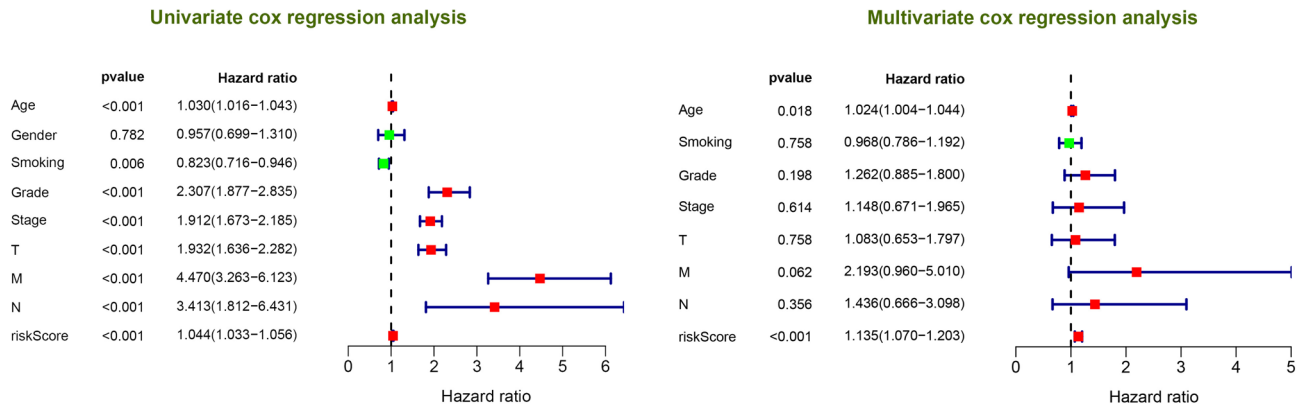
**Figure 3.** Kaplan–Meier curves for the 12-lncRNA signature in the entire (A), training (B), and validation sets (C).



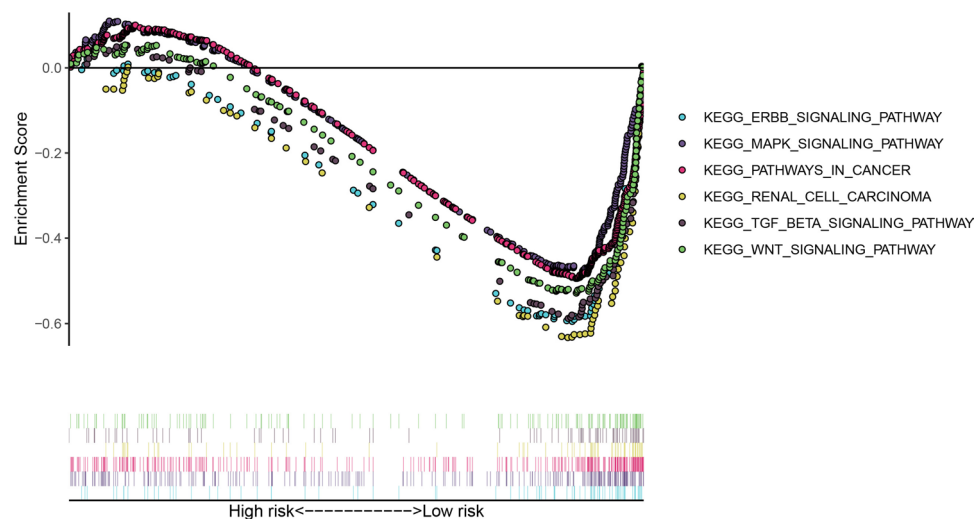
**Figure 4.** Time ROC curves analysis for the 12-lncRNA signature in the training dataset (A) and validation dataset (B).

increased expression level of AC005104.1 and AL360181.2 from stage I to stage IV. On the other hand, the expression levels of AC093278.2, AC098484.1, and EMX2OS were decreased from stage I to stage IV. Moreover, based on the grade, expression levels of AC007637.1, AC093278.2, AC098484.1, AL354733.3, AP001189.3, AP001372.2, and EMX2OS decreased from grade 1 to grade 4 (Supplementary Fig. 5B).

Through GSEA enrichment analysis, we explored the potential underlying pathway in the ccRCC in the high-risk group and low-risk group. Notably, 71 significant pathways were enriched in the low low-risk group following the cut-off: FDR < 0.05 and nominal P < 0.05. Among them, the ERBB signaling pathway, MAPK signaling pathway, pathways of cancer, renal cell carcinoma, TCG beta signaling pathway, and WNT signaling pathway



**Figure 5.** Univariate and multivariate Cox regression analysis for the 12 lncRNAs signature and clinical features.

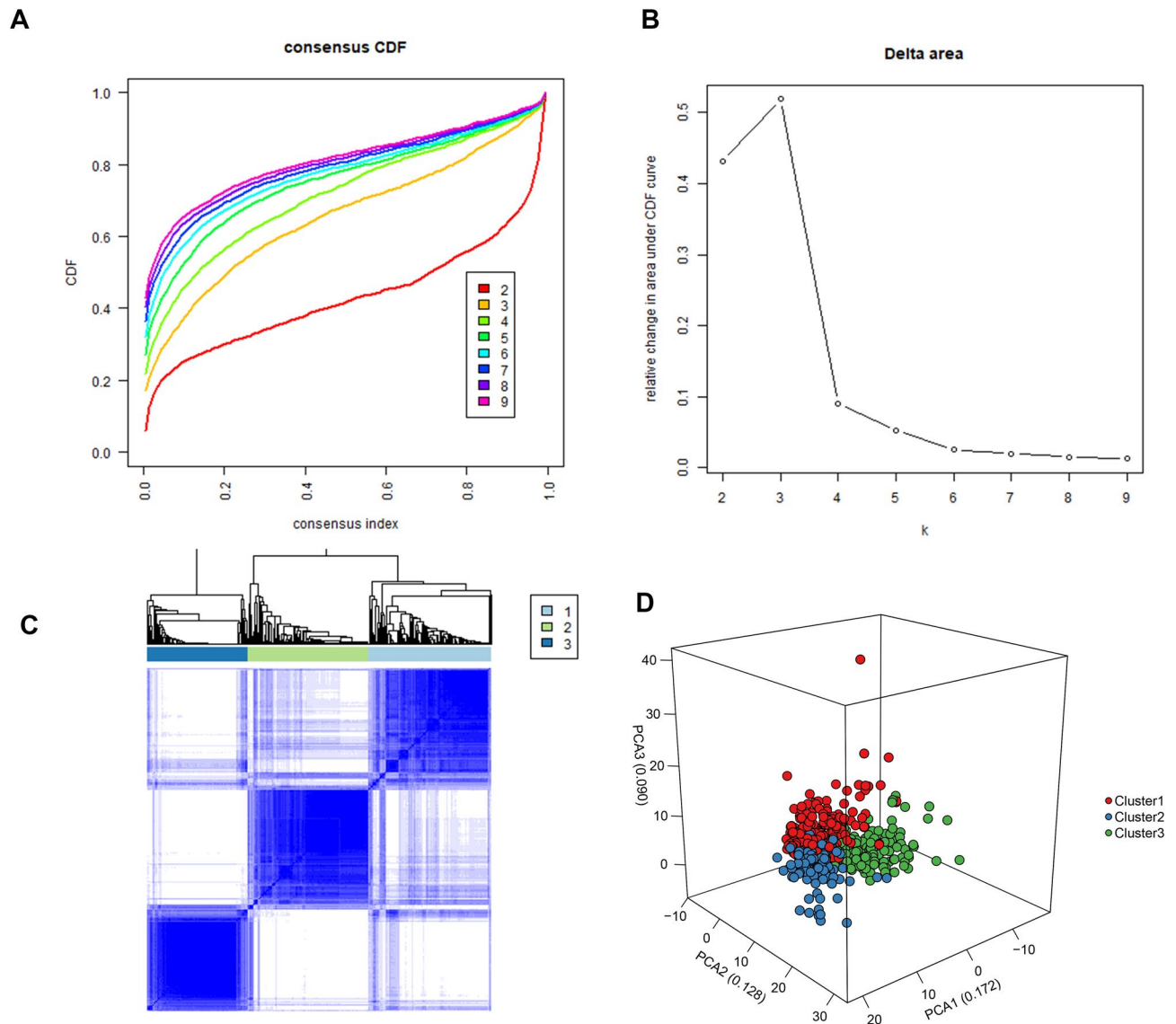


**Figure 6.** Gene Set Enrichment Analysis (GSEA) identified cancer-related pathway for the low risk group.

were enriched in the low-risk group (Fig. 6). In contrast, there was no significantly enriched signal pathway in the high-risk group with a similar cut-off.

**Identifying molecular subtypes of ccRCC.** According to the expression profile of immune genes and immune-related lncRNAs, we employed a consensus clustering algorithm to categorize the ccRCC patients into de novo groups. As indicated in Fig. 7A,B, we performed a consensus clustering algorithm via the Cancersubtype R package and revealed that the area under the cumulative distribution function (CDF) curve maintained a maximal consensus within clusters and a minimal ambiguity rate in cluster assignments when  $k=3$ . Therefore, the dataset was finally partitioned into three groups: cluster 1 ( $N=182$ ), cluster 2 ( $N=179$ ), and cluster 3 ( $N=152$ ) (Fig. 7C). Then, the Principal Component Analysis (PCA) analysis was performed on the expression of immune genes and immune-related lncRNAs on the three clusters. A clear distinction across the three clusters was noted (Fig. 7D). survival analysis showed that cluster 2 had a better survival time compared to cluster 1 and cluster 3 (Fig. 8A). The three clusters with expression data and clinical traits were presented in Fig. 8B. Besides, a consensus clustering algorithm on the ICGC dataset was used to validate the feasibility and reliability of the three molecular subtypes. The area under the cumulative distribution function (CDF) curve was calculated to determine the optimal  $k$  value (Fig. 9). Notably, after a comprehensive consideration,  $k=3$  was selected as the optimal cluster number. Furthermore, PCA results indicated divergence in the three subgroups. However, due to the small sample size ( $N=91$ ), the survival outcome ( $p=0.051$ ) in the subgroups was insignificant (Supplementary Fig. 6).

**Identification of Immunogenomic profiling.** The 29 immune-related gene sets represented immune cell types, functions, and pathways that were fitted into the analysis (Supplementary Table 3). Of note, the ssGSEA score was used to quantify enrichment levels of the immune cells, function, and pathways in ccRCC samples. The immune scores were further classified into three clusters corresponding to the previously partitioned cluster

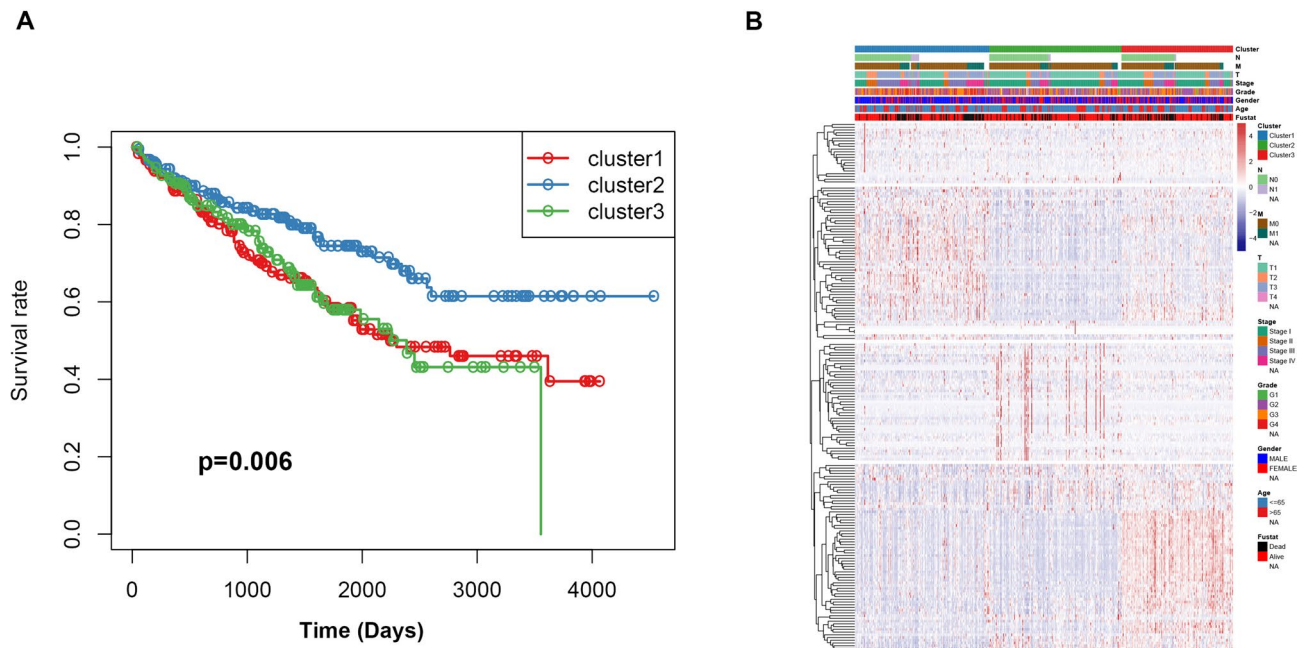


**Figure 7.** Consensus clustering for the ccRCC immune genes and immune related lncRNA expression in the TCGA dataset. **(A)** Consensus clustering cumulative distribution function (CDF) for  $k=2$  to 9. **(B)** The Relative change in area under CDF curve for  $k=2$  to 9. **(C)** Consensus matrix heatmap plots when  $k=3$ . **(D)** Principal component analysis of the immune genes and lncRNAs expression when  $k=3$ .

groups. Interestingly, the high immune infiltration level corresponded to cluster 1, median immune infiltration level corresponded to cluster 3 whereas, low immune infiltration level corresponded to cluster 2 (Fig. 10). Also, the immune score and tumor purity were analyzed using the ESTIMATE R package. The immune scores were significantly higher in cluster 1 while the scores were significantly lower in cluster 2 (Kruskal–Wallis test,  $P < 0.001$ ) (Fig. 11A). However, we noted a reverse trend on tumor purity which was significantly higher in cluster 2 but significantly lower in cluster 1 (Kruskal–Wallis test,  $P < 0.001$ ) (Fig. 11B). Generally, these findings indicated that the cluster was highly associated with immunity. Of note, cluster 2 comprised more tumor cells whereas, cluster 1 had more immune cells.

## Discussion

Unlike protein-coding genes, lncRNAs are involved in many important biological functions linked to human disease, epigenetic and post-transcription regulation<sup>17</sup>. Current literature shows that the non-coding RNAs, specifically lncRNAs, are tightly associated with tumorigenesis and progression of tumors. Also, cumulative evidence reveals that dysregulated lncRNAs may serve as a critical biomarker for many types of cancers, in particular, esophageal squamous cell carcinoma and breast cancer<sup>18,19</sup>. Numerical evidence pronounce that lncRNAs can associate with miRNA to weaken the effects of miRNA on in mediating mRNA expression<sup>20</sup>. Some recent studies have suggested that aberrant lncRNA expression is associated with ccRCC, and many potential biomarkers have been uncovered in ccRCC<sup>21–23</sup>. However, these results rely on a temporary TCGA database that lacks external

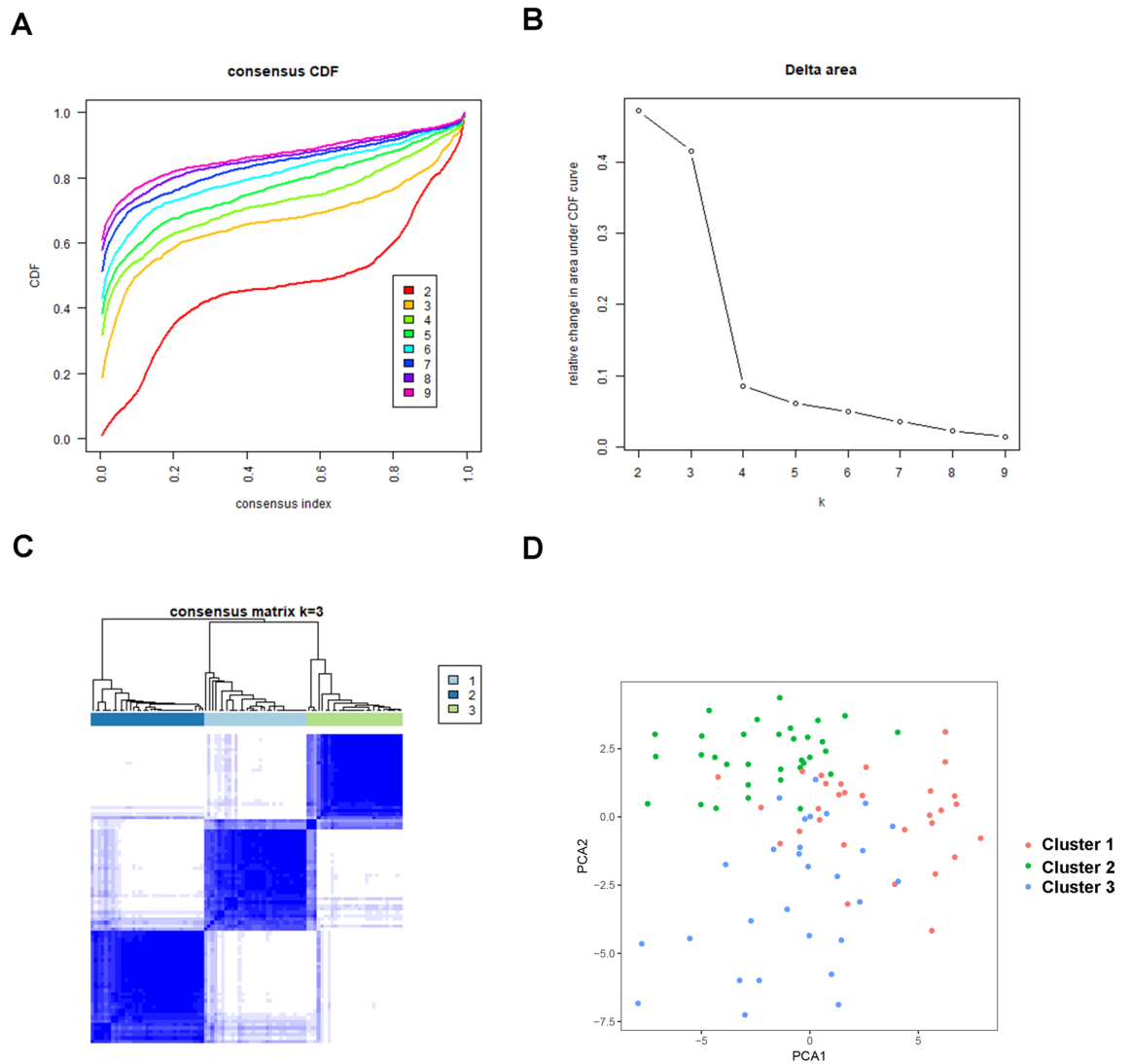


**Figure 8.** Prognostic significance of ccRCC immune subtypes. **(A)** Kaplan–Meier curves for the overall survival of different subtypes. The cluster 2 is associated with better outcomes compared to cluster 1 and cluster 3. **(B)** Heatmap of the immune genes and immune related lncRNAs expression profiles ordered by three subtypes, with clinical features associated with each cluster.

validation sets, hence, it poses difficulty in assessing the accuracy of the model. Moreover, due to the complexity of ccRCC, there is an urgent need to identify a large number of lncRNA markers for the prognosis and diagnosis of ccRCC. Presently, immunotherapy is a commonly used tumor intervention. However, for tumors and ccRCC, it exhibits a remarkable immune infiltration and other immune-related signatures<sup>24</sup>. Nevertheless, an increasing number of immunotherapy compounds such as PD-1/PD-L1 blocking agents have been approved in the treatment of ccRCC and have achieved noticeable progress<sup>25</sup>. However, few patients are unresponsive due to tumor heterogeneity, and cases of drug resistance occur<sup>6</sup>. Therefore, individualized immunotherapy treatment for patients is highly preferable.

In the present study, we included four GEO datasets (GSE46699, GSE36895, GSE15641, and GSE53757). Exactly 49 immune genes associated with ccRCC were screened from the GEO datasets using RRA methods. Through correlation analysis, 140 high confidence lncRNAs were identified in the TCGA dataset. Furthermore, a 12 immune-related lncRNA signature was developed and validated. This was performed through univariate cox regression and LASSO-penalized multivariate Cox proportional hazards modeling analyses in the TCGA and ICGC dataset. K–M curves analysis identified a significant divergence in patients classified in the high-risk group and low-risk group. Besides, GSEA enrichment analysis demonstrated that low-risk group patients exhibit significantly enriched pathways such as the ERBB signaling pathway, MAPK signaling pathway, pathways of cancer, renal cell carcinoma, TCG beta signaling pathway, and WNT signaling pathway. Thus, we suggest that the 12-lncRNA prognostic signature plays a crucial role in molecular pathogenesis, progression, and prognosis of ccRCC. Moreover, we compared our 12-lncRNA signature with clinical traits (Age, Gender, Stage, Grade, Smoking, T, N, and M) through ROC analysis, whereby, we demonstrated that our risk model is optimal (AUC = 0.783). Using cox regression analysis ( $p < 0.001$ ), it was revealed that the 12-lncRNA signature can independently predict clinical traits. Therefore, these results imply that the 12-lncRNA prognostic signature may provide a reliable prognostic marker and a theoretical basis on the mechanism of ccRCC. To explore the potential molecular subtype of ccRCC, we exploited the immune gene expression profile. We further identified three subtypes (cluster 1, cluster 2, and cluster 3) that are linked to the overall survival of ccRCC. Compared to cluster 1 and cluster 3, Log-rank test suggested that cluster 2 exhibited a better survival outcome ( $p < 0.05$ ). Interestingly, we observed that cluster 1 displayed a high immune infiltration level whereas cluster 2 showed a low immune infiltration level. Notably, cluster 3 correlated with the median immune infiltration level. Recent studies identified similar subgroup results through the ssGSEA methods based on the immune gene sets<sup>27</sup>. In addition, 3 lncRNAs including, AC005104.1, AL360181.2, and LINC01011) showed a significantly high expression level in cluster 1 when compared to cluster 2. This is an indication that the 3 lncRNAs are positively correlated with immunity (Supplementary Fig. 7).

We further identified 12 prognostic associated lncRNAs in ccRCC. including AC005104.1, AC093278.2, AC098484.1, AL360181.2, LINC01011, AP001372.2, AC007637.1, AL354733.3, and AP001189.3, which were the first time that served as markers in the current research. The lncRNA EMX2OS, which has been validated and identified to potentially induce proliferation and invasion, and promote sphere formation in ovarian cancer cells by regulating the miR-654-3p/AKT3/PD-L1 Axis<sup>28</sup> Additionally, EMX2OS can be used as a biomarker in laryngeal cancer and papillary thyroid cancer<sup>29,30</sup>. Furthermore, SPINT1-AS1 was identified as a potential marker



**Figure 9.** Consensus clustering for the ccRCC immune genes and immune related lncRNA expression in the ICGC dataset. **(A)** Consensus clustering cumulative distribution function (CDF) for  $k=2$  to 9. **(B)** The Relative change in area under CDF curve for  $k=2$  to 9. **(C)** Consensus matrix heatmap plots when  $k=3$ . **(D)** Principal component analysis of the immune genes and lncRNAs expression when  $k=3$ .

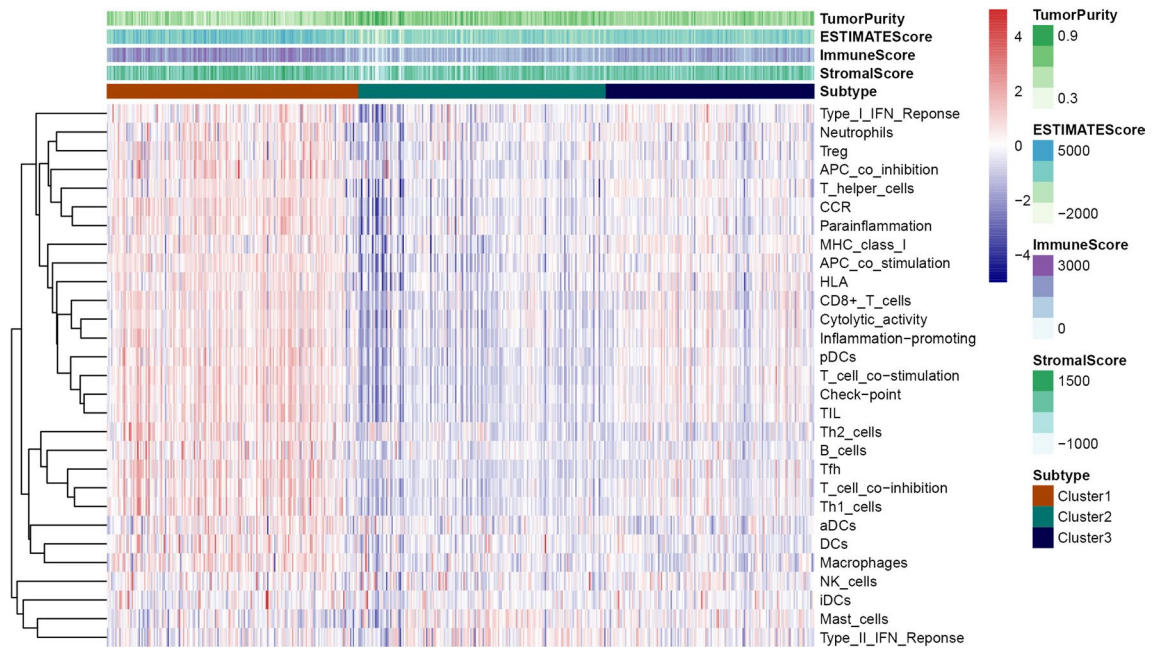
in the prognosis of colorectal cancer. High expression levels of SPINT1-AS1 was reported to correspond with a worse survival outcome<sup>31</sup>. There has been a high expectation on LINC00886 as a novel biomarker in laryngeal carcinoma<sup>32</sup>. However, the lncRNA which is a potential biomarker should be subjected to further experimental validation in future studies.

In conclusion, we identified 12 immune-related lncRNAs that correlated with the survival of ccRCC. Further, we discovered three molecular subtypes that are associated with tumor immunity. Our study has important clinical significance in understanding the underlying mechanism of ccRCC, therefore, can be used as a reference by clinicians for individualized treatment.

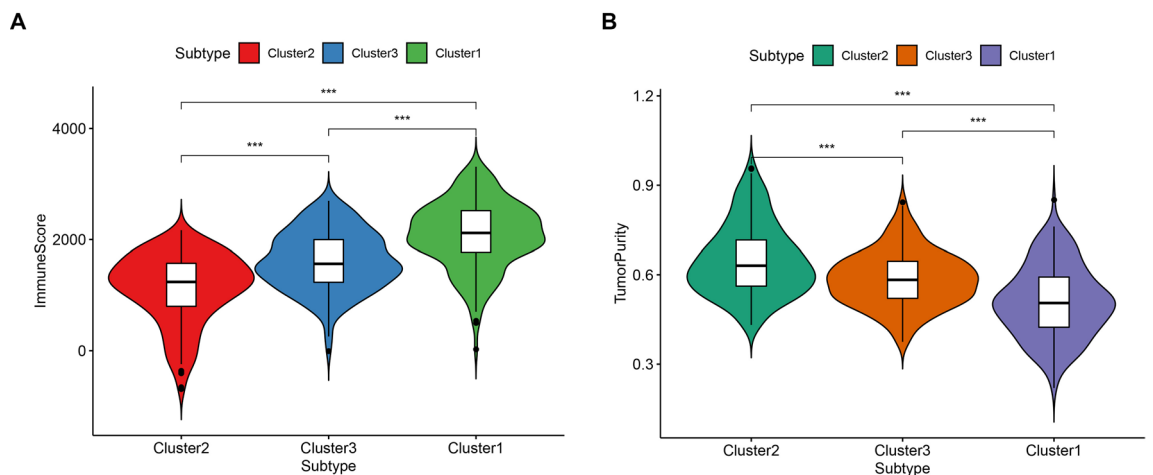
## Methods

**Data collection.** The four microarray datasets (GSE46699, GSE36895, GSE15641, and GSE53757) were downloaded from the GEO database ([www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)) as per the keyword 'renal cell carcinoma'. Notably, the GPL570 version platform was adopted to obtain datasets (GSE46699, GSE36895, and GSE53757). Dataset GSE46699 comprised 63 adjacent normal kidney specimens and 67 ccRCC samples whereas, the GSE36895 dataset comprised 23 adjacent normal kidney samples and 29 ccRCC samples. Dataset GSE53757 composed of 72 adjacent normal kidney samples and 72 ccRCC samples. The platform used to obtain the GSE15641 dataset was based on the GPL96 version which incorporated 23 normal kidney samples and 32 ccRCC samples. All the expression level data were standardized and log<sub>2</sub> transformed. In addition, 539 and 91 ccRCC samples were downloaded with corresponding clinical information from the TCGA database and ICGC database respectively.





**Figure 10.** Heatmap of the immunity ordered by the three subtypes. The cluster 2 corresponding to low immunity, cluster 1 corresponding to high immunity and cluster 3 associated with median immunity. The Tumor\_purity, Stromal\_score, and Immune\_score were estimated by ESTIMATE.



**Figure 11.** Violin plot for the immune score (A) and tumor purity (B) in different subtypes.

**Integration of microarray data.** The differentially expressed genes (DEGs) were identified for the four datasets through Limma analysis<sup>11</sup>. Gene expression analysis of the four microarray datasets was conducted using the RRA R package. A comparison of the expression level of genes was conducted according to their ranks. Screening of the significantly expressed genes were based on the p-adjusted < 0.05. Besides, the immune-related genes were downloaded from the IMMPORT database (<https://www.immport.org/>) and the significantly immune genes in our dataset were obtained from the RRA result<sup>12</sup>.

**Correlation analysis.** Data on the immune genes and lncRNAs expression of ccRCC were extracted from the TCGA dataset. The correlation of lncRNAs with immune genes was determined according to the Pearson coefficient, whereby, we performed cor function using the R package. The immune-related lncRNAs were identified with the criterion: Pearson coefficient > 0.5 and p-value < 0.0001.

**Identification and validation of lncRNA prognostic signature.** To perform survival analysis, we integrated the survival time and the corresponding lncRNA expression. The ccRCC patient samples with immune-related lncRNAs were equally categorized into training and testing datasets. Further, through univariate Cox regression analysis, we selected significantly correlated lncRNAs. For a robust and reliable lncRNA, significantly correlated lncRNA with p-value < 0.05 were selected and fitted into Lasso-penalized multivariate

Cox proportional hazards modeling analysis. After 1,000 iterations using the LASSO analysis, the lncRNAs with frequency > 50 were selected and fitted into multivariate cox regression analysis. The risk score formula was established based on the regression coefficients as follows: Risk score =  $\sum \text{Coef}_{\text{lncRNAs}} \times \text{Exp}_{\text{lncRNAs}}$ . The  $\text{Coef}_{\text{lncRNA}}$  denoted the lncRNA regression coefficient, where  $\text{Exp}_{\text{lncRNA}}$  denoted the expression level of the corresponding lncRNA. Thereafter, the risk score for each patient was calculated and subsequently categorized into high-risk group and low-risk group based on the median risk score in the training and testing datasets, respectively. Kaplan–Meier curve analysis was conducted via the R survival package (<https://cran.r-project.org/web/packages/survival>) to estimate the survival divergence between high-risk and low-risk groups. ROC curve analysis was aided in evaluating the accuracy of the prognostic model via the survival ROC package (<https://cran.r-project.org/web/packages/survivalROC>). Also, the lncRNA prognostic signature was validated in the ICGA dataset. The risk score for each patient in the ICGC dataset was calculated according to the risk model. The survival difference between risk groups was implemented using K–M curve analysis.

**GSEA enrichment analysis.** Gene set enrichment analysis was performed between the high-risk group and low-risk group. The significant pathways were enriched with the NOM p-value < 0.05 and FDR < 0.25. The c2.cp.kegg.v7.0.symbols.gmt was selected as the reference file.

**Molecular subtype identification and validation.** The potential molecular subtypes were identified using the “ExecuteCC” function from the Cancersubtypes R package<sup>13</sup> based on the expression of immune genes and lncRNAs. Principal components analysis (PCA) was performed to validate the subtypes of ccRCC. Further, the survival rate of different subtypes was conducted using K–M curve analysis. The molecular subtypes were further validated in the ICGA dataset based on the expression of immune lncRNA and gene expression using “Cancersubtypes” R package.

**Gene signature obtained and ssGSEA analysis.** The marker genes for immune cell types were referred from a previous study by Bindea et al.<sup>14</sup>. Infiltration levels for each of the immune cell types were estimated by the ssGSEA method from GSVA R package<sup>15</sup>. Then, we transformed each attribute (immune signature or gene set) value (ssGSEA score)  $x_i$  into  $x_i'$  by the equation  $x_i' = (x_i - x_{\min}) / (x_{\max} - x_{\min})$ , where,  $x_{\min}$  and  $x_{\max}$  denoted the minimum and maximum of the ssGSEA scores for the gene set across all ccRCC samples, respectively.

**Evaluation of the immune score, tumor purity, and stromal score in ccRCC.** ESTIMATE algorithm was used to calculating the fraction of immune and stromal cells in tumor tissues based on a gene expression signature. The R script of the ESTIMATE algorithm was downloaded from the public source website (<https://sourceforge.net/projects/estimateproject/>). This was followed by a calculation of the Immune scores, stromal scores, and ESTIMATE scores for each ccRCC patient<sup>16</sup>.

**Statistical analysis.** All the computational and statistical analyses were conducted on the R software (version 3.6.2). Kruskal–Wallis test was used to compare the divergence between multiple groups. Chi-square test or Fisher exact test was used for statistics on clinical information. A Bonferroni test was used to correct the p-value. Kaplan–Meier curves analysis was used to assess survival differences of subtype.

## Data availability

The dataset performed in this study are available from the corresponding author on reasonable requests.

Received: 18 December 2019; Accepted: 20 July 2020

Published online: 02 September 2020

## References

- Hsieh, J. J. et al. Renal cell carcinoma. *Nat. Rev. Dis. Primers* **3**, 17009. <https://doi.org/10.1038/nrdp.2017.9> (2017).
- Cheville, J. C., Lohse, C. M., Zincke, H., Weaver, A. L. & Blute, M. L. Comparisons of outcome and prognostic features among histologic subtypes of renal cell carcinoma. *Am. J. Surg. Pathol.* **27**, 612–624. <https://doi.org/10.1097/0000478-200305000-00005> (2003).
- Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* <https://doi.org/10.3322/caac.21590> (2020).
- Baek, M., Jung, J. Y., Kim, J. J., Park, K. H. & Ryu, D. S. Characteristics and clinical outcomes of renal cell carcinoma in children: a single center experience. *Int. J. Urol.* **17**, 737–740. <https://doi.org/10.1111/j.1442-2042.2010.02588.x> (2010).
- Caceres, W. & Cruz-Chacon, A. Renal cell carcinoma: molecularly targeted therapy. *P. R. Health Sci. J.* **30**, 73–77 (2011).
- Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* **21**, 1253–1261. <https://doi.org/10.1038/nm.3981> (2015).
- Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43**, 904–914. <https://doi.org/10.1016/j.molcel.2011.08.018> (2011).
- Orom, U. A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58. <https://doi.org/10.1016/j.cell.2010.09.001> (2010).
- Fatica, A. & Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21. <https://doi.org/10.1038/nrg3606> (2014).
- Nagano, T. & Fraser, P. No-nonsense functions for long noncoding RNAs. *Cell* **145**, 178–181. <https://doi.org/10.1016/j.cell.2011.03.014> (2011).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* **43**, e47. <https://doi.org/10.1093/nar/gkv007> (2015).
- Bhattacharya, S. et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, 180015. <https://doi.org/10.1038/sdata.2018.15> (2018).

13. Xu, T. *et al.* CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics* **33**, 3131–3133. <https://doi.org/10.1093/bioinformatics/btx378> (2017).
14. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795. <https://doi.org/10.1016/j.immuni.2013.10.003> (2013).
15. Hanzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7. <https://doi.org/10.1186/1471-2105-14-7> (2013).
16. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. <https://doi.org/10.1038/ncomms3612> (2013).
17. Chen, Y. *et al.* Long non-coding RNA ROR promotes radioresistance in hepatocellular carcinoma cells by acting as a ceRNA for microRNA-145 to regulate RAD18 expression. *Arch. Biochem. Biophys.* **645**, 117–125. <https://doi.org/10.1016/j.abb.2018.03.018> (2018).
18. Li, J. *et al.* LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700–1710. <https://doi.org/10.1136/gutjnl-2013-305806> (2014).
19. Meng, J., Li, P., Zhang, Q., Yang, Z. & Fu, S. A four-long non-coding RNA signature in predicting breast cancer survival. *J. Exp. Clin. Cancer Res. CR* **33**, 84. <https://doi.org/10.1186/s13046-014-0084-7> (2014).
20. Zhang, X. F., Ye, Y. & Zhao, S. J. LncRNA Gas5 acts as a ceRNA to regulate PTEN expression by sponging miR-222-3p in papillary thyroid carcinoma. *Oncotarget* **9**, 3519–3530. <https://doi.org/10.18632/oncotarget.23336> (2018).
21. Shi, D. *et al.* A five-long non-coding RNA signature to improve prognosis prediction of clear cell renal cell carcinoma. *Oncotarget* **8**, 58699–58708. <https://doi.org/10.18632/oncotarget.17506> (2017).
22. Yin, H. *et al.* Integrated analysis of long noncoding RNA associated-competing endogenous RNA as prognostic biomarkers in clear cell renal carcinoma. *Cancer Sci.* **109**, 3336–3349. <https://doi.org/10.1111/cas.13778> (2018).
23. Zeng, J. H. *et al.* Prognosis of clear cell renal cell carcinoma (ccRCC) based on a six-lncRNA-based risk score: an investigation based on RNA-sequencing data. *J. Transl. Med.* **17**(1), 281. <https://doi.org/10.1186/s12967-019-2032-y> (2019).
24. Seethapathy, H. *et al.* The incidence, causes, and risk factors of acute kidney injury in patients receiving immune checkpoint inhibitors. *Clin. J. Am. Soc. Nephrol.* **14**(12), 1692–1700. <https://doi.org/10.2215/CJN.00990119> (2019).
25. Hahn, A. W. *et al.* A phase I study of alpha-1,3-galactosyltransferase-expressing allogeneic renal cell carcinoma immunotherapy in patients with refractory metastatic renal cell carcinoma. *Oncologist* **25**, 121–213. <https://doi.org/10.1634/theoncologist.2019-0599> (2020).
26. Stein, J. E. *et al.* Pan-tumor pathologic scoring of response to PD-(L)1 blockade. *Clin. Cancer Res.* **26**, 545–551. <https://doi.org/10.1158/1078-0432> (2020).
27. Senbabaoglu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17**, 231. <https://doi.org/10.1186/s13059-016-1092-z> (2016).
28. Duan, M. *et al.* LncRNA EMX2OS induces proliferation, invasion and sphere formation of ovarian cancer cells via regulating the miR-654-3p/AKT3/PD-L1 axis. *Cancer Manag. Res.* **12**, 2141–2154. <https://doi.org/10.2147/CMAR.S229013> (2020).
29. Tang, W. *et al.* Signature microRNAs and long noncoding RNAs in laryngeal cancer recurrence identified using a competing endogenous RNA network. *Mol. Med. Rep.* **19**(6), 4806–4818. <https://doi.org/10.3892/mmr.2019.10143> (2019).
30. Yi, G. *et al.* The downregulation of lncRNA EMX2OS might independently predict shorter recurrence-free survival of classical papillary thyroid cancer. *PLoS ONE* **13**(12), e0209338. <https://doi.org/10.1371/journal.pone.0209338> (2018).
31. Chen, L. *et al.* Increased expression of antisense lncRNA SPINT1-AS1 predicts a poor prognosis in colorectal cancer and is negatively correlated with its sense transcript. *OncoTargets Ther.* **11**, 3969–3978. <https://doi.org/10.2147/OTT.S163883> (2018).
32. Lan, L. *et al.* Aberrant DNA hypermethylation-silenced LINC00886 gene accelerates malignant progression of laryngeal carcinoma. *Pathol. Res. Pract.* **216**(4), 152877. <https://doi.org/10.1016/j.prp.2020.152877> (2020).

## Acknowledgements

We appreciated TCGA database for providing the original study data. This study was supported by the Health Science Research Personnel Training Program of Fujian Province (2017-CXB-22), Provincial Natural Science Foundation of Fujian, Fujian Science and Technology Plan Guiding Projects (2019D026) and Xiamen Medical Advantage Subspecialty Vascular Access Construction Fund ([2018] 296).

## Author contributions

Y.L. and J.H. designed the study, C.H., J.L., M.C., Y.L., M.Z. and H.Z. collected the clinical information and lncRNA expression data. B.C. and W.Z. analysis data. W.Z. wrote the manuscript and Y.L. and J.H. revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-71150-3>.

**Correspondence** and requests for materials should be addressed to Y.L. or J.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020