Genome Instability-related Long Non-coding RNA in Clear Renal Cell Carcinoma Determined using Computational Biology

Yutao Wang
First Hospital of China Medical University
Kexin Yan
First Hospital of China Medical University
Jianbin Bi (bijianbin@hotmail.com)
First Hospital of China Medical University

Research Article

Keywords: enome instability, Long non-coding RNAs, Computationalbiology,GeneSet VariationAnalysis, Risk signature

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-219885/v1

Genome instability-related long non-coding RNA in clear renal cell carcinoma determined using computational biology

Yutao Wang[1#] , Kexin Yan[2#] , Jianbin Bi[1*]

[1]Department of Urology, China Medical University, The First Hospital of China Medical University, Shenyang, Liaoning, China

[2]Department of Dermatology, China Medical University, The First Hospital of China Medical University, Shenyang, Liaoning, China

#These authors contributed equally to this paper

*** Correspondence: Jianbin Bi**

bijianbin@hotmail.com

Keywords: Genome instability; Long non-coding RNAs; Computational biology; Gene Set Variation Analysis; Risk signature.

Abstract

There is evidence that long non-coding RNA (lncRNA) is related to genetic stability. However, the complex biological functions of these lncRNAs are unclear. In the present study, we applied computational biology to identify genome-related long

noncoding RNA and identified 26 novel genomic instability-associated lncRNAs in clear cell renal cell carcinoma. We identified a genome instability-derived six lncRNA-based gene signature that significantly divided clear renal cell samples into high- and low-risk groups. We validated it in test cohorts. Based on the above analysis, we identified six lncRNAs related to clear cell carcinoma outcomes and genome stability based on computational biology. To further elucidate the role of the six lncRNAs in the model's genome stability, we performed a gene set variation analysis (GSVA) on the matrix. We performed Pearson correlation analysis between the GSVA scores of genomic stability-related pathways and lncRNA. It was determined that LINC00460 and LINC01234 could be used as critical factors in this study. They may influence the genome stability of clear cell carcinoma by participating in mediating critical targets in the base excision repair pathway, the DNA replication pathway, homologous recombination, mismatch repair pathway, and the P53 signaling pathway. These data suggest that LINC00460 and LINC01234 are crucial for the stability of the clear cell renal cell carcinoma genome.

1 Introduction

Clear cell renal cell carcinoma (ccRCC) is the most common subtype of renal cell carcinoma, and clear cell renal cell carcinoma (ccRCC) accounts for 80% to 90% of all renal cell carcinomas. ccRCC is a potentially invasive tumor with an overall progression-free survival rate of 70% and a cancer-specific mortality rate of 24%[1]. It

is 1.5–2.0 times more common in men than in women. Advanced RCC has a five-year survival rate of 11.7%[2]. Risk factors include smoking, obesity, high blood pressure, chronic kidney disease, and exposure to certain chemicals and heavy metals[3]. The diagnosis of ccRCC has been increasing over the past few years. Although surgery is the most common treatment option, early diagnosis is difficult, and many patients have metastatic disease by this time[4]. For patients with advanced ccRCC or relapse, many molecular-targeted drugs have been used as first-line therapies. Nevertheless, outcomes are poor due to the side effects of these agents and individual differences in individual drug sensitivities[5].

It is a fundamental challenge for cells to copy their genetic material for daughter cells accurately. Once this process goes wrong, genomic instability occurs[6]. The level of genomic instability is reflected in nucleotide instability, microsatellite instability, and chromosome instability[7]. DNA damage can be caused by mistakes in DNA replication caused by genotoxic compounds or ultraviolet and ionizing radiation. Incorrect DNA replication can lead to mutations or blocked replication, leading to chromosome breakage, rearrangement, and dislocation[8]. Genomic instability is an essential source of genetic diversity within tumors. Oncogene expression drives proliferation by interfering with regulatory pathways that control cell cycle progression. Genomic instability produces large-scale genetic aberrations but also increases point mutations in protein-coding genes. The estimated mutation rate in tumors is an order of magnitude higher than that of typical healthy tissue. Genomic instability also changes as tumors develop, and this trait

could become a target for treatment[9].

Recent advances in sequencing technology have revealed that only 2% of the human genome codes for proteins[10]. Non-coding RNAs are classified into small non-coding RNAs and long non-coding RNAs according to their size. Long non-coding RNA (lncRNA) predominate. LncRNAs play central roles in many cellular mechanisms, including regulation of cell processes[11]. They also regulate pathophysiological processes through gene imprinting, histone modification, chromatin remodeling, and other mechanisms[12]. LncRNAs also play essential roles in cancer. They are involved in chromatin remodeling and transcriptional and post-transcriptional regulation through various chromatin-based mechanisms and interactions with other RNA species. LncRNA imbalances can alter functions such as cell proliferation, anti-apoptosis, angiogenesis, metastasis, and tumor suppression[13]. Depending on their positions and distribution in the genome, lncRNAs directly or indirectly affect the transcription of various proteins through transcriptional and post-transcriptional changes, some of which may mediate tumor inhibition or promotion[14].

Because chemotherapy, radiation therapy, targeted therapeutic agents, and immune checkpoint inhibitors do not function well in many ccRCC patients, investigators need to develop new treatment options and further identify prognostic biomarkers and therapeutic targets ccRCC. LncRNA screening and model building based on gene instability in ccRCC may represent an important research strategy.

2 Materials and methods

2.1 Data collection

We downloaded clinical information, protein-coding RNA expression data, lncRNA expression data, and somatic mutation information for clear renal cell carcinomas from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/)[15]. We considered 507 ccRCC samples with paired lncRNA and mRNA expression profiles, survival information, and clinical information.

We divided all ccRCC samples into a training set and a test set. The training set included 254 samples for the creation of a clinical outcome lncRNA risk model. The test set included 253 patients, used to validate the predictive ability of the prognostic risk model. We provided detailed data on TCGA clear cell renal carcinoma (Supplementary Table 1).

2.2 Mining lncRNAs related to genetic instability

First, we calculated the number of somatic mutations in each sample. The samples with the number of somatic mutations in the top 25% were defined as the genomic unstable (GU)-like group. The samples with the number of somatic mutations in the bottom 25% were defined as the genomically stable (GS)-like group. We combined the lncRNA expression matrix of TCGA-KIRC with the GU and GS groups and obtained each group's lncRNA expression matrix. We then conducted a difference analysis on these two lncRNAs matrixes; |fold change|> 1 and false discovery rate

adjusted P <0.05 were defined as genome instability-associated lncRNAs. The result of genome instability-associated lncRNAs difference analysis is displayed in Table 1.

2.3 Functional enrichment analysis

We calculated the correlations between each protein-coding gene and the lncRNAs obtained as described above using the Pearson correlation coefficient method[16]. We ranked these protein coding factors in descending order according to the correlation and selected mRNAs with the top 10 correlation coefficients as the co-expression coding genes of lncRNA. Using functional analysis of these co-expressed coding genes, we analyzed the biological functions of these genetically unstable lncRNAs. Gene Ontology (GO) enrichment was performed using the clusterProfiler package in R, version 3.6.3[17].

2.4 Statistical analysis

We used Euclidean distances and Ward's linkage method to perform hierarchical cluster analyses[18]. We used univariate Cox proportional hazard regression analysis to calculate the associations between expression level of genome instability-associated lncRNAs and overall survival. We performed multivariate Cox proportional hazard regression analysis to evaluate the weighting coefficient in the risk signature. The genome instability-related lncRNA (GILncSig) for overall survival was as follows: Log[h(ti)/h0(ti)] = a1X1+ a2X2 + a3X3 + ⋯akXk, where h(ti) is the function hazard, and h0(ti) is the baseline hazard, X1, X2, X3, ⋯Xk are covariates,

and a1, a2, and a3 are the corresponding multivariate Cox proportional hazard regression coefficients. A detailed introduction can be found in our previous articles[19]. We were using the same best cut-off point (the point is determined by the samples, with the maximum sensitivity and specificity in time-dependent receiver operating characteristic (ROC) curve). Hazard ratio (HR) and 95% confidence interval (CI) were calculated using Cox analysis. The Kyoto Encyclopedia of Genes and Genomes (KEGG)[20] pathway of genome instability-related lncRNAs were identified using gene set variation analysis[21]. All statistical analyses were performed using R-version 3.6.3.

3 Results

3.1 Differences in long non-coding RNAs

The flow chart of this paper was shown in Figure 1

To identify non-coding genes related to genome instability, we grouped them according to the number of somatic mutations. We placed the first 25% of somatic mutations (84 samples) into the genetically unstable group and then placed the final 25% of somatic mutations (84 samples) into the genetically stable group. We screened and obtained differential non-coding RNAs using the limma package. We screened a total of 26 non-coding differential RNAs, of which 17 were down-regulated, and nine were up-regulated (Table 1). The levels of differential non-coding RNA expression in both groups are shown in Figure 2A.

3.2 Genome instability-related lncRNA

We performed unsupervised clustering of all samples in KIRC based on the expression levels of these 26 lncRNAs (Figure 2B). We obtained two clustering results, and the number of somatic mutations in the two groups was significantly different (Figure 2C, P = 5.3e-13, Mann–Whitney U-test). Next, we compared the expression levels of the genomic instability driver UBQLN4 in the GS-like and the GU-like groups (Figure 2D). We found that the expression of UBQLN4 was significantly up-regulated in the genetically unstable group. Based on these results, we tested whether samples with different mutation levels could be distinguished based on expression levels of the 26 differential lncRNAs, and indirectly demonstrate that these lncRNAs may be related to genome stability.

3.3 LncRNA-mRNA co-expression network

Based on Pearson correlation coefficients, we determined the top 10 mRNAs that correlated with each lncRNA. We created a co-expression network lncRNAs and mRNAs (Figure 3A). We then analyzed the function of the mRNAs in the co-expression module to determine the associated biological processes. GO enrichment demonstrated that these protein-coding genes are related to biological processes such as monovalent inorganic cation homeostasis and pH regulation (Figure 3B). This analysis suggests that the 26 genomically unstable non-coding RNAs may affect genome stability by regulating their co-expression networks. We

found that these co-expressed protein-coding genes might regulate internal homeostasis processes such as regulation of pH and cations' homeostasis, thereby destroying cell stability. In total, we identified 26 non-coding RNAs related to genome instability.

3.4 The genome instability-related lncRNA risk model

We clarified the lncRNAs and biological processes related to genetic stability. Next, we calculated the correlations between these lncRNAs and clinical survival phenotypes. We randomly divided 507 clear cell carcinoma samples with detailed follow-up information into training groups and validation groups. We constructed a multivariate Cox proportional hazard regression model for ccRCC in the training set based on 26 genomic stable state-related lncRNAs. The coefficients of the risk factors in the model are shown in Table 2. Risk model (GILncSig) = 0.095 * LINC00460 + 0.165 * LINC01234 + 0.152 * AL139351.1 + 0.177 * MIR222HG + 0.123 * AC087636.1-0.027 * LINC02471. We found that LINC00460, LINC01234, AL139351.1, MIR222HG, AC087636.1 were transparent risk factors. The higher their expression, the worse the overall survival of patients with renal cancer. LINC02471 is a protective factor for ccRCC. The higher its expression, the better the overall survival.

lncRNA expression patterns and the distribution of somatic mutation count distribution and UBQLN4 expression for patients in high- and low-risk groups in the training set and testing set are shown in Supplementary Figure 1.

3.5 The verification and evaluation of lncRNA model performance

Risk scores for each sample in the training and test sets were calculated using the GILncSig method. Patients were divided into groups according to the median risk score (0.853); patients in the higher risk group had a risk score >0.853. We then calculated the survival difference between the high- and low-risk groups using survival analysis. In TCGA-KIRC cohort, we found that patients in the low-risk group had better clinical outcomes (Figure 4A, P <0.001). Patients in the low-risk group in the training set (Figure 4B, P <0.001) and validation set (Figure 4C, P <0.001) also had better survival outcomes. The area under the time-dependent ROC curve of TCGA-KIRC cohort was 0.681 (Figure 4D). The area under the time-dependent ROC curve of the training set cohort was 0.726 (Figure 4E). The area under the time-dependent ROC curve of the verification set cohort was 0.642 (Figure 4F). MutS homolog 2 (MSH2) and replication factor C subunit 1 (RFC1) are involved in the process of mismatch recognition. Comparison analysis showed significant differences in MSH2 and RFC1 expression patterns between the samples in the high- and low-risk groups (Figure 5). Expression levels of MSH2 in the low-risk group were significantly higher than those of the low-risk group (P <0.001, Mann–Whitney U-test; Figure 5A-C). RFC1 also showed higher expression levels in low-risk patients than in high-risk patients (P < 0.001, Mann–Whitney U-test; Figure 5D-F).

3.6 Subgroups of the lncRNA model

We then obtained a stable genomic stability-related lnc prognosis model. To further analyze their performance levels in various subgroups, we conducted survival analysis. We found that subgroups of patients in the low-risk group achieve better outcomes (Figure 6).

3.7 Tumor mutation landscapes in high- and low-risk groups

To compare mutations in the high- and low-risk groups, we drew a panorama of mutations in the two groups (Figure 7). A total of 88.24% of the samples had mutations in the low-risk group. The top 10 mutated genes included VHL, PBRM1, TTN, SETD2, BAP1, and MUC16. The high-risk group's mutation frequency (84.62%) was lower than that of the low-risk group (88.24%). The top 10 factors associated with mutations were the same as those of the low-risk group.

3.8 Performance comparison in terms of AUC

To determine the accuracy of clinical predictive models related to genome stability, we performed diagnostic test comparisons. Three recently published lncRNA signatures were involved in comparisons: the three-lncRNA signature derived from Zhang et al. (Zhang Dan)[22], the four-lncRNA signature derived from Liu et al. LiulncSig)[23] and an immune signature derived from Sun et al. (SunlncSig)[24] using the same TCGA patient cohort. As shown in Figure 8, the AUC of overall survival for the GILncSig was 0.681, which was significantly higher than those of SunlncSig (AUC = 0.657) and LiulncSig (AUC = 0.656) (Figure 8). Although our model's AUC was lower than Zhang

Dan's model, our training set score was 0.726.

3.9 GSVA pathway correlation analysis

We obtained genome stability-related lncRNA in various somatic mutation groups; however, we believe that the lncRNA obtained based on differential analysis alone is insufficient to conclude that they are related to genome stability. Therefore, in this section, we obtained genomic stability-related pathway scores of each sample using the GSVA method. We calculated the Pearson correlation coefficients of these genomic stability pathway scores and the differences in lncRNA. We directly explained the pathway that these factors regulated that affected the stability of the genome. Figure 8 shows that the base excision repair pathway, the DNA replication pathway, homologous recombination, the mismatch repair pathway, the p53 signaling pathway, and ubiquitin-mediated proteolysis were related to LINC0460 and LINC01234. The interaction of these pathways appears to ensure the stability of the genome (Figure 9). For these reasons, we believe LINC0460 and LINC01234 affect the stability of the genome by regulating these pathways.

Discussion

The genome structure's relative stability is a prerequisite for the maintenance and continuation of the biological germline. It is crucial to ensure that a set of effective mechanisms is formed in the cell. There is a stable and accurate transmission of

genetic information from generation to generation. Chromosome instability refers to the increased probability of acquiring chromosomal aberrations due to defects in processes such as DNA repair, replication, or chromosome segregation. Genome stability is closely related to the occurrence and progression of cancer[25-27]. Common DNA damage types include DNA base modification, DNA inter-strand, and intra-strand cross-links, and DNA single-strand and double-strand breaks[28]. Such DNA damage often leads to genome instability. Proteins related to DNA damage repair, DNA replication, and cell cycle checkpoints work together to ensure the integrity of the genome and the DNA structure's integrity. However, mutations in these proteins can lead to the accumulation of mutations in chromosomes; as these mutations accumulate, they cause cancer and premature aging[26, 27, 29]. There is no accurate quantitative way to describe genome instability. Various efforts are underway to identify protein-coding genes and microRNAs related to genomic instability that predict outcomes[30-32].

LncRNAs are non-coding RNAs with lengths of more than 200 nucleotides. They participate in biological processes such as dose compensation, epigenetic regulation, cell cycle regulation, and cell differentiation regulation. For these reasons, lncRNA has become a hot area of research in genetics. LncRNA is related to tumor progression, can be used as a prognostic marker of the disease, and affects genome stability[33-35]. Munschauer et al. found that NORAD lncRNA was related to the topoisomerase, critical for genome stability[36]. Hu et al. found that GUARDIN

lncRNA was related to p53-responsive elements and is essential for genomic stability[37].

Although we have made substantial efforts to identify lncRNAs related to genomic instability, whole-genome identification of lncRNA and its clinical research are still in their early stages.

Based on TCGA clear cell cancer cohort and the corresponding number of somatic mutations, we identified 26 differences related to the number of somatic mutations at the computational level. However, the analysis in computational biology is insufficient. Therefore, we combined clinical prognostic phenotype. A clinical predictive lncRNA model was constructed. We found that six lncRNAs in the model could be used as independent prognostic markers for renal cancer. According to our understanding, genome stability is closely related to levels of p53 mutations, DNA repair, and base mismatch repair. On account of the cumulative effect of these factors, normal cells gradually become cancer cells. According to our previous description, the six lncRNAs in the model should be closely related to these processes. Therefore, to verify this point of view, we performed GSVA gene set analysis and obtained the KEGG pathway scores corresponding to each sample. Then, the Pearson correlation coefficient test was performed using these pathways. LINC00460 and LINC01234 are the most relevant to these genomic stability pathways. We demonstrated that this method could screen candidate genome stability-related lncRNAs and identify the relevant pathways involved in these lncRNAs through GSVA analysis.

After a careful literature search, we found that the biological process of LINC00460 and LINC01234 in the GILncSig has not been reported to date. We found that the lncRNA LINC00460 was located on chromosome 13q33.2 and is a prognostic biomarker for esophageal squamous cell carcinoma[38] and renal carcinoma[22]. Another lncRNA, LINC01234, is located on chromosome 12q24.13. LINC01234 was found to regulate proliferation, migration, and invasion of ccRCC cells via the HIF-2α pathway[39]. Although studies have demonstrated the relationship between these two factors and outcomes of RCC, they do not explain the specifically related mechanisms. Finally, by analyzing the GSVA pathway, we found that they have the strongest correlation with the p53 pathway and affect the stability of the genome.

There are some limitations to our study. First, we did not conduct cell or animal experiments. Second, we only identified 26 genomic stability-related lncRNAs; nevertheless, computational biology techniques demonstrated the connection between LINC00460 and LINC01234 and the genome stability pathway. Underlying regulatory mechanisms require further exploration.

In conclusion, we constructed a screening system for genome stability-related lncRNAs, and we identified 26 genomic stability-related lncRNAs. We used these lncRNAs to predict outcomes in patients with ccRCC and found that these lncRNAs can be used as independent predictors. Finally, using GSVA pathway correlation analysis, we found that LNC00460 and LINC01234 are related to genome stability, and we indirectly demonstrated the appropriateness of this strategy.

Author information:

Yutao Wang and Jianbin Bi

Department of Urology, China Medical University, The First Hospital of China Medical University, Shenyang, Liaoning, China

Kexin Yan

Department of Dermatology, China Medical University, The First Hospital of China Medical University, Shenyang, Liaoning, China

Yutao Wang and Kexin Yan contributed equally to this paper

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials: "The datasets analysed during the current study are available in the TCGA repository, [https://portal.gdc.cancer.gov/] and GEO repository [https://www.ncbi.nlm.nih.gov/geo/]"

Competing interests: The authors declare no conflict of interest.

Funding: This work was supported by China Shenyang Science and Technology Plan (20-205-4-015).

Authors' contributions: Yutao Wang, Kexin Yan and Jianbin Bi designed the study; Yutao Wang, and Kexin Yan analyzed and wrote the manuscript. All authors read and agreed to the final version of the manuscript.

Acknowledgements: We appreciate the free use of TCGA and GEO databases

Figure Legends

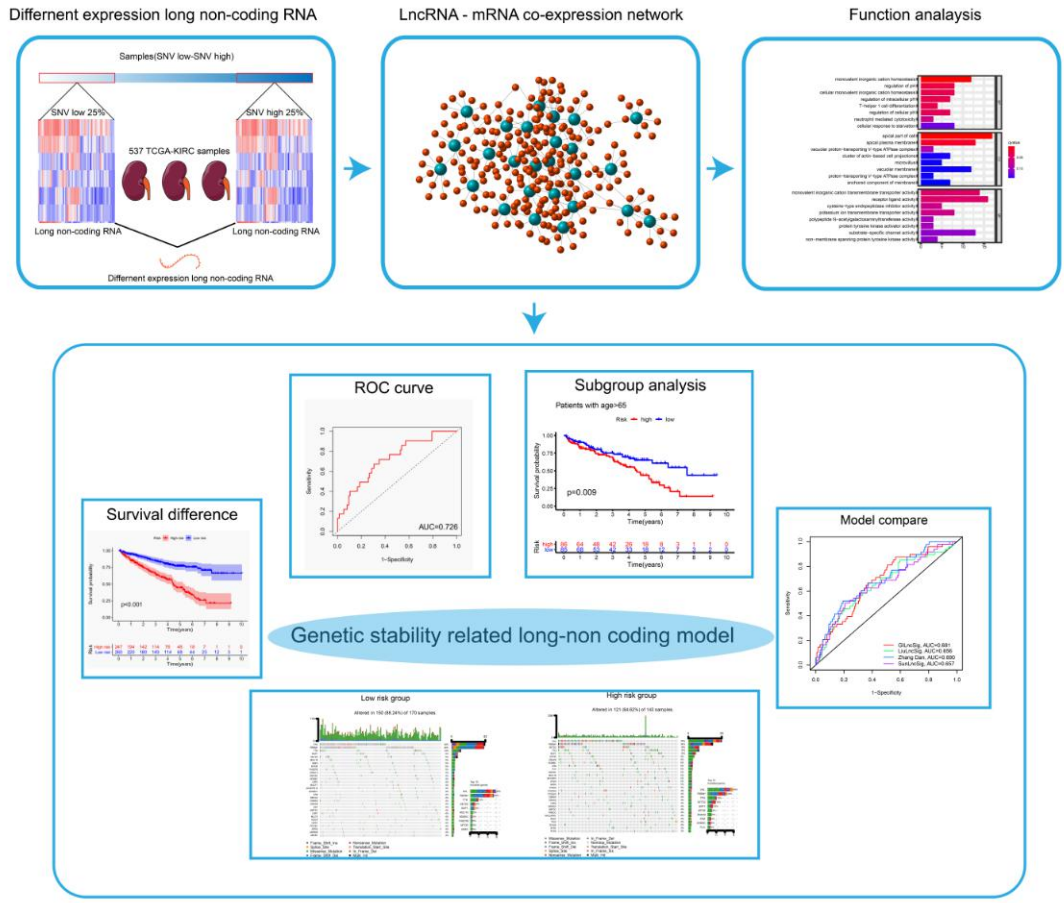

Figure 1: The design flow chart of this study. Clinical follow-up information of renal clear cell carcinoma, protein-coding RNA expression data, long non-coding RNA expression data, and somatic mutation information were downloaded from the TCGA database, and the samples were then divided into training sets and test sets. The

samples were then divided into two groups for difference analysis according to gene mutation. According to the results of difference analysis, the overall samples were divided into gene stable group and gene unstable group by consensus cluster analysis. Then lncRNA-mRNA co-expression network was constructed, and the pathway analysis and GSVA scores were performed for this network. Then a COX regression prognostic model was established, and the model verification processes such as survival analysis, clinical subgroup analysis, tumor mutation burden analysis and model comparison were carried out.

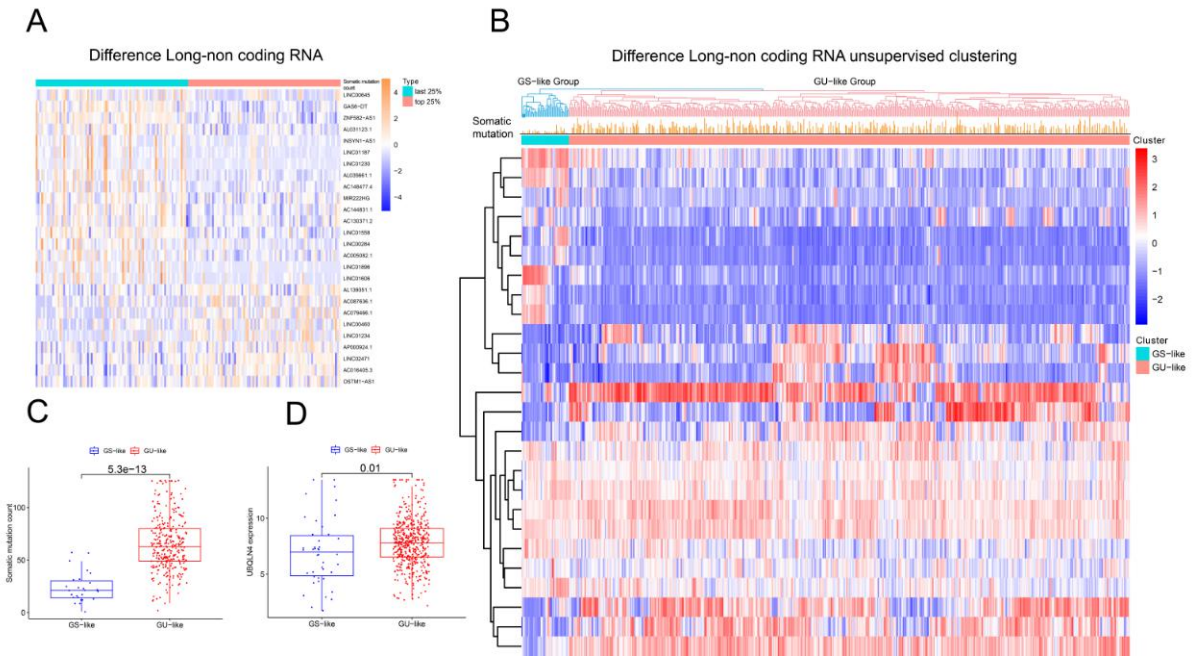

Figure 2: (A) Difference analysis of the group that Somatic cell mutations are in the top 25% between the group that Somatic cell mutations are in the last 25% in RCC. (B) Unsupervised clustering of GS-group and Gu-group. (C) The difference of somatic cell mutation number between GS-group and GU-Group. (D) The different expression of UBQLN4 in GS-Group and GU-Group.

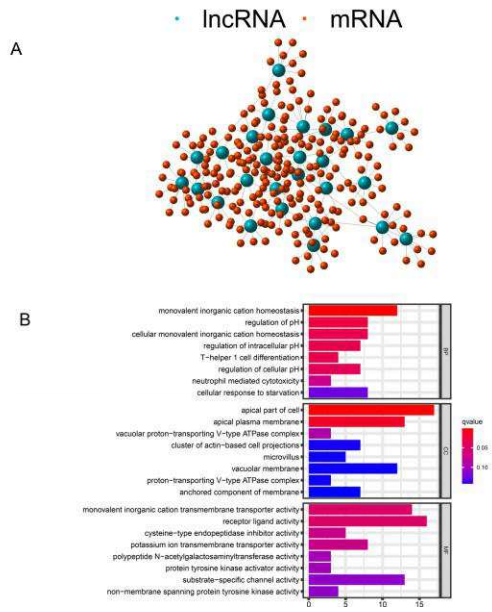

Figure 3: (A) The co-expression network of lncRNA-mRNA. Green stands for LncRNA and red for mRNA. The closer the relationship, the closer the connection. (B) Go analysis of the lncRNA-mRNA network. In the biological process, the network is mainly enriched in the monovalent inorganic homeostasis. In the cellular component, the network is mainly enriched in apical part of cell and apical plasma membrane. In the molecular function, the network is mainly enriched in monovalent inorganic cation transmembrane transporter activity and receptor ligand activity.

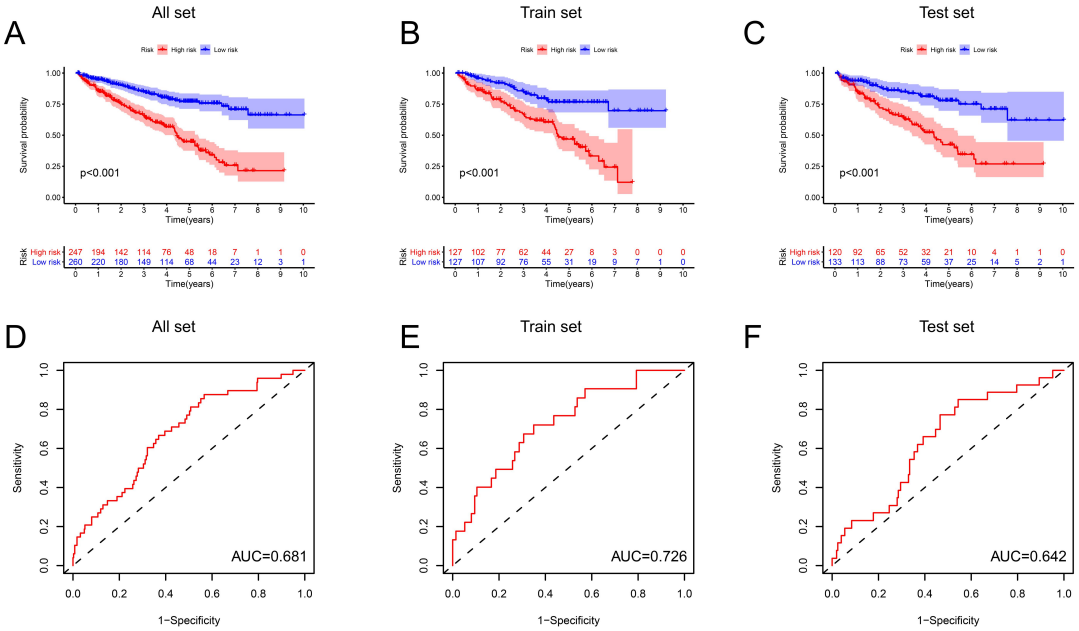

Figure 4: Survival analysis and ROC curve. (A-C) A COX prognostic regression model was established to calculate the scoring threshold, and a survival analysis was performed to assess the difference between the high-risk and low-risk groups. In the all set, train set and test set, patients in the low-risk group had a better prognosis than those in the high-risk group (P<0.01). (D-F) The area under the ROC curve of the all set was 0.681, the area under the ROC curve of the train set was 0.726, and the area under the ROC curve of the test set was 0.642. The model shows good predictive ability.

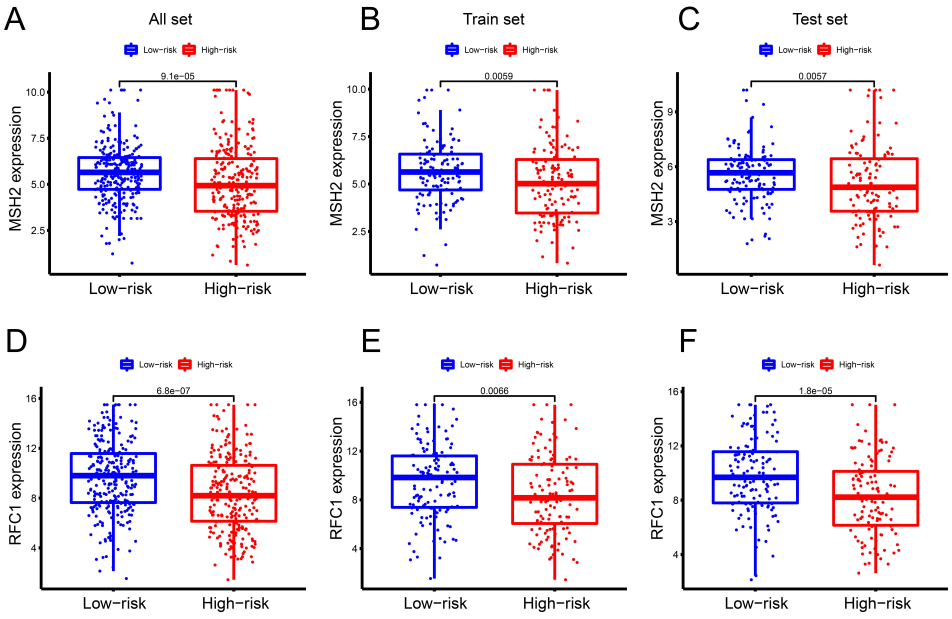

Figure 5: (A-C) The previously reported genetic instability related factor MSH2 showed significant differences in expression patterns between high-risk group and low-risk group in the all set (P = 9.1e-05), train set (P = 0.0059) and test set (P = 0.0057). (D-F) The previously reported genetic instability related factor RFC1 showed significant differences in expression patterns between high-risk group and low-risk group in the all set (P = 6.8e-07), train set (P = 0.0066) and test set (P = 1.8E-05).

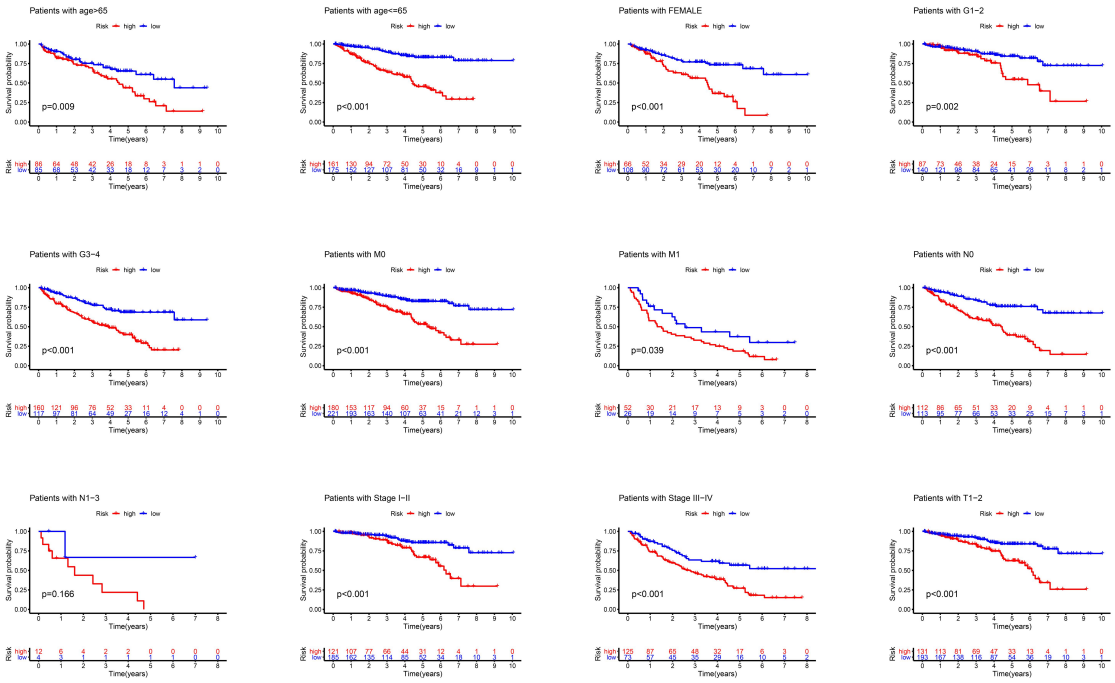

Figure 6: Subgroup analysis.The samples were divided into multiple clinical subgroups according to age, sex, stage, metastasis, and infiltration of lymph nodes. The results showed that in all clinical subgroups, the low-risk group had a better prognosis.

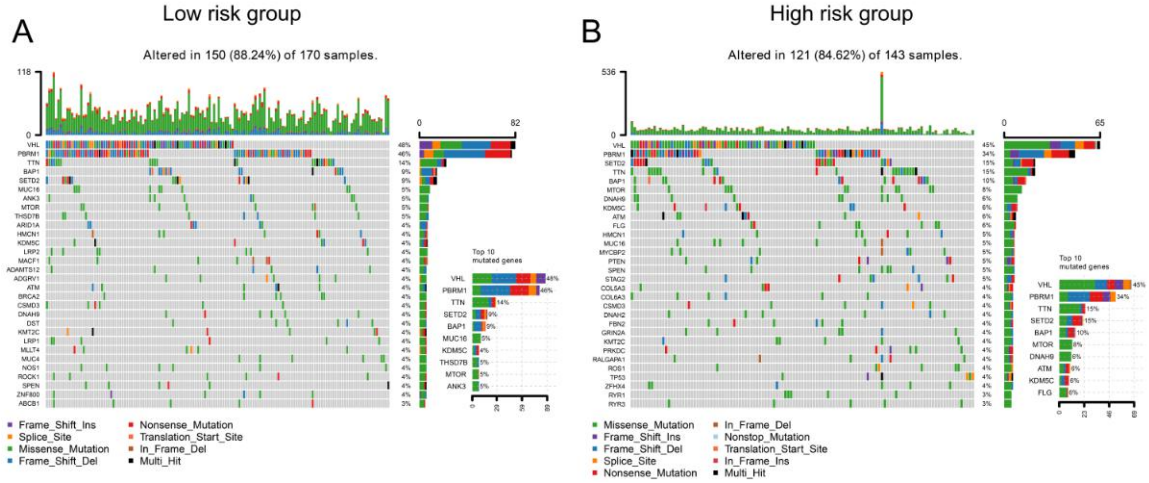

Figure 7: Waterfall map of gene mutation burden. (A) In the low-risk group, the mutation rate was 88.24%. The top three mutated genes were VHL, PBRM1 and TTN. (B) In the high-risk group, the mutation rate was 84.62%. The top five mutated genes were VHL, PBRM1, SETD2, TTN and BAP1.

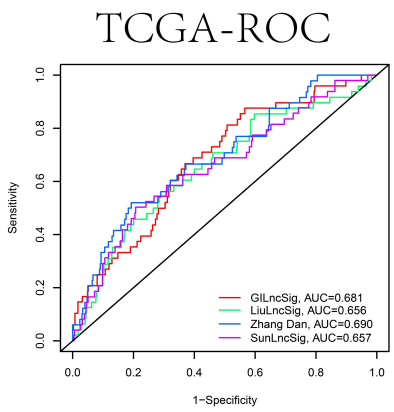

Figure 8: Model comparison. The model proposed in this paper is compared with the model of Liu et al., Sun et al., and Zhang et al., and the model presented in this paper has the highest ROC value, indicating the best evaluation ability.

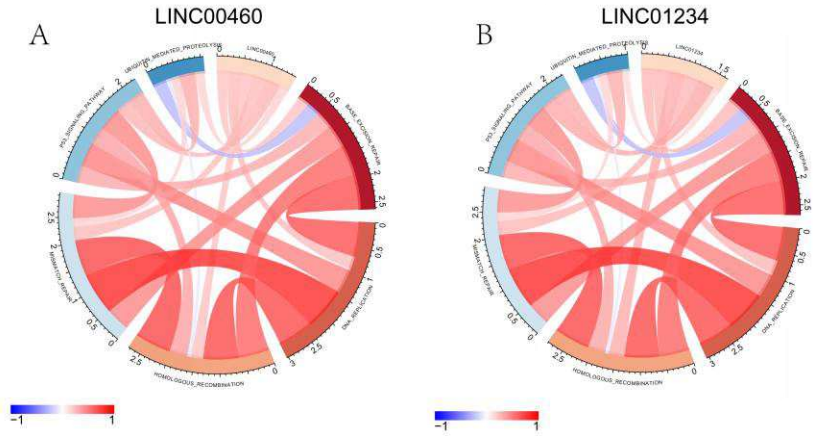

Figure 9: Correlation analysis of lncRNA and genomic instability related pathways. Red represents positive correlation and blue represents negative correlation.

Table 1 lncRNAs related to genetic instability

lncRNA	logFC	p-value	FDR
ZNF582-AS1	-1.067	2.78E-10	1.32E-07

LINC01558	-1.926	3.72E-10	1.32E-07
GAS6-DT	-1.559	1.40E-09	3.30E-07
AL035661.1	-6.144	3.65E-07	5.17E-05
AC016405.3	1.376	6.21E-05	2.20E-03
AC005082.1	-2.200	7.72E-05	2.53E-03
LINC01187	-8.158	8.16E-05	2.53E-03
AL031123.1	-2.978	9.95E-05	2.82E-03
LINC02471	1.040	1.08E-04	2.94E-03
AC079466.1	3.017	1.30E-04	3.18E-03
LINC01606	-4.716	1.41E-04	3.33E-03
LINC01230	-7.363	1.74E-04	3.98E-03
AC148477.4	-4.784	2.06E-04	4.38E-03
LINC01896	-6.526	3.24E-04	6.20E-03
AC144831.1	-1.297	5.42E-04	8.53E-03
LINC00284	-3.520	1.14E-03	1.32E-02
AL139351.1	1.105	1.40E-03	1.52E-02
LINC01234	2.170	1.63E-03	1.67E-02
LINC00460	1.276	1.75E-03	1.75E-02
MIR222HG	-1.371	2.25E-03	1.87E-02
AP000924.1	1.031	2.15E-03	1.87E-02
LINC00645	-2.101	2.24E-03	1.87E-02
OSTM1-AS1	1.425	3.90E-03	2.71E-02

AC130371.2	-1.271	4.05E-03	2.73E-02
INSYN1-AS1	-5.803	6.46E-03	3.94E-02
AC087636.1	1.708	8.52E-03	4.93E-02

lncRNA: Long non-coding RNAs; logFC: log_2Fold Change; FDR: False discovery rate

Table 2. Multivariate Cox proportional hazard regression analysis results

ID	coef	HR	HR.95L	HR.95H	P-value
LINC00460	0.095	1.099	1.010	1.196	0.028
LINC01234	0.165	1.180	0.984	1.414	0.044
AL139351.1	0.152	1.164	0.966	1.402	0.010
MIR222HG	0.177	1.194	1.002	1.422	0.047
AC087636.1	0.123	1.131	1.013	1.263	0.029
LINC02471	-0.027	0.973	0.934	1.014	0.048

Coef: coefficient; HR: hazard rate

References

1. Dagher J, Delahunt B, Rioux-Leclercq N, Egevad L, Srigley JR, Coughlin G, Dunglinson N, Gianduzzo T, Kua B, Malone G, Martin B, Preston J, Pokorny M, Wood S, Yaxley J, Samaratunga H. Clear cell renal cell carcinoma: validation of World Health Organization/International Society of Urological Pathology grading. Histopathology. 2017;71:918-25.
2. Makhov P, Joshi S, Ghatalia P, Kutikov A, Uzzo RG, Kolenko VM. Resistance to Systemic Therapies in Clear Cell Renal Cell Carcinoma: Mechanisms and Management Strategies. Mol Cancer Ther. 2018;17:1355-64.
3. Mehdi A, Riazalhosseini Y. Epigenome Aberrations: Emerging Driving Factors of the Clear Cell Renal Cell Carcinoma. Int J Mol Sci. 2017;18.
4. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic Performance of DWI for Differentiating High- From Low-Grade Clear Cell Renal Cell Carcinoma: A Systematic Review and Meta-Analysis. AJR Am J Roentgenol. 2017;209:W374-374W381.
5. Vera-Badillo FE, Templeton AJ, Duran I, Ocana A, de Gouveia P, Aneja P, Knox JJ, Tannock IF, Escudier B, Amir E. Systemic therapy for non-clear cell renal cell carcinomas: a systematic review and meta-analysis. Eur Urol. 2015;67:740-9.
6. Abbas T, Keaton MA, Dutta A. Genomic instability in cancer. Cold Spring Harb Perspect Biol. 2013;5:a012914.

7. Pikor L, Thu K, Vucic E, Lam W. The detection and implication of genome instability in cancer. Cancer Metastasis Rev. 2013;32:341-52.
8. Gaillard H, García-Muse T, Aguilera A. Replication stress and cancer. Nat Rev Cancer. 2015;15:276-89.
9. Reis AH, Vargas FR, Lemos B. Biomarkers of genome instability and cancer epigenetics. Tumour Biol. 2016;37:13029-38.
10. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. Landscape of transcription in human cells. Nature. 2012;489:101-8.
11. Renganathan A, Felley-Bosco E. Long Noncoding RNAs in Cancer and Therapeutic Potential. Adv Exp Med Biol. 2017;1008:199-222.
12. Thin KZ, Liu X, Feng X, Raveendran S, Tu JC. LncRNA-DANCR: A valuable cancer related long non-coding RNA for human cancers. Pathol Res Pract. 2018;214:801-5.
13. Fang Y, Fullwood MJ. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. Genomics Proteomics Bioinformatics. 2016;14:42-54.
14. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer Res. 2017;77:3965-81.
15. Wang Z, Jensen MA, Zenklusen JC. A Practical Guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol. 2016;1418:111-41.
16. Pripp AH. [Pearson's or Spearman's correlation coefficients]. Tidsskr Nor Laegeforen. 2018;138.
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25-9.
18. Vogt W, Nagel D. Cluster analysis in diagnosis. Clin Chem. 1992;38:182-98.
19. Wang Y, Lin J, Yan K, Wang J. Identification of a Robust Five-Gene Risk Model in Prostate Cancer: A Robust Likelihood-Based Survival Analysis. Int J Genomics. 2020;2020:1097602.
20. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353-353D361.
21. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
22. Zhang D, Zeng S, Hu X. Identification of a three-long noncoding RNA prognostic model involved competitive endogenous RNA in kidney renal clear cell carcinoma. Cancer Cell Int. 2020;20:319.
23. Liu Y, Gou X, Wei Z, Yu H, Zhou X, Li X. Bioinformatics profiling integrating a four immune-related long non-coding RNAs signature as a prognostic model for papillary renal cell carcinoma. Aging (Albany NY). 2020;12:15359-73.
24. Sun Z, Jing C, Xiao C, Li T. Long Non-Coding RNA Profile Study Identifies an Immune-Related lncRNA Prognostic Signature for Kidney Renal Clear Cell Carcinoma. Front Oncol. 2020;10:1430.
25. Aguilera A, García-Muse T. Causes of genome instability. Annu Rev Genet. 2013;47:1-32.

26. Cha HJ, Yim H. The accumulation of DNA repair defects is the molecular origin of carcinogenesis. Tumour Biol. 2013;34:3293-302.
27. Basu AK. DNA Damage, Mutagenesis and Cancer. Int J Mol Sci. 2018;19.
28. Kramara J, Osia B, Malkova A. Break-Induced Replication: The Where, The Why, and The How. Trends Genet. 2018;34:518-31.
29. Cotterill S. Diseases Associated with Mutation of Replication and Repair Proteins. Adv Exp Med Biol. 2018;1076:215-34.
30. Mettu RK, Wan YW, Habermann JK, Ried T, Guo NL. A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types. Int J Biol Markers. 2010;25:219-28.
31. Ferguson LR, Chen H, Collins AR, Connell M, Damia G, Dasgupta S, Malhotra M, Meeker AK, Amedei A, Amin A, Ashraf SS, Aquilano K, Azmi AS, Bhakta D, Bilsland A, Boosani CS, Chen S, Ciriolo MR, Fujii H, Guha G, Halicka D, Helferich WG, Keith WN, Mohammed SI, Niccolai E, Yang X, Honoki K, Parslow VR, Prakash S, Rezazadeh S, Shackelford RE, Sidransky D, Tran PT, Yang ES, Maxwell CA. Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. Semin Cancer Biol. 2015;35 Suppl:S5-5S24.
32. Habermann JK, Doering J, Hautaniemi S, Roblick UJ, Bündgen NK, Nicorici D, Kronenwett U, Rathnagiriswaran S, Mettu RK, Ma Y, Krüger S, Bruch HP, Auer G, Guo NL, Ried T. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Int J Cancer. 2009;124:1552-64.
33. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464:1071-6.
34. Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer. Hum Mol Genet. 2010;19:R152-61.
35. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011;29:742-9.
36. Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, Ulirsch JC, Fulco CP, Subramanian V, Chen J, Schenone M, Guttman M, Carr SA, Lander ES. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. Nature. 2018;561:132-6.
37. Hu WL, Jin L, Xu A, Wang YF, Thorne RF, Zhang XD, Wu M. GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability. Nat Cell Biol. 2018;20:492-502.
38. Liang Y, Wu Y, Chen X, Zhang S, Wang K, Guan X, Yang K, Li J, Bai Y. A novel long noncoding RNA linc00460 up-regulated by CBP/P300 promotes carcinogenesis in esophageal squamous cell carcinoma. Biosci Rep. 2017;37.
39. Yang F, Liu C, Zhao G, Ge L, Song Y, Chen Z, Liu Z, Hong K, Ma L. Long non-coding RNA LINC01234 regulates proliferation, migration and invasion via HIF-2α pathways in clear cell renal cell carcinoma cells. PeerJ. 2020;8:e10149.

Figures

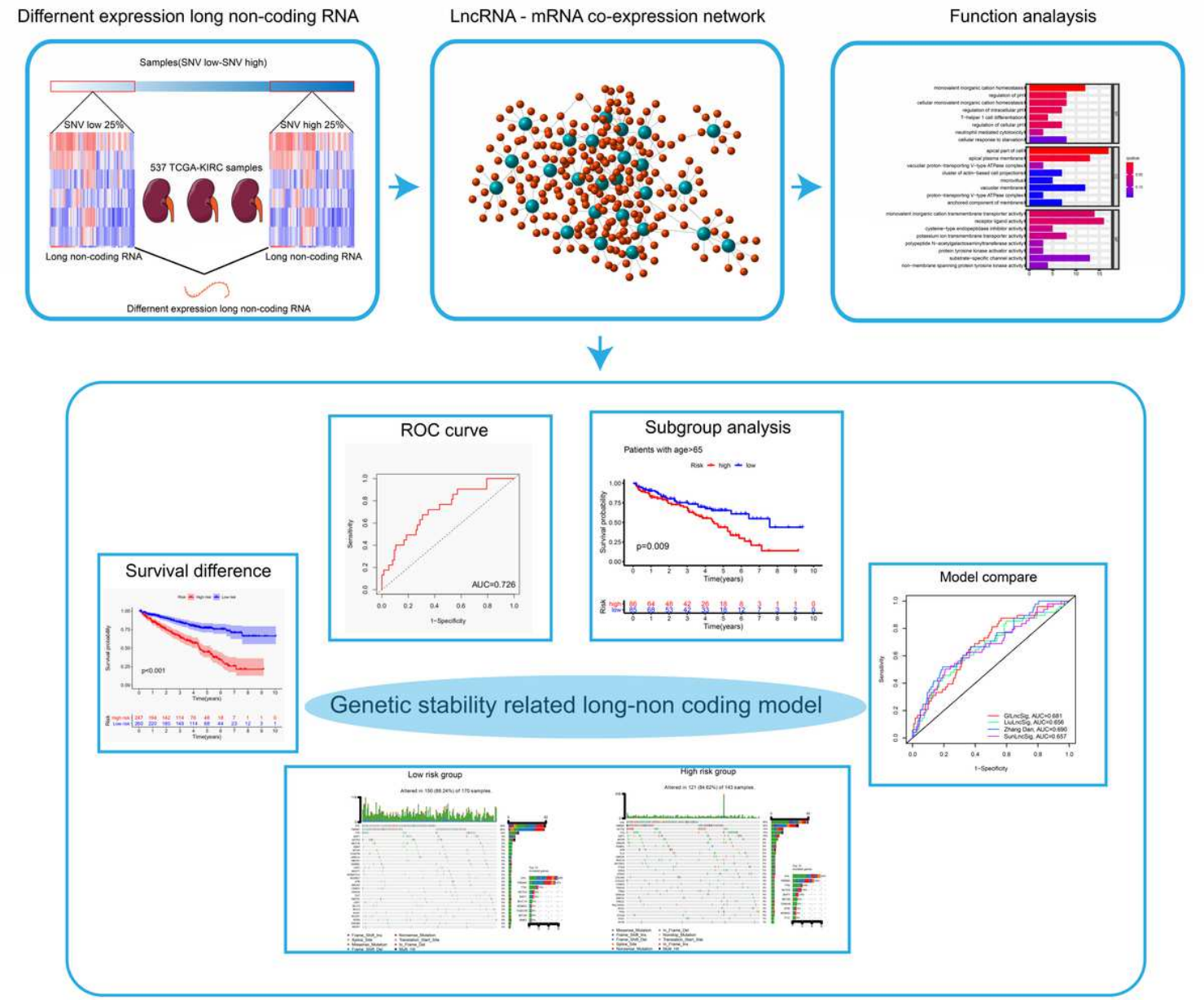

Figure 1

The design flow chart of this study. Clinical follow-up information of renal clear cell carcinoma, protein-coding RNA expression data, long non-coding RNA expression data, and somatic mutation information were downloaded from the TCGA database, and the samples were then divided into training sets and test sets. The samples were then divided into two groups for difference analysis according to gene mutation. According to the results of difference analysis, the overall samples were divided into gene stable group and gene unstable group by consensus cluster analysis. Then lncRNA-mRNA co-expression network was constructed, and the pathway analysis and GSVA scores were performed for this network. Then a COX regression prognostic model was established, and the model verification processes such as survival

analysis, clinical subgroup analysis, tumor mutation burden analysis and model comparison were carried out.

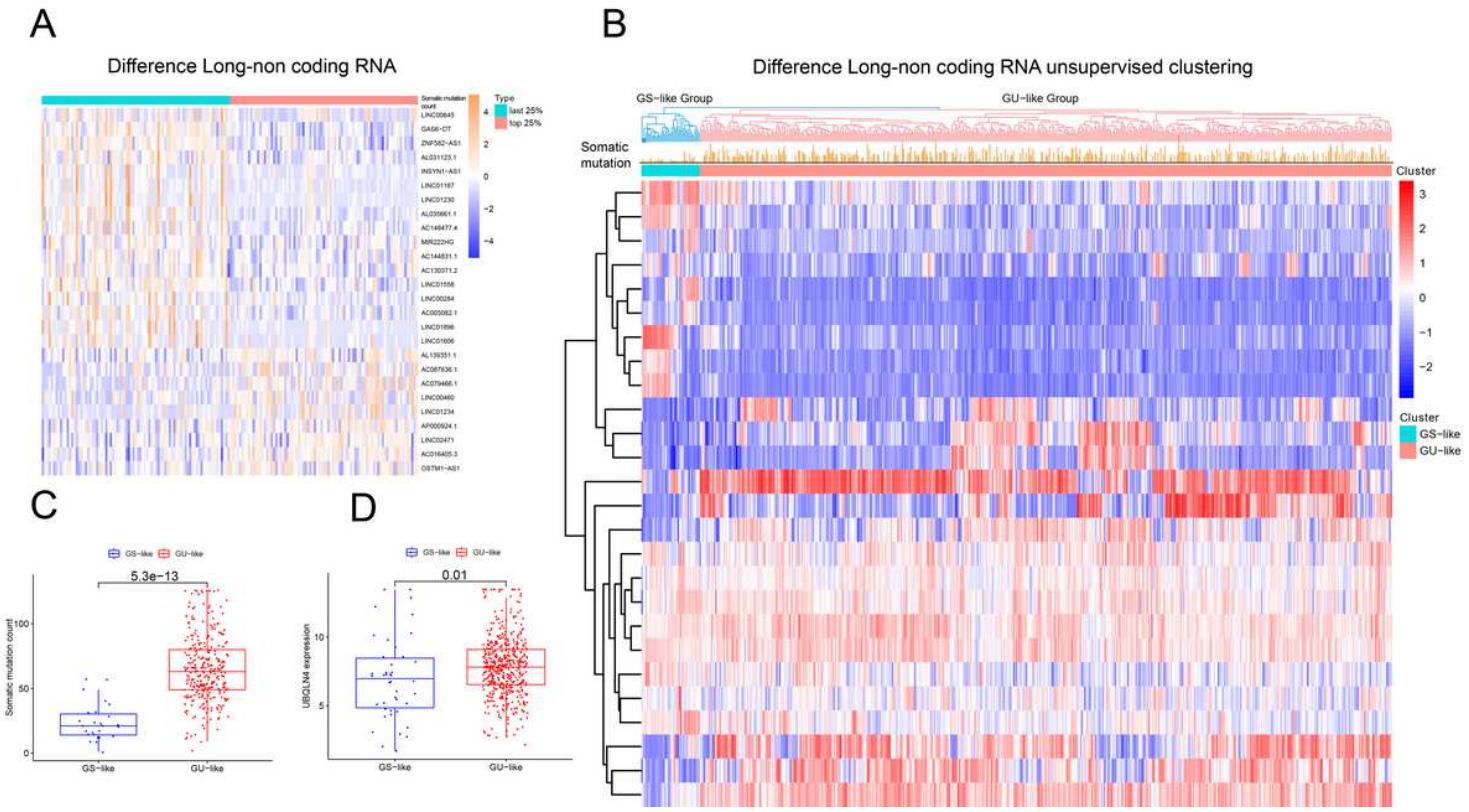

Figure 2

(A) Difference analysis of the group that Somatic cell mutations are in the top 25% between the group that Somatic cell mutations are in the last 25% in RCC. (B)Unsupervised clustering of GS-groupandGu-group. (C)Thedifference ofsomatic cell mutation number between GS-group and GU-Group. (D) The different expression of UBQLN4 in GS-Group andGU-Group.

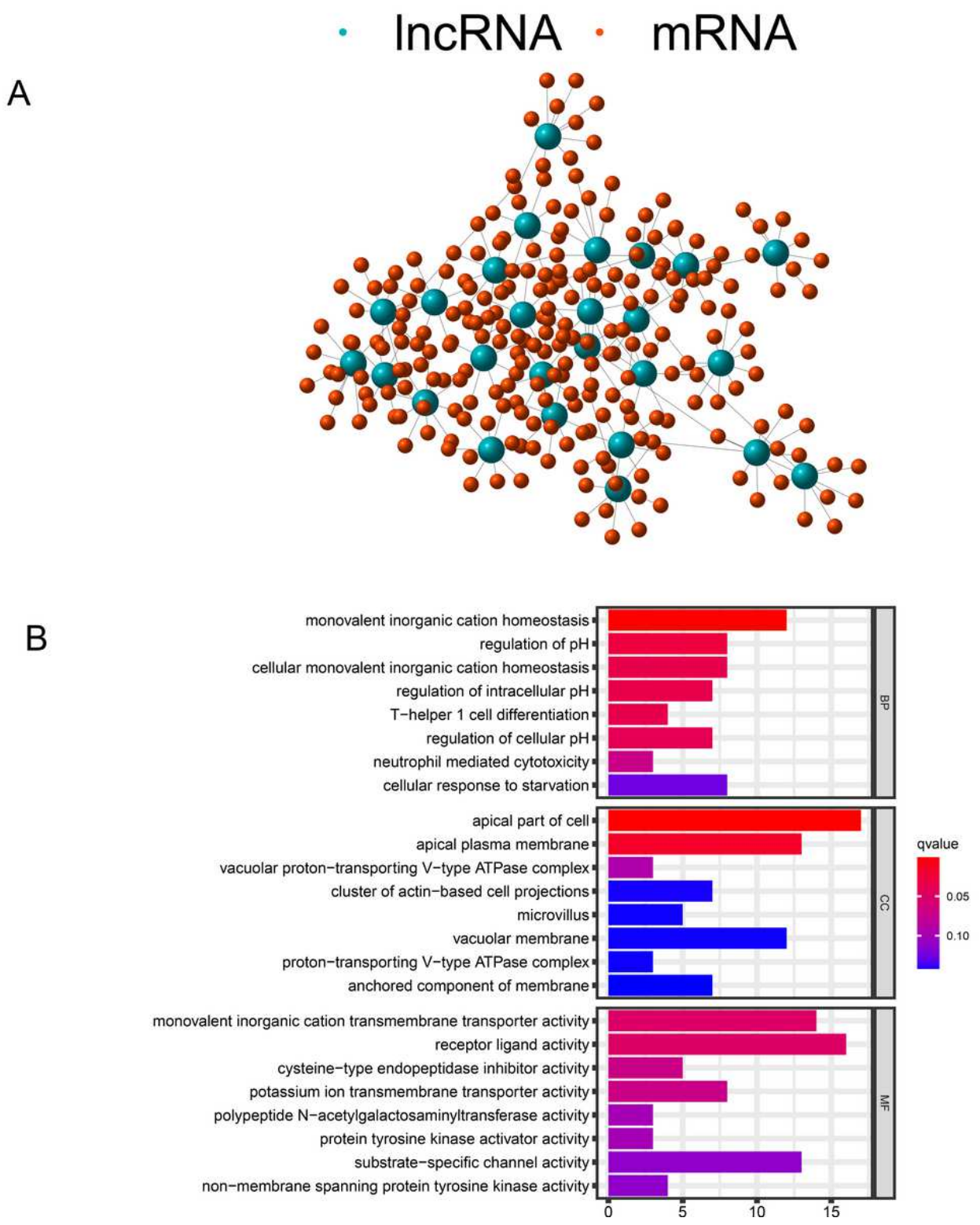

Figure 3

(A)The co-expression network of lncRNA-mRNA. Green standsforLncRNA and red for mRNA. The closer the relationship, the closer the connection. (B) Go analysis of the lncRNA-mRNA network. In the biological process, the network is mainly enriched in the monovalent inorganic homeostasis. In the cellular component, the network is mainly enriched in apical part of cell and apical plasma membrane. In the

molecular function, the network is mainly enriched in monovalent inorganic cation transmembrane transporteractivity and receptor ligand activity.

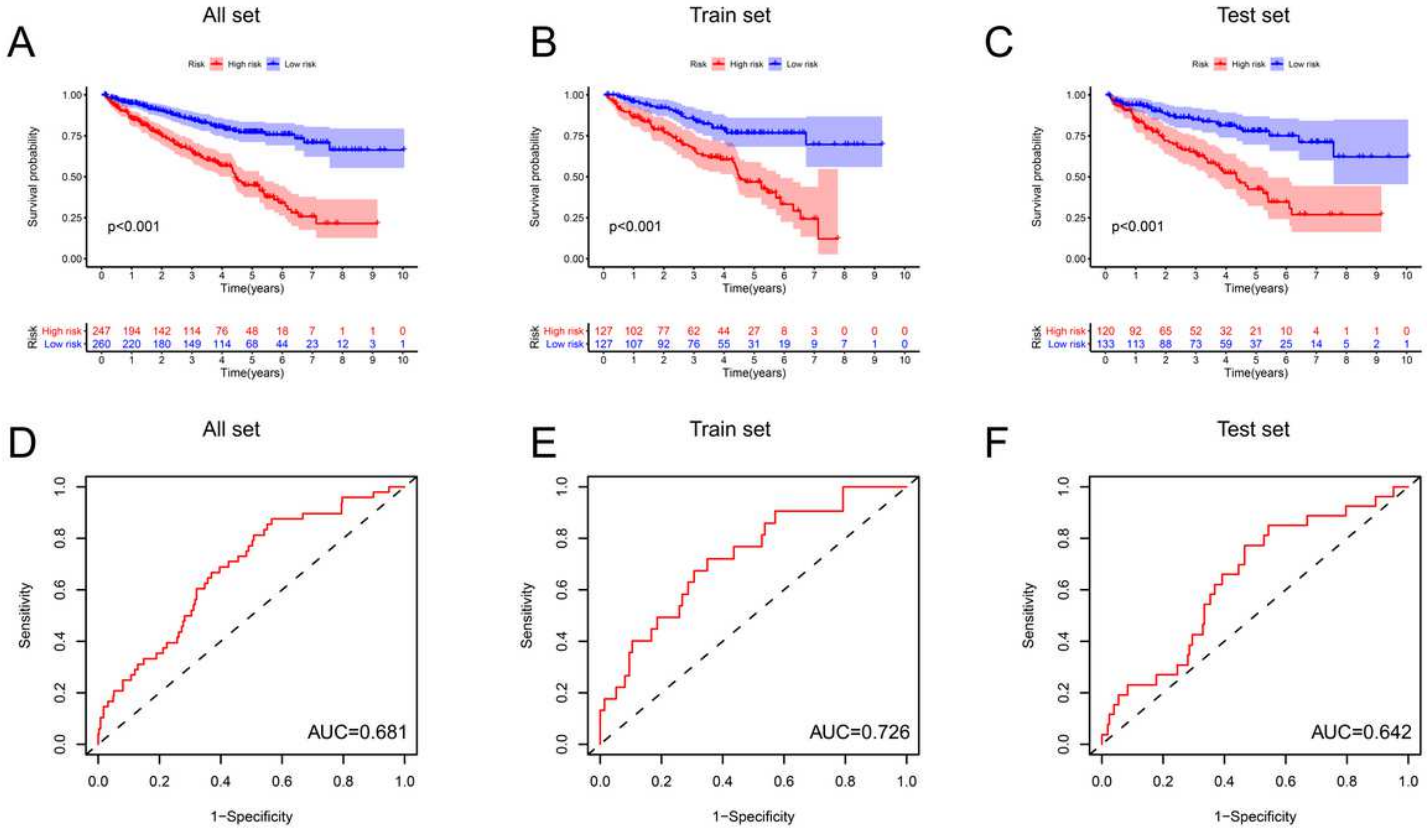

Figure 4

SurvivalanalysisandROCcurve. (A-C)ACOXprognosticregressionmodel was established to calculate the scoring threshold, and a survival analysis was performed to assess the difference between the high-risk and low-risk groups. In the allset,trainset andtestset,patientsin thelow-risk grouphada betterprognosisthan those in the high-risk group (P<0.01). (D-F) The area under the ROC curve of the all set was 0.681, the area underthe ROC curve of the train set was 0.726, and the area under the ROC curve of the test set was 0.642. The model shows good predictive ability.

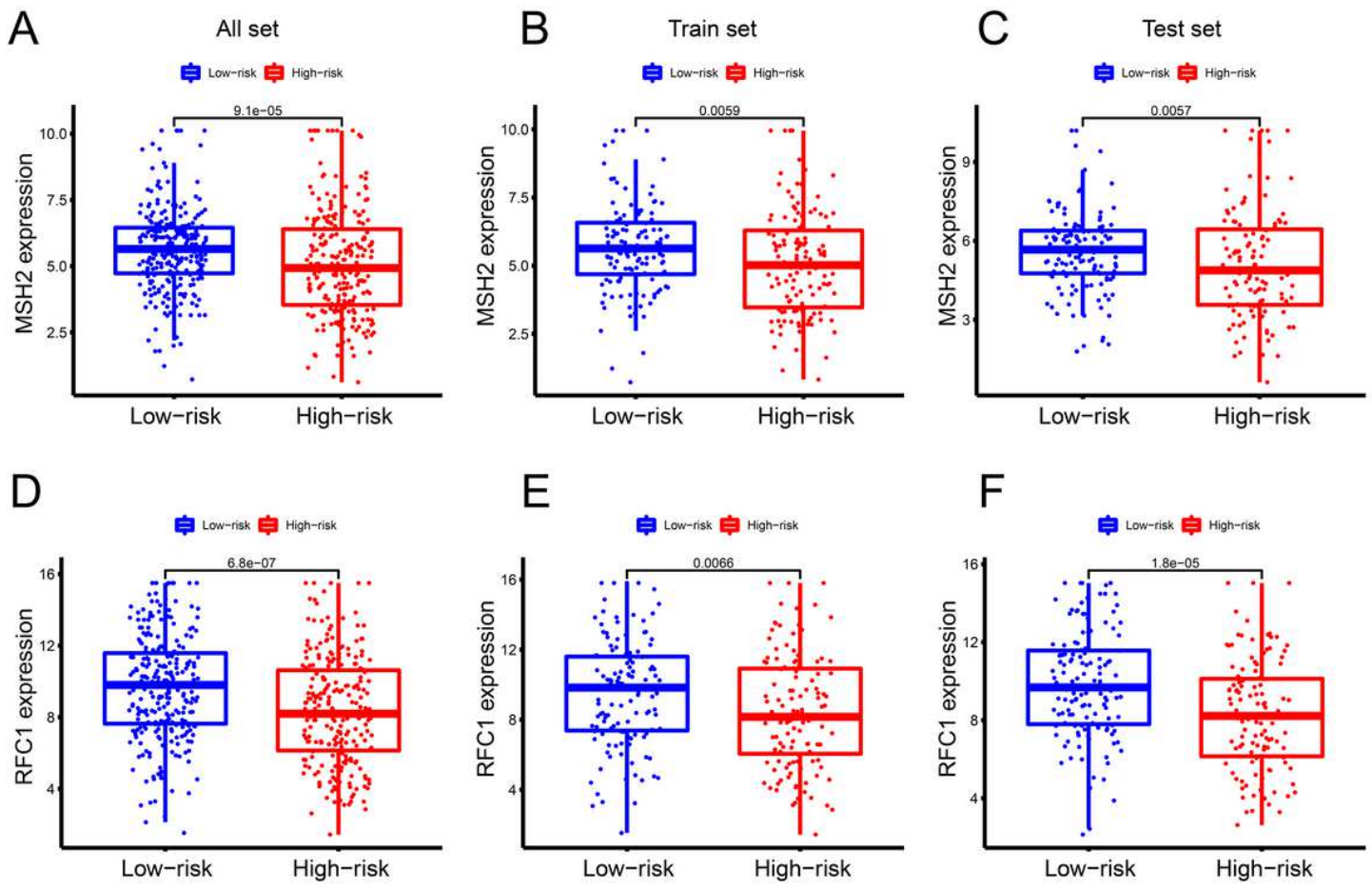

Figure 5

(A-C) The previously reported genetic instability related factor MSH2 showed significant differences in expression patterns between high-risk group and low-risk group in the all set (P = 9.1e-05), train set (P = 0.0059) and test set (P = 0.0057). (D-F) The previously reported genetic instability related factor RFC1 showed significant differences in expression patterns between high-risk group and low-risk group in the all set (P= 6.8e-07), train set (P =0.0066) and test set (P= 1.8E-05).

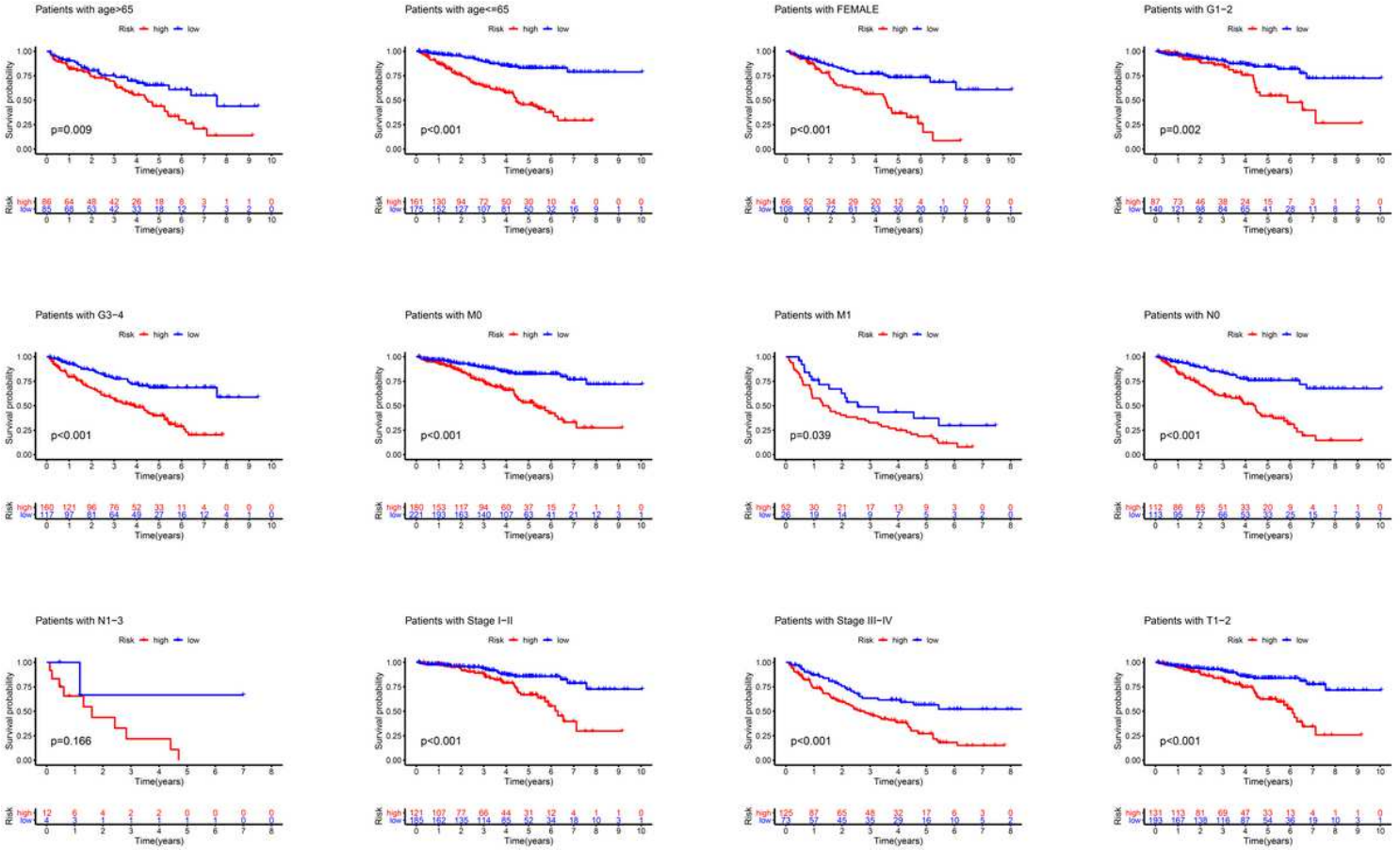

Figure 6

Subgroup analysis.The samples were divided into multiple clinical subgroups according to age, sex, stage, metastasis, and infiltration of lymph nodes. The results showed that in all clinical subgroups, the low-risk group had a better prognosis.

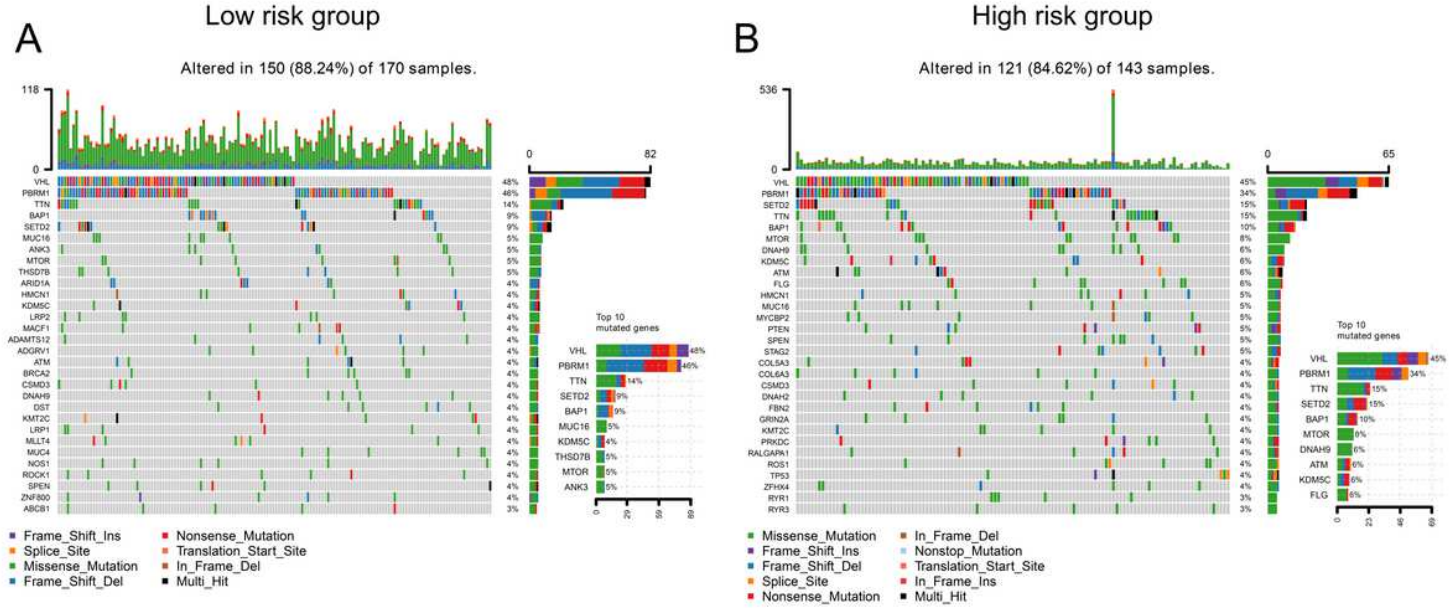

Figure 7

Waterfall map of gene mutation burden. (A) In the low-risk group, the mutationrate was88.24%.Thetopthree mutatedgenes were VHL, PBRM1 andTTN. (B) In the high-risk group, the mutation rate was 84.62%. The top five mutated genes were VHL, PBRM1, SETD2,TTN andBAP1.

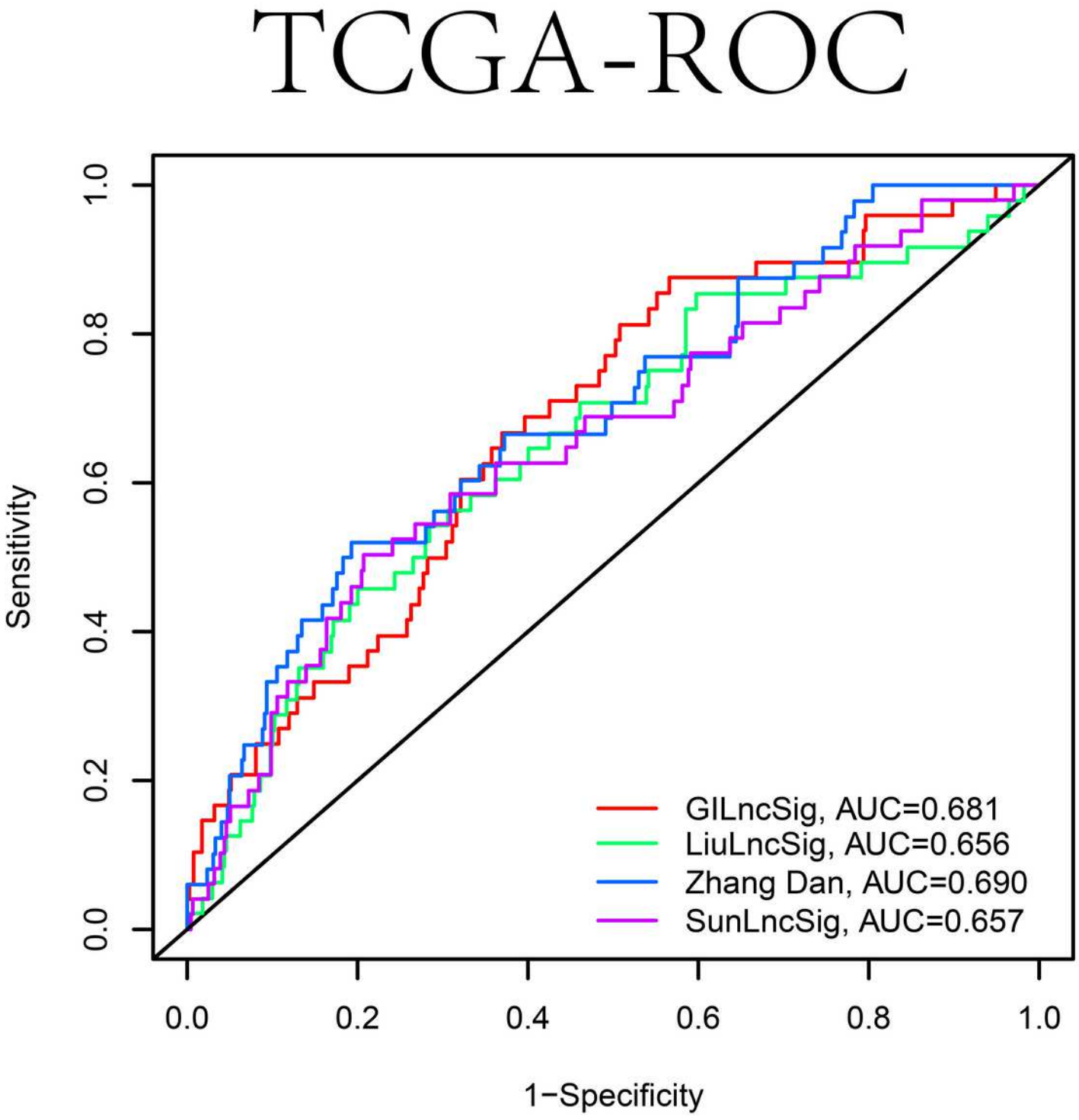

Figure 8

Model comparison. The model proposed in this paper is compared with the modelof Liu etal., Sunet al., andZhang etal., and the modelpresented in thispaper hasthehighest ROC value, indicating thebest

evaluation ability.

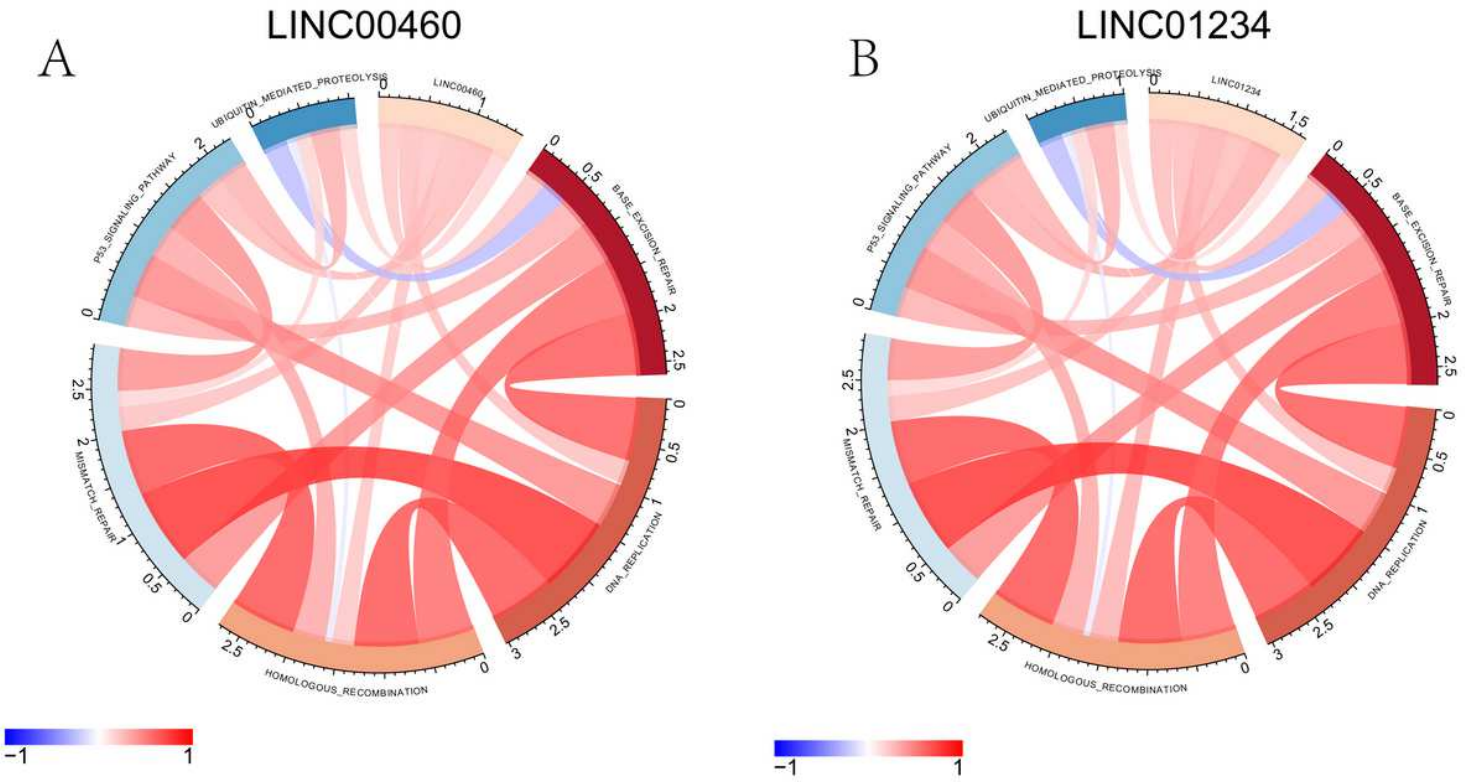

Figure 9

Correlation analysis of lncRNA and genomic instability related pathways. Red representspositive correlation andblue represents negative correlation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigure1.tif
- Supplementarytable1.xlsx