




OPEN

PBX1, EMCN and ERG are associated with the sub-clusters and the prognosis of VHL mutant clear cell renal cell carcinoma

Haiwei Wang^{1,3}, Xinrui Wang^{1,3}, Liangpu Xu¹ & Ji Zhang²

The molecular heterogeneity of primary clear cell renal cell carcinoma (ccRCC) has been reported. However, the classifications of Von Hippel–Lindau (VHL) mutant ccRCC are unclear. Here, VHL mutant ccRCC from The Cancer Genome Atlas and E-MTAB-1980 datasets were divided into two sub-clusters through non-negative matrix factorization algorithm. Most VHL mutant ccRCC patients in sub-cluster2 were with pathological T1 stage and VHL mutant ccRCC patients in sub-cluster1 were with decreased overall survival. DNA replication and homologous recombination scores were higher, while, WNT signaling pathway and regulation of autophagy scores were lower in sub-cluster1 VHL mutant ccRCC. Moreover, PBX1 transcriptional scores and mRNA expressions were lower in sub-cluster1 VHL mutant ccRCC patients and were associated with the overall survival of VHL mutant ccRCC. Furthermore, PBX1 associated genes EMCN and ERG were down-regulated in sub-cluster1 VHL mutant ccRCC and overall survival was decreased in EMCN or ERG lowly expressed VHL mutant ccRCC patients. Also, PBX1 and EMCN were down-regulated in ccRCC tissues, compared with normal kidney tissues. At last, we constructed risk models based on PBX1, EMCN and EGR expression features. With the increase of the risk score, the number of death of VHL mutant ccRCC patients was increased.

Renal cell carcinoma (RCC) originating from the renal nephron is a heterogeneous disease. RCC includes clear cell, papillary and chromophobe subtypes characterized by distinct genetic alterations, different histological features and varied clinical response to therapies^{1,2}. Clear cell renal cell carcinoma (ccRCC) represents the most common type of RCC^{3,4}. Despite the improvements of clinical management, the 5 years overall survival of ccRCC is still unsatisfied⁵. Genetics studies show that nearly 80% sporadic ccRCC tumors contain genetic mutations of Von Hippel–Lindau (VHL)^{6,7}. Individuals with inherit VHL mutations are with the increased risks for the development of ccRCC^{8,9}. Renal epithelial cells with combined deletion of VHL, TP53 and Rb1 in mouse model share similar molecular profiles and therapeutic responses with ccRCC¹⁰. Moreover, ccRCC patients with abnormal VHL are correlated with high metastatic risk¹¹ and poor prognosis¹².

VHL is a part of E3 ubiquitin ligase complex, mediating the degradation of hypoxia-inducible transcription factors (HIFs)¹³. Inactivation of VHL induces the stabilization and accumulation of HIFs. The constitutive signaling of HIFs up-regulates its target genes, like vascular endothelial growth factor (VEGF), epidermal growth factor (EGF) and platelet-derived growth factor (PDGF) to promote angiogenesis, proliferation and migration¹⁴. Over the past decade, multiple HIF2a antagonists are developed and achieved effective targeted therapies in VHL mutant ccRCC patients^{15,16}. Also, tyrosine kinase inhibitors (TKIs), such as sunitinib, pazopanib and axitinib achieve successful targeted therapies in ccRCC by targeting on the VEGF^{17,18}. All those results highlight the importance of VHL-HIFs-VEGF axis in the development and therapy of ccRCC.

ccRCC is also a heterogeneous disease¹⁹. Consensus clustering algorithm is extensively used to reveal the subtypes of ccRCC. Using consensus clustering, ccRCC is divided into ccA and ccB subtypes with different clinical outcomes^{20–22}. ccRCC patients in The Cancer Genome Atlas (TCGA) dataset are classified into four subsets based on the mRNA and microRNA expression using unsupervised consensus clustering method⁷. Transcriptional analysis suggests that metastatic ccRCC also contains four subtypes associated with different responses to sunitinib treatment²³. However, another work using genome, transcriptome and methylation data

¹Fujian Maternity and Child Health Hospital, Fujian Medical University, Fuzhou, Fujian, China. ²Shanghai Institute of Hematology, Rui-Jin Hospital Affiliated to School of Medicine, Shanghai Jiao Tong University, Shanghai, China. ³These authors contributed equally: Haiwei Wang and Xinrui Wang. ✉email: hwwang@sibs.ac.cn; Zj11222@rjh.com.cn

in TCGA shows that ccRCC only includes three subtypes²⁴. An unbiased proteomic analysis also suggests three major proteomic ccRCC subgroups discriminated by seven major protein sub-clusters²⁵. Moreover, radiomic profiling of ccRCC reveals three ccRCC subtypes with distinct genetic alterations, pathological characteristics and prognoses²⁶. Those results provide deep understanding of the genetic heterogeneity of ccRCC. However, those analyses are focused on the whole ccRCC patients, the further classifications of VHL mutant ccRCC is not clear.

Our and previous results suggest that non-negative matrix factorization (NMF) is a robust clustering method in colon cancer^{27,28}, liver cancer²⁹, lung cancer³⁰ and lower grade glioma³¹. Here, using NMF algorithm, VHL mutant ccRCC from TCGA⁷ and E-MTAB-1980³² datasets was divided into two sub-clusters. We further analyzed the clinical outcomes, signaling pathways and transcription factors associated with the different sub-clusters of VHL mutant ccRCC. Our results provided insights of the molecular heterogeneity of VHL mutant ccRCC and suggested that PBX1, EMCN and ERG were prognostic makers associated with the overall survival of VHL mutant ccRCC.

Results

Two molecular sub-clusters of VHL mutant ccRCC with different clinical outcomes. Somatic alterations analysis of 354 ccRCC patients in TCGA Kidney Clear Cell Carcinoma (KIRC) dataset showed that 170 patients were with VHL mutations. Based on the RNA-seq data, those 170 VHL mutant ccRCC patients were classified into two distinctive sub-clusters using “NMF” algorithm, as demonstrated in the consensus heatmaps (Fig. 1a). 62 VHL mutant ccRCC patients were in sub-cluster1 and 108 VHL mutant ccRCC patients were divided into sub-cluster2, respectively (Fig. 1b). The clinical profiling was significantly different between sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients. Most of VHL mutant ccRCC patients in sub-cluster2 were with pathological T1 stage (Fig. 1b). Also, more than 70% VHL mutant ccRCC patients in sub-cluster1 were male, while, only 51% VHL mutant ccRCC patients in sub-cluster2 were male (Fig. 1b). Moreover, VHL mutant ccRCC patients in sub-cluster1 had decreased overall survival than VHL mutant ccRCC patients in sub-cluster2 in TCGA dataset (Fig. 1c).

The classifications of VHL mutant ccRCC were validated using independent E-MTAB-1980 cohort. 101 ccRCC patients were collected in E-MTAB-1980 dataset and 67 ccRCC patients were with VHL mutations. Similarly, using “NMF” algorithm, the 67 VHL mutant ccRCC patients were classified into two sub-clusters (Fig. 1d). 24 VHL mutant ccRCC patients were in sub-cluster1 and 43 VHL mutant ccRCC patients were in sub-cluster2, respectively. Most VHL mutant ccRCC patients in sub-cluster2 were with pathological T1 stage in E-MTAB-1980 dataset (Fig. 1e). However, age or gender difference was not significant between the two sub-clusters of VHL mutant ccRCC patients in E-MTAB-1980 dataset (Fig. 1e). Moreover, VHL mutant ccRCC patients in sub-cluster1 had decreased overall survival than VHL mutant ccRCC patients in sub-cluster2 in E-MTAB-1980 dataset (Fig. 1f).

Except VHL mutations, alterations of chromatin remodeling protein PBRM1³³ and histone methyltransferase SETD2³⁴ were also detected in ccRCC patients. However, the PBRM1 or SETD2 alterations in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA dataset were not significantly different (Fig. 1g). Those results suggested that VHL mutant ccRCC was a heterogeneous disease and could be further classified into two sub-clusters with different prognosis.

Two molecular sub-clusters of VHL mutant ccRCC with different immune infiltrations. RCC represents one of the most immune infiltrated tumor types in a pan-cancer analysis^{35,36}. Immune related genes were associated with the clinical overall survival and subtypes of ccRCC³⁷. However, the immune infiltrations in the tumor microenvironment of VHL mutant ccRCC were not clear. The stromal scores and immune scores of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets were calculated using “ESTIMATE” algorithm. Although, ccRCC patients with lower immune scores had prolonged overall survival than ccRCC patients with higher immune scores in TCGA dataset³⁸, the immune scores was not associated with the clinical overall survival of VHL mutant ccRCC in TCGA (Supplementary Fig. 1a) and E-MTAB-1980 (Supplementary Fig. 1b) datasets. Moreover, the stromal scores were not correlated with the prognosis of VHL mutant ccRCC in TCGA (Supplementary Fig. 1a) and E-MTAB-1980 (Supplementary Fig. 1b) datasets.

However, compared with VHL mutant ccRCC patients in sub-cluster1, VHL mutant ccRCC patients in sub-cluster2 were with higher stromal scores in TCGA and E-MTAB-1980 datasets (Fig. 1h). On the contrary, VHL mutant ccRCC patients in sub-cluster2 were with lower immune scores in TCGA dataset (Fig. 1h). The immune scores in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in E-MTAB-1980 dataset was not significantly different (Fig. 1h).

Transcriptional characteristics of the sub-clusters of VHL mutant ccRCC. Next, we determined the differentially expressed genes between sub-cluster1 and sub-cluster2 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. Based on the threshold of fold changes > 1.5 and *P* values < 0.001, 408 genes were commonly up-regulated in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 2a). 289 genes were commonly down-regulated in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 2a). The differentially expressed 697 genes were further shown in the heatmaps and those genes distinguished the sub-cluster1 from the sub-cluster2 VHL mutant ccRCC (Fig. 2b).

Using ssGSEA, we determined the scores of 186 KEGG signaling pathways in VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. The relative scores of DNA replication and homologous recombination were higher in sub-cluster1 VHL mutant ccRCC, compared with sub-cluster2 VHL mutant ccRCC in TCGA (Fig. 2c) and E-MTAB-1980 (Fig. 2d) datasets. On the contrary, the relative scores of WNT signaling pathway and regulation of autophagy were lower in sub-cluster1 VHL mutant ccRCC in TCGA (Fig. 2c) and E-MTAB-1980

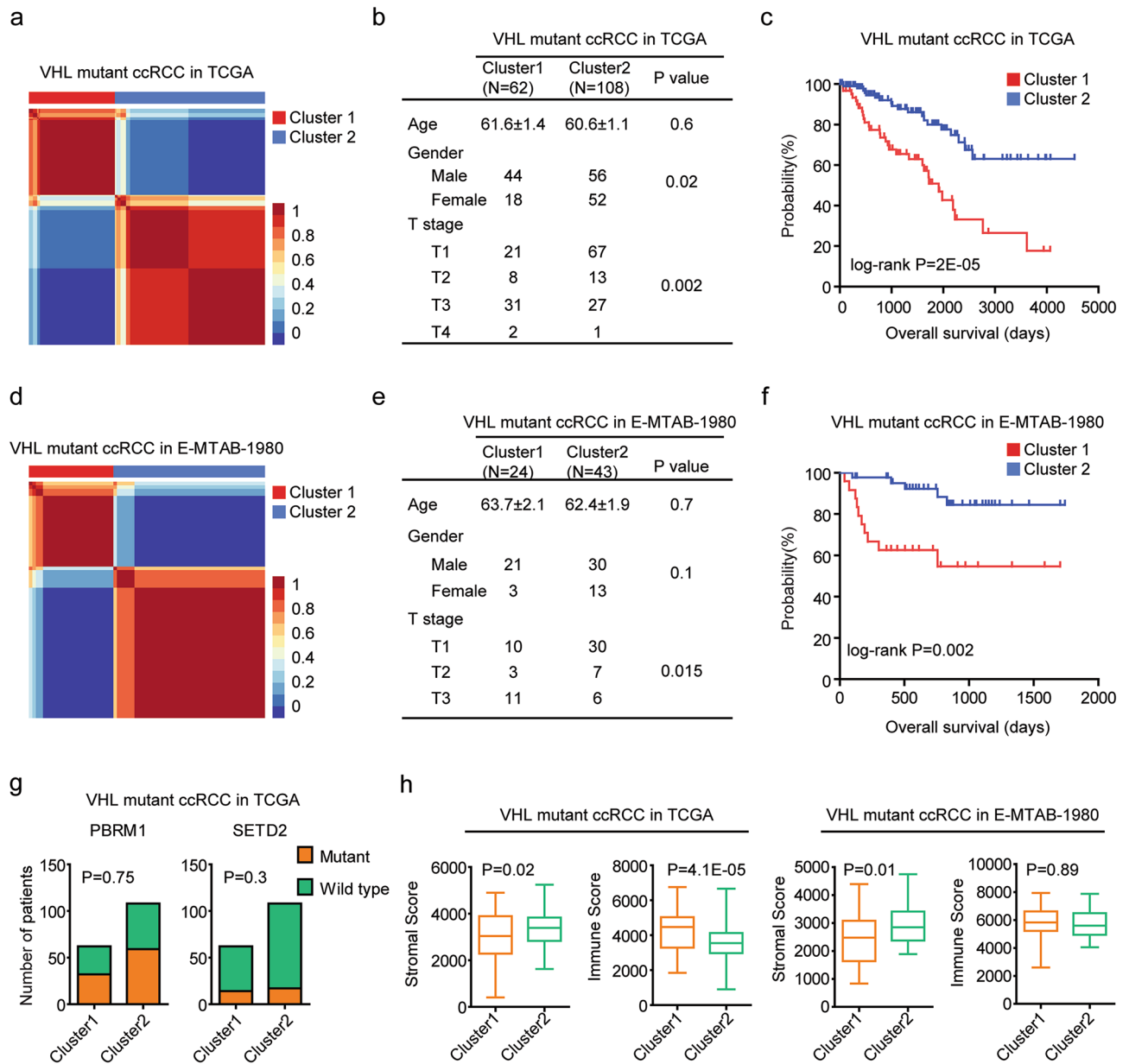


Figure 1. Two molecular sub-clusters of VHL mutant ccRCC patients with different clinical outcomes. **(a)** Consensus-map showed that primary ccRCC patients with VHL mutations from TCGA dataset were divided into two sub-clusters using NMF method. **(b)** Table showed the different clinical characteristics of VHL mutant ccRCC patients in sub-cluster1 and sub-cluster2 in TCGA dataset. **(c)** Kaplan–Meier survival plot showed the different overall survival of VHL mutant ccRCC patients in sub-cluster1 and sub-cluster2. *P* value was determined by log-rank test. **(d)** Primary ccRCC patients with VHL mutations from E-MTAB-1980 dataset were divided into two sub-clusters using NMF method. **(e)** Table showed the different clinical characteristics of VHL mutant ccRCC patients in sub-cluster1 and sub-cluster2 in E-MTAB-1980 dataset. **(f)** Kaplan–Meier survival plot showed the clinical outcomes of sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in E-MTAB-1980 dataset. **(g)** The number of VHL mutant ccRCC patients with or without PBRM1 or SETD2 alterations in each sub-cluster of TCGA dataset. *P* values were calculated by Chi-square test. **(h)** Box plots showed the stromal scores and immune scores in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets.

(Fig. 2d) datasets. Moreover, the higher scores of DNA replication and homologous recombination were associated with the worse prognosis of VHL mutant ccRCC in TCGA (Fig. 2e) and E-MTAB-1980 (Supplementary Fig. 2a) datasets, while, the higher scores of WNT signaling pathway was associated with the better prognosis of VHL mutant ccRCC patients in TCGA (Fig. 2e) and E-MTAB-1980 (Supplementary Fig. 2a) datasets. Furthermore, regulation of autophagy was associated with the prognosis of VHL mutant ccRCC in E-MTAB-1980 (Supplementary Fig. 2a), but not in TCGA dataset (Fig. 2e).

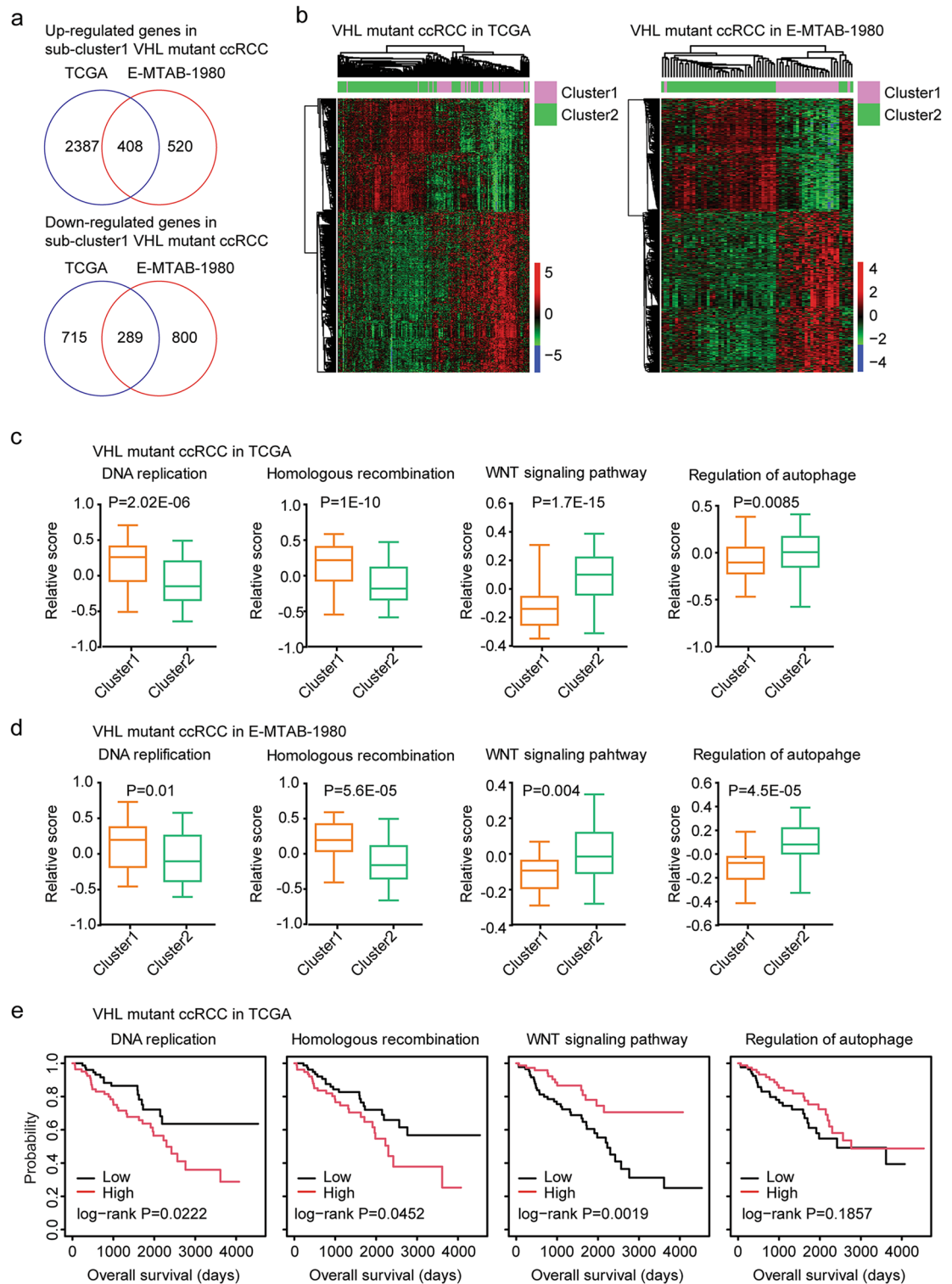


Figure 2. Transcriptional characteristics of the different sub-clusters of VHL mutant ccRCC. (a) The overlapped up-regulated or down-regulated genes in sub-cluster1 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (b) Heatmaps demonstrated the overlapped differentially expressed genes in sub-cluster1 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (c) The scores of signaling pathways in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA dataset. (d) The scores of signaling pathways in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in E-MTAB-1980 dataset. (e) Kaplan–Meier Plotters showed the correlations of the signaling pathways and the overall survival of VHL mutant ccRCC patients in TCGA dataset.

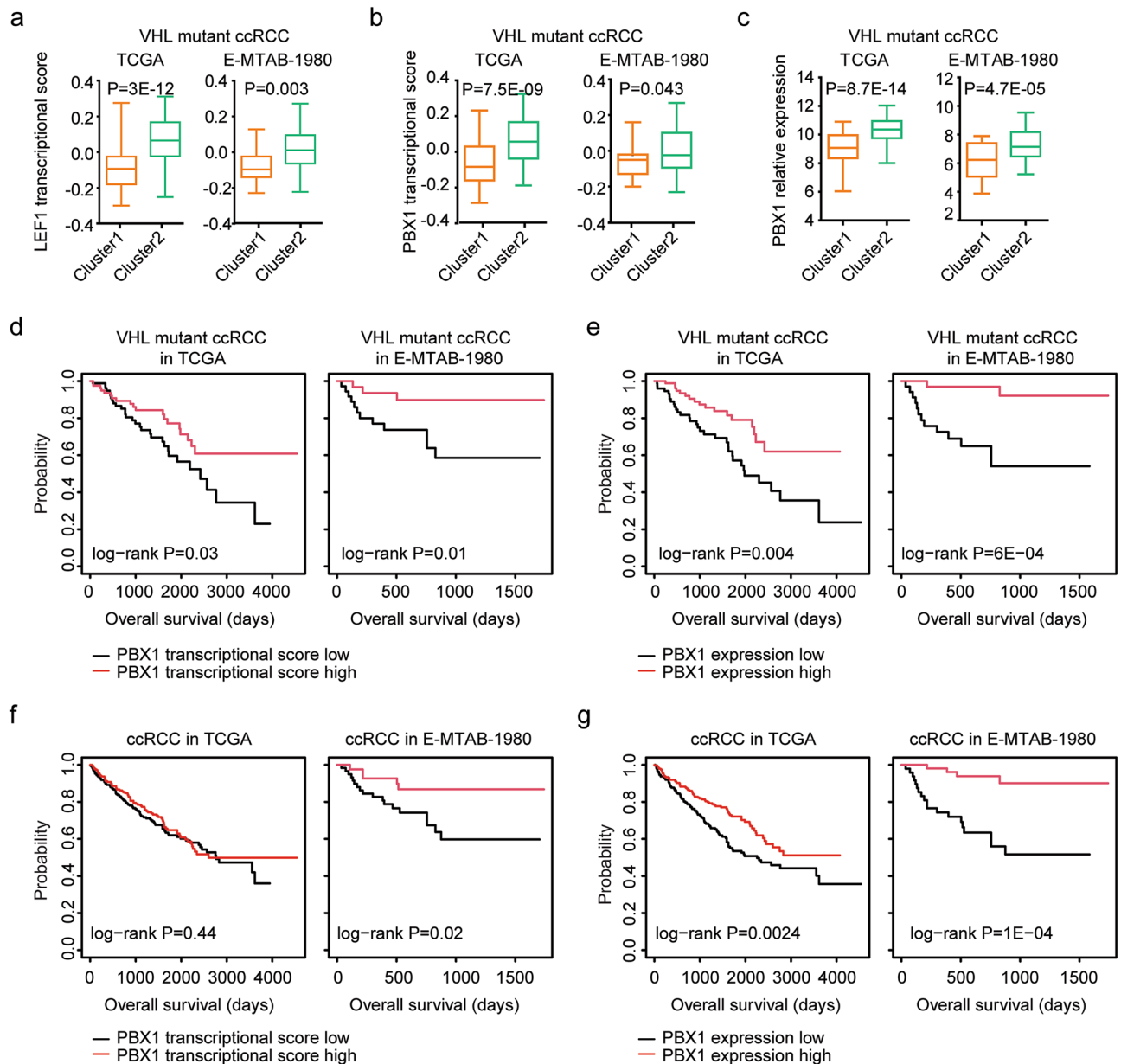


Figure 3. Transcription factor PBX1 is associated with the classification and prognosis of VHL mutant ccRCC. **(a)** The transcriptional scores of LEF1 in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. **(b)** The transcriptional scores of PBX1 in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. **(c)** The expression levels of PBX1 in sub-cluster1 and sub-cluster2 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. **(d)** Kaplan–Meier Plotters showed the associations of the transcriptional scores of PBX1 and the overall survival of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. **(e)** The associations of the expression levels of PBX1 and the overall survival of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. **(f)** The associations of the transcriptional scores of PBX1 and the overall survival of ccRCC in TCGA and E-MTAB-1980 datasets. **(g)** The associations of the expression levels of PBX1 and the overall survival of ccRCC in TCGA and E-MTAB-1980 datasets.

Transcription factor PBX1 is associated with the classification and prognosis of VHL mutant ccRCC. Also, scores of 958 transcriptional gene sets in VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets were identified using ssGSEA. LEF1 is a downstream transcription factor of WNT signaling pathway. Consistent with the lower scores of WNT signaling pathway, LEF1 transcriptional scores were lower in sub-cluster1 VHL mutant ccRCC (Fig. 3a). Moreover, the higher transcriptional scores of LEF1 were associated with the better prognosis of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Supplementary Fig. 2b). However, the mRNA expression levels of LEF1 were not associated with the prognosis of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Supplementary Fig. 2c).

PBX1 is a homeobox transcription factor. Insufficiency PBX1 activity is associated with congenital anomalies of kidney^{39,40}. However, the prognosis of PBX1 in ccRCC, particularly in VHL mutant ccRCC patients is

unknown. Lower PBX1 transcriptional scores were observed in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 3b). PBX1 mRNA expression levels were also lower in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 3c). Moreover, PBX1 transcriptional scores and PBX1 mRNA expression levels were both associated with the overall survival of VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. VHL mutant ccRCC patients with higher PBX1 regulatory scores were with prolonged overall survival in TCGA and E-MTAB-1980 datasets (Fig. 3d). And compared with PBX1 highly expressed VHL mutant ccRCC, PBX1 lowly expressed VHL mutant ccRCC were with worse prognosis in TCGA and E-MTAB-1980 datasets (Fig. 3e).

Interestingly, PBX1 transcriptional scores and PBX1 mRNA expression levels were also correlated with the clinical outcomes of ccRCC. Higher PBX1 transcriptional scores were correlated with the better overall survival of ccRCC in E-MTAB-1980 dataset, but not in TCGA dataset (Fig. 3f). Moreover, overall survival was increased in PBX1 highly expressed ccRCC patients, compared with PBX1 lowly expressed ccRCC patients in TCGA and E-MTAB-1980 datasets (Fig. 3g).

EMCN and ERG are associated with the classification and prognosis of VHL mutant ccRCC. In GSEA database, 37 genes were associated with PBX1. Among those genes, ANGPT1, EMCN, EMX2, ERG, HOXB9, ONECUT2 and SSTR1 were differentially expressed in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4a). Similar to PBX1, EMX2, SSTR1, ANGPT1, ERG and EMCN were all down-regulated in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4b). On the contrary, ONECUT2 and HOXB9 were up-regulated in sub-cluster1 VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4b).

The prognostic effects of those genes were determined in VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. Univariate cox regression analysis suggested that all those genes were associated with the overall survival of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4c). However, in Kaplan–Meier survival analysis, only two genes EMCN and ERG had significant prognostic effects in VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. Overall survival was increased in EMCN or ERG highly expressed VHL mutant ccRCC, compared with EMCN or ERG lowly expressed VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4d).

Moreover, ANGPT1, EMCN, EMX2, ERG, HOXB9, ONECUT2 and SSTR1 were not only correlated with the clinical outcomes of VHL mutant ccRCC, but also were correlated with the clinical outcomes of all ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4e). Furthermore, higher EMCN or ERG mRNA expression levels were correlated with the better overall survival of ccRCC in TCGA and E-MTAB-1980 datasets (Fig. 4f).

PBX1, EMCN and ERG are associated with the stromal score of VHL mutant ccRCC. We had shown the different immune infiltrations between sub-cluster1 and sub-cluster2 VHL mutant ccRCC (Fig. 1h). Next, we determined the correlations of PBX1, EMCN and ERG with the immune infiltrations in VHL mutant ccRCC. PBX1, EMCN and ERG were up-regulated in VHL mutant ccRCC with lower immune scores, compared with VHL mutant ccRCC with higher immune scores in TCGA dataset (Supplementary Fig. 3a), but not in E-MTAB-1980 dataset (Supplementary Fig. 3b). Moreover, PBX1 and EMCN were negatively correlated with the immune scores of VHL mutant ccRCC in TCGA dataset (Supplementary Fig. 3c).

Also, compared with VHL mutant ccRCC with lower stromal scores, the expression levels of PBX1 and ERG were higher in VHL mutant ccRCC with higher stromal scores in TCGA (Fig. 5a) and E-MTAB-1980 (Fig. 5b) datasets. Moreover, PBX1, EMCN and ERG were positively correlated with the stromal scores of VHL mutant ccRCC patients in TCGA (Fig. 5c) and E-MTAB-1980 (Fig. 5d) datasets.

PBX1 and EMCN are down-regulated in ccRCC tissues. Next, we analyzed the expression levels of PBX1, EMCN and EGR in normal kidney and ccRCC tissues. First, 64 normal kidney samples and matched ccRCC samples in TCGA dataset were studied (Supplementary Fig. 4a). The expression levels of PBX1 were lower in ccRCC samples, compared with the normal kidney samples in TCGA dataset (Fig. 6a). Also, EMCN was down-regulated in ccRCC tissues in TCGA dataset (Fig. 6b). On the contrary, ERG was up-regulated in ccRCC tissues in TCGA dataset (Supplementary Fig. 4b).

The expressions of PBX1, EMCN and EGR in normal kidney and ccRCC tissues were further analyzed using published GEO datasets. Totally, 313 normal kidney samples and 354 ccRCC samples were collected from eight independent datasets based on different microarray platforms (Supplementary Fig. 4a). Consistent with the down-regulation of PBX1 in ccRCC tissues in TCGA dataset, in GSE36895, GSE40435, GSE46699, GSE47032, GSE53757, GSE66270, GSE68417 and GSE71963 datasets, PBX1 was down-regulated in ccRCC tissues, compared with the normal kidney tissues (Fig. 6a).

Similarly, compared with the normal kidney tissues, EMCN was lowly expressed in ccRCC tissues in GSE36895, GSE40435, GSE46699, GSE53757, GSE68417 and GSE71963 datasets (Fig. 6b). However, in GSE47032 and GSE66270 datasets, the EMCN expression levels were not significantly different in normal kidney and ccRCC tissues (Fig. 6b). Moreover, in GSE36895, GSE40435, GSE46699, GSE53757 and GSE66270 datasets, ERG was up-regulated in ccRCC tissues (Supplementary Fig. 4b).

Construction of risk models of VHL mutant ccRCC based on PBX1, EMCN and EGR expressions. Our results suggested that in TCGA and E-MTAB-1980 datasets, PBX1, EMCN and EGR were all associated with the overall survival of VHL mutant ccRCC. We then assessed the associations of PBX1, EMCN and EGR using multivariate cox regression analysis in VHL mutant ccRCC. Age was an independent prognostic

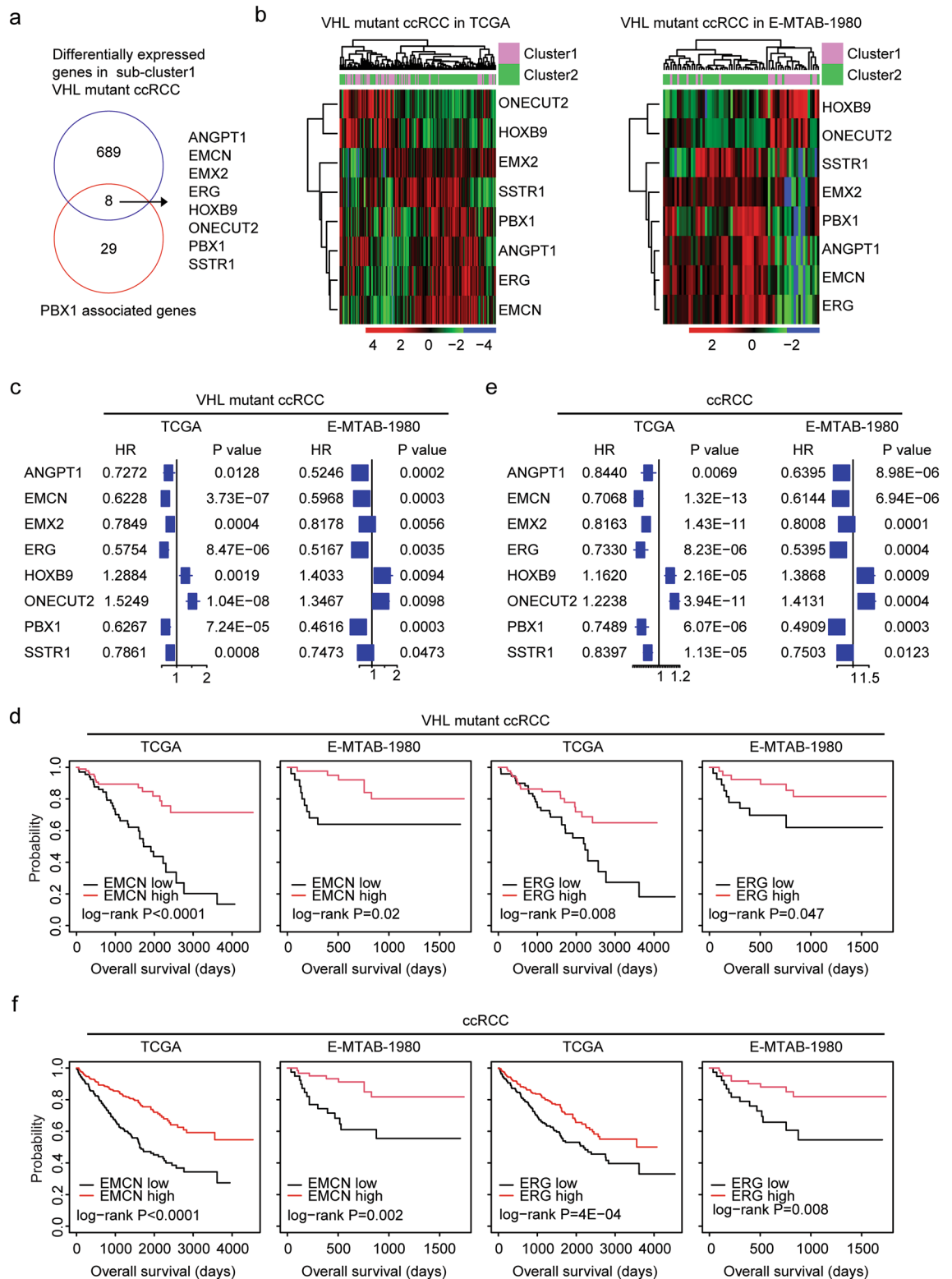


Figure 4. EMCN and ERG are associated with the classification and prognosis of VHL mutant ccRCC. (a) Eight PBX1 associated genes were differentially expressed in sub-cluster1 VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (b) Un-supervised clustering heatmaps demonstrated the expression levels of PBX1 associated genes in VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (c) Forest plots showed the prognosis of PBX1 associated genes in the predication of the clinical overall survival of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets. Hazard ratio (HR) and P values were determined by univariate cox regression assay. (d) Kaplan–Meier Plotters showed the prognostic effects of EMCN and ERG in VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (e) The prognosis of PBX1 associated genes in the predication of the clinical overall survival of ccRCC in TCGA and E-MTAB-1980 datasets. (f) The prognostic effects of EMCN and ERG in ccRCC patients in TCGA and E-MTAB-1980 datasets.

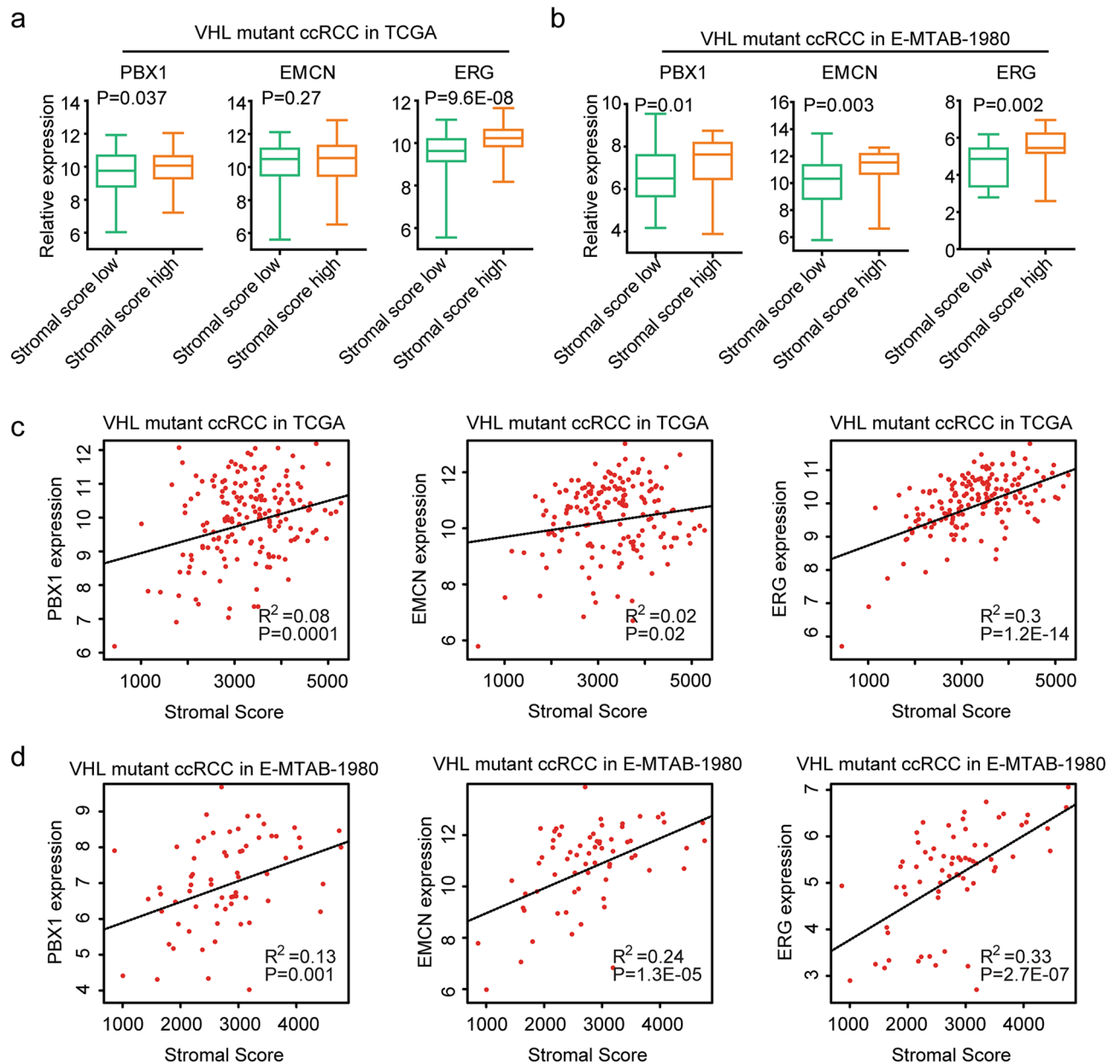


Figure 5. PBX1, EMCN and ERG are associated with the stromal score of VHL mutant ccRCC. **(a)** Box plots showed the PBX1, EMCN and ERG expression levels in VHL mutant ccRCC patients with higher stromal scores or with lower stromal scores in TCGA dataset. **(b)** PBX1, EMCN and ERG expression levels in VHL mutant ccRCC patients with higher stromal scores or with lower stromal scores in E-MTAB-1980 dataset. **(c)** Spearman correlations between PBX1, EMCN and ERG expression levels and stromal scores in VHL mutant ccRCC patients in TCGA dataset. **(d)** Spearman correlations between PBX1, EMCN and ERG expression levels and the stromal scores in VHL mutant ccRCC patients in E-MTAB-1980 dataset.

factor in VHL mutant ccRCC in TCGA dataset (Fig. 7a). However, in both TCGA and E-MTAB-1980 datasets, PBX1, EMCN and EGR were not independent prognostic factors (Fig. 7a).

We constructed risk models based on PBX1, EMCN and EGR expression features to predict the prognosis of VHL mutant ccRCC. The risk score of each patient in TCGA and E-MTAB-1980 datasets was obtained using RiskScore calculation formula. High risk or low risk subgroup was classified based on the median values of the risk score. The risk score distribution of each patient in TCGA and E-MTAB-1980 datasets was shown in Fig. 7b. With the increase of the risk score, the number of death of VHL mutant ccRCC patients was increased in TCGA and E-MTAB-1980 datasets (Fig. 7b). Moreover, lower expression levels of PBX1, EMCN and EGR were positively correlated with the risk score in VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets (Fig. 7b).

Construction of risk models of ccRCC based on PBX1, EMCN and EGR expressions. PBX1, EMCN and EGR were also associated with the overall survival of ccRCC. We determined the associations of

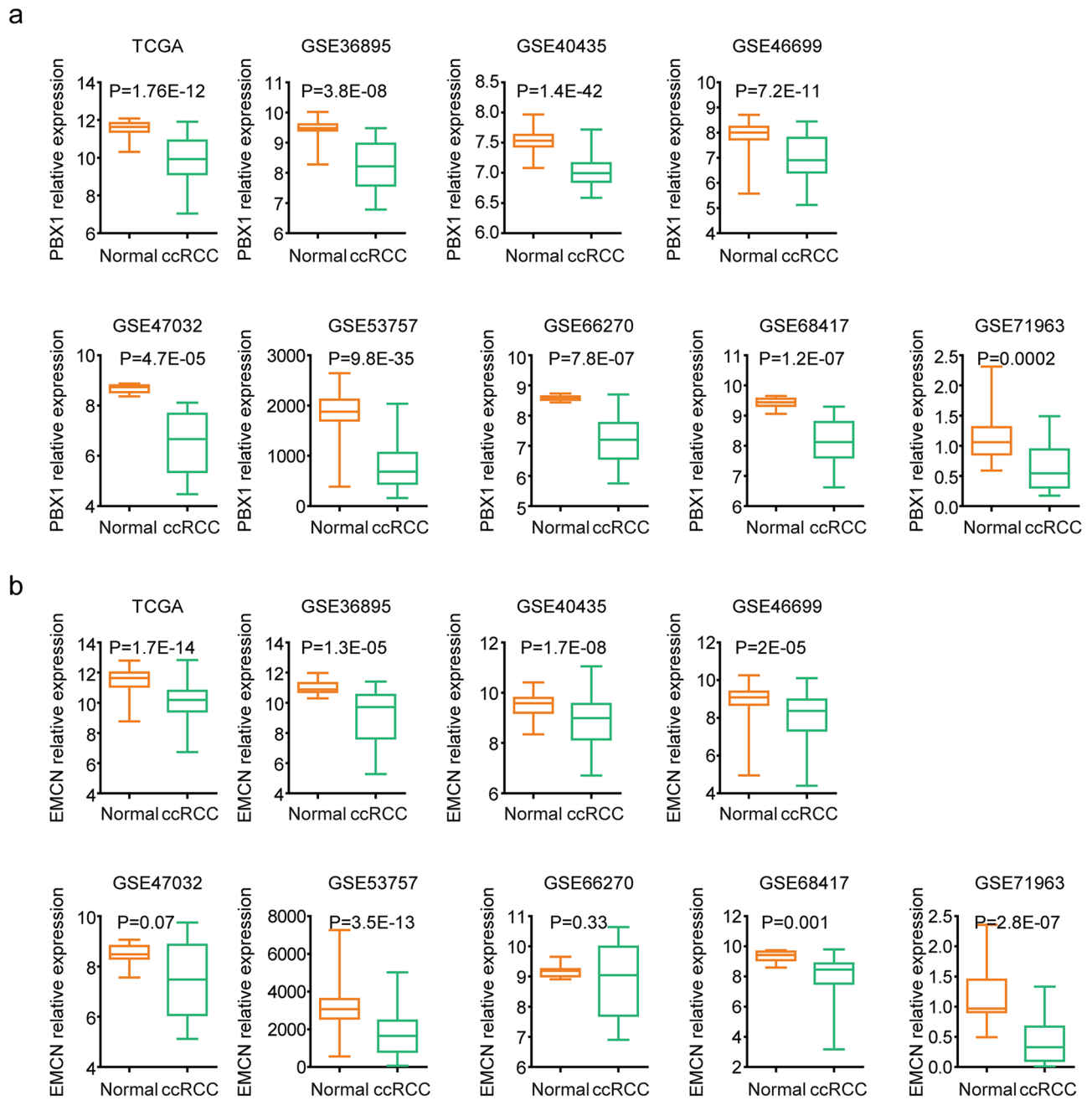


Figure 6. PBX1 and EMCN are down-regulated in ccRCC tissues. (a) Box plots showed the PBX1 expression levels in normal kidney and ccRCC tissues in TCGA and GEO datasets. (b) Expression levels of EMCN in normal kidney and ccRCC tissues in TCGA and GEO datasets.

PBX1, EMCN and EGR in ccRCC using multivariate cox regression analysis. The forest plots showed that age and EMCN expression were independent prognostic factors of ccRCC in TCGA dataset (Fig. 8a). However, in E-MTAB-1980 dataset, age, gender, PBX1, EMCN and EGR were not independent prognostic factors (Fig. 8a).

Similarly, with the increase of the risk score, the number of death of ccRCC patients was increased in TCGA and E-MTAB-1980 datasets (Fig. 8b). Moreover, lower expression levels of PBX1, EMCN and EGR were positively correlated with the risk score in ccRCC patients in TCGA and E-MTAB-1980 datasets (Fig. 8b).

Discussion

Our analysis suggested that VHL mutant ccRCC was a heterogeneous disease. Based on the mRNA expression profiling, we identified two sub-clusters of VHL mutant ccRCCs with different transcriptional characteristics, clinical outcomes and immune infiltrations. VHL mutant ccRCC patients in sub-cluster2 had prolonged overall survival and high stromal scores. DNA replication, homologous recombination, WNT signaling pathway and regulation of autophagy were associated with the classifications of VHL mutant ccRCC. Moreover, PBX1, EMCN

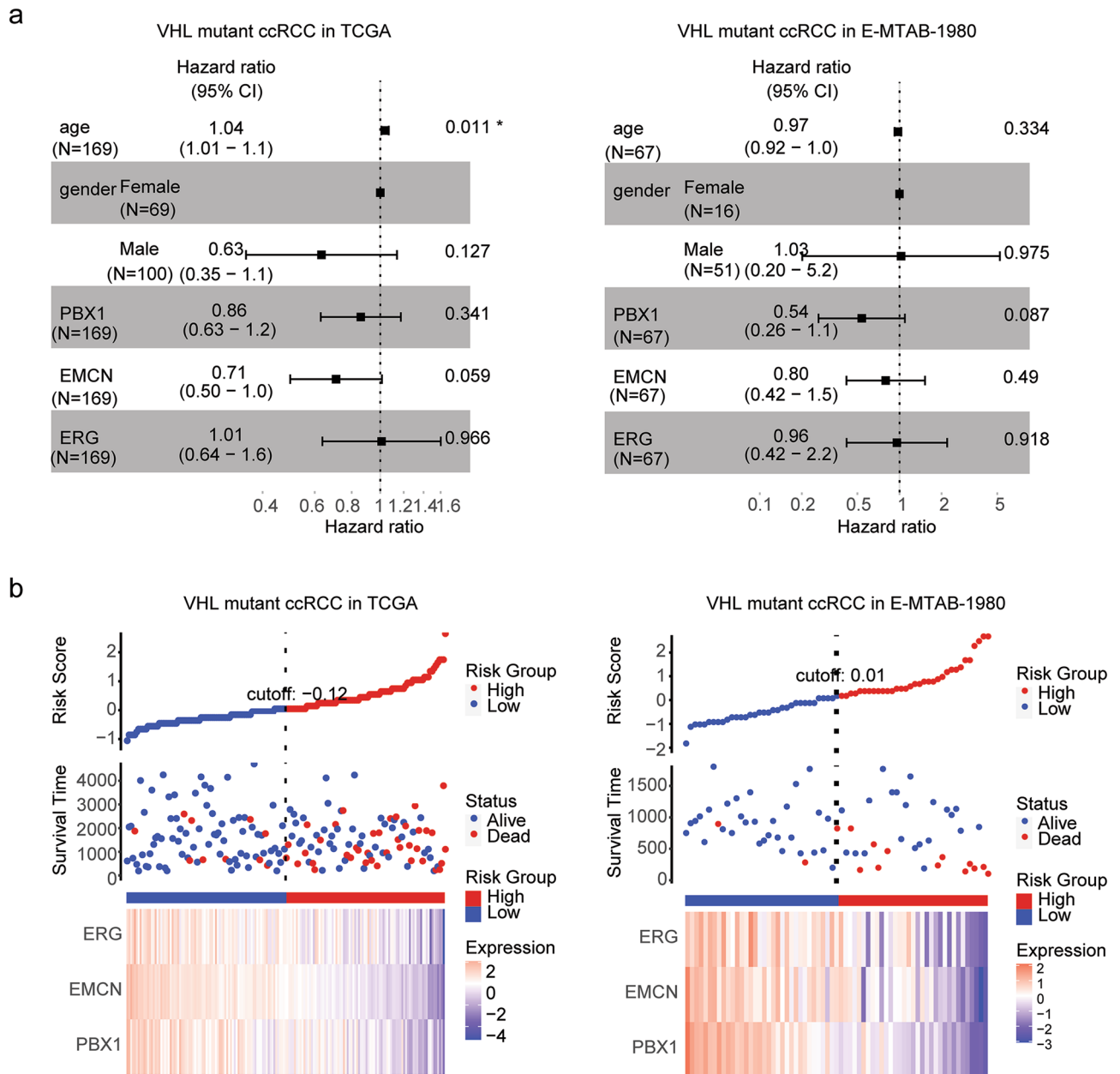


Figure 7. Construction of risk models of VHL mutant ccRCC based on PBX1, EMCN and EGR expressions. (a) Forest plots showed the associations of age, gender, PBX1, EMCN and EGR expressions with the clinical overall survival of VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets. (b) The distribution of risk score, survival status and expression levels of PBX1, EMCN and EGR between low-risk group and high-risk group of VHL mutant ccRCC in TCGA and E-MTAB-1980 datasets.

and ERG were down-regulated in sub-cluster1 VHL mutant ccRCC patients and associated with the overall survival of VHL mutant ccRCC. The risk models suggested that PBX1, EMCN and ERG were prognostic makers associated with the overall survival of VHL mutant ccRCC.

PBX1 belongs to the homeobox family of transcription factors and modulates the transcriptional levels of multiple genes involved in kidney development^{39,40}, stem cell differentiation⁴¹ and skeleton patterning⁴². In B cell lineage acute lymphoblastic leukemia (ALL), PBX1 is fused to E2A, forming an oncoprotein⁴³. PBX1 is also identified as a pioneer factor mediated the aggressiveness of estrogen receptor positive breast cancer⁴⁴. However the functions of PBX1 in ccRCC are controversial. Reports had suggested that PBX1 was up-regulated in ccRCC tissues and inhibition of PBX1 decreased the ccRCC cell proliferation through JAK2/STAT3 signaling⁴⁵. On the contrary, our results showed that PBX1 was down-regulated in ccRCC tissue and lower PBX1 expression was associated with the worse prognosis of ccRCC and VHL mutant ccRCC. So, the roles of PBX1 in ccRCC and VHL mutant ccRCC should be further studied.

EMCN is a glycoprotein, expressed in the endothelial cells⁴⁶. Previously reports showed that EMCN had predictive values in liver cancer⁴⁷ and gastric cancer⁴⁸. However, the prognostic effects of EMCN in ccRCC or

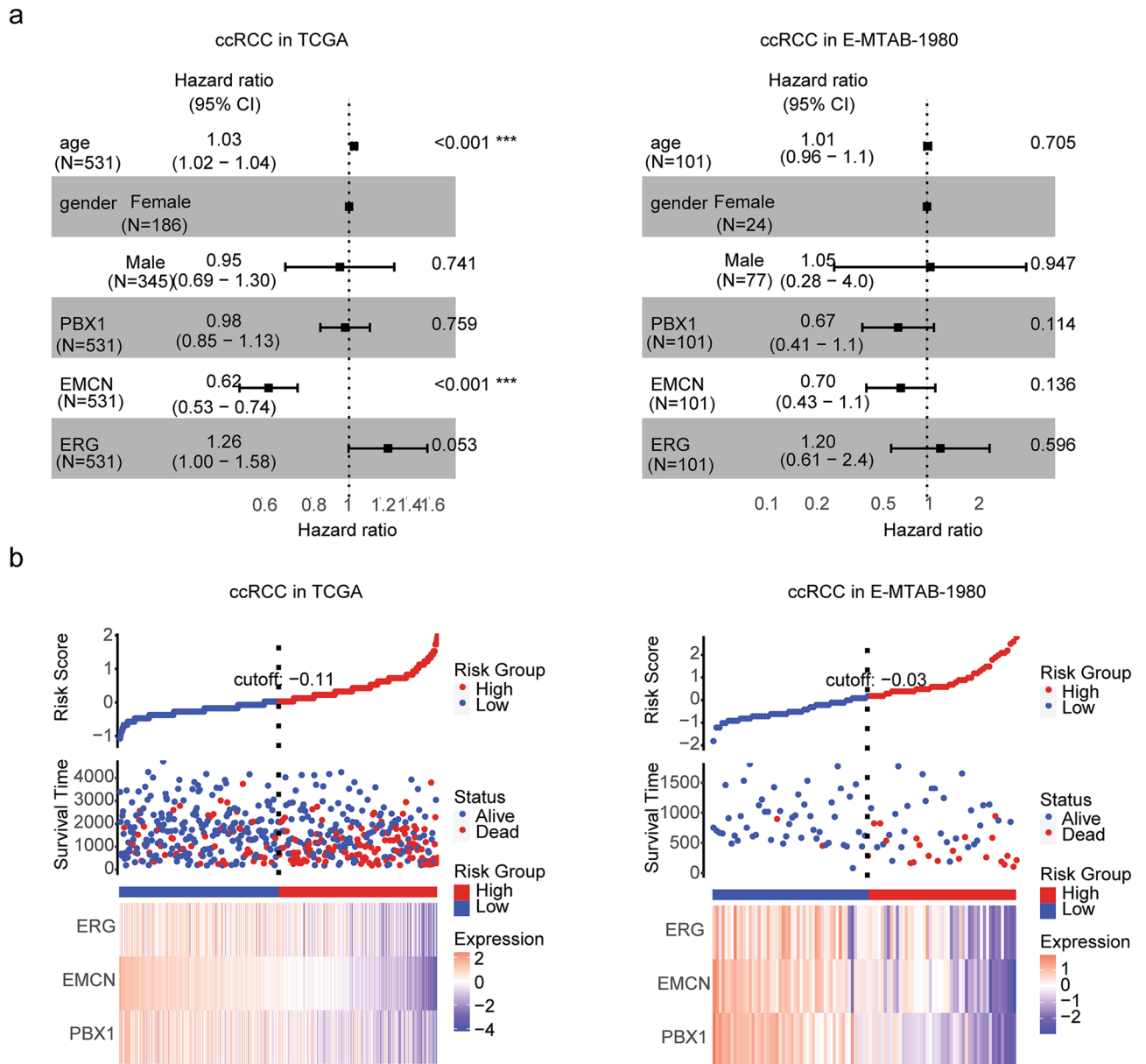


Figure 8. Construction of risk models of ccRCC based on PBX1, EMCN and EGR expressions. **(a)** Forest plots showed the associations of age, gender, PBX1, EMCN and EGR expressions with the clinical overall survival of ccRCC patients in TCGA and E-MTAB-1980 datasets. **(b)** The distribution of risk score, survival status and expression levels of PBX1, EMCN and EGR between low-risk group and high-risk group of ccRCC in TCGA and E-MTAB-1980 datasets.

VHL mutant ccRCC are unclear. Like PBX1, EMCN was also down-regulated in ccRCC tissues and lower EMCN expression was associated the worse prognosis of ccRCC and VHL mutant ccRCC. Moreover, EMCN represents a new target for angiogenesis related diseases by regulation of VEGFR2⁴⁹. And, ccRCC is particularly response to angiogenesis inhibitors, such as sunitinib^{17,18}. So, it is interesting to test whether the down-regulation of EMCN in ccRCC or VHL mutant ccRCC is conferring the drug resistance of sunitinib.

ERG belongs to the erythroblast transformation-specific (ETS) transcription factors, involved in cell proliferation, cell differentiation and angiogenesis⁵⁰. Different ERG chromosomal translocations were detected in different tumor types, such as ERG-TMPSSR2 translocations in prostate cancer⁵¹, ERG-EWS translocations in Ewing's sarcoma⁵² and ERG-FUS translocations in acute myeloid leukemia⁵³. However, ERG chromosomal translocations or mutations were barely detected in ccRCC⁷. Although, we showed that in some ccRCC cohorts, ERG was up-regulated in ccRCC tissues, the lower expression of EGR was associated with the worse prognosis of ccRCC and VHL mutant ccRCC.

Overall, our analysis provided insights of the molecular heterogeneity within VHL mutant ccRCC subgroup and suggested new prognostic factors of PBX1, EMCN and ERG in VHL mutant ccRCC. However, those conclusions were derived from published TCGA and E-MTAB-1980 datasets and the expression and prognosis of PBX1,

EMCN and ERG in VHL mutant ccRCC should be further validated using clinical data. Also the functions of PBX1, EMCN and ERG in ccRCC development and sunitinib drug resistance should be studied.

Materials and methods

Data collection. The RNA-seq data along with the clinical characteristics of 354 ccRCC patients in TCGA KIRC dataset were downloaded from TCGA hub (<https://tcga.xenahubs.net>)⁷. The gene expressions along with the clinical characteristics of 101 ccRCC samples in E-MTAB-1980 dataset were downloaded from <https://www.ebi.ac.uk/arrayexpress/> website³². GSE36895⁵⁴, GSE40435⁵⁵, GSE46699⁵⁶, GSE47032⁵⁷, GSE53757⁵⁸, GSE66270⁵⁹, GSE68417⁶⁰ and GSE71963⁶¹ datasets were downloaded from the gene expression omnibus (GEO) website (www.ncbi.nlm.nih.gov/geo).

Non-negative matrix factorization (NMF) classification. VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets were classified into two sub-clusters using “NMF” package in R software⁶². Kaplan–Meier estimator tested the clinical overall survival of VHL mutant ccRCC patients. Log-rank test was used to determine the *P* values.

Box and contingency plots. Box plots and contingency plots were generated using GraphPad Prism. *P* values were determined by two tails paired student's *t* test or Chi-square test, respectively.

Heatmap presentation. The differentially expressed genes in sub-cluster1 of VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets were clustered using “pheatmap” package in R software. The differentially expressed genes were determined by the threshold of fold changes > 1.5 and *P* values < 0.001.

Estimation of the immune score and stromal score. The immune scores and stromal scores of VHL mutant ccRCC patients in TCGA and E-MTAB-1980 datasets were determined by “ESTIMATE” package in R software⁶³. The classification of “high” and “low” immune or stromal scores was determined using “scale” method in R software.

Single sample gene set enrichment analysis (ssGSEA). One hundred and eighty-sixth Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathways^{64,65} and 958 transcriptional gene datasets were downloaded from GSEA website (www.broad.mit.edu/gsea/index.html). The scores of signaling pathways and transcriptional gene datasets were determined using “GSVA” package in R software.

Survival analysis. The prognosis of PBX1, EMCN and ERG was determined using “survival” package in R software. The “high” and “low” gene expression subgroups were classified based on the mean expression values. *P* values were determined by log-rank test.

Forest plot. The forest plots were generated using “survival” and “survminer” packages “ggforest” method in R software. The Hazard ratio (HR) and *P* values were determined using univariate cox regression or multivariate cox regression survival analysis.

Risk score plot. The risk score plots were generated using “ggrisk” and “rms” packages in R software. The risk score was based on the cox regression in “survival” package. The cutoff was determined by the median of risk score.

Received: 13 September 2021; Accepted: 20 May 2022

Published online: 27 May 2022

References

- Linehan, W. M., Srinivasan, R. & Schmidt, L. S. The genetic basis of kidney cancer: A metabolic disease. *Nat. Rev. Urol.* **7**, 277–285. <https://doi.org/10.1038/nrurol.2010.47> (2010).
- Chen, F. *et al.* Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep.* **14**, 2476–2489. <https://doi.org/10.1016/j.celrep.2016.02.024> (2016).
- Jonasch, E., Gao, J. & Rathmell, W. K. Renal cell carcinoma. *BMJ* **349**, g4797. <https://doi.org/10.1136/bmj.g4797> (2014).
- Hsieh, J. J. *et al.* Renal cell carcinoma. *Nat. Rev. Dis. Prim.* **3**, 17009. <https://doi.org/10.1038/nrdp.2017.9> (2017).
- Choueiri, T. K. & Motzer, R. J. Systemic therapy for metastatic renal-cell carcinoma. *N. Engl. J. Med.* **376**, 354–366. <https://doi.org/10.1056/NEJMra1601333> (2017).
- Gnarra, J. R. *et al.* Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat. Genet.* **7**, 85–90. <https://doi.org/10.1038/ng0594-85> (1994).
- Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49. <https://doi.org/10.1038/nature12222> (2013).
- Moore, L. E. *et al.* Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: Associations with germline VHL polymorphisms and etiologic risk factors. *PLoS Genet* **7**, e1002312. <https://doi.org/10.1371/journal.pgen.1002312> (2011).
- Latif, F. *et al.* Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* **260**, 1317–1320. <https://doi.org/10.1126/science.8493574> (1993).
- Harlander, S. *et al.* Combined mutation in Vhl, Trp53 and Rb1 causes clear cell renal cell carcinoma in mice. *Nat. Med.* **23**, 869–877. <https://doi.org/10.1038/nm.4343> (2017).

11. Vanharanta, S. *et al.* Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer. *Nat. Med.* **19**, 50–56. <https://doi.org/10.1038/nm.3029> (2013).
12. Melendez-Rodriguez, F., Roche, O., Sanchez-Prieto, R. & Aragonés, J. Hypoxia-inducible factor 2-dependent pathways driving von Hippel-Lindau-deficient renal cancer. *Front. Oncol.* **8**, 214. <https://doi.org/10.3389/fonc.2018.00214> (2018).
13. Kondo, K., Klcó, J., Nakamura, E., Lechpammer, M. & Kaelin, W. G. Jr. Inhibition of HIF is necessary for tumor suppression by the von Hippel-Lindau protein. *Cancer Cell* **1**, 237–246. [https://doi.org/10.1016/s1535-6108\(02\)00043-0](https://doi.org/10.1016/s1535-6108(02)00043-0) (2002).
14. Choueiri, T. K. & Kaelin, W. G. Jr. Targeting the HIF2-VEGF axis in renal cell carcinoma. *Nat. Med.* **26**, 1519–1530. <https://doi.org/10.1038/s41591-020-1093-z> (2020).
15. Cho, H. *et al.* On-target efficacy of a HIF-2alpha antagonist in preclinical kidney cancer models. *Nature* **539**, 107–111. <https://doi.org/10.1038/nature19795> (2016).
16. Chen, W. *et al.* Targeting renal cell carcinoma with a HIF-2 antagonist. *Nature* **539**, 112–117. <https://doi.org/10.1038/nature19796> (2016).
17. Ravaud, A. *et al.* Adjuvant sunitinib in high-risk renal-cell carcinoma after nephrectomy. *N. Engl. J. Med.* **375**, 2246–2254. <https://doi.org/10.1056/NEJMoa1611406> (2016).
18. Motzer, R. J. *et al.* Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. *N. Engl. J. Med.* **356**, 115–124. <https://doi.org/10.1056/NEJMoa065044> (2007).
19. Obradovic, A. *et al.* Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. *Cell* **184**, 2988–3005 e2916. <https://doi.org/10.1016/j.cell.2021.04.038> (2021).
20. Brannon, A. R. *et al.* Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes Cancer* **1**, 152–163. <https://doi.org/10.1177/1947601909359929> (2010).
21. Brannon, A. R. *et al.* Meta-analysis of clear cell renal cell carcinoma gene expression defines a variant subgroup and identifies gender influences on tumor biology. *Eur. Urol.* **61**, 258–268. <https://doi.org/10.1016/j.eururo.2011.10.007> (2012).
22. Serie, D. J. *et al.* Clear cell type a and b molecular subtypes in metastatic clear cell renal cell carcinoma: Tumor heterogeneity and aggressiveness. *Eur. Urol.* **71**, 979–985. <https://doi.org/10.1016/j.eururo.2016.11.018> (2017).
23. Beuselinck, B. *et al.* Molecular subtypes of clear cell renal cell carcinoma are associated with sunitinib response in the metastatic setting. *Clin. Cancer Res.* **21**, 1329–1339. <https://doi.org/10.1158/1078-0432.CCR-14-1128> (2015).
24. Wu, P. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of renal clear cell carcinoma. *BMC Cancer* **18**, 287. <https://doi.org/10.1186/s12885-018-4176-1> (2018).
25. Clark, D. J. *et al.* Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* **179**, 964–983 e931. <https://doi.org/10.1016/j.cell.2019.10.007> (2019).
26. Lin, P. *et al.* Radiomic profiling of clear cell renal cell carcinoma reveals subtypes with distinct prognoses and molecular pathways. *Transl. Oncol.* **14**, 101078. <https://doi.org/10.1016/j.tranon.2021.101078> (2021).
27. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619–625. <https://doi.org/10.1038/nm.3175> (2013).
28. Wang, H., Wang, X., Xu, L., Zhang, J. & Cao, H. A molecular sub-cluster of colon cancer cells with low VDR expression is sensitive to chemotherapy, BRAF inhibitors and PI3K-mTOR inhibitors treatment. *Aging* **11**, 8587–8603. <https://doi.org/10.18632/aging.102349> (2019).
29. Ke, K. *et al.* Evaluation and prediction of hepatocellular carcinoma prognosis based on molecular classification. *Cancer Manag. Res.* **10**, 5291–5302. <https://doi.org/10.2147/CMAR.S178579> (2018).
30. Wang, H., Wang, X., Xu, L., Cao, H. & Zhang, J. Nonnegative matrix factorization-based bioinformatics analysis reveals that TPX2 and SELENBP1 are two predictors of the inner sub-consensuses of lung adenocarcinoma. *Cancer Med.* **10**, 9058–9077. <https://doi.org/10.1002/cam4.4386> (2021).
31. Wang, H., Wang, X., Xu, L., Zhang, J. & Cao, H. RUNX1 and REXO2 are associated with the heterogeneity and prognosis of IDH wild type lower grade glioma. *Sci. Rep.* **11**, 11836. <https://doi.org/10.1038/s41598-021-91382-1> (2021).
32. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867. <https://doi.org/10.1038/ng.2699> (2013).
33. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542. <https://doi.org/10.1038/nature09639> (2011).
34. Dalglish, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360–363. <https://doi.org/10.1038/nature08672> (2010).
35. Vuong, L., Kotecha, R. R., Voss, M. H. & Hakimi, A. A. Tumor microenvironment dynamics in clear-cell renal cell carcinoma. *Cancer Discov.* **9**, 1349–1357. <https://doi.org/10.1158/2159-8290.CD-19-0499> (2019).
36. Chevrier, S. *et al.* An immune atlas of clear cell renal cell carcinoma. *Cell* **169**, 736–749 e718. <https://doi.org/10.1016/j.cell.2017.04.016> (2017).
37. Hakimi, A. A. *et al.* Transcriptomic profiling of the tumor microenvironment reveals distinct subgroups of clear cell renal cell cancer: Data from a randomized phase III trial. *Cancer Discov.* **9**, 510–525. <https://doi.org/10.1158/2159-8290.CD-18-0957> (2019).
38. Luo, J. *et al.* Comprehensive insights on pivotal prognostic signature involved in clear cell renal cell carcinoma microenvironment using the ESTIMATE algorithm. *Cancer Med.* **9**, 4310–4323. <https://doi.org/10.1002/cam4.2983> (2020).
39. Le Tanno, P. *et al.* PBX1 haploinsufficiency leads to syndromic congenital anomalies of the kidney and urinary tract (CAKUT) in humans. *J. Med. Genet.* **54**, 502–510. <https://doi.org/10.1136/jmedgenet-2016-104435> (2017).
40. Heidet, L. *et al.* Targeted exome sequencing identifies PBX1 as involved in monogenic congenital anomalies of the kidney and urinary tract. *J. Am. Soc. Nephrol.* **28**, 2901–2914. <https://doi.org/10.1681/ASN.2017010043> (2017).
41. Smale, S. T. Pioneer factors in embryonic stem cells and differentiation. *Curr. Opin. Genet. Dev.* **20**, 519–526. <https://doi.org/10.1016/j.gde.2010.06.010> (2010).
42. Selleri, L. *et al.* Requirement for Pbx1 in skeletal patterning and programming chondrocyte proliferation and differentiation. *Development* **128**, 3543–3557 (2001).
43. Kamps, M. P., Murre, C., Sun, X. H. & Baltimore, D. A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in pre-B ALL. *Cell* **60**, 547–555. [https://doi.org/10.1016/0092-8674\(90\)90658-2](https://doi.org/10.1016/0092-8674(90)90658-2) (1990).
44. Magnani, L., Ballantyne, E. B., Zhang, X. & Lupien, M. PBX1 genomic pioneer function drives ERalpha signaling underlying progression in breast cancer. *PLoS Genet.* **7**, e1002368. <https://doi.org/10.1371/journal.pgen.1002368> (2011).
45. Wei, X., Yu, L. & Li, Y. PBX1 promotes the cell proliferation via JAK2/STAT3 signaling in clear cell renal carcinoma. *Biochem. Biophys. Res. Commun.* **500**, 650–657. <https://doi.org/10.1016/j.bbrc.2018.04.127> (2018).
46. Zahr, A. *et al.* Endomucin prevents leukocyte-endothelial cell adhesion and has a critical role under resting and inflammatory conditions. *Nat. Commun.* **7**, 10363. <https://doi.org/10.1038/ncomms10363> (2016).
47. Zhang, Q. *et al.* Weighted correlation gene network analysis reveals a new stemness index-related survival model for prognostic prediction in hepatocellular carcinoma. *Aging* **12**, 13502–13517. <https://doi.org/10.18632/aging.103454> (2020).
48. Dai, W. *et al.* Systematical analysis of the cancer genome atlas database reveals EMCN/MUC15 combination as a prognostic signature for gastric cancer. *Front. Mol. Biosci.* **7**, 19. <https://doi.org/10.3389/fmolb.2020.00019> (2020).
49. Hu, Z. *et al.* Elements of the endomucin extracellular domain essential for VEGF-induced VEGFR2 activity. *Cells* <https://doi.org/10.3390/cells9061413> (2020).

50. Fry, E. A. & Inoue, K. Aberrant expression of ETS1 and ETS2 proteins in cancer. *Cancer Rep. Rev.* <https://doi.org/10.15761/CRR.1000151> (2018).
51. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648. <https://doi.org/10.1126/science.1117679> (2005).
52. Sorensen, P. H. *et al.* A second Ewing's sarcoma translocation, t(21;22), fuses the EWS gene to another ETS-family transcription factor ERG. *Nat. Genet.* **6**, 146–151. <https://doi.org/10.1038/ng0294-146> (1994).
53. Kong, X. T. *et al.* Consistent detection of TLS/FUS-ERG chimeric transcripts in acute myeloid leukemia with t(16;21)(p11;q22) and identification of a novel transcript. *Blood* **90**, 1192–1199 (1997).
54. Pena-Llopis, S. *et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nat. Genet.* **44**, 751–759. <https://doi.org/10.1038/ng.2323> (2012).
55. Wozniak, M. B. *et al.* Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. *PLoS ONE* **8**, e57886. <https://doi.org/10.1371/journal.pone.0057886> (2013).
56. Eckel-Passow, J. E. *et al.* ANKS1B is a smoking-related molecular alteration in clear cell renal cell carcinoma. *BMC Urol.* **14**, 14. <https://doi.org/10.1186/1471-2490-14-14> (2014).
57. Valletti, A. *et al.* Genome-wide analysis of differentially expressed genes and splicing isoforms in clear cell renal cell carcinoma. *PLoS ONE* **8**, e78452. <https://doi.org/10.1371/journal.pone.0078452> (2013).
58. von Roemeling, C. A. *et al.* Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the AMPA-selective glutamate receptor-4. *Cancer Res.* **74**, 4796–4810. <https://doi.org/10.1158/0008-5472.CAN-14-0210> (2014).
59. Wotschovsky, Z. *et al.* Integrated microRNA and mRNA signature associated with the transition from the locally confined to the metastasized clear cell renal cell carcinoma exemplified by miR-146-5p. *PLoS ONE* **11**, e0148746. <https://doi.org/10.1371/journal.pone.0148746> (2016).
60. Thibodeau, B. J. *et al.* Characterization of clear cell renal cell carcinoma by gene expression profiling. *Urol. Oncol.* **34**(168), e161–169. <https://doi.org/10.1016/j.urolonc.2015.11.001> (2016).
61. Takahashi, M. *et al.* Downregulation of WDR20 due to loss of 14q is involved in the malignant transformation of clear cell renal cell carcinoma. *Cancer Sci.* **107**, 417–423. <https://doi.org/10.1111/cas.12892> (2016).
62. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **11**, 367. <https://doi.org/10.1186/1471-2105-11-367> (2010).
63. 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).
64. Kanehisa, M., Sato, Y. & Kawashima, M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.* **31**, 47–53. <https://doi.org/10.1002/pro.4172> (2022).
65. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361. <https://doi.org/10.1093/nar/gkw1092> (2017).

Author contributions

H.W. designed the study and wrote the manuscript. H.W. and X.W. performed the data analysis. L.X. and J.Z. supervised the work.

Funding

This study was supported by Natural Science Foundation of Fujian province (Grant Nos. 2020J01337).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-13148-7>.

Correspondence and requests for materials should be addressed to H.W. or J.Z.

Reprints and permissions information is available at 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) 2022