Supplementary Materials for

**Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis**

Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, Kun Huang

**Methods**

**Aggregation of cell-level features into patient-level features**

After 10 types of cell-level features were extracted for each segmented nucleus, we needed to aggregate the cell-level features extracted from a patient into patient-level features. To this end, histograms and distribution statistics were employed. For constructing histogram features, a bag-of-visual-words model was utilized. Specifically, for each type of cell-level feature, a large set of cell-level features were collected across patients and fed into K-means algorithm to learn words (i.e. clustering centers). The number of clusters in K-means algorithm was fixed to ten based on the sensitivity analysis (described in the Results section and Fig. S3). After that, cell-level features extracted from the same patient were assigned to their nearest words using Euclidean distance, which would result in a histogram of word counts for each patient and for each type of cell-level features. Histograms were L1-normalized to eliminate the impact of patients having different number of nuclei. As for distribution statistics, five parameters were calculated for each type of cell-level features, namely, mean, standard deviation, skewness, kurtosis, and entropy. The entropy was computed based on the foregoing normalized histograms. Note that since there were images with different magnifications, all measures concerning to size and length were converted to real size.

To evaluate the performance of the nucleus segmentation algorithm, ten image patches containing 4409 nuclei were randomly selected. The segmentation algorithm segmented 4019 nuclei, among which 112 were false positives (recall, (4019-112)/4409 = 88.6%; precision, (4019-112)/4019 = 97.2%). A qualitative example of the segmentation results is shown in Fig. S1.

**Training and prediction process of lasso-Cox model**

A two-level cross validation strategy was used to validate our method. The first level was leave-one-out CV (LOOCV). Namely, a single patient was chosen as test set, with the rest as training set. The second level was a 10-fold CV performed in the training set to select the best regularization parameter. The readers may notice that for a sample with size *n*, LOOCV is *n*-fold CV, so the time for LOOCV is roughly *n/10* times longer than that for 10-fold CV. Based on this consideration, we chose the frequently-used 10-fold CV for the training process.

Regularized Cox proportional hazards model was built on the training set using the selected parameter, and based on the model, risk indices of all patients were calculated. After that, the median of risk indices of the training examples was used as a threshold to split patients into low-risk and high-risk groups. The same threshold was applied to classify the single held-out patient into one of the two groups. After each patient was used as a test sample and classified, all patients were divided into a low-risk or high-risk group. This process does not necessarily result in two equal-sized groups. Finally, we tested if these two groups had distinct survival outcome using Kaplan-Meier estimator and log-rank test.

**Table S1**. The 15 co-expressed gene modules generated by gene co-expression network analysis.

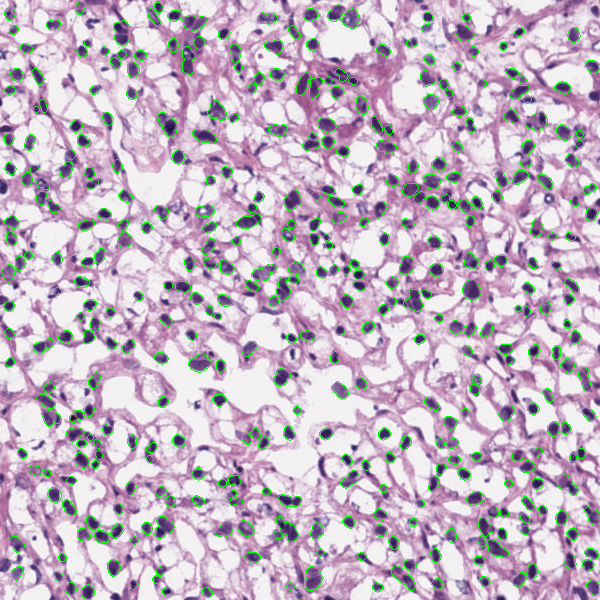
|  |  |
| --- | --- |
| **Module name** | **Genes** |
| Module1 | AFG3L1 AHSA2 AMY2B ANKRD36 ARGLU1 ATXN2 BZRAP1 C12orf51 C17orf68 CABIN1 CALCOCO1 CAPRIN2 CBFA2T2 CCAR1 CCNL1 CELF6 CEP290 CRAMP1L CREBZF CROCCL1 CSAD CUL9 DDX17 DHX16 DIDO1 DMTF1 DNAH1 DNHD1 EP400 FAM48A FAM73B FLJ45340 FNBP4 GABBR1 GCFC1 GGA3 GOLGA1 GOLGA6L9 GOLGA8B HCFC1 HERC2P2 HNRNPA2B1 IKBKB ING5 KCTD7 KIAA0495 KIAA0556 KIAA0907 KIAA1407 KIAA1530 KIAA1731 LENG8 LOC200030 LOC338799 LOC440944 LOC91316 LUC7L3 MALAT1 MAN2A2 MAPKBP1 MAVS MDM4 METTL3 MINK1 MIS12 MLL2 MLLT6 NCRNA00105 NEAT1 NEURL4 NFATC2IP NFKBIZ NICN1 NISCH NKTR NSUN5P1 NSUN5P2 NXF1 OGT PAN2 PAN3 PCNT PDXDC2 PI4KAP1 PIP5K1A PLA2G6 PNN POLI PRPF39 RANBP10 RBM25 RBM39 RBM5 RBM6 RCOR3 RNPC3 RPAP1 RSRC2 RXRB SDHAP1 SDHAP2 SEC31B SETD1B SFRS11 SFRS12 SFRS13A SFRS14 SFRS18 SFRS1 SFRS5 SFRS6 SFRS7 SGSM2 SLC25A27 SLC35E2 SLC9A8 SRCAP SRRM2 STAG3L3 TBRG1 TIA1 TNRC6A TRIM66 TSC1 TTC14 TTLL3 TUBGCP6 VAMP1 VPS39 WDR27 WDR6 WSB1 ZBED5 ZC3H12A ZFC3H1 ZMAT1 ZNF224 ZNF226 ZNF37B ZNF397OS ZNF493 ZNF506 ZNF514 ZNF529 ZNF587 ZNF605 ZNF621 ZNF700 ZNF75D ZNF767 ZNF841 ZNF862 ZRANB2 |
| Module2 | ACTC1 ADAM12 ADH1A ADH1B ADRA2A AOC3 APCDD1L ASPN C1R C1S C7 CACNA1H CAMK1G CCDC3 CCDC80 CCL21 CFH CNN1 COL14A1 COL16A1 COL1A2 COL3A1 COMP CSRP1 DACT1 DACT3 DCN DENND2A DES DIO2 EBF1 ECM2 EDNRA ELN ELOVL4 EMILIN1 F3 FAM101A FGF13 FLNA FMOD FNDC1 GALNT5 GGT5 GLDN GPR124 GREM1 HEPH HS3ST3A1 HSPB6 ISLR ITGA11 ITGBL1 LAMC3 LHFP LMOD1 LOC145820 LTBP2 LTBP4 LUM MFAP4 MFAP5 MGP MRGPRF MRVI1 MYH11 MYH9 NEXN NOTCH3 OGN OLFML1 PCDH7 PDGFRA PDGFRB PLAU PLN PLTP PODN PRELP PTGIS PVRL1 RASL10B ROR2 RSPO3 SCRG1 SELP SERPINF1 SGCD SLC24A3 SLC2A10 SLIT3 SRPX2 SYNPO2 TAGLN THBS2 TLN1 TMEM119 TNFAIP8L3 WFDC1 |
| Module3 | ABCC6P1 ABCC6 ABP1 ACE2 ACMSD ACSM2A ACSM2B AGMAT AGXT2 AKR1C4 ALPI ALPL AMACR ANKS4B ANPEP APOM ARSB ASPDH BBOX1 C22orf45 C9orf66 CABP1 CDHR2 CHDH CMBL CRAT CUBN CYP4A11 CYP8B1 DAB2 DAO DDC DENND1A DPYS ENPP7 FAM151A FCAMR FMO1 FUT6 GALNT11 GC GLYATL1 GLYAT GRHPR HAO2 HNF4A IL17RB KCNH6 KHK LRP2 MIOX MSRA MYO7B MYOM3 NAT8B NAT8 NGEF PGPEP1 PIPOX PKLR PLA2G12B PRAP1 PRODH2 RBP4 SERPINA7 SLC22A11 SLC22A12 SLC22A4 SLC22A5 SLC22A6 SLC22A7 SLC23A1 SLC23A3 SLC25A42 SLC39A5 SLC5A10 SLC5A12 SLC7A9 TINAG TMEM27 TMEM82 TREH UPB1 VIL1 |
| Module4 | ADRBK2 ALOX5AP AOAH ATP8B4 BTK C1QA C1QB C1QC C3AR1 C5AR1 CCR1 CCR2 CD180 CD38 CD53 CD84 CD86 CECR1 CMKLR1 CSF1R CSF2RA CTSS CYBB DOCK2 EVI2B FAM105A FAM49B FCGR2A FCGR2B FCGR2C FPR1 FPR3 FYB GAPT GAS7 GPR183 GPR65 HLA-DMB HLA-DOA HLA-DPA1 HLA-DPB1 HLA-DQA2 HLA-DQB2 HLA-DRA HLA-DRB6 IGF1 IGSF6 IKZF1 IRF8 LCP1 LYZ MARCH1 MNDA MPEG1 MS4A6A MSR1 NCF2 NCKAP1L P2RY13 PLD4 PLEK PRKCB PTAFR PTPLAD2 PTPRC RNASE6 SAMSN1 TAGAP TLR7 TLR8 WDFY4 |
| Module5 | ACVRL1 ARHGEF15 BCL6B C20orf160 CCL14 CD34 CD93 CDH5 CLDN5 CLEC14A COL15A1 COL4A1 COL4A2 CPNE2 CXorf36 CYYR1 ECSCR EDNRB ELFN1 ENG EPHB4 ERG ESAM EXOC3L2 FAM167B FLT4 GIPC3 GJA5 GSN HSPA12B HSPG2 HYAL2 JAM3 LDB2 LIMS2 LMO2 MMRN2 MYCT1 NID1 NID2 NOS3 OAZ2 PALMD PCDH12 PDGFB PDGFD PPAP2A RAMP2 RAMP3 RASIP1 RHOJ ROBO4 S1PR1 SLC9A3R2 SOX7 SPARC TBXA2R THSD1 TIE1 TMEM204 TMEM88 TP53I11 |
| Module6 | B2M BATF2 BTN3A3 CCR2 CCR5 CD38 CD8A CD96 CECR1 CXCL10 CXCL11 CXCL9 CXCR2P1 EOMES EVI2B FYB GBP1 GBP5 GZMK HCP5 HLA-DMB HLA-DOA HLA-DPA1 HLA-DRA IKZF1 IRF4 ITK LCP1 PRKCB PTPRC RASSF5 RHOH SH2D1A SLAMF6 SLAMF7 SPN TAGAP TIGIT WARS |
| Module7 | ATF3 CSRNP1 CYR61 DUSP5 EGR1 EGR2 EGR3 EPHA2 EREG FOSB FOS HBEGF IER2 IER3 IL6 IL8 JUNB MAFF MIDN NR4A1 NR4A2 RGS2 SERTAD1 SOCS3 SOX9 SRF TESC ZFP36 |
| Module8 | BFAR CSNK2A2 EEF1A1P9 EEF1A1 EEF2 EIF2A EIF3D EIF3E EIF3H FXR1 HNRNPA1L2 HSP90AB1 IGBP1 NCRNA00188 NONO RPL10A RPL17 RPL3 RPL7A RPL7 RPS12 RPS3A RPS6 RSL24D1 |
| Module9 | SHCBP1 KIF23 CDK1 KIF11 HMMR TOP2A MELK ZWINT NCAPG ARHGAP11A CENPF MKI67 ASPM RRM2 CEP55 BUB1 TPX2 CDCA5 PRC1 |
| Module10 | APOH C8orf84 CFI CPN2 CYP4F2 CYP4F3 CYP8B1 DIO1 GFRA1 HECW1 MUC13 PIPOX SLC13A3 SLC23A1 SLC36A2 TMEM90B |
| Module11 | WNT5B TMEM130 CLIC6 FMOD ROR2 NXN FNDC4 IGDCC4 MMP7 ZBTB7C IL1R2 PDE10A KRT19 |
| Module12 | DCTN3 STOML2 MRPS2 C9orf23 NDUFB6 TRUB2 SURF1 TMEM203 C9orf123 ATP6V1G1 CLTA APTX |
| Module13 | UBR7 GOLGA5 SNW1 MED6 VIPAR FNTB APEX1 TIMM9 PSMC6 NGDN |
| Module14 | FOLR2 CD209 F13A1 CD163L1 IL2RA CD163 CMKLR1 VSIG4 MS4A4A SLCO2B1 |
| Module15 | CCT6A ABCF2 GARS CCT5 NIP7 PNO1 RAN COPS3 CCT2 ILF2 |

**Table S2**. Gene set enrichment analysis of survival-associated gene modules using ToppGene.

|  |  |  |
| --- | --- | --- |
| **Module name** | **Top molecular function** | **Top biological process** |
| Module1 | RNA binding; poly(A RNA) binding | mRNA processing; RNA splicing |
| Module2 | Glycosaminoglycan binding; extracellular matrix structural constituent | Extracellular matrix organization; extracellular structure organization |
| Module3 | Organic anion transmembrane transporter activity; anion transmembrane transporter activity | Organic acid metabolic process; carboxylic acid metabolic process |
| Module9 | ATP binding; adenyl ribonucleotide binding | Mitotic cell cycle process; mitotic cell cycle |
| Module11 | frizzled binding; G-protein coupled receptor binding | -- |
| Module13 | Transcription factor binding; transcription coactivator activity | RNA polymerase II transcriptional preinitiation complex assembly; transcription initiation from RNA polymerase II promoter |

**Table S3**. Survival-associated image features and eigengenes, identified by Kaplan-Meier estimator and log-rank test in early-stage (stage I and II) patients. For each variable, patients were stratified into low and high groups using the median as cut-off point. For P/N, P means positive relation to survival (i.e., patients with high feature values have good prognosis), whereas N means negative relation to survival.

|  |  |  |
| --- | --- | --- |
| **Feature** | **P value** | **P/N** |
| rMean\_mean | 0.0069 | N |
| rMean\_bin2 | 0.0124 | P |
| rMean\_bin4 | 0.0126 | P |
| rMean\_bin10 | 0.0181 | N |
| area\_bin9 | 0.0205 | N |
| distMean\_bin10 | 0.0212 | N |
| distMin\_std | 0.0230 | N |
| distMin\_entropy | 0.0247 | N |
| distMin\_bin10 | 0.0264 | N |
| distMean\_std | 0.0297 | N |
| rMean\_bin8 | 0.0358 | N |
| rMean\_bin9 | 0.0391 | N |
| distMean\_bin9 | 0.0401 | N |
| eigengene3 | 0.0024 | P |
| eigengene13 | 0.0438 | P |



**Fig. S1**. An example of nucleus segmentation results.

D:\Research\my publications\cancer research\fig_s2.tif

**Fig S2**: (A) Mutual information values between 5 eigengene expression features and 8 image features selected by lasso. (B) Mutual information values between relatively highly correlated (Spearman correlation > 0.3) eigengene expression features and image features.

D:\Research\my publications\cancer research\fig_s3.tif

**Fig. S3**. Log-rank test p value of lasso-Cox prognostic model as a function of the number of clusters in K-means algorithm.

D:\Research\my publications\cancer research\fig_s4.tif

**Fig. S4**. Log-rank test p value of lasso-Cox prognostic model as a function of k in k-fold cross validation. The sample size of our dataset is 410, so 410-fold cross validation is actually leave-one-out cross validation.