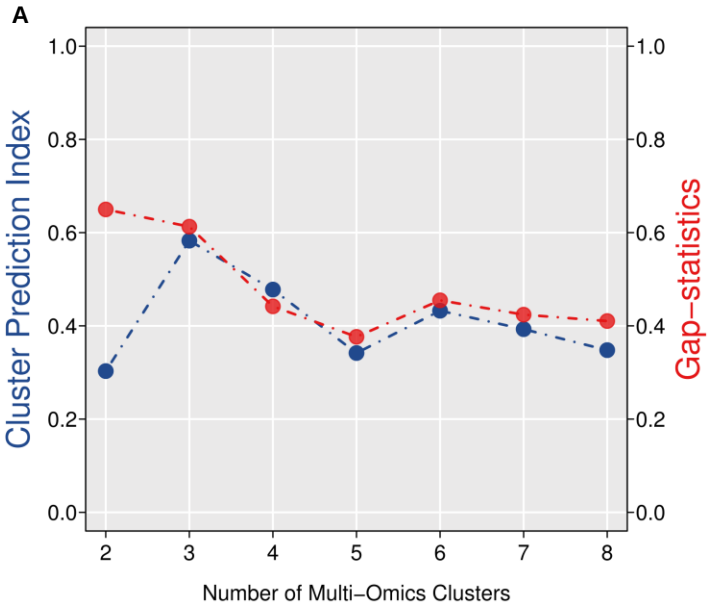
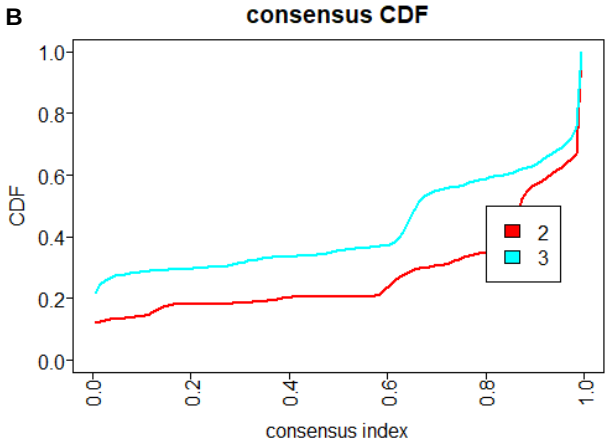
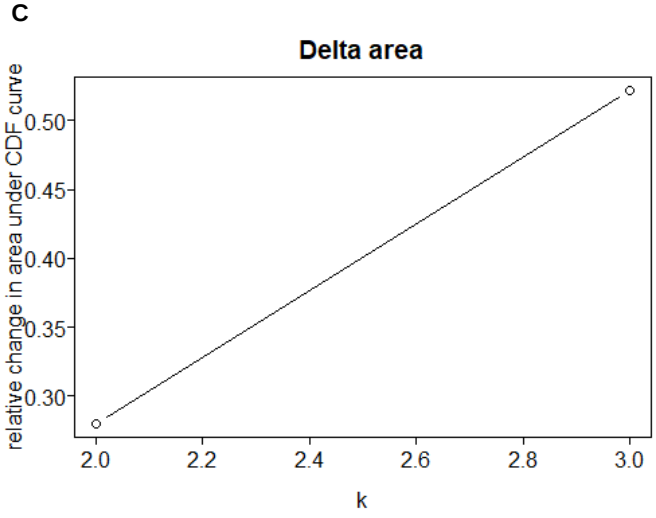
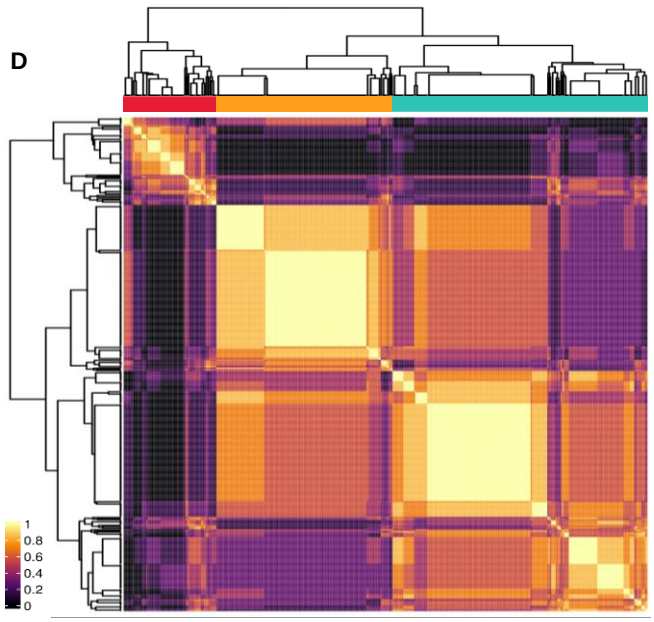
Supplemental information

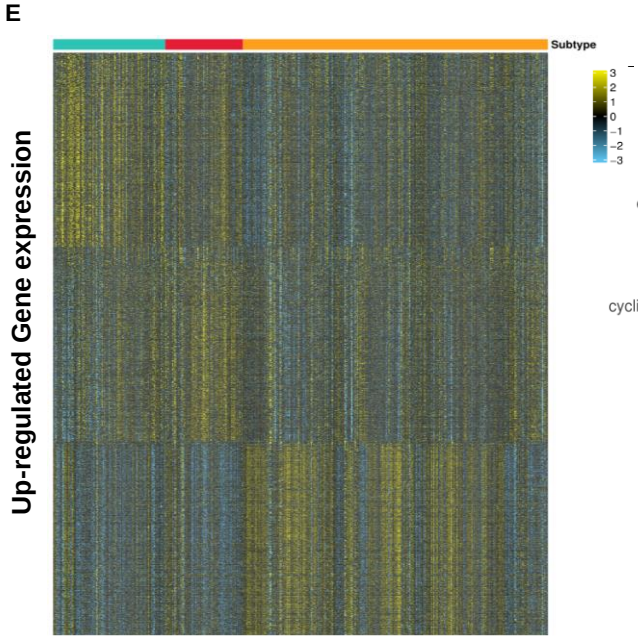
Microbiome subsets determine tumor prognosis and molecular characteristics of clear-cell renal cell carcinoma: A multi-center integrated analysis of microbiome, metabolome and transcriptome data


A
Cluster Prediction Index
Gap−statistics
Number of Multi−Omics Clusters


B
consensus CDF
CDF
consensus index
2
3


C
Delta area
relative change in area under CDF curve
k


D


E
Subtype
Up-regulated Gene expression

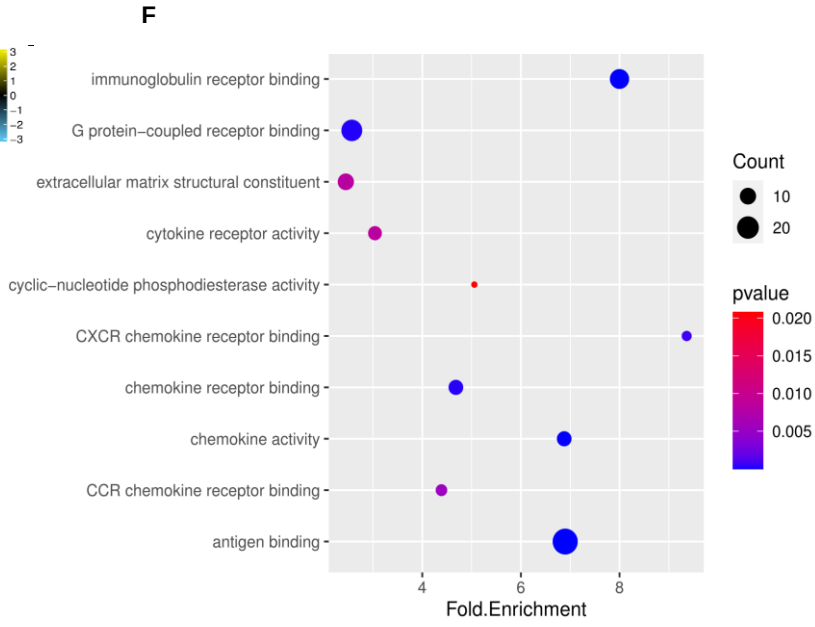

F
immunoglobulin receptor binding
G protein−coupled receptor binding
extracellular matrix structural constituent
cytokine receptor activity
cyclic−nucleotide phosphodiesterase activity
CXCR chemokine receptor binding
chemokine receptor binding
chemokine activity
CCR chemokine receptor binding
antigen binding
Fold.Enrichment
Count
pvalue

Fig.S1

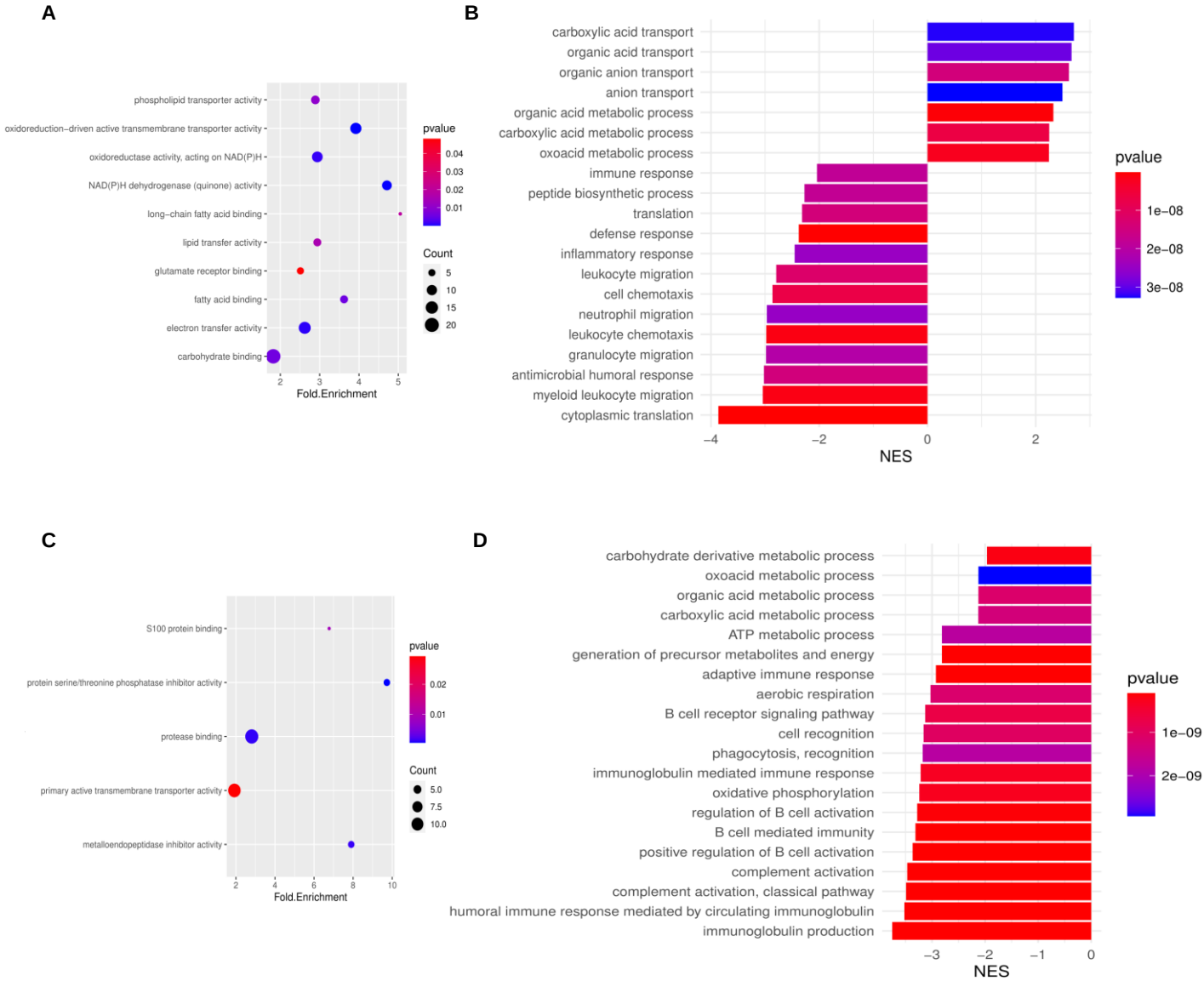
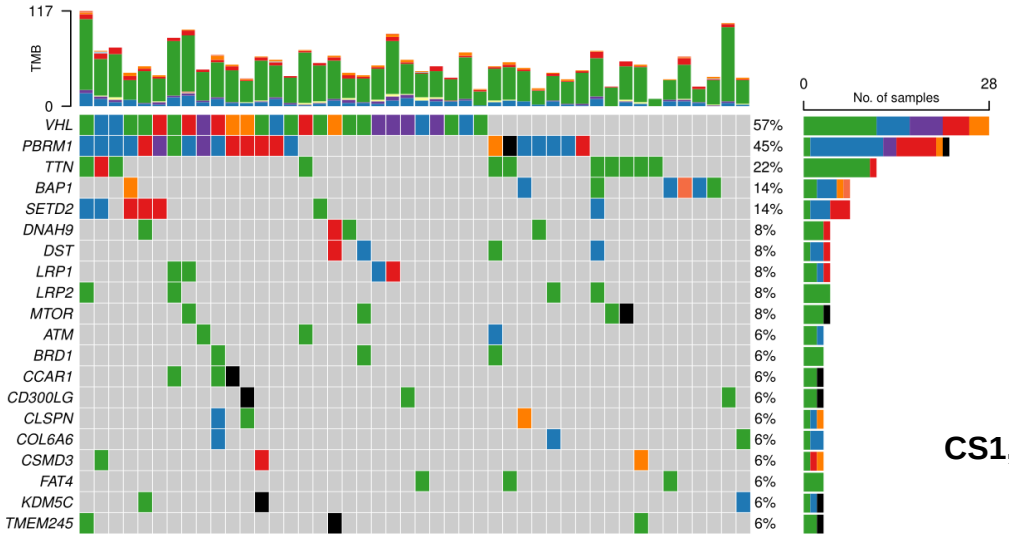
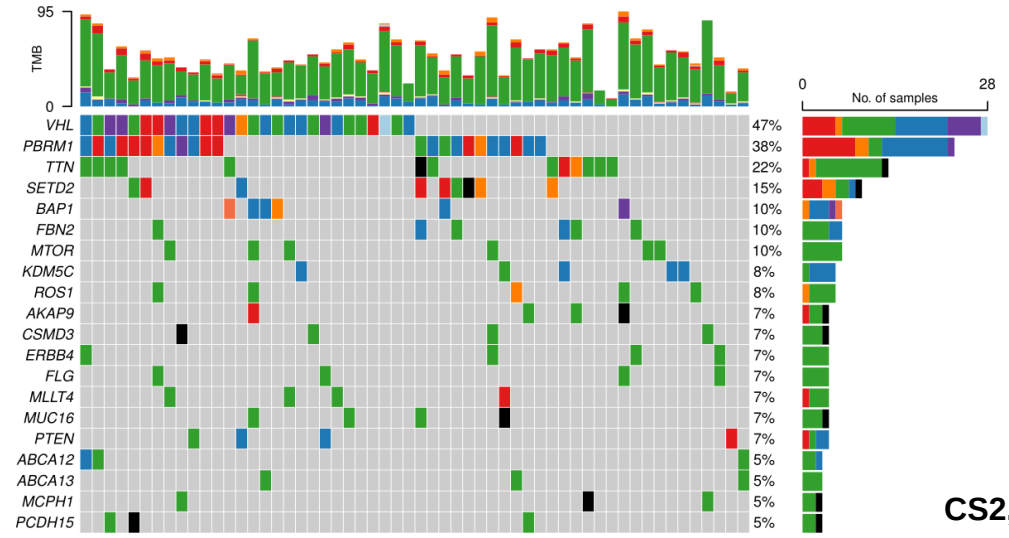

A
phospholipid transporter activity
oxidoreduction−driven active transmembrane transporter activity
oxidoreductase activity, acting on NAD(P)H
NAD(P)H dehydrogenase (quinone) activity
long−chain fatty acid binding
lipid transfer activity
glutamate receptor binding
fatty acid binding
electron transfer activity
carbohydrate binding
2
3
4
5
Fold.Enrichment
pvalue
0.04
0.03
0.02
0.01
Count
5
10
15
20
B
carboxylic acid transport
organic acid transport
organic anion transport
anion transport
organic acid metabolic process
carboxylic acid metabolic process
oxoacid metabolic process
immune response
peptide biosynthetic process
translation
defense response
inflammatory response
leukocyte migration
cell chemotaxis
neutrophil migration
leukocyte chemotaxis
granulocyte migration
antimicrobial humoral response
myeloid leukocyte migration
cytoplasmic translation
−4
−2
0
2
NES
pvalue
1e−08
2e−08
3e−08
C
S100 protein binding
protein serine/threonine phosphatase inhibitor activity
protease binding
primary active transmembrane transporter activity
metalloendopeptidase inhibitor activity
2
4
6
8
10
Fold.Enrichment
pvalue
0.02
0.01
Count
5.0
7.5
10.0
D
carbohydrate derivative metabolic process
oxoacid metabolic process
organic acid metabolic process
carboxylic acid metabolic process
ATP metabolic process
generation of precursor metabolites and energy
adaptive immune response
aerobic respiration
B cell receptor signaling pathway
cell recognition
phagocytosis, recognition
immunoglobulin mediated immune response
oxidative phosphorylation
regulation of B cell activation
B cell mediated immunity
positive regulation of B cell activation
complement activation
complement activation, classical pathway
humoral immune response mediated by circulating immunoglobulin
immunoglobulin production
−3
−2
−1
0
NES
pvalue
1e−09
2e−09

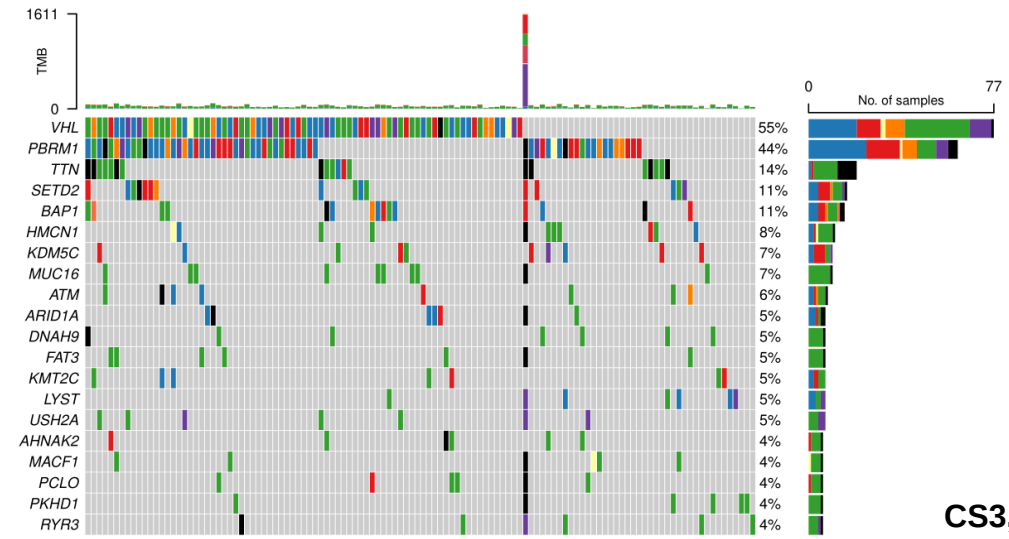
Fig.S2



CS1, N=49



CS2, N=60



CS3, N=141

Fig.S3

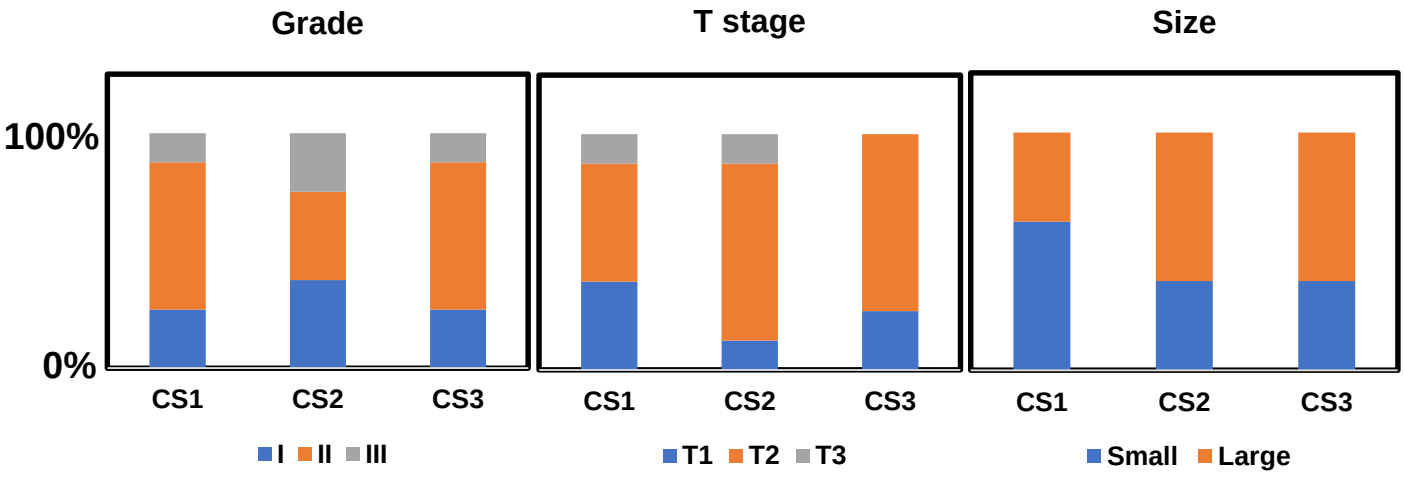

Grade
T stage
Size
100%
0%
CS1
CS2
CS3
CS1
CS2
CS3
CS1
CS2
CS3
I
II
III
T1
T2
T3
Small
Large

Fig.S4

Table S1 Clinical characteristics of ccRCC cohort in our institution.

Characteristic		Total (n = 31)
Age (Y)	< 60	19 (61.3%)
	≥ 60	12 (38.7%)
Gender	Male	21 (67.7)
	Female	10 (32.3%)
Grade	I-II	24 (77.4%)
	III-IV	7 (22.6%)
pT_Stage	T1	6 (19.4%)
	T2	23 (74.2%)
	T3-T4	2 (6.5%)
Size	Small (≤ 4cm)	13 (41.9%)
	Large (< 4cm)	18 (58.1%)

Supplementary figure legends

Fig. S1. **Identifying the bacteria clusters**. A-C. Prediction of optimal cluster number of multi-omics clusters by cluster prediction index and Gap-statistics; D. Consensus heatmap based on the 10 integrative clustering algorithms to refine the clusters; E. The upregulated mRNA expression in each group by TCGA KIRC bulk-seq; F. The dot plot of GSEA enriched result in Group 2. The size of dots represents the pathways count; The color of dots represents the P value.

Fig. S2. **GSEA results in each group**. A. The dot plot of GSEA enriched result in Group 1. The size of dots represents the pathways count; The color of dots represents the P value;

B. The bar plot presents the comparison between Group 1 and Group 2. The color of bar represents the P value. C. The dot plot of GSEA enriched result in Group 3; The size of dots represents the pathways count; The color of dots represents the P value. D. The bar plot presents the comparison between Group 1 and Group 2. The color of bar represents the P value.

Fig. S3. **Oncoplot of top 20 mutation genes in each group**. From top to bottom is Group 1 (N=49), 2 (N=60), 3 (N=141). TMB, Tumor mutation burden.

Fig. S4. **Clinical charateristics of patients in each group from our cohort.** From left to right is for Tumor Grade, T-stage and size.

Methods

Patients and tissue samples

The clinical characteristics and demographics of thirty-one patients pathologically diagnosed as ccRCC were presented in Table 1. Tumor and para-tumor tissues were both obtained from Xinhua hospital and Third affiliated hospital of Naval Medical University. All samples were collected and analyzed after informed consent was obtained from the patients and the approval of Ethical Committee of these two institutions.

Bacteria DNA extraction and 16S rRNA sequencing

Frozen tissue was extracted with the TGuide S96 Magnetic Soil/Stool DNA Kit (Tiangen Biotech (Beijing) Co.,Ltd, China) according to the manufacturer's instructions. PCR amplification was performed with 16S rRNA universal primers 338F (5'- ACTCCTACGGGAGGCAGCA-3') and 806R (5'- GGACTACHVGGGTWTCTAAT-3') from V3 to V4 region. The amplified products were purified, quantified and homogenized to form sequencing libraries. The established libraries were inspected first, and the qualified libraries were double-ended sequenced with Illumina Novaseq 6000 (Illumina，USA). All the sequencing data could be accessed in the Genome Sequence Archive of Human in the BIG Data Center, Chinese Academy of Sciences under accession codes HRA002590 (https://bigd.big.ac.cn/gsa-human).

16S data analysis

Contamination correction was performed as previous studies [1, 2]. The mean and standard errors of each cluster abundance are expressed as bar-PLO The Alpha diversity matrix and analysis are calculated by "-alpha \u div" and "-alpha \u div\u rare" in Usearch [3], and the

contamination correctly abundance was used. For phenotypes prediction, BugBase first normalized OTU with a predicted 16S copy number and then predicted microbial phenotypes using the provided pre-calculated files [4]. After setting thresholds (identity cutoff 0.97), BugBase generates a final biology-level trait prediction table containing the relative abundance of predicted traits for each sample. Wilcoxon test (paired) was used to assess the significance of differences in Alpha diversity or prediction between the two groups, which were presented as bar plot. Contamination-corrected abundance data are used to analyze of differences between samples from two or more sources by to edgeR [2, 5]. Classifications with $P<0.05$ are considered enriched or deficient. Volcano map was taken from edgeR output.

Gram staining and 16S RNA FISH

To visualize bacteria, gram staining was performed on both the tumor and normal tissues using Gram staining kit for tissue (Sigma, St. Louis, MO) according to the manufacture's instruction.

The tissue slices were dewaxed, rehydrated and incubated in 70% ethanol at 4˚C for 2 h. Slides were rinsed in 2X SSC (Ambion #AM9765) and incubated with protease K(10μg/mL, Ambion #AM2546 in 2X SSC) for 10 min. Wash twice with 2XSSC and then twice with 15% formamide (Ambion #AM9342) in 2X SSC. The cy5-labeled probe was referenced from previous study: EUB338 (37) - GCTGCCTCCCGTAGGAGT and nonspecific complement probe – CGACGGAGGGCATCCTCA [1], hybridized overnight at 30˚C 2XSSC with 10% dexan sulfate (Sigma #D8906), 1mg/ml Escherichia coli tRNA (Sigma #R4251), 0.02% BSA (Ambion #AM2616), 2mM vanadyl-RIbonucleic acid (New England Biolabs #S1402S) and 15% formamide. The slices are washed at 2XSSC with 15% formamide 30˚C for 30 min and then incubated with 2XSSC, 15% formamide and DAPI 30˚C for 30 min.

The Cancer Microbiome Atlas (TCMA)

Due to the procedural controls are rarely implemented in cancer genomics projects, TCGA database are limited by the potential for sample contamination during collection, processing, and sequencing. Microbial-based diagnostics can be rationally developed using recently developed tools to minimize the contribution of contaminants to microbial signatures. To

characterize cancer-related microbiome, we re-examined the whole transcriptome sequencing in 428 patients (both have the clinical information and bacteria) from TCGA KIRC as the Poore et al reported [2]. The details information about this cohort was in **Supplementary Table. TCGA train cohort information.**

Identification of subtypes by multiple clustering algorithms

The bacteria (genus level) abundance was finally used to perform clustering due to the species annotation accuracy of the 16S rRNA sequencing. The abundance count of these 428-patient cohort was in **Supplementary Table. TCMA train cohort abundance.** After normalization by GDCRNATools [2], the analysis of cluster prediction index (CPI) and gap statistics were evaluated to find the optimal number of subtypes by MOVICS [3]. Then, ten clustering algorithms were used to classify patients according to different subtypes. Finally, combinatorial classification is obtained from consistent sets and subtypes are identified with high robustness. The ten clustering algorithms include iClusterBayes, moCluster, CIMLR, IntNMF, ConsensusClustering, COCA, NEMO, PINSPlus, SNF and LRA. Sample similarity in subtypes was quantified using contour scores.

Matched bulk RNA-seq expression and whole-exome sequencing (WES) analysis

The "TCGAbiolinks" package in R was used to obtain the transcriptome expression data and clinicopathological information of TCGA-KIRC cohort. GENCODE27 annotations file and was used to match mRNA gene symbol. The transcriptomic count data were were also normalized by GDCRNATools [2] for consistency when normalized data were needed. The differentially expressed genes (DEGs) were calculated by the “runDEA” function based on EdgeR in the MOVICS package. The heatmap of DEGs in three groups was plotted by “runMarker” function. The DEGs ($P < 0.05$) were subsequently used to Gene Set Enrichment Analysis (GSEA) by clusterProfiler package [3]. Only the pathways with enrichment P value < 0.05 were regarded significant.

The WES data of TCGA KIRC were acquired and analyzed by maftools package [4]. The numbers of matched patients in three groups were 49, 60, 141 (Fig. S3). The oncoplots

presented the top 20 mutation genes in each group.

Metabolites detection and analysis

Frozen samples are freeze-dried using a vacuum freeze-dryer (ScientZ-100F). The lyophilized samples were pulverized for 1.5 min at 30 Hz using a mixed mill (MM 400, Retsch) with zirconia beads. Dissolve 100mg freeze-dried powder in 1.2ml 70% methanol solution, vortex every 30 minutes for 30 seconds, 6 times in total, and place the sample in a 4°C refrigerator overnight. After centrifugation at 12000rpm for 10 min, the extract was filtered (SCAA-104, 0.22μm aperture; ANPEL, Shanghai, China, http://www.anpel.com.cn/), and uplC-MS /MS analysis was performed.

Use UPLC-ESI-MS /MS system (UPLC, SHIMADZU Nexera X2, www.shimadzu.com.cn/; MS, Applied Biosystems 4500 Q TRAP, extract the www.appliedbiosystems.com.cn/) samples. The analysis conditions were as follows: UPLC column, Agilent SB-C18 (1.8 μm, 2.1 mm*100 mm); The mobile phase consists of solvent A, pure water containing 0.1% formic acid, and solvent B, acetonitrile containing 0.1% formic acid. Sample measurements were made using A gradient program, starting at 95% A and 5% B. For 9 minutes, the program was set to A linear gradient of 5% A and 95% B, and the composition of 5% A and 95% B remained for 1 minute. Subsequently, the composition of 95% A and 5.0% B was adjusted and maintained for 2.9 minutes in 1.10 minutes. The flow rate was set to 0.35 mL /min. The column temperature box is set to 40°C; The injection volume was 4μ L.

LIT and Triple Quadrupole (QQQ) scans were obtained on a triple quadrupole Linear Ion TRAP Mass Spectrometer (Q TRAP), AB4500 Q TRAP UPLC/MS/MS system equipped with ESI Turbo ion spray interface, Ion mode is run in positive and negative mode and controlled by Analyst 1.6.3 software (AB Sciex). Operating parameters of ESI source are as follows: ion source, turbine spray; Source temperature 550°C; Ion spray voltage (IS) 5500 V (positive ion mode) /-4500 V (negative ion mode); Ion source gas I (GSI), gas II (GSII) and curtain gas (CUR) are set at 50 psi, 60 psi and 25.0 psi, respectively. Collision-activated dissociation (CAD) is high. Instrument tuning and mass calibration were performed using 10 and 100 μmol/L polypropylene

glycol solution in QQQ and LIT modes, respectively. The QQQ scan was obtained as an MRM experiment with the collision gas (nitrogen) set to medium. DP and CE of a single MRM transformation is completed by further optimizing DP and CE. A specific set of MRM ion pairs for each period was monitored based on the eluted metabolites during the period. Finally, the metabolites abundance matrix (both negative and positive) were used to perform differential metabolites analysis ($P < 0.05$) by the “runDEA” function based on EdgeR in the MOVICS package. All the differential metabolites ($P < 0.05$) of three groups were applied to perform pathways analysis on https://www.metaboanalyst.ca/. Only the pathways with enrichment P value < 0.05 were regarded significant [2].

References

[1] Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science. 2020;368:973-80.

[2] Fu A, Yao B, Dong T, Chen Y, Yao J, Liu Y, et al. Tumor-resident intracellular microbiota promotes metastatic colonization in breast cancer. Cell. 2022;185.

[3] Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10:996-8.

[4] Ward T, Larson J, Meulemans J, Hillmann B, Lynch J, Sidiropoulos D, et al. BugBase predicts organism-level microbiome phenotypes. bioRxiv. 2017:133462.

[5] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-40.

[6] Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2020;579:567-74.

[7] Li R, Qu H, Wang S, Wei J, Zhang L, Ma R, et al. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics. 2018;34:2515-7.

[8] Lu X, Meng J, Zhou Y, Jiang L, Yan F. MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. Bioinformatics. 2020.

[9] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological

themes among gene clusters. OMICS. 2012;16:284-7.

[10] Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747-56.

[11] Pang Z, Zhou G, Ewald J, Chang L, Hacariz O, Basu N, et al. Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nat Protoc. 2022.