SUPPLEMENTARY INFORMATION

Single Cell Analyses of Renal Cell Cancers Reveal Insights into Tumor Microenvironment, Cell of Origin, and Therapy Response

Yuping Zhang[1,#], Sathiya P. Narayanan[1,#], Rahul Mannan[1], Gregory Raskind[1], Xiaoming Wang[1], Pankaj Vats[1], Fengyun Su[1], Noshad Hosseini[1,2], Xuhong Cao[1,3,4], Chandan Kumar-Sinha[1,3], Stephanie J. Ellison[1], Thomas J. Giordano[3], Todd M. Morgan[1,5,6], Sethuramasundaram Pitchiaya[1,3], Ajjai Alva[1,5,7], Rohit Mehra[1,3,6], Marcin Cieslik[1,3], Saravana M. Dhanasekaran[1,3,@], and Arul M. Chinnaiyan[1,3,4,5,6,@,*]

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan 48109
[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109
[3]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109
[4]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan 48109
[5]Department of Urology, University of Michigan, Ann Arbor, Michigan 48109
[6]Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan 48109
[7]Department of Internal Medicine, Division of Hematology/Oncology, University of Michigan, Ann Arbor, Michigan 48109
[#]These authors contributed equally
[@]Senior authors

***Correspondence to:**
Arul M. Chinnaiyan, M.D., Ph.D.
arul@umich.edu

This PDF includes:
Figures S1 to S9
Supplementary Methods
Supplementary Information References

Other supplementary materials for this manuscript include the following (separate Excel files):
Dataset S1. Patient clinical characteristics and sequencing library information
Dataset S2. Cell clusters and markers from benign kidney
Dataset S3. Cell clusters from ccRCC and chRCC single cell analyses
Dataset S4. Concept analysis of differentially expressed genes

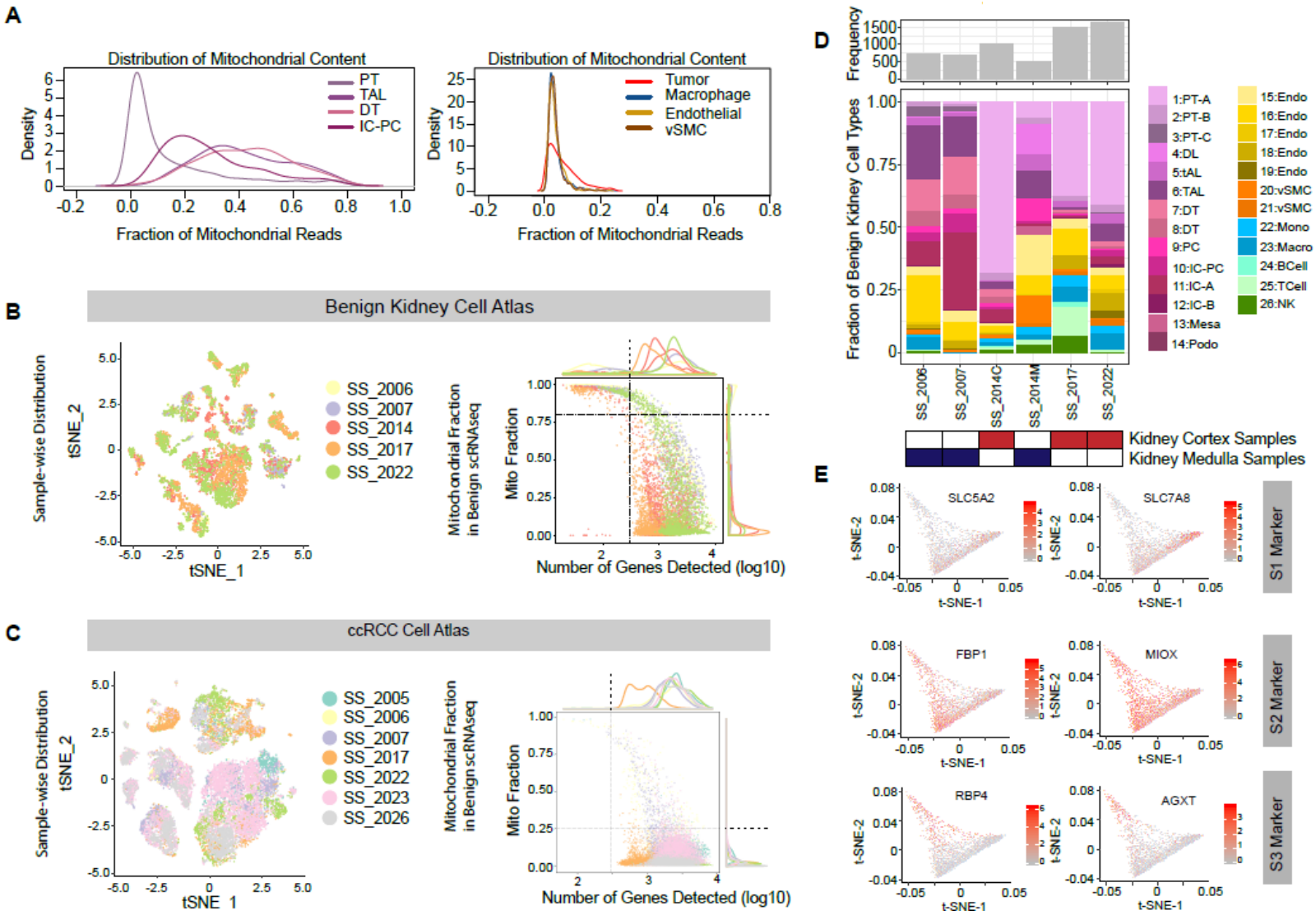

Figure S1. Association between benign kidney cell type composition and anatomic location. A) Density plot shows the distribution of mitochondrial reads among renal tubular epithelial cell types (left) and tumor epithelia/microenvironment cell types (right). **B)** t-SNE plot (left) shows benign tissue cell type clusters presented in **Figure 1A** colored by sample (left), and the waterfall plot again colored by sample (right) shows the mitochondrial threshold employed for benign tissue data filtering. **C)** t-SNE plot (left) shows tumor tissue cell type clusters presented in **Figure S2A** colored by sample (left), and the waterfall plot colored by sample (right) shows the mitochondrial threshold employed for tumor tissue data filtering. **D)** Stacked bar plots depict the proportion of the different cell clusters as identified by single cell sequencing analysis in each benign cortex (red) or medulla (blue) sample as indicated in the annotation box below the stacked bar plot. For example, PT cell clusters were predominantly derived from benign cortex samples (SS_2014C, SS_2017, and SS_2022). **E)** t-SNE plots showing expression of known proximal tubule S1, S2, and S3 marker genes among the PT-A subclusters identified by Slingshot trajectory analysis.

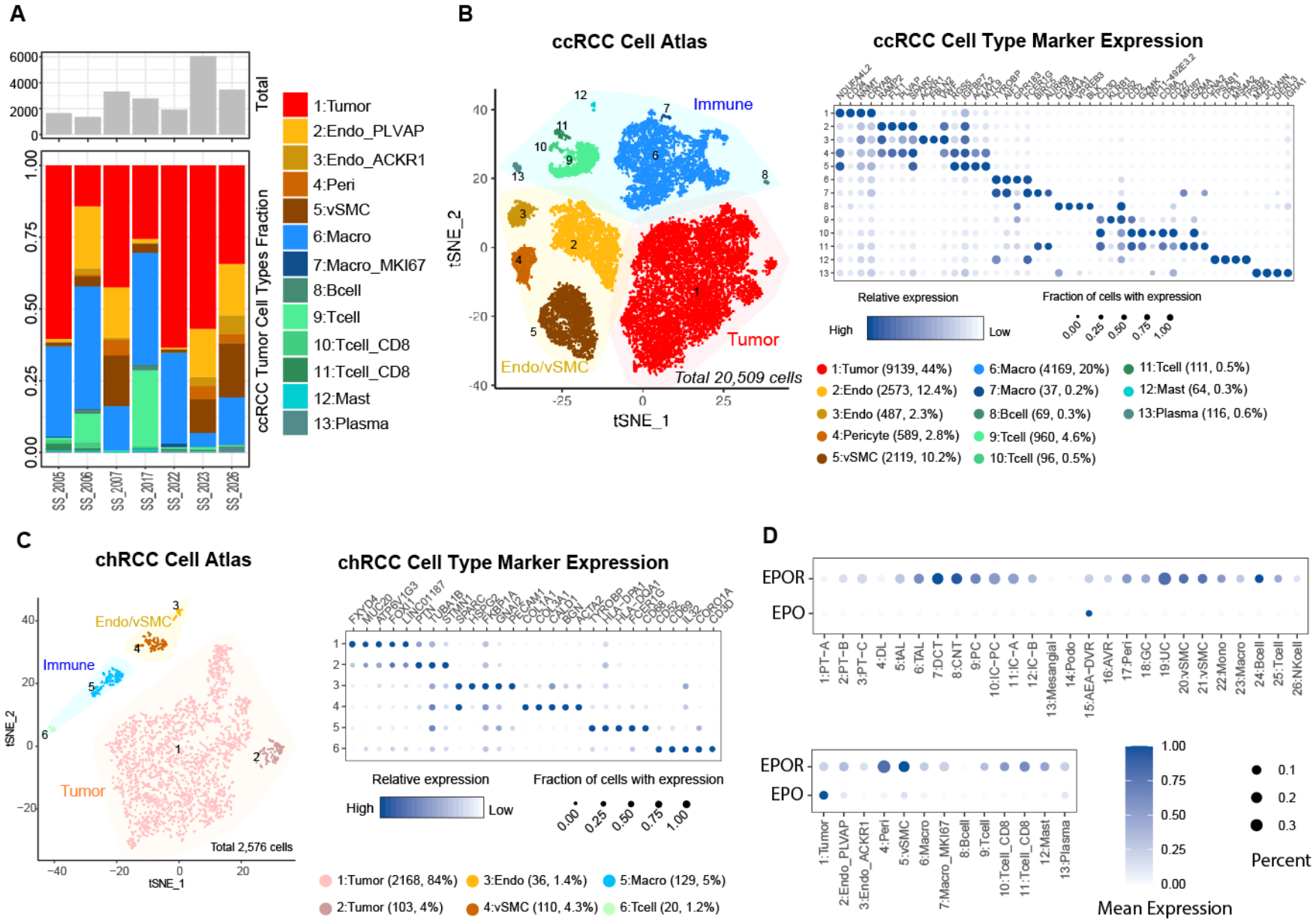

Figure S2. Clear cell (ccRCC) and chromophobe renal cell carcinoma (chRCC) tumor cell atlases. A) Stacked bar plots depict the proportion of different cell clusters as identified by single cell sequencing analysis in the various ccRCC samples. **B)** t-SNE plot (left panel) shows 13 different cell clusters identified from a total of 20,509 cells from seven different ccRCC samples. Number of cells in each cluster and their percentage is listed. Bubble plot (right panel) shows the expression of top cell type-specific markers, where the diameter and color of the bubbles are proportional to the percentage of cells that express a given marker and the expression level, respectively. **C)** Likewise, six cell clusters from 2,576 cells from one chRCC single cell sequencing (t-SNE plot: left panel). Bubble plot shows expression of top cell type-specific markers (right panel). **D)** *EPO* and *EPOR* gene expression in benign kidney and ccRCC tissue cell types.

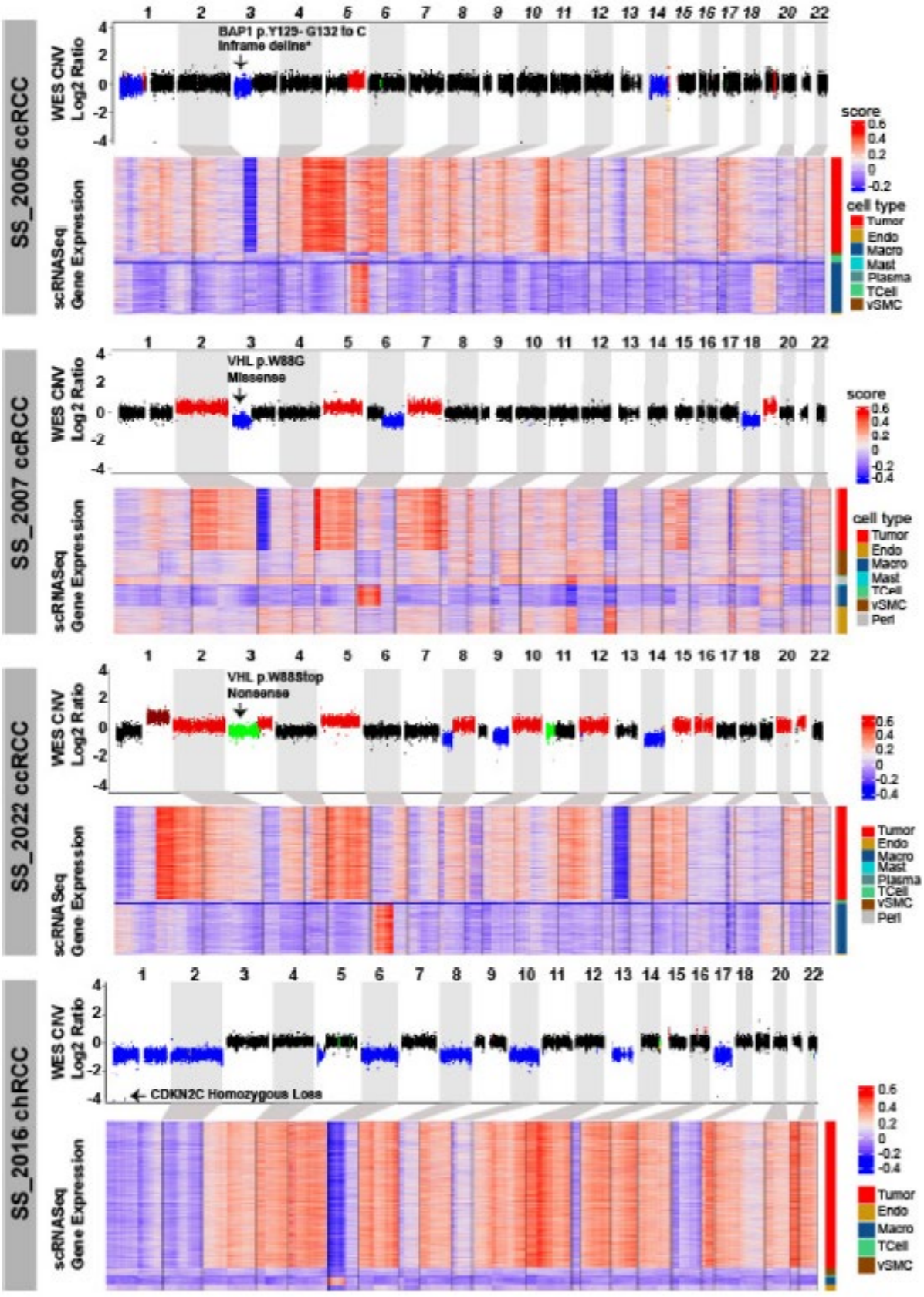

Figure S3. Association between copy-number variation (CNV) and gene expression. Top panels: CNV assessed from whole-exome capture sequencing data analysis in ccRCC (SS_2005, SS_2007, and SS_2022) and chRCC (SS_2016) samples. In the copy-number log ratio plot: blue - copy loss, red - copy gain, black - diploid, green - copy neutral loss of heterozygosity. Bottom panels: Heat maps show relative expression of every cell in each sample along the genome. The average expression of benign cell clusters within the same sample was used as reference. Genes were ordered according to chromosomal location.

A

No	Comparison	Data Type	Genes Up (Fraction %)	Genes Down (Fraction %)	Not Significant (Fraction %)
1	ccRCC / PT-A	scRNAseq	2609 (19.5)	2322 (17.3)	8460 (63.2)
2	ccRCC / PT-B	scRNAseq	907 (6.8)	809 (6.0)	11675 (87.2)
3	ccRCC / PT-C	scRNAseq	534 (4.0)	638 (4.8)	12219 (91.2)
4	ccRCC / NAT	Bulk RNAseq	3257 (22.7)	3958 (26.5)	7696 (51.6)
5	chRCC / IC-A	scRNAseq	377 (2.8)	365 (2.7)	12649 (94.5)
6	chRCC / NAT	Bulk RNAseq	4559 (30.6)	3716 (25.9)	6083 (42.4)

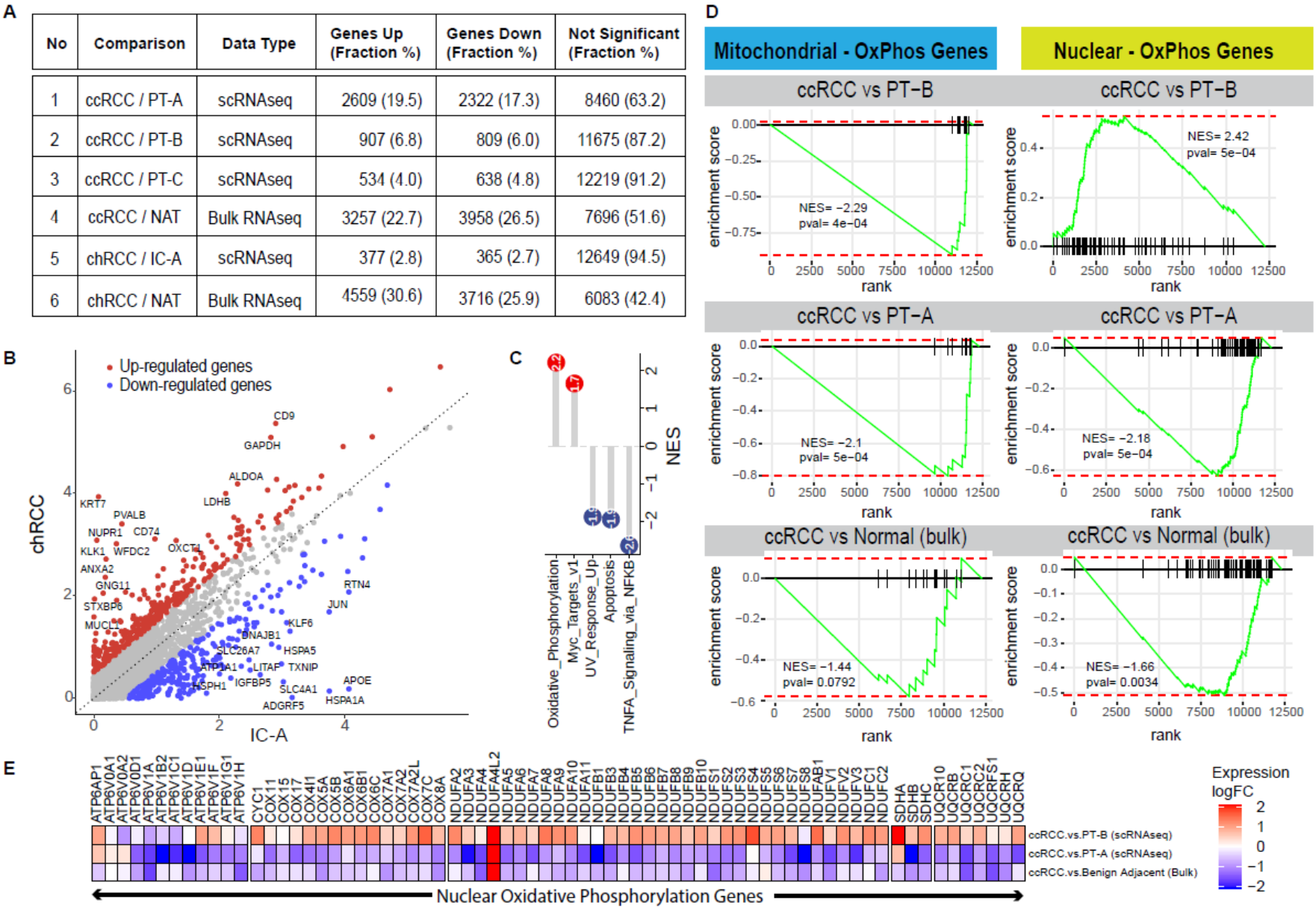

Figure S4. Differential gene expression analysis. A) Table provides the number and fraction of significantly up- or downregulated genes in various differential expression analyses. ccRCC tumor epithelial cells vs. benign kidney proximal tubule cell clusters including 1) PT-A, 2) PT-B, and 3) PT-C; 4) Bulk ccRCC vs. bulk normal adjacent tissues (NAT); 5) chRCC tumor epithelial cells vs. intercalated cells (IC-A); 6) chRCC bulk tumor versus bulk NAT comparison. **B)** Scatter plot shows differentially expressed genes (DEGs) identified in chRCC versus IC-A scRNA-seq data analysis. **C)** Hallmark pathway enrichment analysis of DEGs identified from panel **B**. **D)** Gene set enrichment analysis (GSEA) plots for nuclear and mitochondrial oxidative phosphorylation (OxPhos) genes. **E)** Gene expression pattern of nuclear OxPhos genes in ccRCC vs PT-A, ccRCC vs PT-B, and ccRCC vs NAT (bulk).

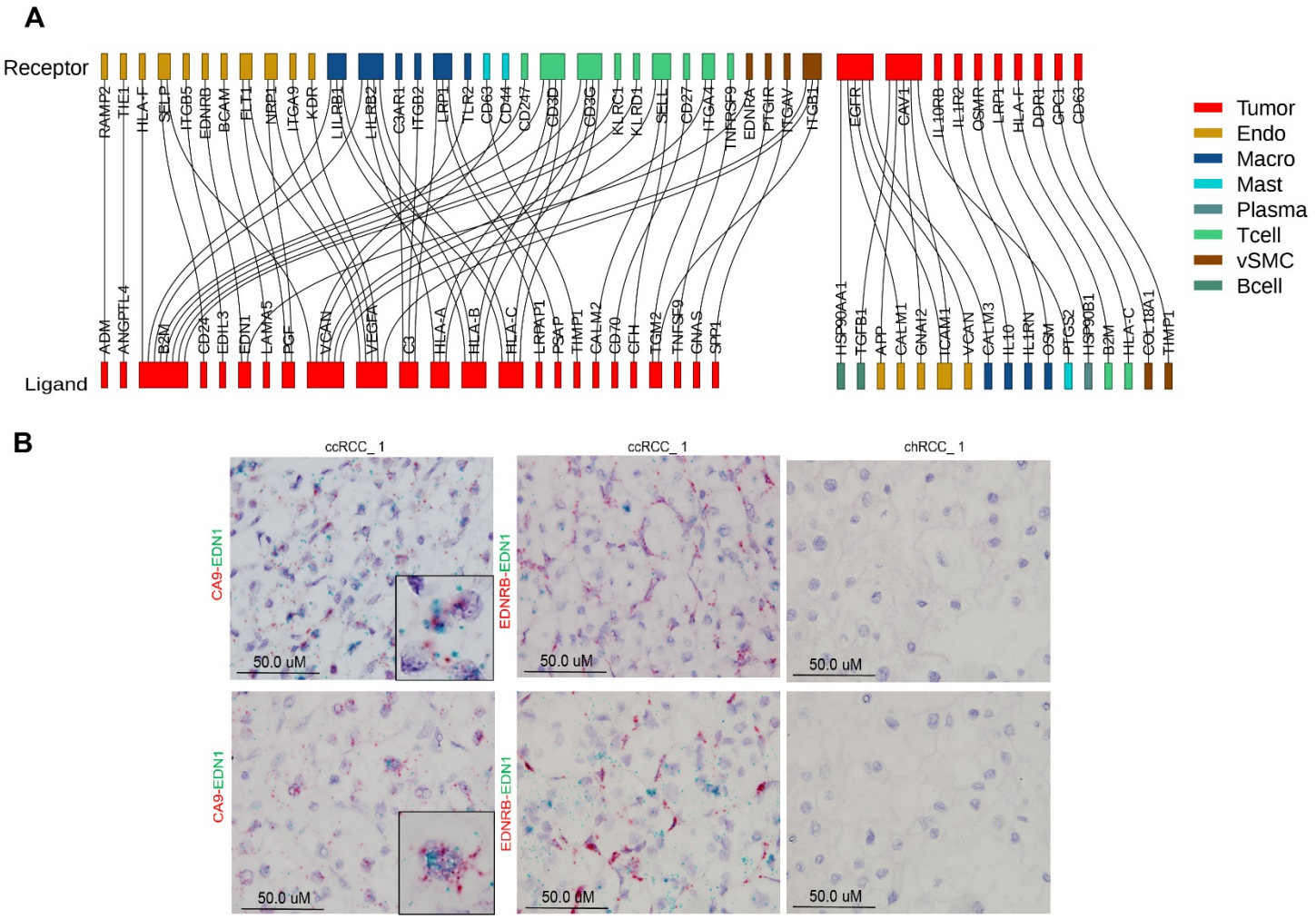

Figure S5. Ligand receptor analysis in the tumor epithelia and the microenvironment. A) Cognate ligand receptor expression analysis in the tumor epithelia and the microenvironment identified from scRNA-seq. **B)** Dual RNA *in situ* hybridization (RNA-ISH) validation - Left panel: ccRCC tumor epithelia marker *CA9* (red) / *EDN1* (cyan) in ccRCC tissues; Middle panel: *EDNRB* (red) / *EDN1* (cyan) in ccRCC tissues; Right panel: *EDNRB* (red) / *EDN1* (cyan) in chRCC tissues.

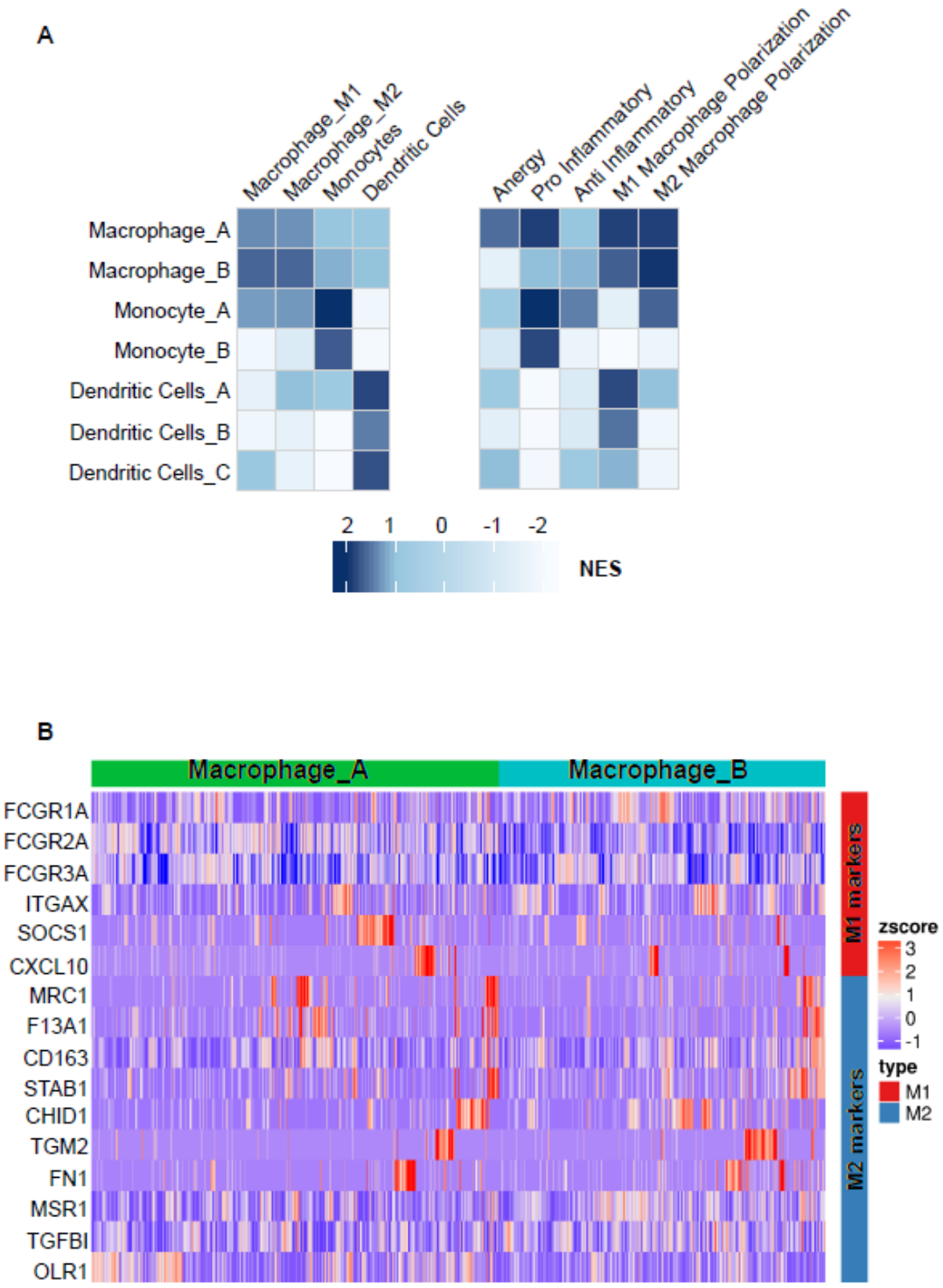

Figure S6. Macrophages in the tumor microenvironment. A) Enrichment of myeloid lineage signatures- LM22 (left) and Azizi et al. (1) (right) in the seven major myeloid populations identified in this study. **B)** Heatmap shows the known M1/M2 marker gene expression in the macrophage-A and macrophage-B populations.

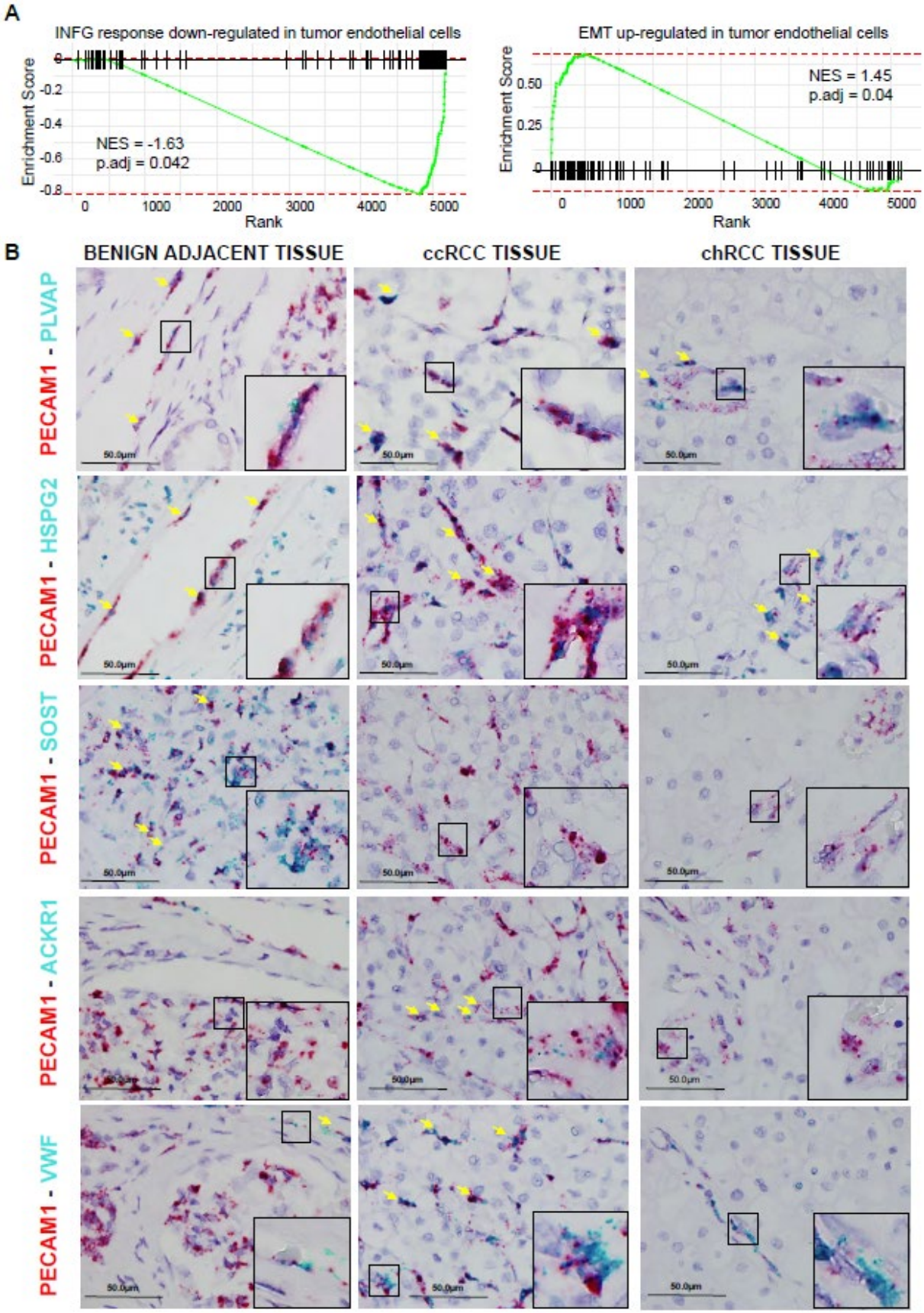

Figure S7. Endothelial cell types in ccRCC. A) Hallmark pathways enriched in comparison of endothelial cells between "Normal Endo cluster 16" (AVR-1) and "ccRCC tumor endothelial cluster 2 (AVR-1)". **B)** RNA-ISH dual staining of indicated markers with respective colors employed. A representative stain on benign adjacent tissue (left), ccRCC (middle), and chRCC tumors (right) is shown. Pan-endothelial marker *PECAM1* (red probe) and various endothelial cell type-specific maker genes, including *HSPG2*, *PLVAP*, *SOST*, *ACKR1* and *VWF* (blue probes), are shown.

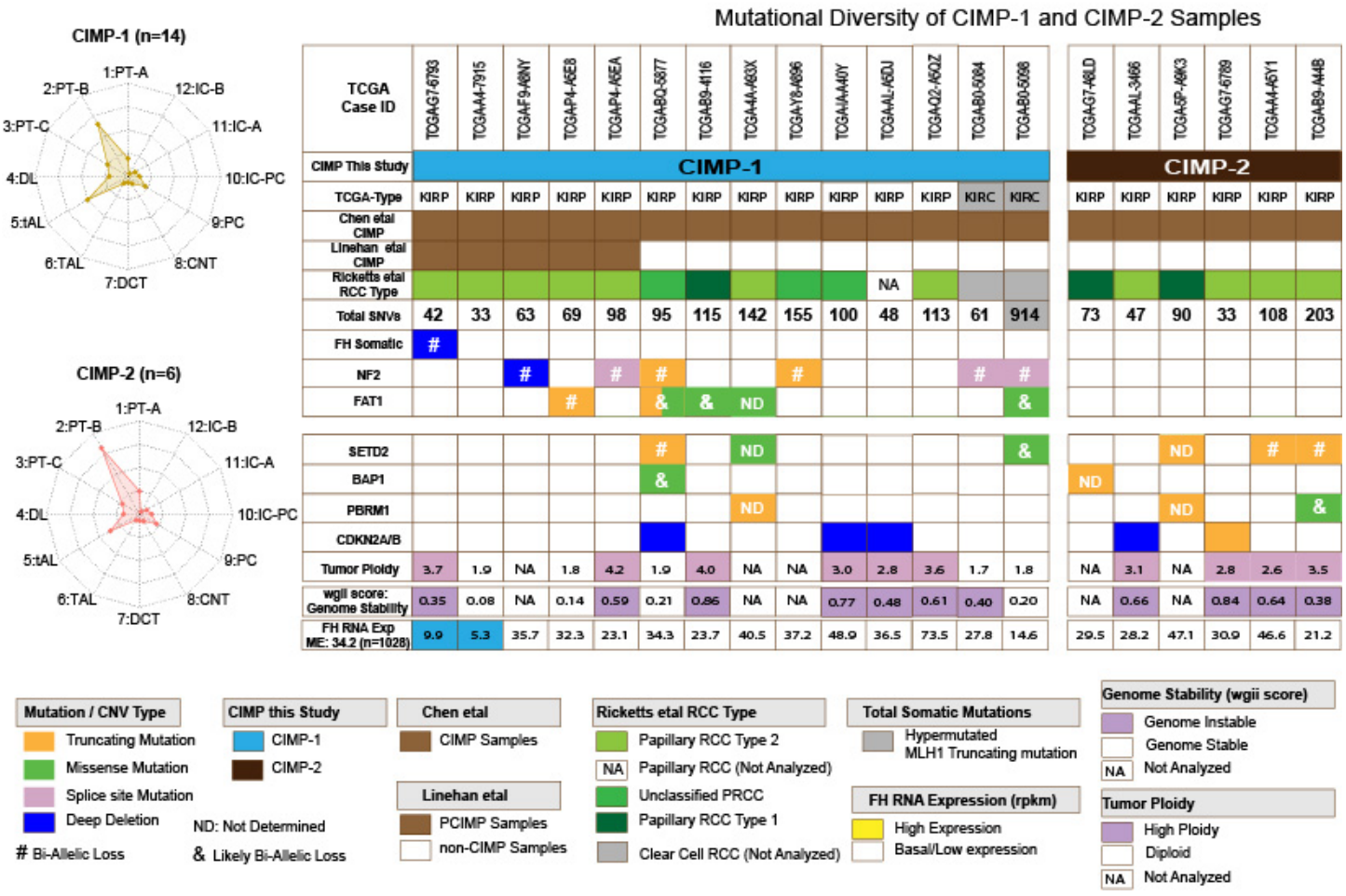

Figure S8. Schematic representation of the genomic aberrations in the two molecular subtypes of CpG island methylator phenotype (CIMP) RCC tumors (CIMP-1 and CIMP-2) identified in the current study. Left panel: Cell of origin as presented in **Figure 2C** for CIMP tumors. Right panel: The schematic summarizes the recurrent genomic aberrations found in CIMP-1 and CIMP-2 tumors. Genes such as *NF2* and *FAT1* were frequently mutated in CIMP-1 compared to CIMP-2. Rows in the following order represent 1) TCGA sample identification numbers; 2) CIMP annotation from the single cell RNA sequencing in this study (blue: CIMP-1; brown: CIMP-2); 3) TCGA RCC cohort abbreviation, KIRC- clear cell RCC, KIRP- papillary RCC; 4) CIMP annotation from Chen et al. (2); 5) CIMP annotation by Linehan et al. (3) (brown- CIMP; white- non-CIMP); 6) Revised RCC histology by Ricketts et al. (4) (green (lite): papillary RCC (pRCC) type-2; N/A: pRCC not-analyzed; green: unclassified pRCC; green (dark)- pRCC type-1; grey- ccRCC not-analyzed); 7) total number of SNVs for each case available from GDC portal; 8-14) aberrations in *FH*, *NF2*, *FAT1*, *SETD2*, *BAP1*, *PBRM1*, and *CDKN2A/B* genes (orange- truncating mutations, green- missense mutations, pink- splice site mutations, blue- homozygous deletion / deep deletion, #- biallelic loss, &- likely biallelic loss, ND- not determined); 15) tumor ploidy estimates; 16) wgii score as a measure of genome stability/ copy-number burden; 17) *FH* mRNA expression levels; ME- median *FH* mRNA expression across 1028 samples in the TCGA RCC cohort.

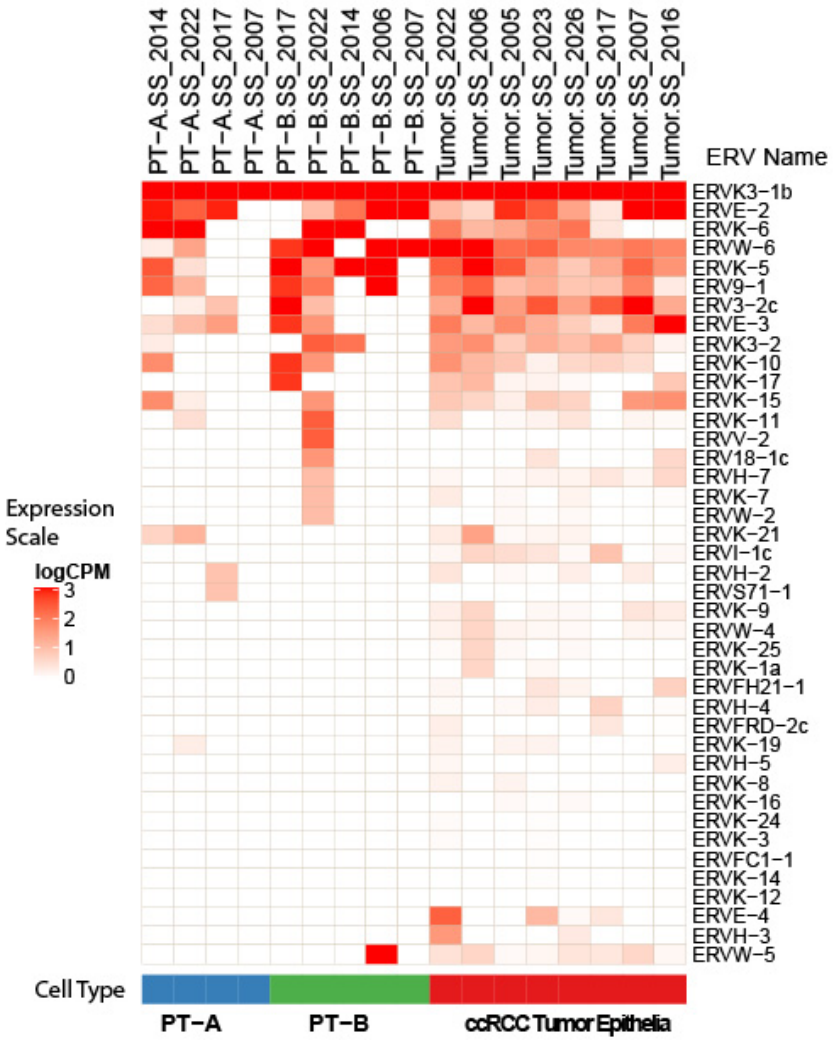

Figure S9. Expression pattern of endogenous retroviruses (ERVs) in kidney benign (PT-A and PT-B) and tumor epithelial cell types. Heatmap shows expression pattern of 41 ERVs detected in the benign kidney and ccRCC scRNA-seq data. Log CPM values for each ERV are plotted (red-high to white-low) for the three different cell types- PT-A (blue), PT-B (green), and ccRCC tumor epithelia (red).

SUPPLEMENTARY METHODS

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD31 (JC70) Mouse Monoclonal Antibody	Ventana Medical Systems / Roche Diagnostics	760-4378
Biological Samples		
Tumor/normal tissues from renal cancer patients	University of Michigan	See **Dataset S1**
Chemicals and Other Reagents		
Collagenase Type II	Thermo Scientific	17101-015
DNAse	Sigma-Aldrich	10104159001
Cell Strainers 40 μm	Flowmi	BAH136800040
Cell Strainers 70 μm	Flowmi	BAH136800070
Critical Commercial Assays		
Chromium Single Cell 3' Library & Gel Bead Kit v2	10X Genomics	120267
Chromium Single Cell A Chip Kit	10X Genomics	1000009
Chromium Multiplex Kit	10X Genomics	120262
SPRIselect	Beckman Coulter	B23318
All Prep DNA/RNA/miRNA Universal Kit	Qiagen	80224
KAPA Hyper Prep Kit for Illumina	Kapa Biosystems	KK8504
SureSelect XT Human All Exon V4 library	Agilent Technologies	5190–4632
SureSelectXT Reagent kit	Agilent Technologies	G9611B

RNA 6000 Nano kit	Agilent Technologies	5067–1511
DNA 1000 kit	Agilent Technologies	5067–1504
QIAGEN Multiplex PCR Kit	Qiagen	206143

ISH Probes	**Cat. no. (ACD)**	**Start/End position**
ITGB8	506871	72-1015
PDZK1IP1	573511-C2	4-763
ALPK2	416181	6057 - 6989
CREB5	477351	4558 - 5572
CA9	559341-C2	326 - 1528
CALB1	422161	250-1589
FOXI1	476351-C2	1057 - 1973
PECAM1	455931-C2	220 - 1269
PLVAP	437461	647-2039
HSPG2	573501	141-1139
SOST	452941	2-1269

Deposited Data

Database	**Accession Number**	**Data type**
GEO	GSE159115	Single cell gene expression count matrix

Software and Algorithms

Cell Ranger v2.1	10X Genomics	https://www.10xgenomics.com/

R v3.5	The Comprehensive R Archive Network	https://cran.r-project.org/
Scrublet	GitHub	https://github.com/AllonKleinLab/scrublet
Scran	Bioconductor	https://bioconductor.org/packages/release/bioc/html/scran.html
DBSCAN	GitHub	https://github.com/mhahsler/dbscan
PhenoGraph	GitHub	https://github.com/JinmiaoChenLab/Rphenograph
Limma	Bioconductor	http://bioconductor.org/packages/release/bioc/html/limma.html
Slingshot	GitHub	https://github.com/kstreet13/slingshot
randomForest	R Project	https://cran.r-project.org/web/packages/randomForest/index.html
EPIC	GitHub	https://github.com/GfellerLab/EPIC
UMAP	GitHub	https://github.com/jlmelville/uwot
tSNE	GitHub	https://github.com/jkrijthe/Rtsne
ComplexHeatmap	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html
survminer	GitHub	https://github.com/kassambara/survminer
fgsea	Bioconductor	https://bioconductor.org/packages/release/bioc/html/fgsea.html

RNA *in situ* hybridization (RNA-ISH)

Dual color RNA-ISH was performed on 4-micron formalin-fixed paraffin-embedded (FFPE) tissue sections using the RNAscope 2.5 duplex detection kit (Advanced Cell Diagnostics). A full list of target probe pairs is summarized in the **Key Resources Table** above. Sample RNA quality was evaluated using positive control probe pairs against the human *PPIB* gene in channel 1 (green) and *POLR2A* gene in channel 2 (red). Assay background was monitored by negative control probes against the bacterial *DapB* gene in both channels. After deparaffinization and target retrieval, tissue sections were permeabilized using protease and hybridized with a mixture of two target probes for 2 hours at 40°C, followed by a series of signal amplifications. Chromogenic detection for channel 1 and channel 2 utilized HRP-based Green and AP-based Fast Red chromogens respectively, followed by counterstaining with 50% Gill's Hematoxylin I (Fisher Scientific). Stained slides were examined under 100x and 200x magnification for RNA-ISH signals in tumor cells and adjacent benign kidneys.

Immunohistochemistry (IHC)

IHC studies for immune profiling were performed on 4-µm-thick FFPE sections. Slides were

stained on the Ventana Medical Systems automatic staining platform for endothelial cell CD31 expression assessment using an anti-human mouse monoclonal CD31 antibody (clone JC70, for 32 min; CC1 antigen retrieval for 30 minutes, catalog number 760–4378, Ventana Medical Systems). In all experimental runs, appropriate positive and negative control samples were also included. Finally, antigen detection was performed using both UltraView and Optiview DAB IHC Detection Kit (Ventana Medical Systems) with diaminobenzidine as the chromogen to detect antigen expression. Tissue sections were counterstained with Mayer's hematoxylin. The stained slides were reviewed by two pathologists (R. Mannan and R. Mehra) for membranous immunoexpression. Five CD31 high power field (HPF) areas showing the most vascularized areas ("hot spots") were analyzed for endothelial expression, and blood vessels were counted manually.

Whole-exome sequencing data analysis

The sequence files from whole-exome libraries were processed through an in-house certified mutation and copy-number calling pipeline to carry out the analysis from matched tumor/normal pairs as described previously (5). The sequencing files were aligned to the GRCh37 reference genome built using Novoalign, and bam files were sorted and indexed as described in (5, 6). SNV analysis was performed using freebayes software (version 1.0.1), and for indel analysis, pindel (version 0.2.5b9) was used. The mutations were called as somatic if they were present with at least six variant reads and 5% allelic fraction in the tumor sample, and present at no more than 2% allelic fraction in the normal sample with at least 20X coverage; additionally, the ratio of variant allelic fractions between tumor and normal samples was required to be at least six in order to avoid sequencing and alignment artifacts at low allelic fractions. Germline variants were called using ten variant reads and 20% allelic fraction as minimum thresholds and were classified as rare if they had less than 1% observed population frequency in both the 1000 Genomes and ExAC databases. Exome data was analyzed for copy-number aberrations and loss of heterozygosity by jointly segmenting B-allele frequencies and log2-transformed tumor/normal coverage ratios across targeted regions using the DNAcopy (version 1.48.0) implementation of the Circular Binary Segmentation algorithm previously described (5).

To perform copy-number analysis of TCGA CIMP RCC tumors (**Figure S8**), we downloaded the available aligned BAM files for these samples, removed duplicate reads, and ran the germline somatic caller DNAscope (7). Files were also utilized to run CNVEX for copy-number analysis, estimate tumor ploidy, and to calculate Weighted Genomic Instability Index (wGII). Briefly, copy-number analysis was performed for each patient using DNA whole-exome sequencing data of the tumor and matched normal. Germline DNA variants were assessed by Sentieon Genomics Tools-DNAscope. We then input these data into a copy-number analysis tool called CNVEX (8). We next used CNVEX, a genomic tool leveraging aligned read files (BAMs) and germline variant calls (VCF), to calculate GC-adjusted log2 coverage ratios and B-allele frequencies (BAF) to perform accurate genome segmentation. After segmentation, CNVEX performs an optimization step to identify the tumor ploidy, tumor purity, and tumor absolute minor and major allele copy-number profile. Finally, using the inferred copy-number profiles, CNVEX annotates the segments as "copy-number gain", "copy-number loss", or "loss of heterozygosity (LOH)". To estimate the chromosomal instability of each sample, we used a modified version of Genome Instability Index (GII) (9). We calculated GII scores as the portion of autosome that has an absolute copy-number unequal to the weighted median absolute copy-number across the autosomal chromosomes. To account for the variation in chromosome size and avoid the overrepresentation of larger chromosomes in the CIN estimation, we used a modified version of GII called weighted Genome Instability Index (wGII) (10). To generate wGII, we first calculated the GII for each autosomal chromosome, then took the mean of all the GII scores for all 22 chromosomes.

Single cell 3'mRNA sequencing data analysis

Read alignment and quantification were conducted with Cell Ranger (2.1.1) and the pre-built reference genome (GRCh38). Cells with high mitochondrial content (25% for tumor libraries and 80% for normal libraries) and low number of genes detected (<300) were considered as low quality cells and discarded. Potential doublets identified by scrublet (11) were also removed from further analyses. Low expressed genes (detected in less than five cells) and a list of uninformative genes were removed including mitochondrial, ribosomal, and sex genes before the total number of UMI was standardized to 5,000 per cell and log-transformed [log2(X+1)]. When different samples were pooled, highly variable genes (HVGs) were identified and batch correction with fastMNN (12) was applied based on HVGs to remove batch effect before clustering. Cells were projected into a 2-D map with t-distributed stochastic neighbor embedding (t-SNE) (13) and UMAP (14) for visualization, and DBSCAN (15) was applied on t-SNE dimensions to assign cells into clusters. For pooled libraries of normal kidney tissues, clusters with obvious substructures were further divided into smaller clusters by running DBSCAN on those clusters separately. Markers of each cell cluster were identified with findMarkers function from scran (16). Batch-corrected data was only used for dimension reduction; marker identification was conducted using log-transformed UMI and with sample as block factor. Known markers for different renal cell types (**Dataset S2**) were used to annotate identified cell clusters. Gene set enrichment analysis (GSEA) was based on a gene list ranked by logFC and conducted with R package fgsea (17). Trajectory inference tool slingshot (18) was applied to cell populations from proximal tubule (PT-A, PT-B and PT-C) and from collecting duct (IC, PC, and IC-PC) separately following dimension reduction with diffusion map (19). Differential analysis between cell populations in tumor tissues and benign adjacent were based on summed UMI counts of all cells of the same cluster from the same patient (pseudo-bulk) (20) and implemented with limma package (21) following voom transformation (22), which is a well-established procedure for differential expression analysis of bulk RNA-seq. To analyze myeloid populations separately, all cells annotated as the myeloid lineage from both normal and tumor cell atlases were pooled, and batch correction was applied with fastMNN before dimension reduction with tSNE and clustering with Phenograph (23). Analyses of endothelial cells followed the same procedure.

Putative cell of origin (P-CO) for RCCs

A random forest model was trained with single cell data of normal kidney epithelial cells (only HVGs were used; 200 cells were randomly selected for overrepresented clusters like the proximal tubule (PT) cluster to minimize bias due to an unbalanced sample size). The model was then applied to single cell data of RCC tumor cells (seven ccRCC and one chRCC cases) to predict their closest normal cell type (P-CO). We also made use of bulk RNA-seq data from The Cancer Genome Atlas (TCGA) and an in-house collection of several rare RCC subtypes to predict P-COs of each RCC subtype. Because of the inherent differences in data structure of the two data types (single cell and bulk), we first applied rank-based inverse normal transformation (24) on both data sets to force them to have the same data distribution. A random forest classifier (25) was then applied to transformed single cell data of normal kidney epithelial cells, and transformed bulk data of RCCs were used to predict P-COs.

Nivolumab-treated RCC RNA-seq dataset description

Complete results from this project are being prepared for an independent submission. Only bulk RNA-seq data from this project were used in this study for integrative analysis. Bulk RNA-seq data were obtained from pretreatment tumor samples of 27 ccRCC patients (stage IV). These patients later received tyrosine kinase inhibitors followed by nivolumab treatment (**Figure 6B**). Nivolumab response/disease progression was evaluated and recorded following treatment. Patients with progressive and stable diseases were grouped into "no clinical benefit" (NCB) and patients with partial/complete response were grouped into "clinical benefit" (CB) category.

Differential expression analysis was performed between CB and NCB groups using the limma-voom procedure (21, 22). An external dataset that included 16 bulk RNA-seq samples from a similar cohort of metastatic ccRCC reported by Miao et. al. (26) was analyzed, where we noted similar association between immunotherapy response and cell type fraction.

Copy-number variation (CNV) estimation from scRNA-seq

CNV of tumor cells was estimated following previously published methods (27) with modifications. Pooled PT cells and IC-A cells were used as common references for ccRCC and chRCC, respectively. To reduce noise from low expression genes, only genes detected in >25% cells were kept for this analysis. Expression of tumor cells relative to the average of reference cells was calculated and then smoothed with a sliding window of 100 genes within each chromosome arm.

Ligand-receptor analysis of ccRCC

Ligand-receptor pairs were obtained from a curated database (28). For each gene, the mean expression of each cell cluster was calculated and rescaled to the range of 0 to 1 by dividing by the maximum mean expression across cell clusters. For each ligand-receptor pair, the interaction score between two cell types was defined as the product of mean expression values (scaled) of the ligand in one cell type and the receptor in the other cell type. To define the significance of the interaction score, we randomly selected 1000 pairs of genes and calculated their interaction scores; the 95% percentile of the scores was then used as the cutoff. We identified significant ligand-receptor pairs between tumor cells and other coexisting cell types for each tumor sample individually; only pairs appearing in at least two samples were retained. In addition, we filtered out pairs with low correlation (Pearson correlation coefficient <0.2) in TCGA kidney renal cell carcinoma (KIRC) bulk data. The rationale was that tumors with higher numbers of “sender” cells may also contain a larger number of “receiver” cells, which would be indicated by good correlation of the abundance of transcripts of ligands and receptors (29).

Deconvolution of TCGA bulk RNA-seq

Bulk RNA-seq data of 505 KIRC tumor samples was obtained from TCGA. Transcripts per million (TPM) without log-transformation was used for this task. The framework provided by EPIC (30) was used to deconvolute the bulk samples, with a custom deconvolution reference built from our pooled ccRCC scRNA-seq data. To build the reference expression profiles, raw unique molecular identifier (UMI) counts of each cell type from the same sample were summed and normalized to 10^6 total UMI; for each cell type, the average of normalized UMI counts across samples was used as the reference expression profile of that cell type, and the standard deviation of normalized UMI counts across samples was used to represent gene variability of that cell type. Signature genes for each cell type used in the reference were chosen based on markers identified by findMarkers (16) and manually inspected for specificity. Cell types included in the reference were tumor cells, endothelial cells, macrophages, T cells, CD8 T cells, B cells, mast cells, and plasma cells. Pericytes were not included in the reference because pericytes share signatures with both endothelial cells and vSMC, which makes it difficult to deconvolute correctly. The final gene list is available in **Dataset S2**. Cases with the highest endothelial fractions (>90% percentiles) were selected as endothelial outliers, and cases with highest CD8 T-cell fractions (>90% percentiles) were selected as CD8 T-cell outliers for survival analysis using survminer (31). Survival data were downloaded from UCSC Xena browser (32).

Endogenous retrovirus (ERV) expression

Sequences of 66 known transcribed ERV species were downloaded from GeneBank, and each species was treated as one additional chromosome and added to the human genome reference provided by CellRanger. In addition, the human genome reference was masked at ERV proviral loci based on a curated list to avoid multi-mapping. Multiple sequences of the same species were

considered as different exons of the same gene in the gene annotation file so that all reads mapped to the same species were counted for that species. The custom CellRanger reference was then built using the ERV appended fasta and gtf files with the mkref function. Finally, all genes including added ERV species were quantified with the cellranger count function. Cell type-specific ERV expression was evaluated in a "pseudo-bulk" way, which means for each ERV species, UMI counts from all cells of the same cell type were summed and scaled to count per million to normalize by library size. In the heatmap of **Figure S9**, we show normalized UMI counts (log Count Per Million) of each ERV species in PT-A/PT-B/tumor cells, respectively, from different scRNA-seq sample libraries.

SUPPLEMENTARY INFORMATION REFERENCES

1. Azizi E*, et al.* (2018) Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* 174(5):1293-1308 e1236.
2. Chen F*, et al.* (2016) Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Rep* 14(10):2476-2489.
3. Cancer Genome Atlas Research Network*, et al.* (2016) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* 374(2):135-145.
4. Ricketts CJ*, et al.* (2018) The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep* 23(1):313-326 e315.
5. Wu YM*, et al.* (2018) Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. *Cell* 173(7):1770-1782.e1714.
6. Robinson DR*, et al.* (2017) Integrative clinical genomics of metastatic cancer. *Nature* 548(7667):297-303.
7. Donald Freed RA, Jessica A. Weber, Jeremy S. Edwards (2017) The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *BioRxiv*.
8. Clark DJ*, et al.* (2019) Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* 179(4):964-983 e931.
9. Chin SF*, et al.* (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* 8(10):R215.
10. Burrell RA*, et al.* (2013) Replication stress links structural and numerical cancer chromosomal instability. *Nature* 494(7438):492-496.
11. Wolock SL, Lopez R, & Klein AM (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* 8(4):281-291 e289.
12. Haghverdi L, Lun ATL, Morgan MD, & Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36(5):421-427.
13. Maaten Lvd, and Geoffrey Hinton (2008) Visualizing Data Using T-SNE. *Journal of Machine Learning Research* 9:2579-2605.
14. Leland McInnes JH, Nathaniel Saul, and Lukas Großberger (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3(29):861.
15. Hahsler M, Matthew Piekenbrock, and Derek Doran (2019) Dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software* 91(1):1-30.
16. Lun AT, Bach K, & Marioni JC (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 17:75.
17. Sergushichev A (2016) An Algorithm for Fast Preranked Gene Set Enrichment Analysis Using Cumulative Statistic Calculation. *bioRxiv*.
18. Street K*, et al.* (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19(1):477.

19. Angerer P*, et al.* (2016) destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32(8):1241-1243.
20. Crowell HL, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D. Robinson. (2020) On the Discovery of Subpopulation-Specific State Transitions from Multi-Sample Multi-Condition Single-Cell RNA Sequencing Data. *bioRxiv*.
21. Ritchie ME*, et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.
22. Law CW, Chen Y, Shi W, & Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15(2):R29.
23. Levine JH*, et al.* (2015) Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162(1):184-197.
24. Aulchenko YS, Stephan Ripke, Aaron Isaacs, and Cornelia M. van Duijn (2007) GenABEL: An R Library for Genome-Wide Association Analysis. *Bioinformatics* 23(10):1294-1296.
25. Liaw A, Matthew Wiener (2002) Classification and Regression by randomForest. *R News* 2(3):18-22.
26. Miao D*, et al.* (2018) Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* 359(6377):801-806.
27. Puram SV*, et al.* (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 171(7):1611-1624 e1624.
28. Ramilowski JA*, et al.* (2015) A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* 6:7866.
29. Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, & Huang S (2017) Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes. *Sci Rep* 7(1):8815.
30. Racle J, de Jonge K, Baumgaertner P, Speiser DE, & Gfeller D (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* 6.
31. Kassambara AK, M. (2018) Survminer: Drawing Survival Curves Using 'ggplot2'.
32. Goldman MJ*, et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 38(6):675-678.