

Single-cell multi-omic and spatial profiling of human kidneys implicates the fibrotic microenvironment in kidney disease progression

In the format provided by the
authors and unedited

Supplementary Note, Table of Contents

Supplementary Note, Bioinformatic Analysis.....	2
Supplementary Note, Supplementary Methods	10
Supplementary Figure 1.....	11
Supplementary Figure 2.....	13
Supplementary Figure 3.....	15
Supplementary Figure 4.....	17
Supplementary Figure 5.....	19
Supplementary Figure 6.....	21
Supplementary Figure 7.....	23
Supplementary Figure 8.....	25
Supplementary Figure 9.....	27
Supplementary Figure 10.....	29
Supplementary Figure 11.....	31
Supplementary Figure 12.....	33
Supplementary Figure 13.....	35
Supplementary Figure 14.....	37
Supplementary Figure 15.....	39
Supplementary Figure 16.....	41
Supplementary Figure 17.....	43
Supplementary Figure 18.....	45
Supplementary Figure 19.....	47
Supplementary Figure 20.....	49
Supplementary Figure 21.....	51
Supplementary Figure 22.....	53
Supplementary Figure 23.....	55
Supplementary Figure 24.....	57
Supplementary Figure 25.....	59
Supplementary Figure 26.....	61
Supplementary Figure 27.....	63

Bioinformatic analysis

Primary single nuclei and cell RNAseq data processing

FASTQ files from each 10X single nuclei run were processed using Cell Ranger v6.0.1 (10X Genomics). Gene expression matrices for each cell were produced using the human genome reference GRCh38.

Data Processing and Computational Analyses

Ambient RNA was corrected using SoupX⁵⁸ (v1.5.0) and doublets were removed by DoubletFinder⁵⁹ (v2.0) with default parameters. Seurat objects were created from the aligned outputs of multiple samples, retaining genes expressed in more than three cells and cells with at least 300 genes. Further, a merged Seurat object was obtained using the merge function in Seurat v4.0.3. The QC filters applied were: (a) cells with n_feature counts more than 3000 or less than 200, (b) cells with more than 15% mitochondrial counts (for snRNAseq data) and more than 50% mitochondrial counts (for scRNAseq data), and (c) cells with more than 15% ribosomal gene counts.

Data Normalization and Cell Population Identification

Initially, highly variable genes were identified using the vst method. The data was then natural log-transformed and scaled. The scaled values underwent principle component analysis (PCA) for linear dimension reduction. We used the RunHarmony function in the harmony⁶⁰ (v1.0) package for batch effect correction. A shared nearest neighbor network was created based on Euclidean distances between cells in a multidimensional PC space (the first 50 PCs were used for both snRNAseq and scRNAseq). This network was used to generate a 2-dimensional Uniform Manifold Approximation and Projection (UMAP) for visualization.

To identify cell-type markers, we used the FindAllMarkers function in Seurat. This method calculates log fold changes, percentages of expression within and outside a group, and p-values of the Wilcoxon-Rank Sum test comparing a group to all cells outside that specific group, including adjustment for multiple testing. A log-fold-change threshold of 0.25 and FDR<0.05 was considered significant. These steps were performed on the snRNAseq and scRNAseq datasets

separately. Clusters expressing multiple cell types of specific marker genes were excluded as potential doublets. The dataset was further refined: first, clusters were annotated based on well-known marker genes, then the dataset was projected to a published human kidney single cell⁴³ reference using the CCA method. The FindTransferAnchors function was used to discover matching genes between the two datasets by using shared correlation patterns in the gene activity matrix and snRNAseq dataset. Predicted labels within the two datasets were identified using the TransferData method. Cells were considered doublets and eliminated from the dataset if the original annotation was incompatible with the predicted cell type.

Single nuclei ATACseq analysis

Raw FASTQ files were aligned to the GRCh38 reference genome and quantified using Cell Ranger ATAC (v. 1.1.0). Outputs from 19 snATACseq datasets were embedded using Signac (v.1.3.0)⁶¹ to generate Signac object. Low-quality cells were removed from each snATAC object using the following criteria: `peak_region_fragments < 3000 & peak_region_fragments > 20000 & pct_reads_in_peaks < 15 & nucleosome_signal > 4 & TSS.enrichment < 2`). The filtered cells in 19 objects were merged in Seurat. Dimension reduction was performed by singular value decomposition (SVD) of the TFIDF matrix and UMAP. Batch effect was corrected using Harmony⁶⁰ via the RunHarmony function in Seurat. A KNN graph was created to cluster cells using the Louvain algorithm. Peaks observed in at least 20% of cells were evaluated for differentially accessible chromatin regions (DARs) between different cell types using a likelihood ratio test, a log-fold-change threshold of 0.25, and an FDR of 0.05 with the Signac FindMarkers function. To annotate the genomic regions harboring snATACseq peaks, ChIPSeeker (v1.24.0) 70 was used⁶².

Removing the doublets from snATACseq dataset

The GeneActivity tool in Signac was used to create a gene activity matrix after clustering the twenty integrated snATACseq datasets using protein-coding genes annotated in the Ensembl database (this technique counts the ATAC peaks inside the gene body and 2 kb upstream of TSS). Log normalization was applied to the gene activity matrix.

A published snRNAseq⁴³ dataset was utilized as a reference, and the FindTransferAnchors function was used to discover matching genes between the snRNAseq and snATACseq datasets by using shared correlation patterns in the gene activity matrix and snRNAseq dataset. Subsequently, the predicted labels within two datasets were identified using the TransferData method. If the original annotation was incompatible with the predicted cell type, the cell was considered a doublet and was eliminated from the dataset.

Motif Enrichment Analysis and Motif Activities

First, we used the AddMotifs function of Signac to perform motif enrichment analysis. This was done by creating a matrix of positional weights for motif candidates from JASPAR2020. To identify potentially important cell-type-specific regulatory sequences and motifs, we evaluated the differentially accessible peaks between each cell type versus other cell types using the 'LR' test and min.pct = 0.05. Subsequently, a hypergeometric test was used to determine the likelihood of observing a specific motif at a particular frequency by chance. We compared this observation with a group of peaks that have the same GC content as the motif, which we considered the background set. This analysis was performed using the Find Motifs function. We used the RunChromVAR function in conjunction with chromVAR (v.1.6.0)⁶³ to ascertain transcription factor activity.

DARs between groups

We used the FindMarkers function after selecting DefaultAssay as peaks to identify DARs in each cell type and diseased and healthy conditions, with a log-fold-change threshold of 0.25 and FDR<0.05. Peaks translated to related genes using ChIPSeeker (v1.21.1)⁶².

Integration of snRNAseq, scRNAseq, and snATACseq datasets

We used SCVI^{29, 64} (v0.20.3) to integrate scRNAseq, snRNAseq, and snATACseq data from multiple samples. Briefly, we used the preprocessed and clean datasets previously described and then combined them into a single AnnData object using the Scanpy (v1.4.4 package). After normalization and identification of the 3000 highly variable genes in the integrated dataset, we

ran SCVI on the combined dataset to learn a joint latent representation of the data, using the sample, nFeature_RNA, nCount_RNA, and tech factors as covariates. We utilized the scanorama setting of SCVI to account for batch effects across samples. We incorporated four layers in the model and reduced the dimensionality of the data to 45 dimensions using the SCVI's dimensionality reduction method. Finally, we used the Scanpy package to perform downstream analysis on the integrated dataset, including clustering, differential expression analysis, and visualization.

Ligand–receptor interactions in integrated sc/snRNAseq and snATACseq

The CellChat⁴¹ repository (v1.4.0) was used to assess cellular interactions between different cell types and to infer cell–cell communication networks from integrated sc/snRNAseq and snATACseq data. Only receptors and ligands expressed in more than 10 cells in each cluster were considered. Probability and *P* values were measured for each interaction.

DEGs between diseased and healthy groups

To identify DEGs between groups, we utilized the FindMarkers tool for each cell type and condition, a log-fold-change threshold of 0.25, and an FDR 0.05. DEGs in the integrated dataset of snRNAseq, scRNAseq, and snATACseq were calculated using model.differential_expression of SCVI tool with the default parameters.

Single nuclei RNAseq trajectory analysis

Proximal tubules (PT) and Injured PT cells were subclustered for the trajectory analysis. The trajectory analysis was done in two steps. Different sub-types of iPT cells with equal numbers were randomly subsampled and cell dataset object (CDS) was generated using Monocle3^{65,66}(v1.3.1). After preprocessing and batch effects correction, the dataset was embedded for dimension reduction and pseudotemporal ordering. We used the order_cell function and indicated the PT as start point for pseudotime analysis. The track genes algorithm was used to identify the DEGs along the trajectories, and genes with *q* values of 0.05 or higher were considered significant.

MetaNeighbor analysis

MetaNeighbor³⁶ (v1.10) is a versatile computational tool designed to quantify the degree of cell type replicability between datasets by assessing cell type-specific gene expression patterns. The method involves constructing a cell-cell similarity network using a shared, reduced gene space and subsequently calculating the area under the receiver operating characteristic curve (AUROC) to estimate cell type-specific gene expression conservation. High AUROC values indicate a high degree of cell type replicability, whereas low values suggest limited conservation. We followed the original protocol and used default settings for parameter selection, including the 50 nearest neighbors for network construction and 1,000 most variable genes for gene space reduction.

Single nuclei WGCNA

We used the R packages hdWGCNA and WGCNA (v1.72)³¹. Cell2location is a Bayesian probabilistic model that integrates transcriptomic and spatial information to estimate cell type-specific gene expression patterns and their spatial distribution. We first preprocessed the spatial transcriptomics data by filtering out lowly expressed genes and normalizing the expression values. We then mapped the spatial locations of the cells onto a two-dimensional coordinate space using the cell coordinates provided by the dataset. We subsequently trained the Cell2location model on the preprocessed data using our integrated sc/snRNAseq and snATACseq data and prior information on cell type co-localization. This prior information, obtained from imaging data, was used to constrain the spatial distribution of each cell type.

Finally, we used the trained model to estimate cell type-specific gene expression patterns and their spatial distribution in the tissue. The resulting cell type-specific expression maps and spatial distributions were visualized on a spatial map of the tissue to identify the spatial organization of different cell types.

Deconvolution of Visium Dataset

To deconvolve the spots in our spatial transcriptomics dataset, we utilized the Cell2location method. We began by preprocessing the raw data using standard quality control and filtering steps to remove low-quality spots and genes with low expression. We then normalized the data using the total-count normalization method implemented in the Cell2location package. Next, we

trained the Cell2location model using our integrated sc/snRNAseq and snATACseq atlas and our spatial transcriptomics data. The single-cell RNA sequencing data were used to define the cell types present in the tissue and to estimate their gene expression signatures, while the spatial transcriptomics data were used to estimate the spatial distribution of the cell types within the tissue. We estimated the model parameters using the variational Bayesian inference algorithm implemented in the Cell2location package, running the algorithm for 5,000 iterations with a convergence threshold of 0.01 for the change in the lower bound on the marginal likelihood of the observed data.

Once the model parameters were estimated, we used the Cell2location package to predict the expression of each gene in each cell type at each spot in the tissue. This allowed us to identify the cell types present in each spot and visualize their spatial distribution within the tissue. The Cell2location method proved highly effective in deconvolving the spots in our spatial transcriptomics dataset, providing insights into the spatial organization of cells in the tissue.

Mapping sn/scRNAseq to spatial location

To map back the cell types identified in the dissociated data (sn/scRNAseq datasets), we used the Celltrek³² package. We first downsampled the integrated sc/snRNAseq and snATACseq datasets to 20,000 cells. Then, using the traint function, sn/scRNAseq datasets were co-embedded with spRNAseq datasets. Subsequently, using the random forest model, single cells were mapped to their spatial locations. This analysis was performed by merging snRNAseq and immune cell types to enrich the dataset for immune cells. Regarding colocalization, the sColoc function of CellTrek was used. To identify the different cell type modules in the spRNAseq, we performed a spatial-weighted gene co-expression analysis.

Ligand–Receptor Interactions in Spatial Transcriptomics Data

We used the CellChat v1.4.0 package to predict cell type-specific ligand–receptor interactions (1939 interactions) between cells in the spRNAseq data. Only receptors and ligands expressed in more than 10 cells in each cluster were considered. To compute the interactions, we used the `computeCommunProb` function with the following parameters: `distance.use = TRUE`, `interaction.length = 200`, and `scale.distance = 0.01`. Probability and P-values were measured for each interaction.

In addition, we utilized the COMMOT package (v0.03), which provides a spatially informed approach to account for cell type locations in our data. After preprocessing each spRNAseq data, we used the `ct.tl.spatial_communication` function with default parameters to calculate the communication scores between each pair of cell types based on the expression levels of ligand and receptor genes. We visualized the results using the `plotComm` functions to generate a network.

Finding microenvironments in spRNAseq

To identify microenvironments on the merged spatial dataset after SC transformation (SCT) we performed an NMF reduction using the STUtility package (v1.1)³⁹. The dataset was then clustered using the first 20 factors from the NMF reduction. To identify MEs specific markers, we used Seurat's `FindAllMarkers` function. Identified MEs were then annotated based on marker genes. To find correlations between the number of spots of each MEs and clinical characteristics, we extracted the number of spots for each MEs per sample and then performed a Pearson correlation analysis with clinical findings.

ECM production score

To calculate the extracellular matrix production (ECM) score, we calculated the proportion of the expressions of collagen, proteoglycan, and glycoprotein³⁵ genes in each cell types.

CosMx Data Analysis

CosMx data was exported after AtoMx preprocessing (v.1.1) using the export function and analyzed in python using the scanpy (v1.9.4) and squidpy (v.1.2.3) packages. Probe counts were normalized and log transformed. Cells with fewer than 30 reads were removed. Principal components were calculated which were used to determine 15 neighbors for each cell. These neighbors were used to create a UMAP and Leiden clusters (resolution 1.0). Annotation of leiden clusters was manually performed—the iPT/PEC cluster was subclustered and annotated to resolve these two populations. Annotation locations were manually inspected after annotation based on gene expression. Neighborhoods were calculated using the squidpy function, using a neighborhood size of 20 microns. The data was integrated with our single nuclei data using SCVI (v.1.0.3), using 4 layers and 30 latent variables.

Bulk RNAseq Analysis

The human bulk kidney RNAseq data were downloaded from GEO ([GSE115098](#) and [GSE173343](#)). **We used previously published pipeline for bulk RNAseq analysis** ⁶⁷. Briefly, the FASTQC was used to check the QC of the sequencing results. The adapters and low-quality bases were trimmed using TrimGalore (v0.4.5). The trimmed FASTQ files were aligned to the human genome (hg19/GRCh37) using STAR (v2.7.3a)^{68,69} based on GENCODE v19 annotations^{68,69}. Gene expression was quantified using RSEM by calculating uniquely mapped reads as transcripts per million (TPM). To ensure accurate microdissection, we projected the TPM matrix of the tubule and glomerular compartments onto a PCA plot and samples that did not separate clearly were removed.

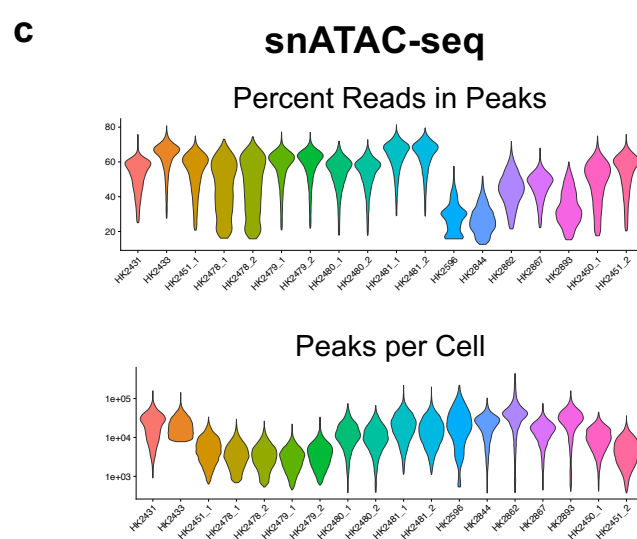
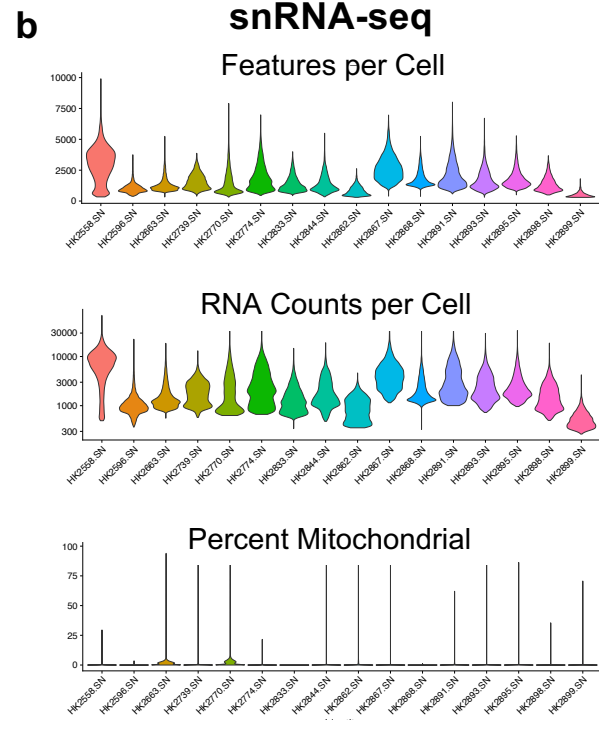
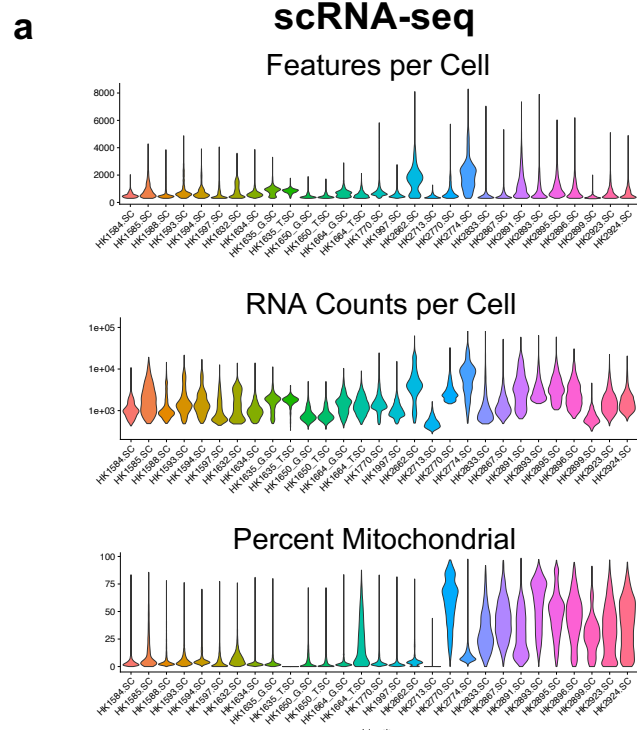
Hierarchical clustering of microdissected human kidney tubule samples based on FME-gene signature

Hierarchical clustering was performed on the scaled TPM matrix of microdissected human tubules datasets using the FME-GS list as the base. Ward's method with Euclidean distances was

used to cluster the datasets. The optimal number of clusters was determined by average silhouette method.

Methods Reference

65. Young, M.D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**(2020).
66. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4 (2019).
67. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).
68. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *BioRxiv* (2020).
69. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383 (2015).
70. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).
71. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).
72. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
73. Trapnell, C. *et al.* Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology* **32**, 381 (2014).
74. Ni, Z. *et al.* SpotClean adjusts for spot swapping in spatial transcriptomics data. *Nature Communications* **13**, 2971 (2022).
75. Sheng, X. *et al.* Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nature genetics* **53**, 1322-1333 (2021).
76. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
77. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774 (2012).



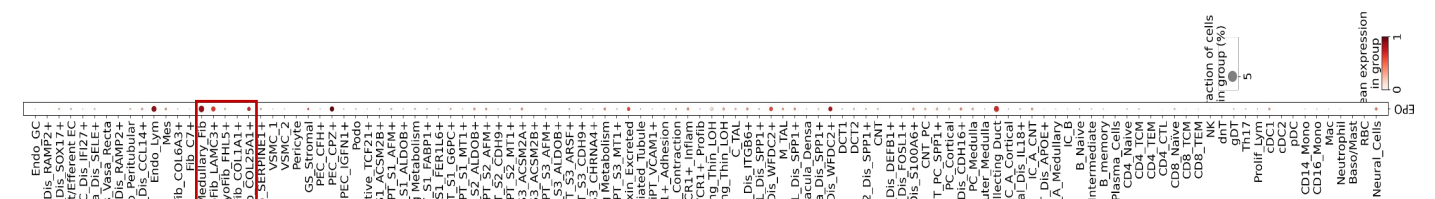
Supplementary Fig. 1

Supplementary Fig. 1. Quality metrics of the sn/scRNAseq, snATACseq and spatial transcriptomics data.

(a) QC parameters of the scRNAseq. The plots depict the distribution of nFeature_Count (number of genes per cell), nCount_RNA (number of reads per cell), and mitochondrial percentage (fraction of mitochondrial genes per cell) in each sample. X axis indicates the individual sample and y-axis shows the values.

(b) QC parameters of the snRNAseq. The plots depict the distribution of nFeature_Count (number of genes per cell), nCount_RNA (number of reads per cell), and mitochondrial percentage (fraction of mitochondrial genes per cell) in each sample. X axis indicates the individual sample and y-axis shows the values.

(c) QC parameters of the snATACseq. The plots depict the distribution of pct_reads in peaks, nFeature_ATAC per sample. X-axis indicates the individual sample and y-axis shows the values.



Supplementary Fig. 2

Supplementary Fig. 2. High resolution clustering of the integrated data.

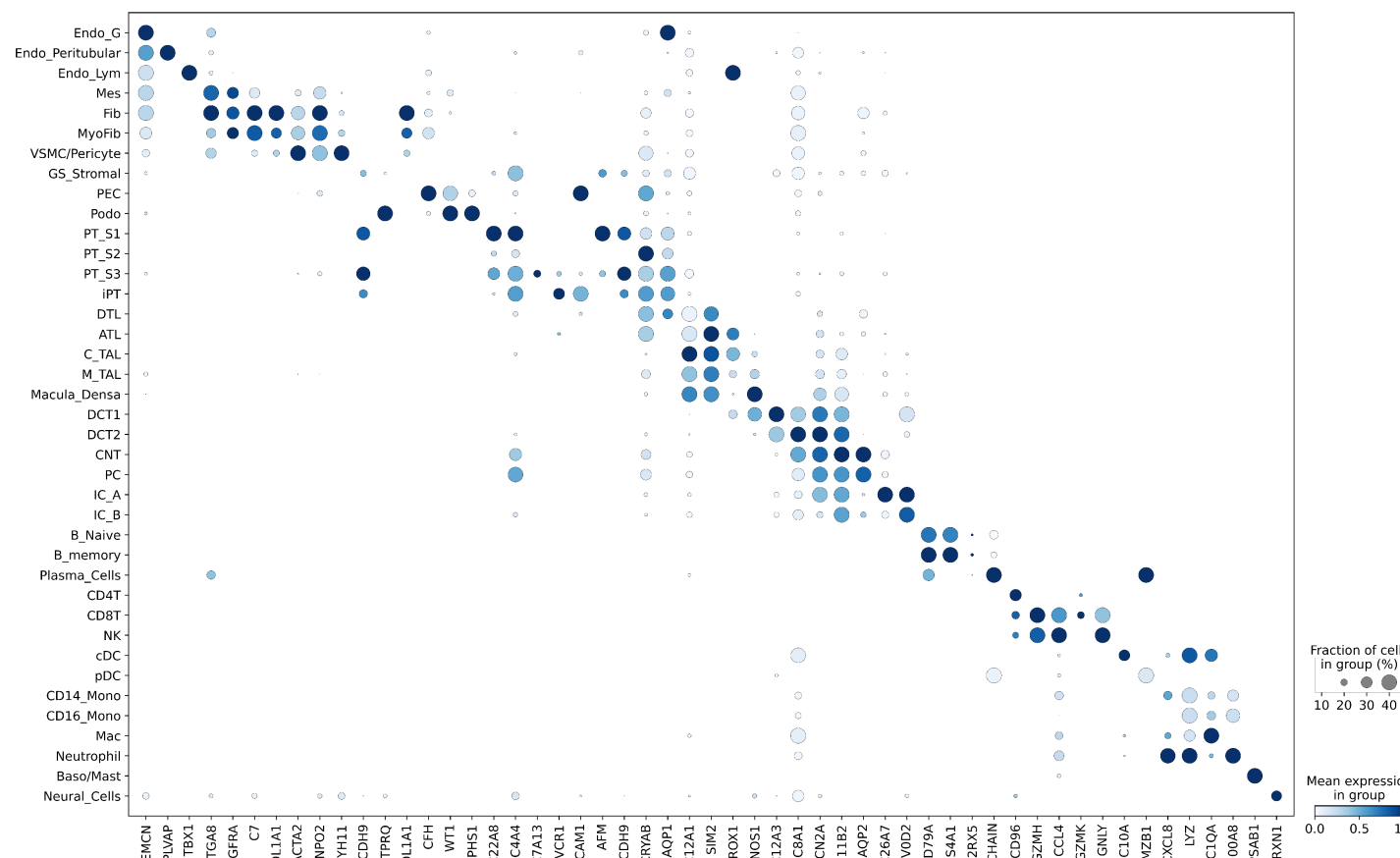
(a) UMAP of integrated snRNAseq, scRNAseq, and snATACseq datasets of 338,565 cells and nuclei and identification of 114 clusters. Right panel shows the original source data.

(b) The dot plots of marker genes used for the annotation of 144 fine clusters in the integrated dataset. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

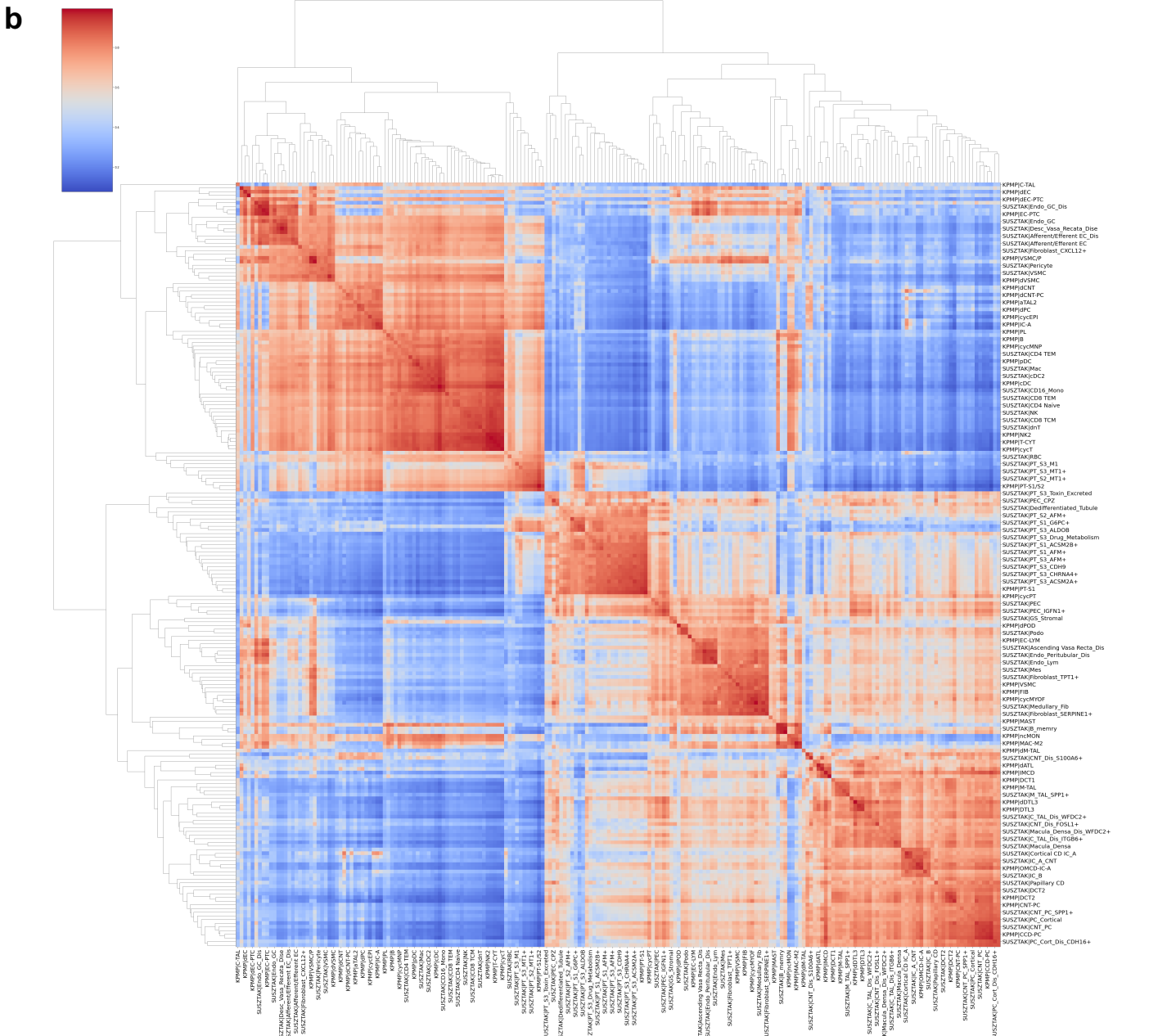
(c) The dot plot of erythropoietin (*EPO*) expression in the integrated snRNAseq, scRNAseq, and snATACseq datasets. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression. Stromal cells are the main source of *EPO* in the kidney.

Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels, Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naive; Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

a



b



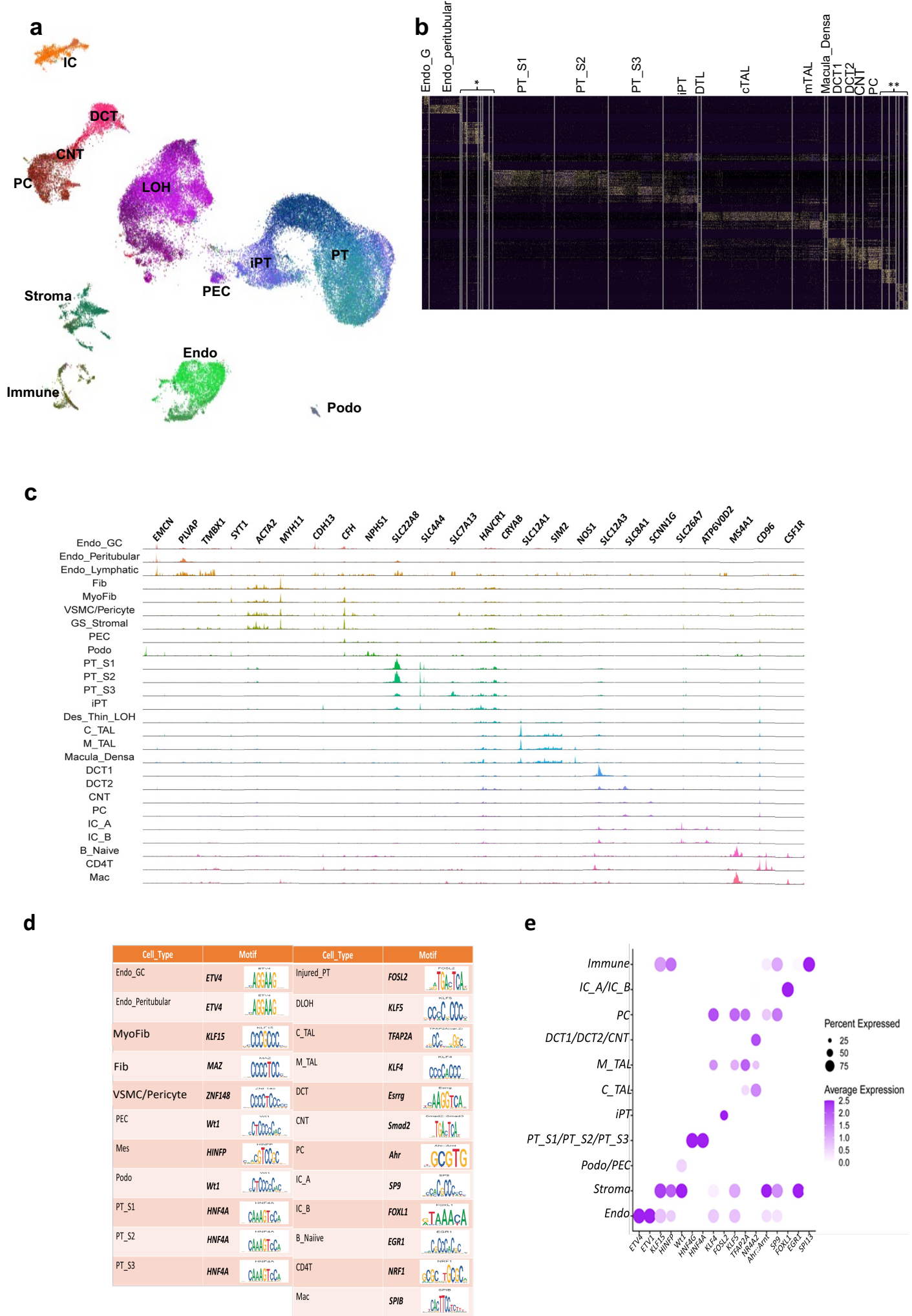
Supplementary Fig. 3

Supplementary Fig. 3. Integrations of snRNAseq, scRNAseq, and snATACseq datasets from multiple sources

(a) Dot plots of marker genes used for the annotation of 39 main cell types in the integrated dataset. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

(b) Cell type consistency of the different datasets using MetaNeighbor. The heatmap displays the Area Under the Receiver Operating Characteristic (AUROC) scores among different cell subtypes in two different data sources (this study vs KPMP). Dendrograms were created via hierarchical clustering, based on Euclidean distances, and used average linkage for this purpose. Each row and each column are a cell type in one of the source data.

Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels, Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific stromal cells, VSMC/Pericyte; vascular smooth muscle cells/pericyte, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, dnT; double negative T cells, Prolif_Lym; proliferative lymphocyte, Th17; T helper 17 lymphocyte, B_Naive; Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.



Supplementary Fig. 4

Supplementary Fig. 4. Human kidney single nuclear open chromatin (snATACseq) data.

(a) UMAP of 58,155 nuclei after QC filtering in snATACseq dataset. 26 clusters were identified.

(b) Heatmap of top 10 differentially accessible regions per cells in the snATACseq dataset. Rows indicate the top 10 differentially accessible regions and their degree of accessibility pre cluster and columns show the cell types. Populations indicated within the * region are Endo_lym, Fib, MyoFib, VSMC/Pericyte, GS_stromal, PEC, Podo (labeled left to right). Populations indicated in the ** region are IC_A, IC_B, B_naive, CD4T, and Mac, (labeled left to right).

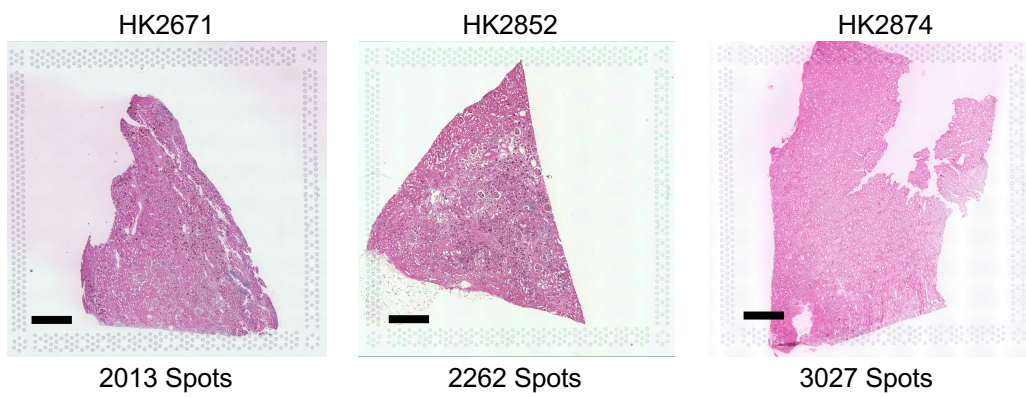
(c) Accessibility of marker genes across annotations. Reads at each locus are shown as a with a higher peak indicating more open chromatin.

(d) Enriched motif and their activity per cluster in snATACseq dataset. The table shows the enriched motif in each cell type.

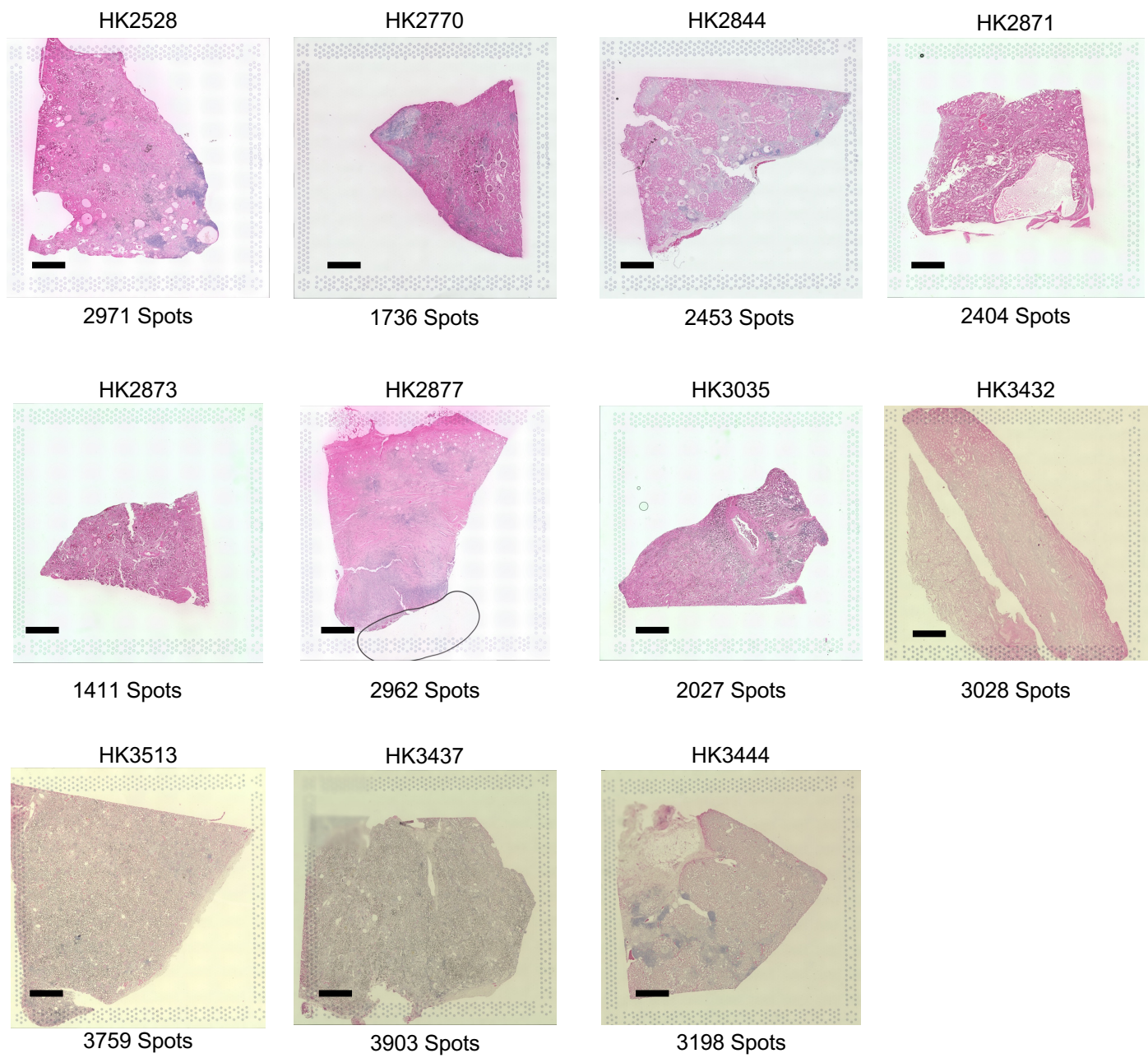
(e) The dot plot indicates the motifs activity in main clusters of snATACseq dataset. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average activity.

Endo_GC; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels, Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; meseangial cells, GS_Stromal; glomerulosclerosis-specific stromal cells, VSMC/Pericyte; vascular smooth muscle cells/pericyte, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, dnT; double negative T cells, Prolif_Lym; proliferative lymphocyte, Th17; T helper 17 lymphocyte, B_Naive; Naive B lymphocytes, Mac; macrophage

Control



Disease

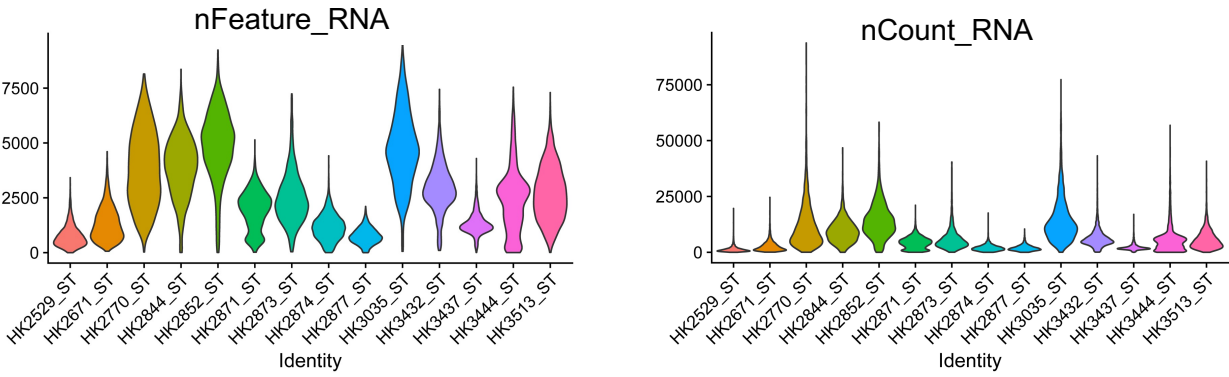


Supplementary Fig. 5. Histological images of the Visium data.

H&E imaging of the 3 control and 11 diseased samples. Each sample is H&E stained and the number of Visium spots with significant RNA quantity is indicated below. Scale bar is 1 mm.

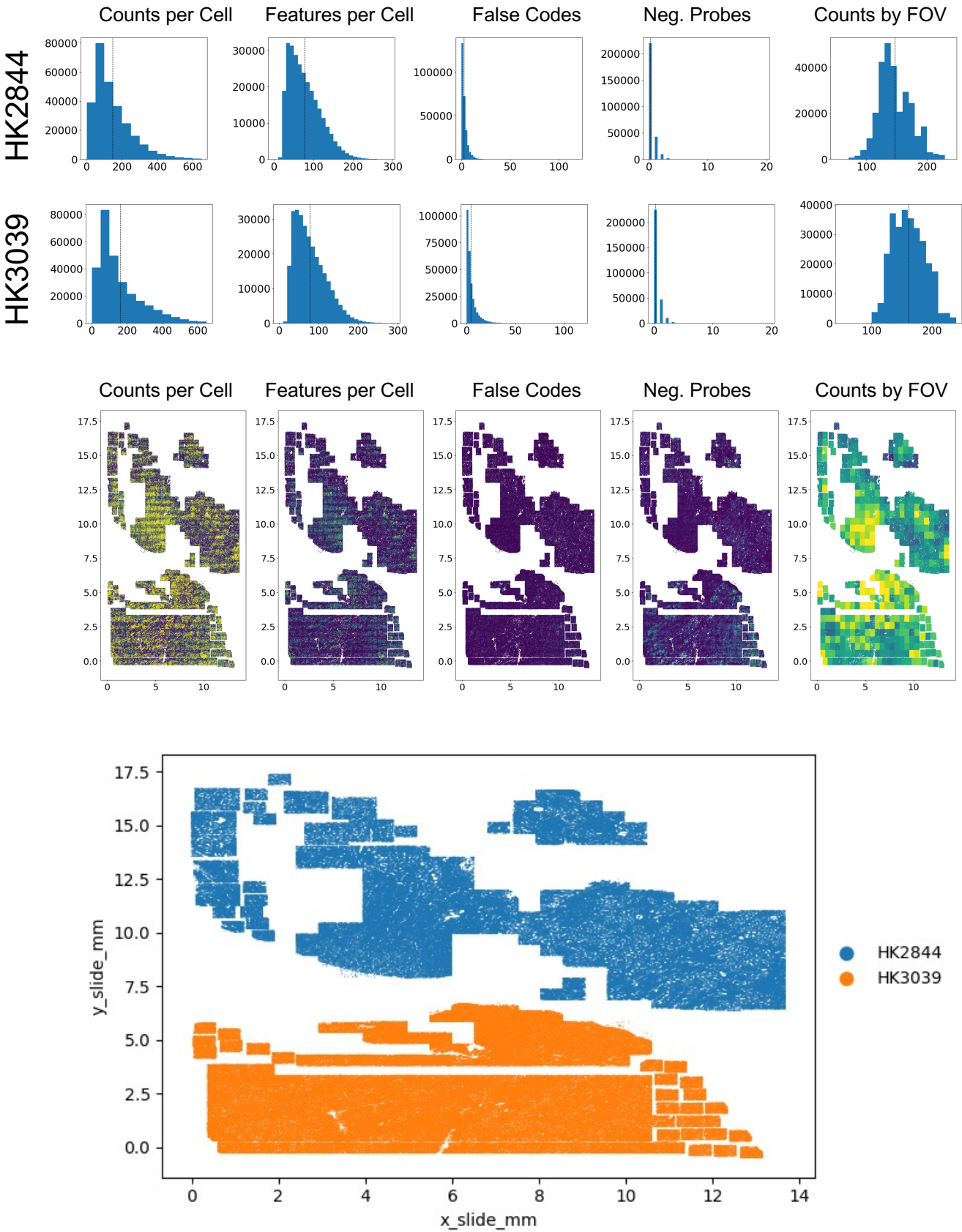
a

Visium



b

CosMx

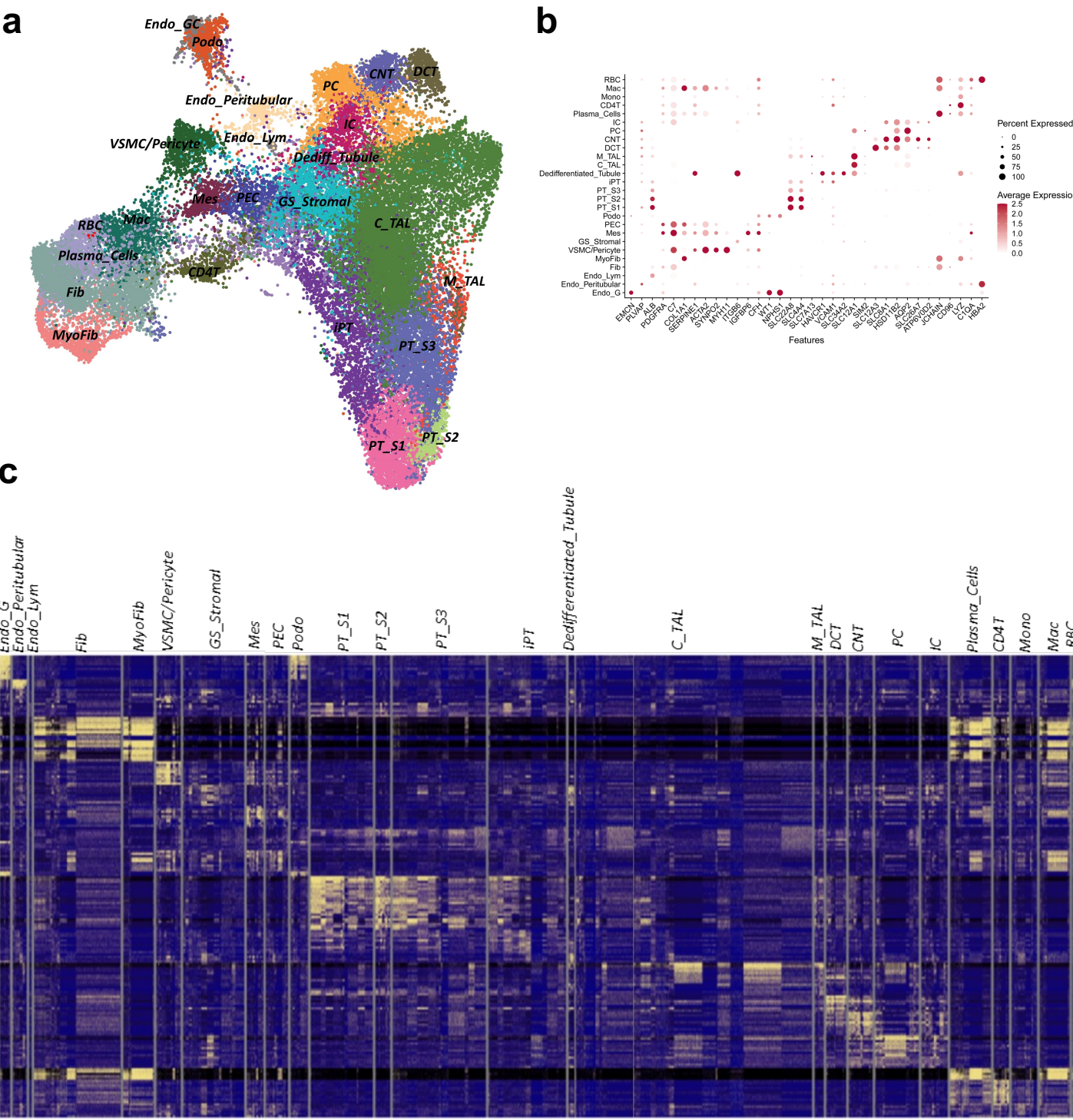


Supplementary Fig. 6

Supplementary Fig 6. Quality Control metrics of spatial transcriptomic data.

(a) QC parameters of the Visium spatial RNA assay. We show number of features and number of RNA counts for each individual sample.

(b) QC parameters for the CosMx analysis represented as histograms, which are parsed by sample. False codes indicates reads that do not correspond to a known gene barcode, and negative probe counts corresponds to reads to a probe that is designed not to anneal to a human RNA. Counts were aggregated by field of view (FOV) to inspect if a particular region of tissue performed poorly. These metrics are then shown in space on the slide. Lastly, we also show the location of each sample on the CosMx slide (below).



Supplementary Fig. 7

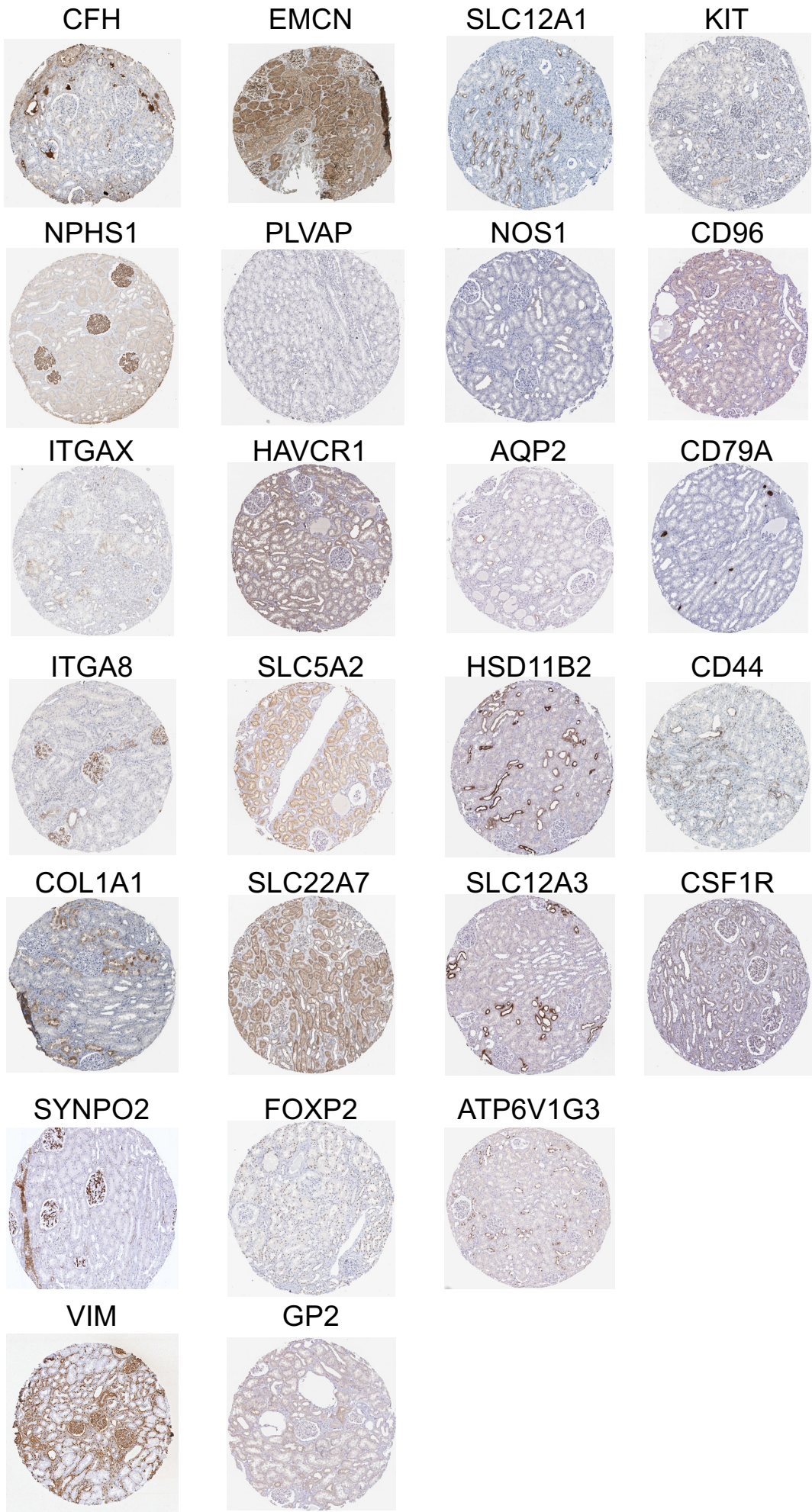
Supplementary Fig. 7. Human kidney Visium spatial transcriptomics data

(a) UMAP of the 37,436 spots, across all 14 samples. Spots were annotated based on markers shown in B, though multiple cell types likely exist within each spot.

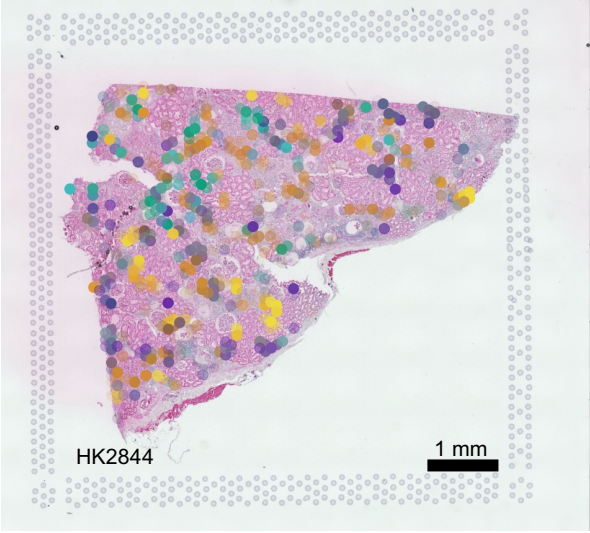
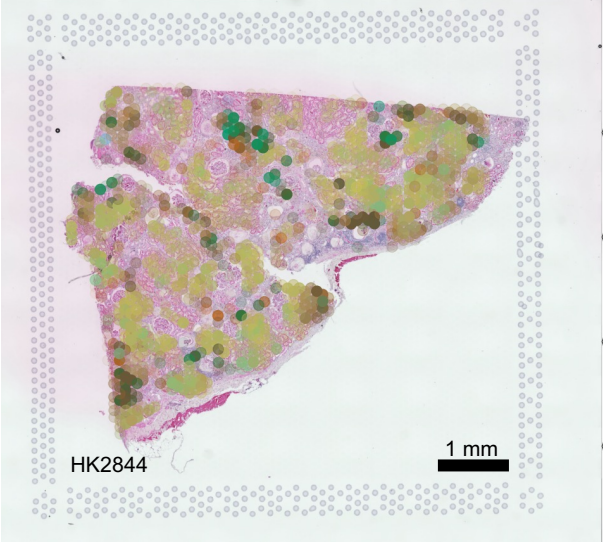
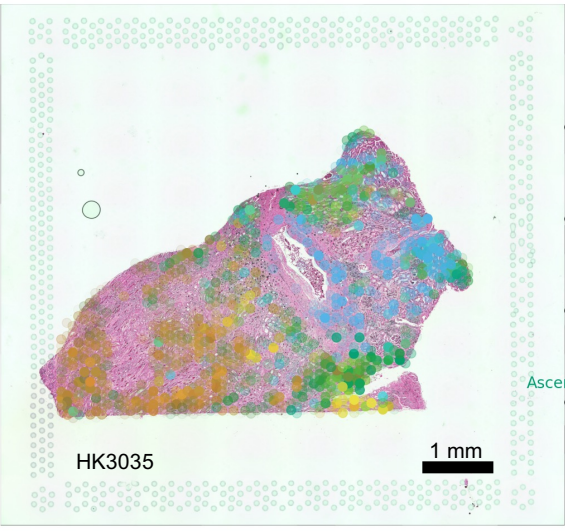
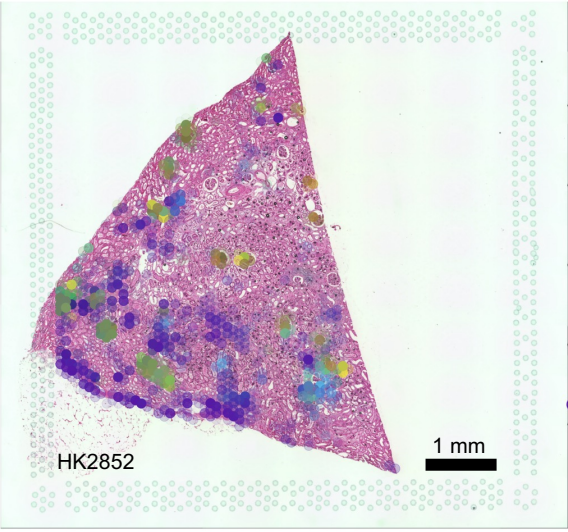
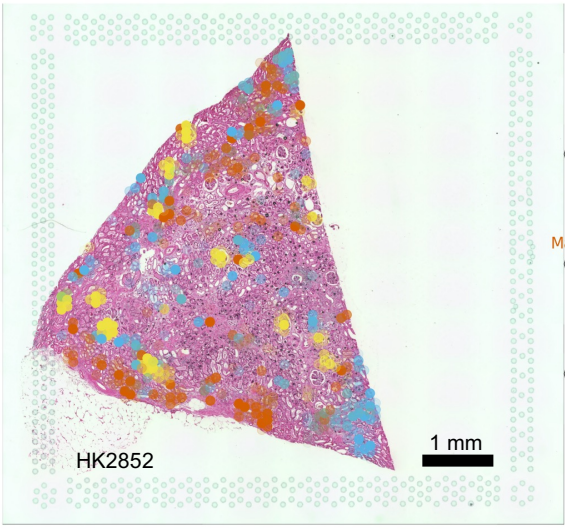
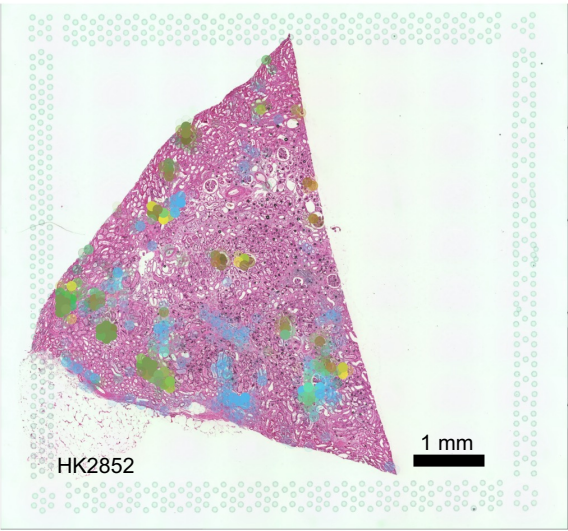
(b) The dot plots of marker genes used for spots annotation in the spRNAseq. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

(c) The heatmap of the top 10 differentially expressed genes in the identified spots in spRNAseq. Rows indicate the top 10 differentially expressed genes and their degree of accessibility pre cluster and columns show the cell types.

Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels, Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific stromal cells, Fib; Fibroblast, MyoFib; Myofibroblast, VSMC/Pericyte; vascular smooth muscle cells/Pericyte, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naive; Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

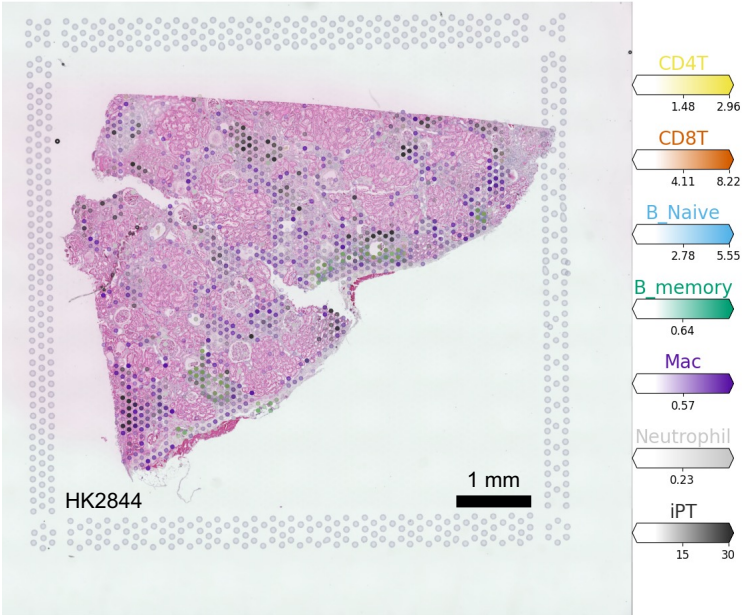
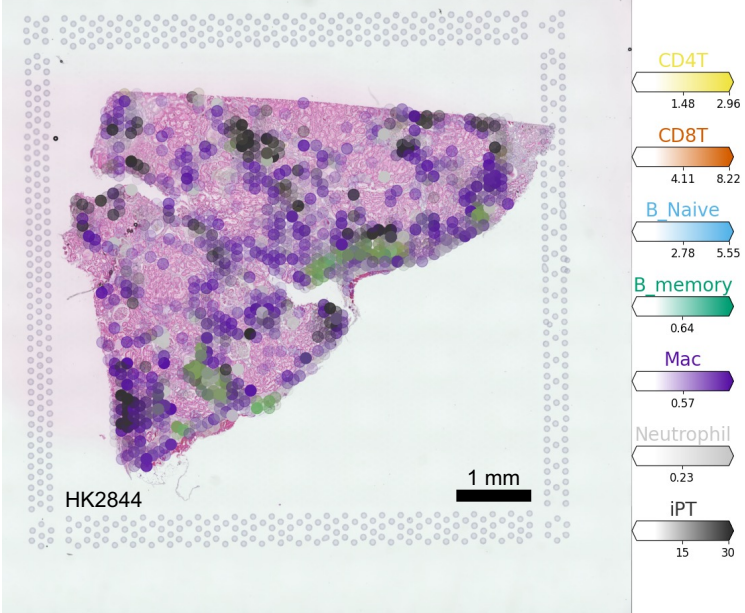
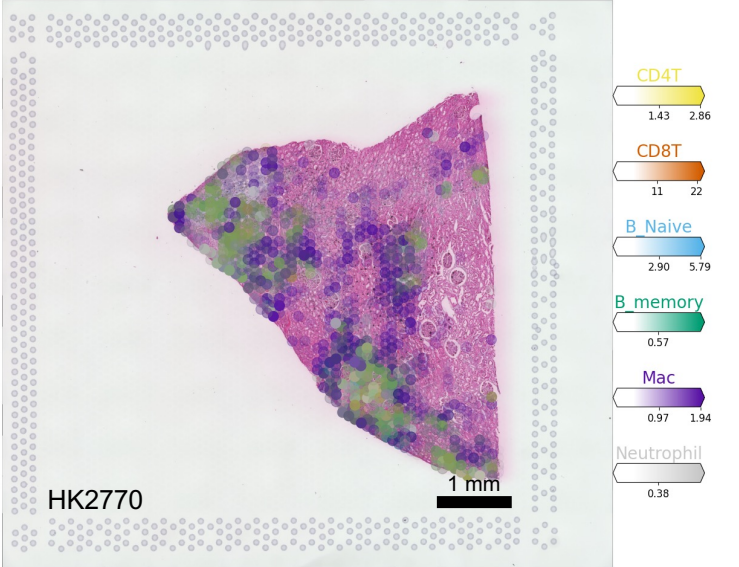


Supplementary Fig. 8. Protein expression of human kidney cell type markers in the Human Protein Atlas
Images were taken <https://www.proteinatlas.org/>.



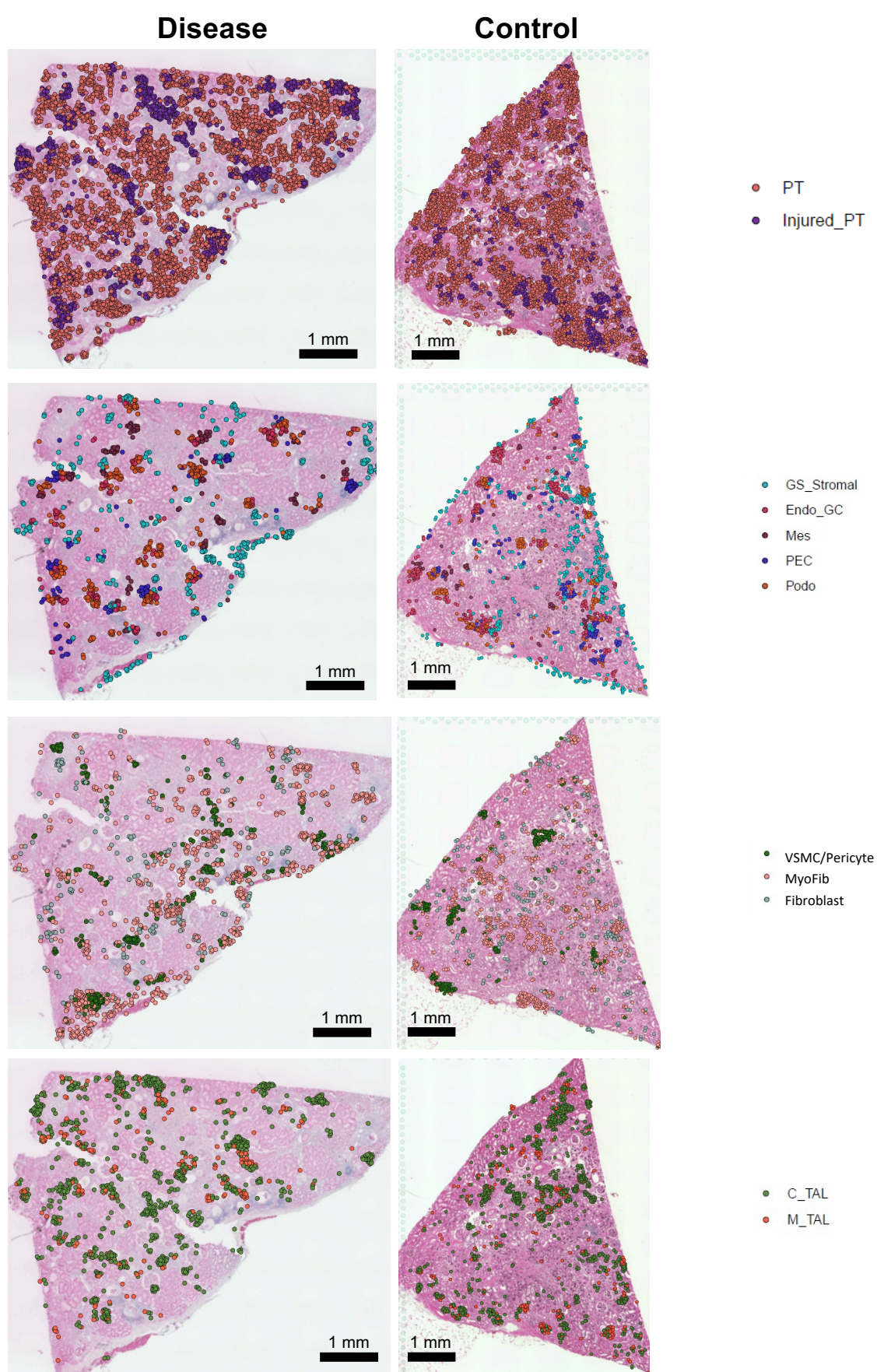
Supplementary Fig. 9

Supplementary Fig. 9. Glomerular, proximal and distal tubule cells in each kidney sample
(Cell2Location was used to map cells).

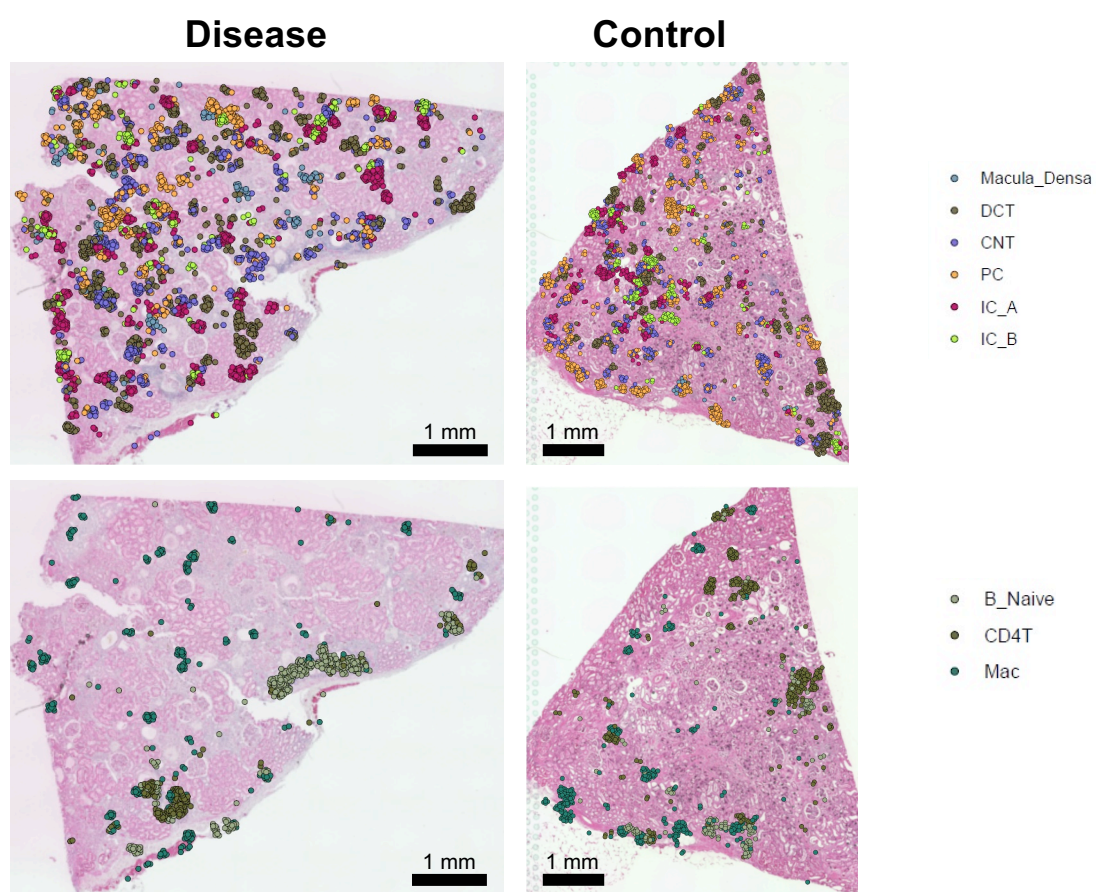


Supplementary Fig. 10

Supplementary Fig. 10. Immune and stromal cells in each kidney sample (Cell2Location was used to map cells).

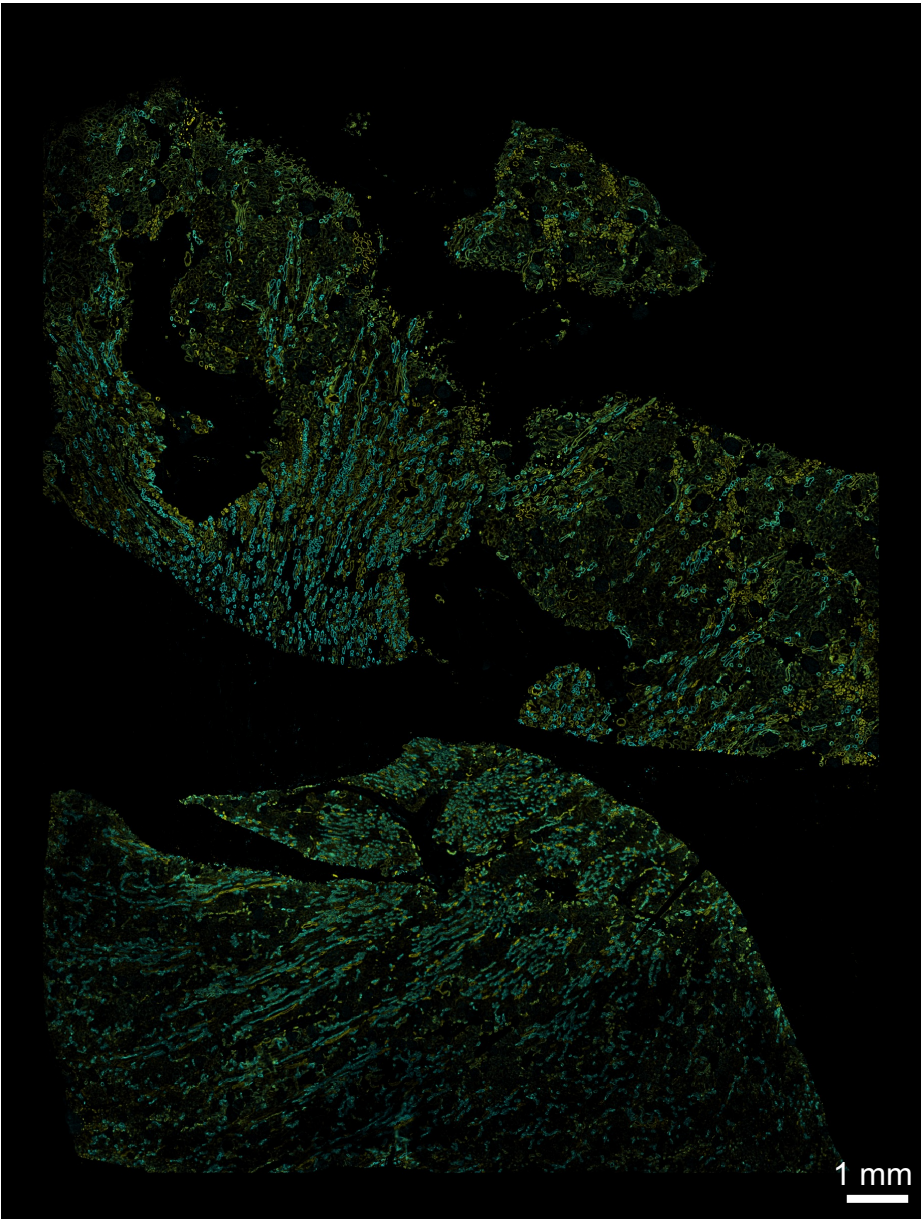


Supplementary Fig. 11. Proximal tubule, glomerular, stromal and LOH cells in each kidney sample
(CellTrek was used to map cells).

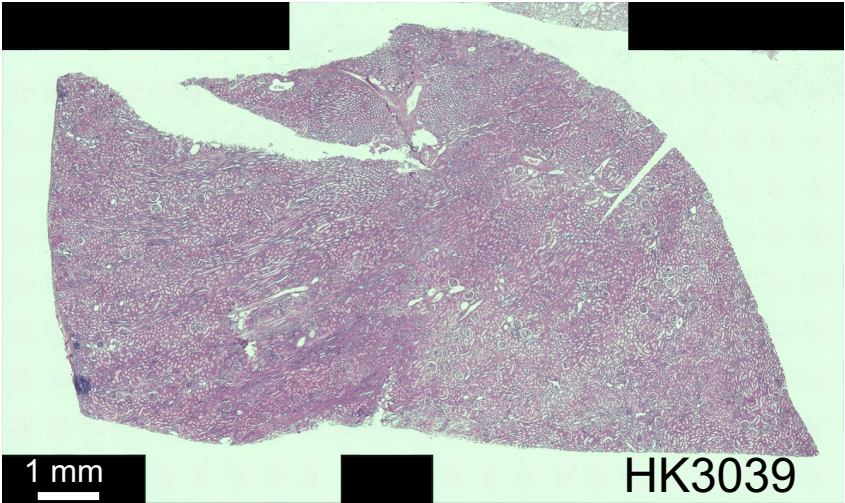
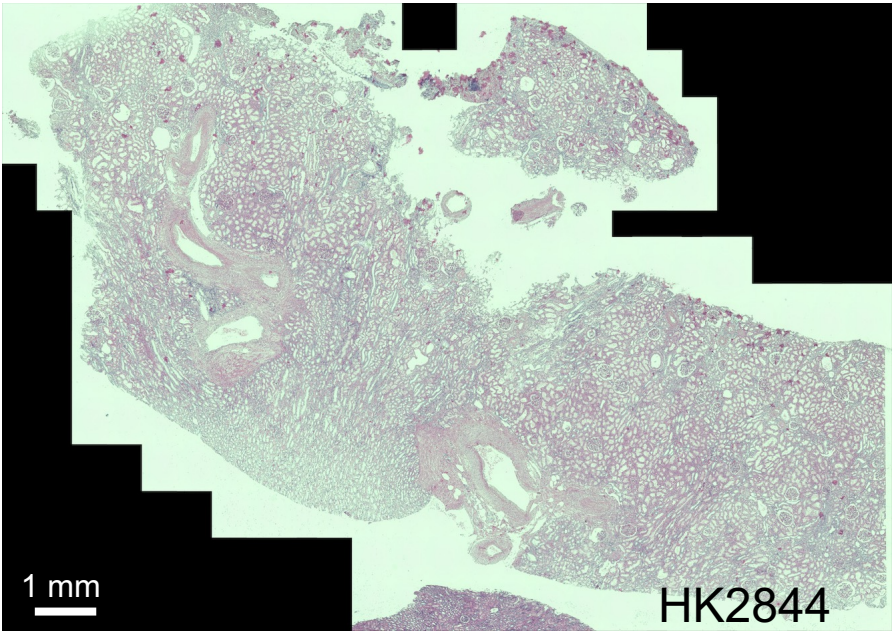


Supplementary Fig. 12. Distal tubule and immune cells in each kidney sample (CellTrek was used to map cells).

a



b

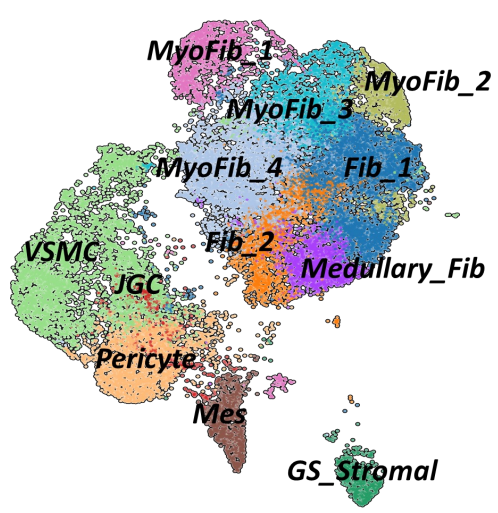


Supplementary Fig 13. Tissue staining of the CosMx slides.

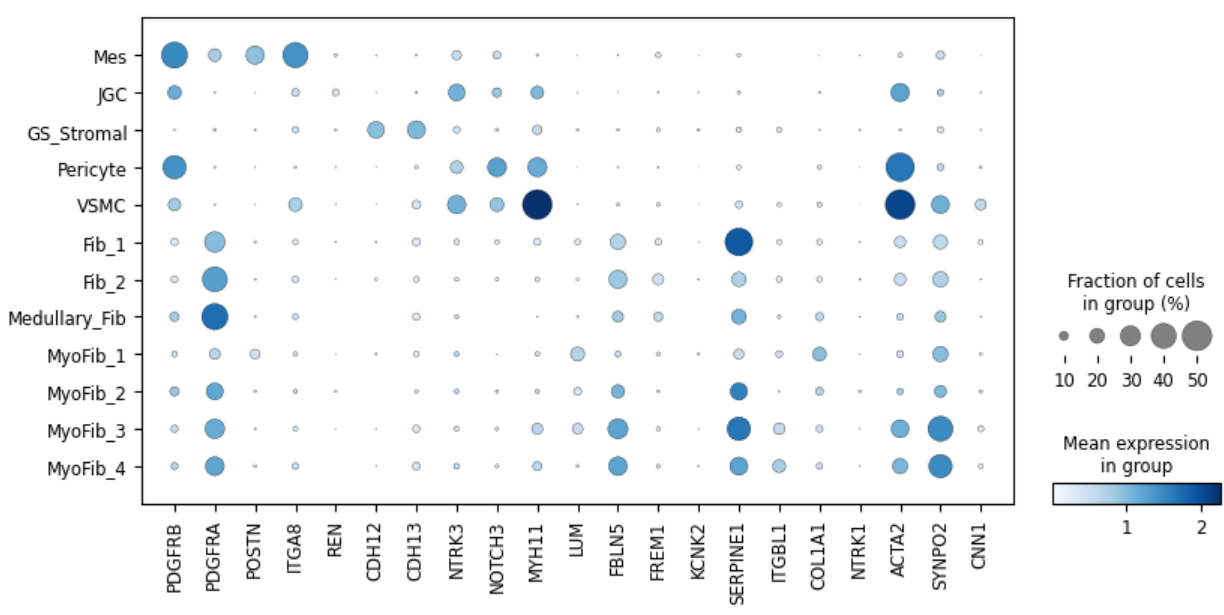
(a) Immunostained CosMx slide. CosMx slides were stained with DAPI, Pan-Cytokeratin, CD-45, and membrane stain.

(b) After imaging the slide, the flowcell (sample) was removed and the slides were then H&E stained and re-imaged.

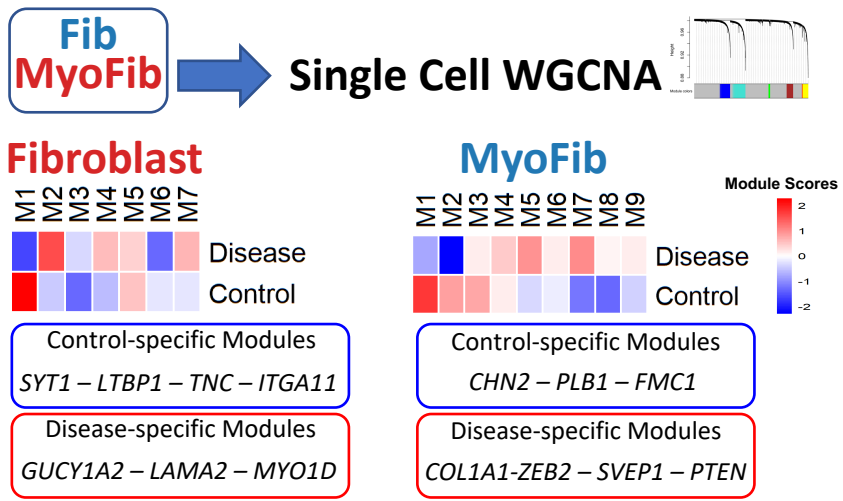
a



b



c



Supplementary Fig. 14. Sub-clustering of stromal cells in the integrated dataset

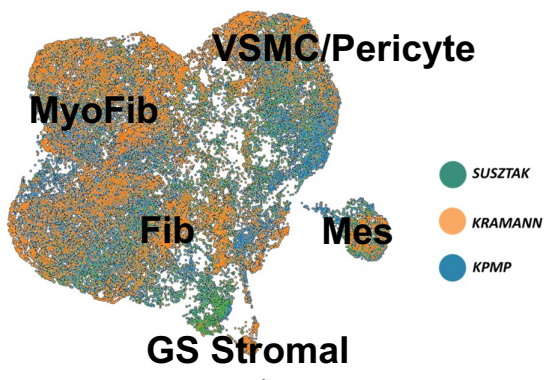
(a) UMAP of 41,422 stromal cells in the integrated dataset.

(b) The dot plots of marker genes used for cell type annotation. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

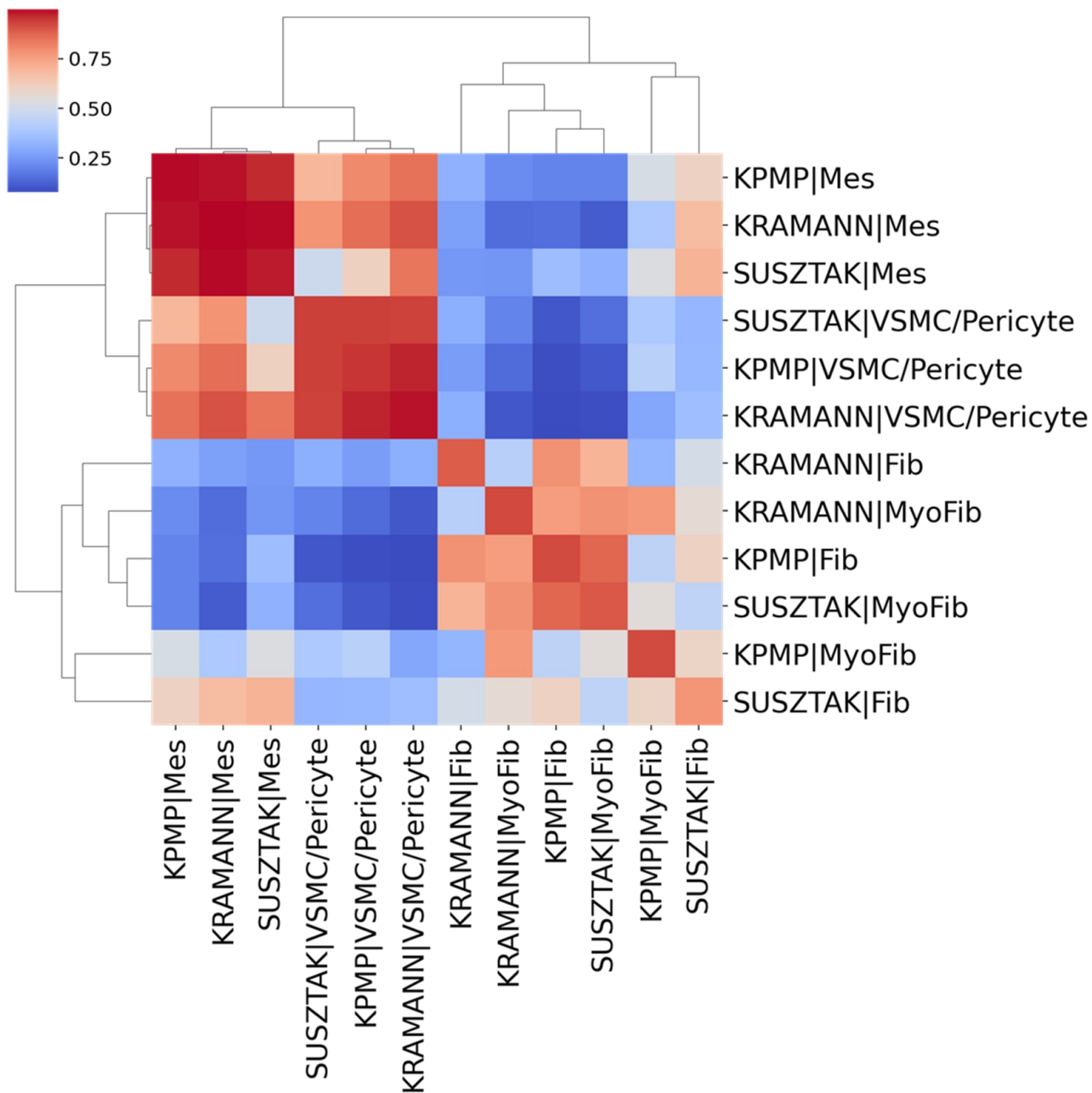
(c) The single cell Weighted correlation network analysis (WGCNA) on fibroblast and myofibroblast identified specific modules in control and disease samples. The heatmaps shows the identified modules in each cell types and their scores. The lower panel shows the representative genes with high scores in each cell type.

Mes; Mesangial cells, GS_Stromal; glomerulosclerosis associated stromal cells, JGC; juxta-glomerular cells, VSMC; vascular smooth muscle cell, Fib; fibroblast, MyoFib; myofibroblast.

a



b

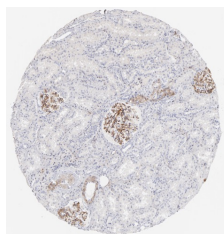


Supplementary Fig. 15. The integration of stromal cells with prior studies

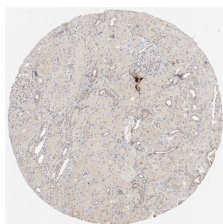
(a) UMAP of integration of 86,481 stromal cells of current study prior papers by KPMP and Kramann *et al.* The color indicates the data source.

(b) Heatmap display of MetaNeighbor analysis of the similarities of the identified stromal cell types. The heatmap displays the Area Under the Receiver Operating Characteristic (AUROC) scores among different cell subtypes in three different sources of datasets, which were determined using a set of highly variable genes (HVG). Dendrograms were created via hierarchical clustering, based on Euclidean distances, and used average linkage for this purpose. Each row and column is a cell type in the source data.

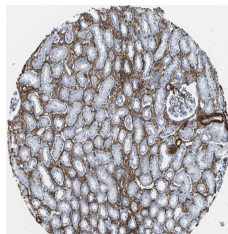
ITGA8



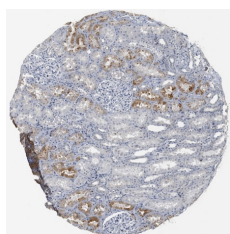
REN



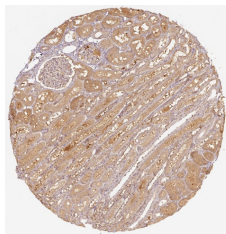
ACTA2



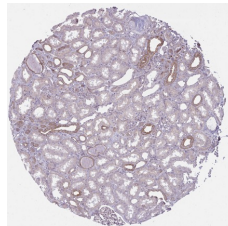
COL1A1



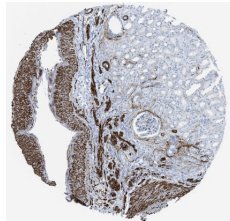
SERPINE1



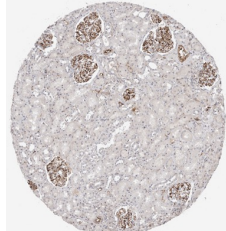
FAP



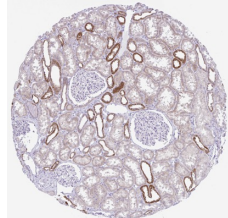
MYH11



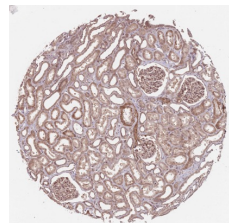
CDH13



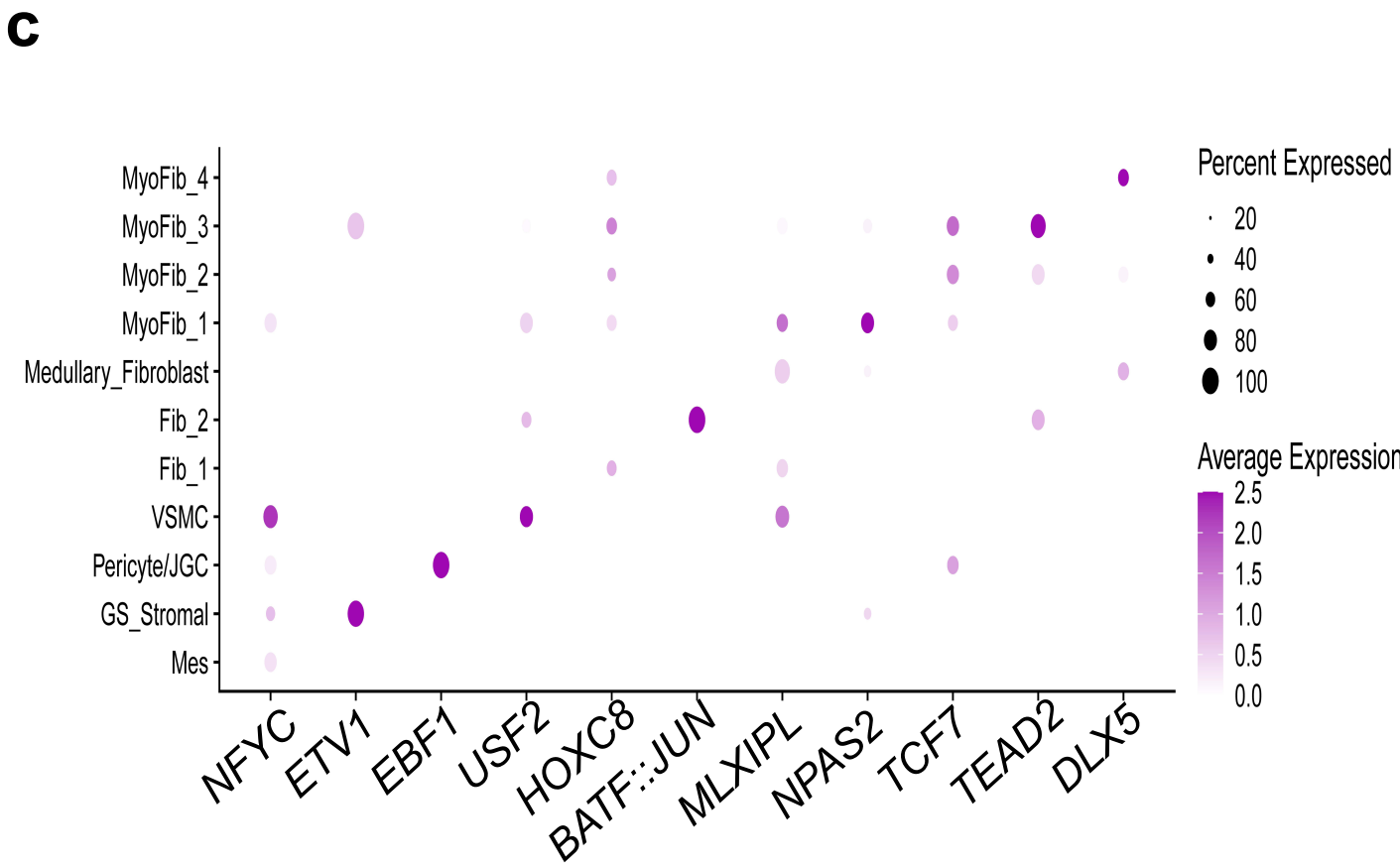
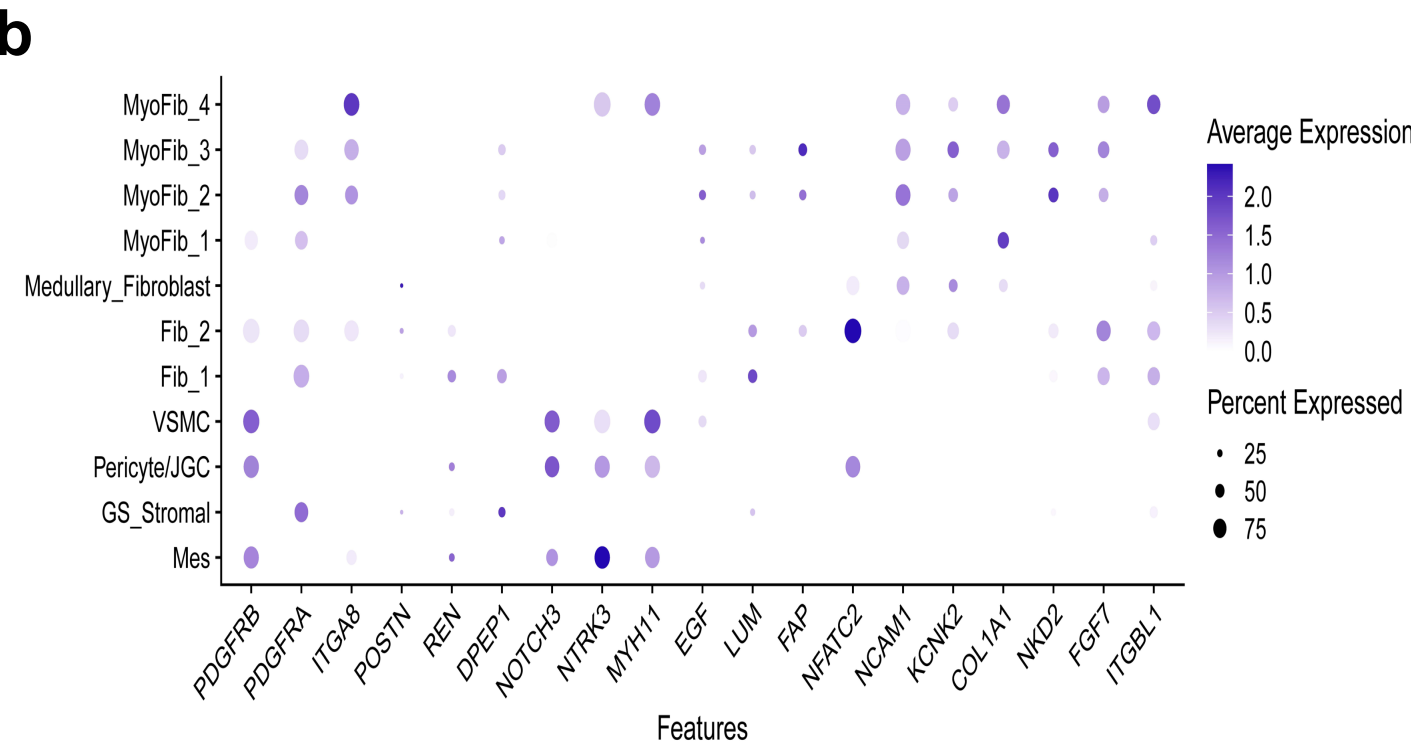
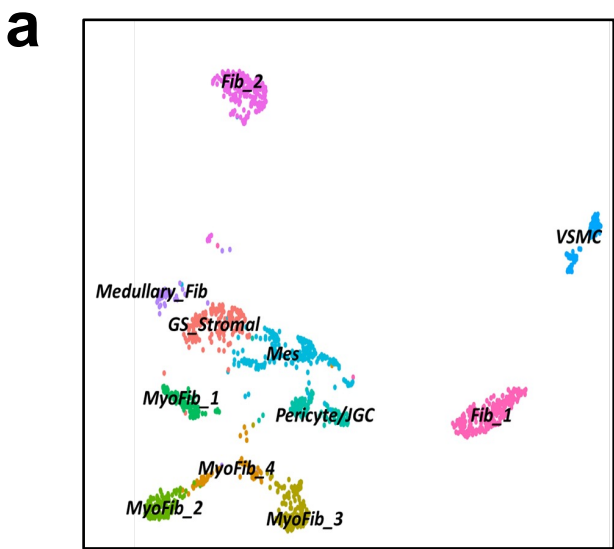
SYT1



NOTCH3



Supplementary Fig. 16. Expression of stromal markers in the Human Protein Atlas kidney samples. Images were taken from <https://www.proteinatlas.org/>.



Supplementary Fig. 17

Supplementary Fig. 17. Stromal sub-clustering in human snATACseq data.

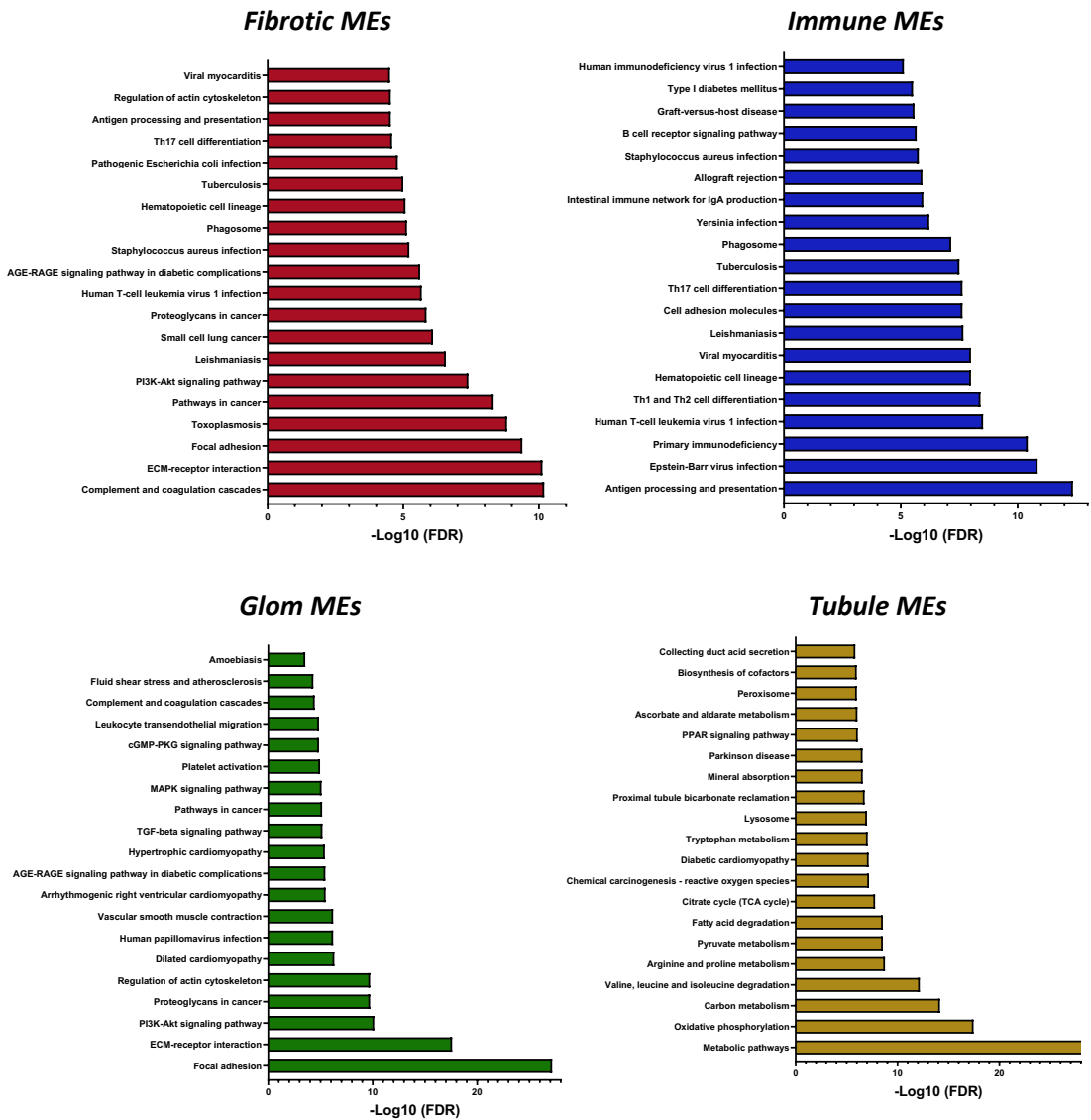
(a) UMAP of 5,822 stromal cells in human snATACseq dataset.

(b) The dot plot of gene markers used to annotate the different stromal cell types in snATACseq dataset. The dot plot shows the gene activity in different cell types. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

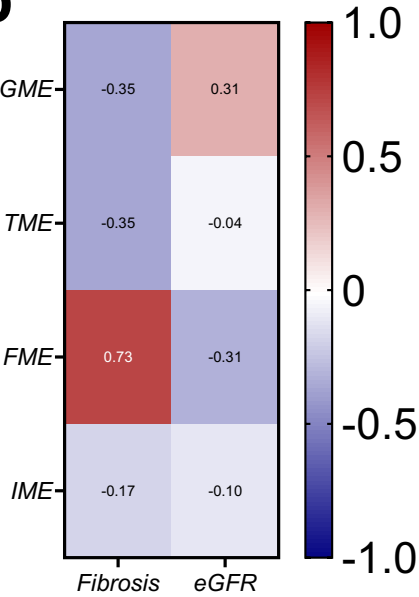
(c) The dot plot of motif activity of enriched motifs in each identified stromal cluster. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

Mes; Mesangial cells, GS_Stromal; glomerulosclerosis associated stromal cells, JGC; juxta-glomerular cells, VSMC; vascular smooth muscle cell, Fib; fibroblast, MyoFib; myofibroblast.

a



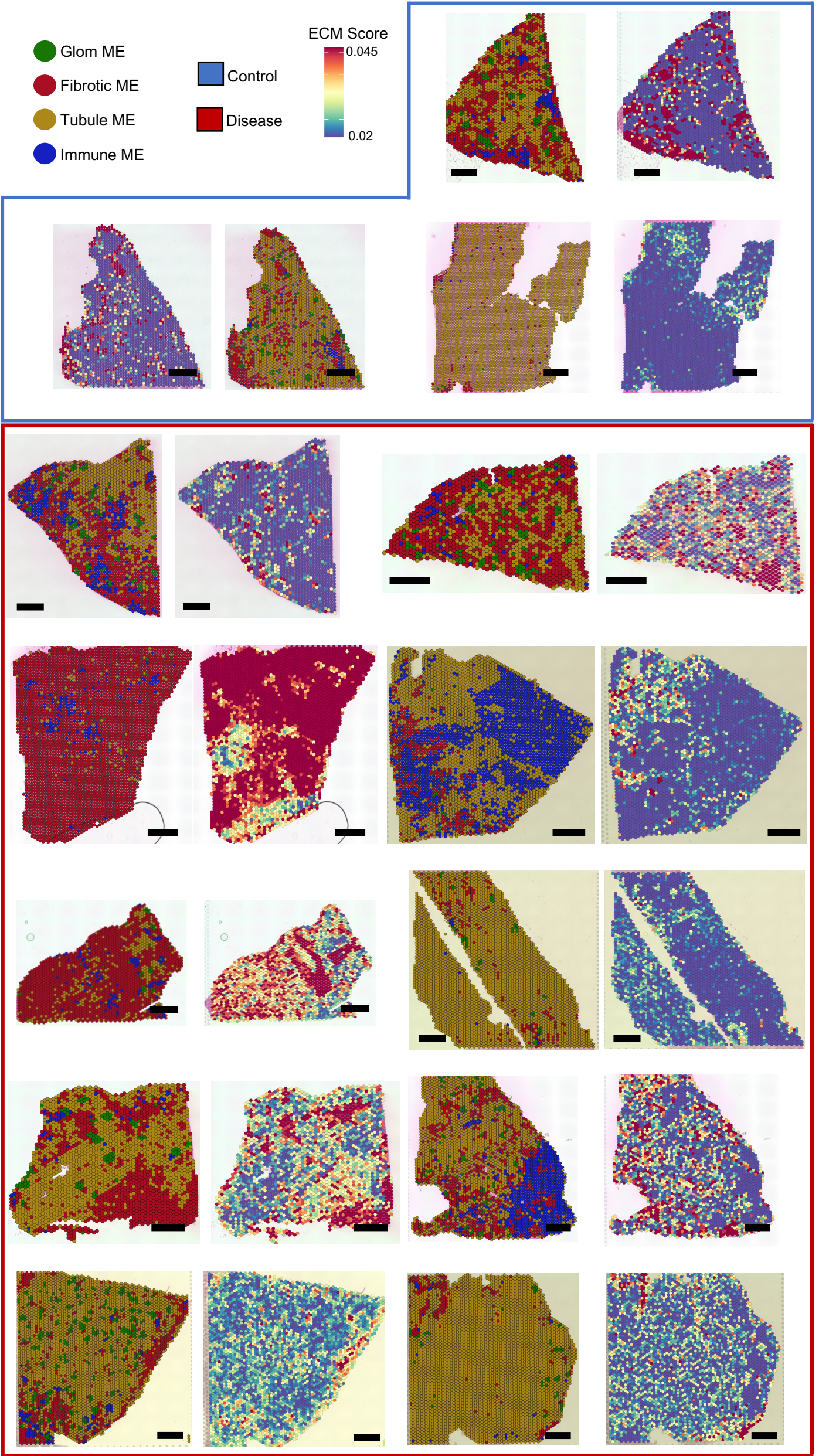
b



Supplementary Fig. 18. Gene ontology analysis of microenvironment gene signature

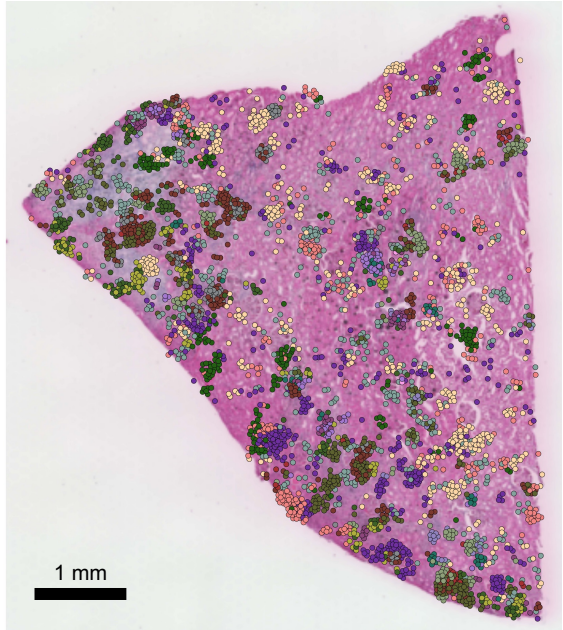
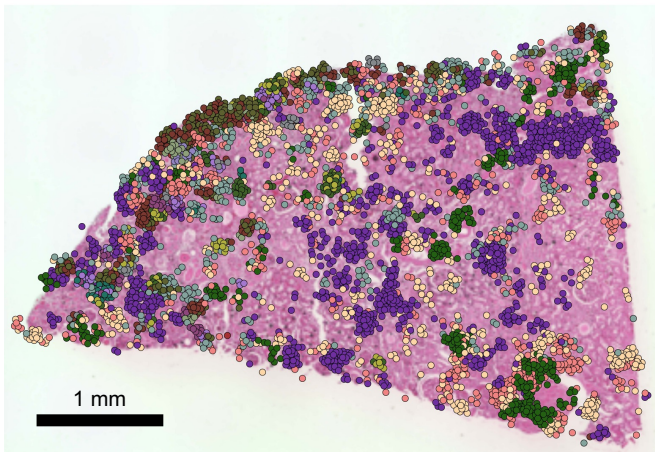
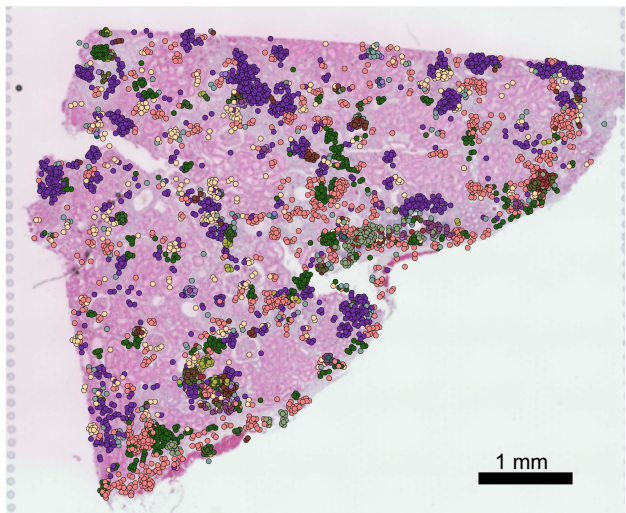
(a) Gene ontology analysis (DAVID) of the microenvironment gene signature. The X axis shows the $-\text{Log}_{10}(\text{FDR})$ of the identified pathways and Y axis lists the pathways.

(b) Person correlation of the number of spots assigned to each microenvironment and eGFR and fibrosis.



Supplementary Fig. 19

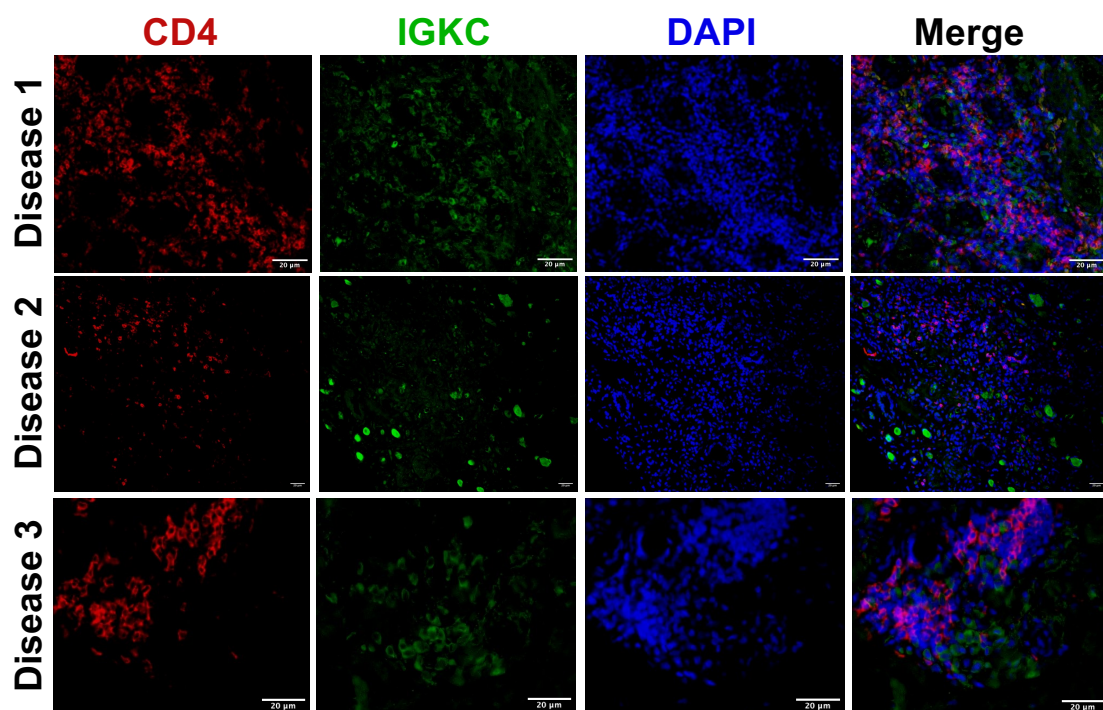
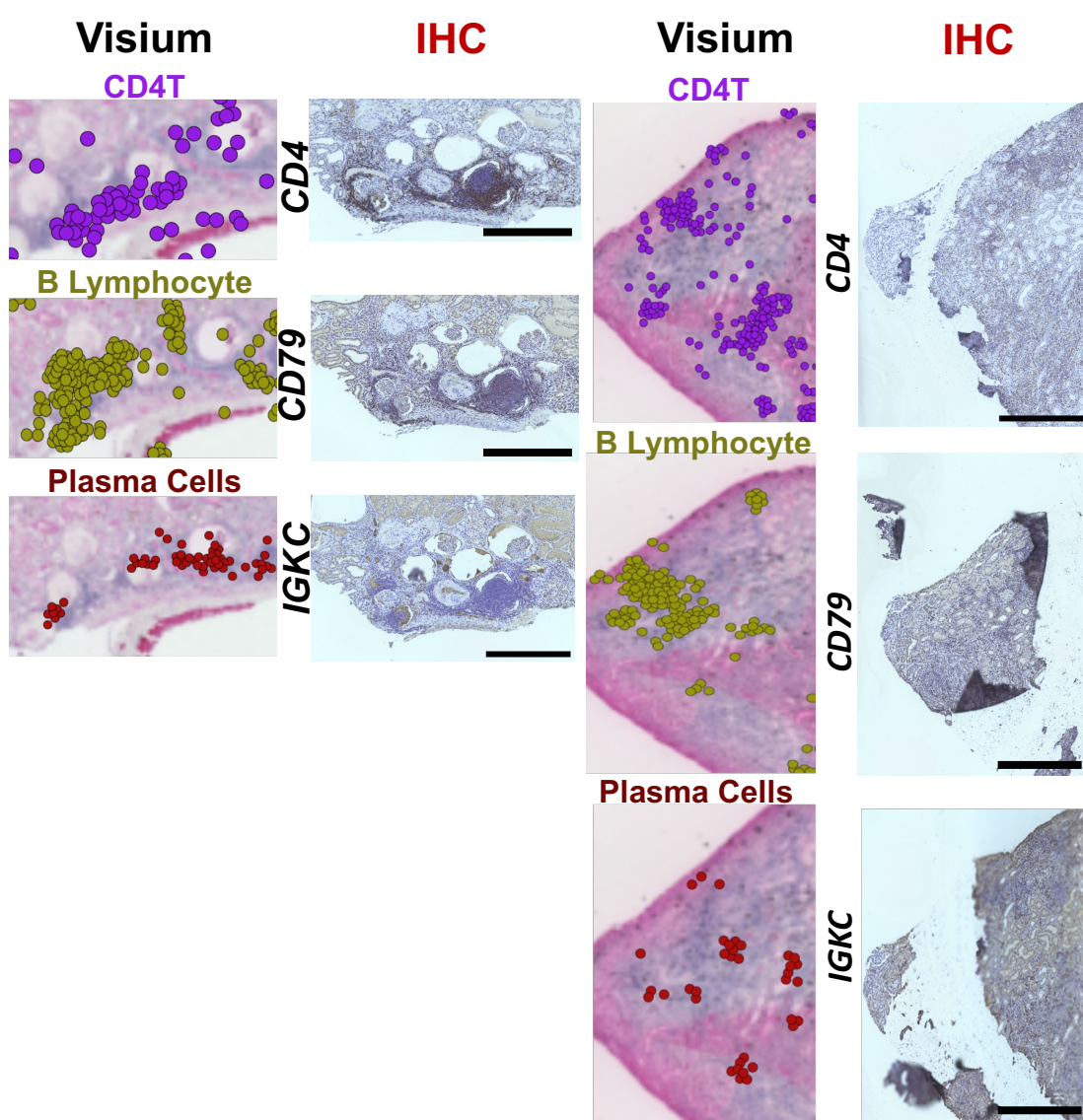
Supplementary Fig. 19. The spatial microenvironments in the human kidney. Four spatial microenvironments were identified in human kidney samples using matrix factorization: glomerular (green), tubule (brown), fibrotic (red) and immune (blue). The ECM score (calculated by aggregate expression collagen and other glycoprotein, and proteoglycan) in Visium data shown next to the tissue microenvironment information.



Cell Types

- Injured_PT
- CD16_Mono
- Neutrophil
- Endo_Peritubular
- MyoFib
- VSMC/Pericyte
- CD14_Mono
- cDC
- CD8T
- NK
- CD4T
- B_Naive
- Fibroblast
- pDC
- Plasma_Cells
- Mac
- B_memory

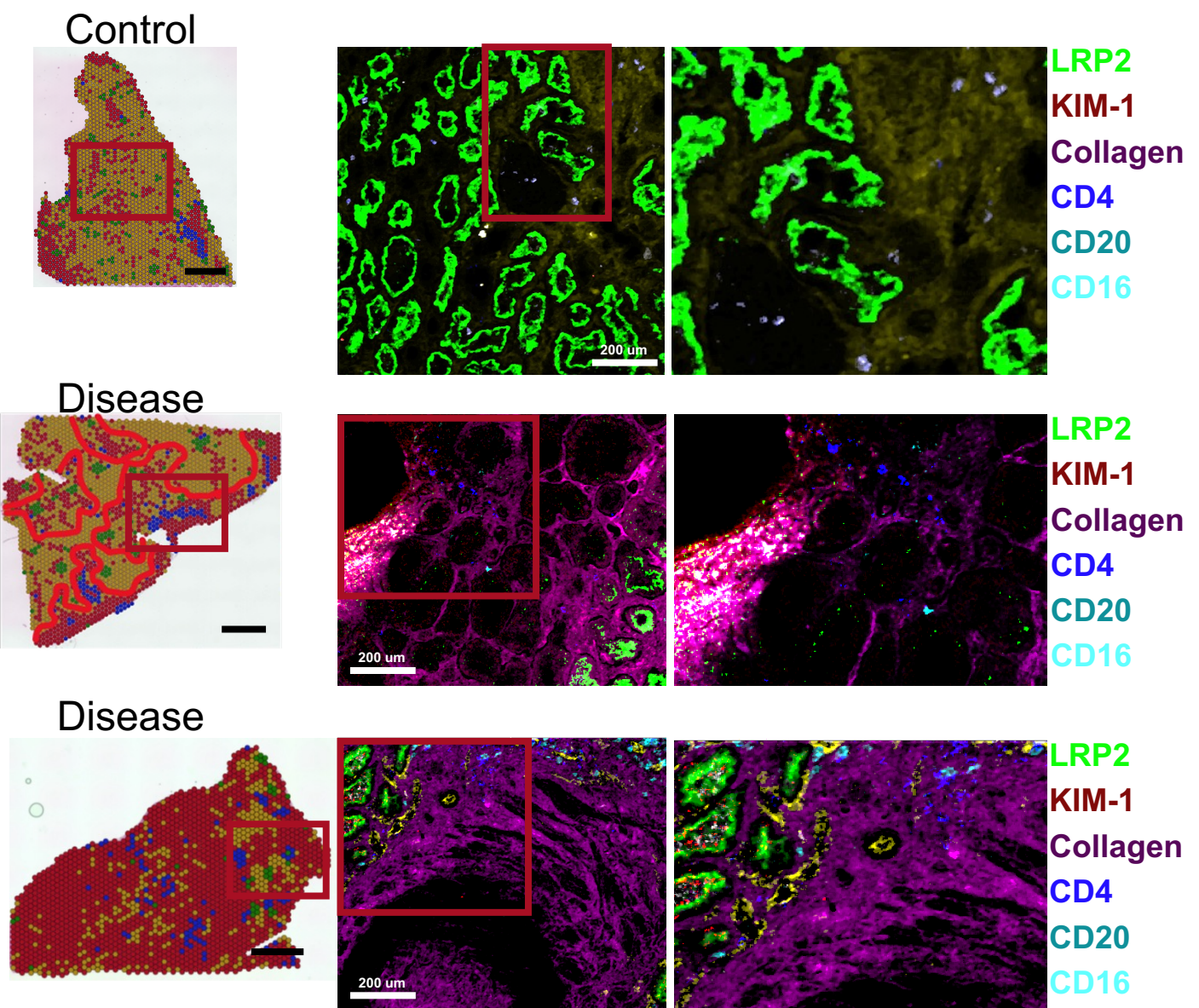
Supplementary Fig. 20. Different cell types in kidney microenvironment. Left panel kidney microenvironment in tissue samples, right panel shows cell abundance mapped using Cell2location (endothelial cell, injured_PT cells, stromal cells and immune cells)

a**b**

Supplementary Fig. 21. Immunostaining of the lymphocyte and plasma cells in regions of fibrosis-associated immune cells aggregates.

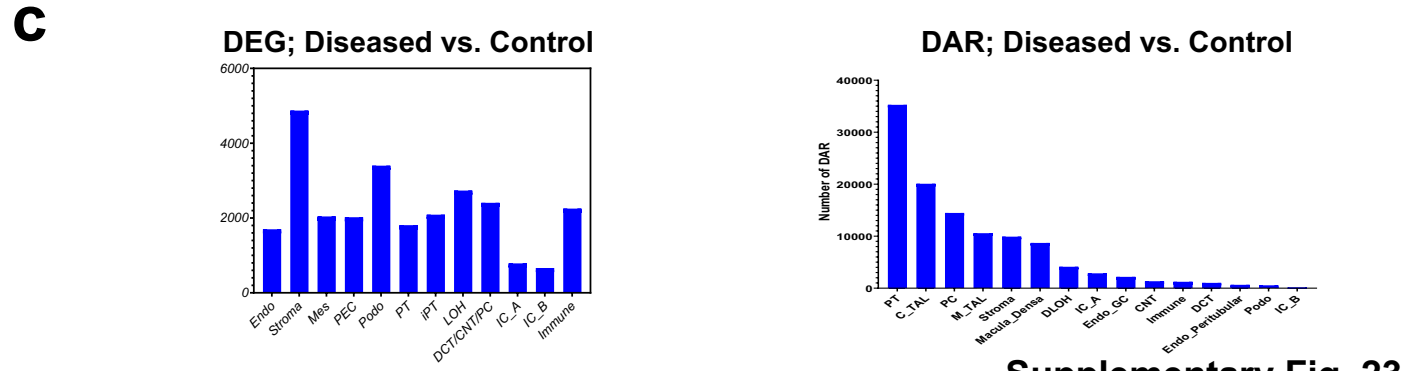
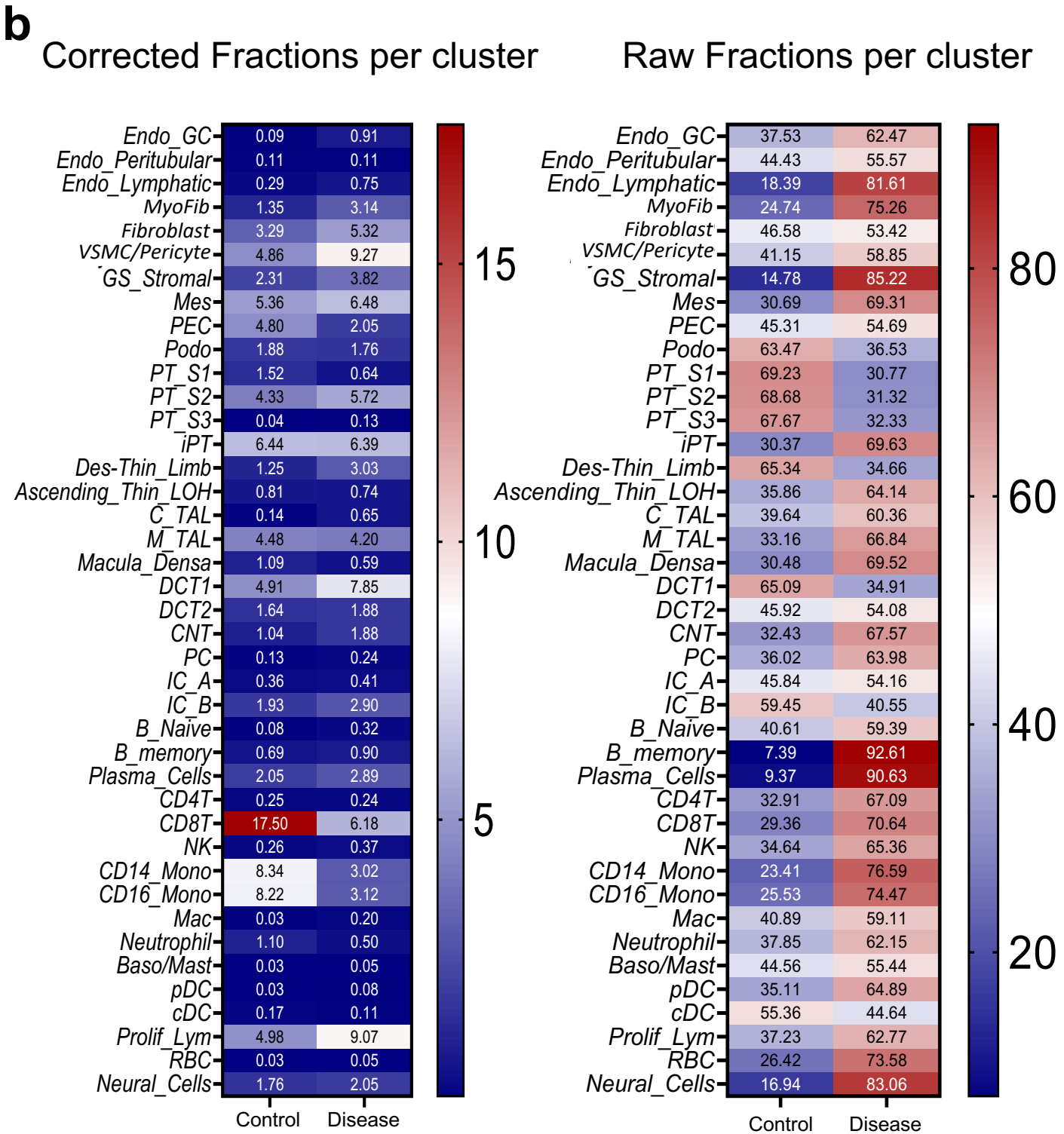
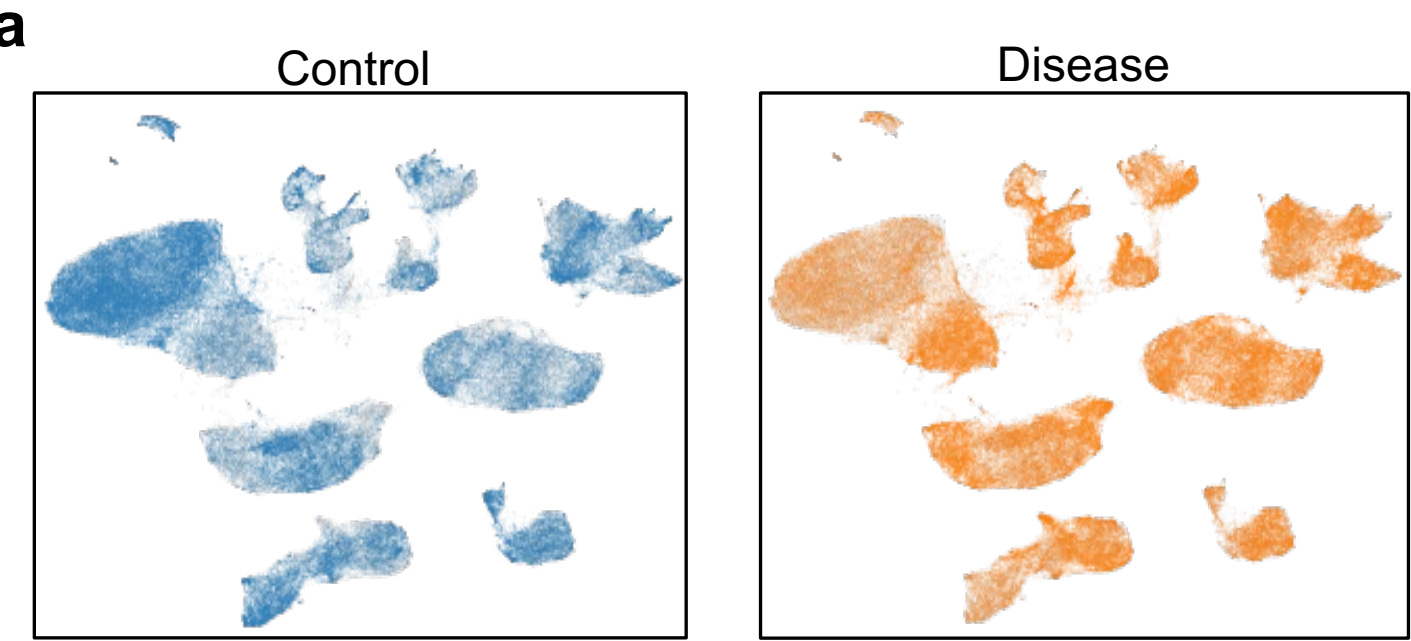
(a) Immunofluorescent staining of T-cell markers (CD4) and plasma cells (IGKC) in diseased human kidney samples.

(b) Spatial RNAseq (10x Visium with CellTrek imputation) and immunohistochemistry staining of T-cell, B-cell, and plasma cells in diseased human kidney samples in the regions of fibrosis-associated immune cells aggregates.



Supplementary Fig. 22. *In situ* mass spectrometry of the control, and diseased samples to verify cell types in fibrotic microenvironments.

The left panel shows the Visium 10x data cellular microenvironment. The right panel shows the IMC staining of the area, markers and colors are listed on the right.



Supplementary Fig. 23

Supplementary Fig. 23. Comparison of control and diseased samples in the integrated sn/scRNAseq and snATACseq.

(a) UMAP distribution of control and diseased samples in the integrated snRNA/scRNAseq dataset.

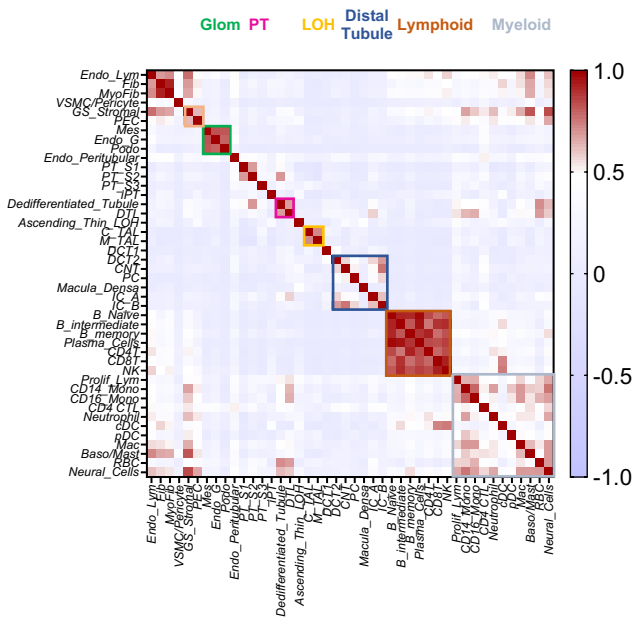
(b) Heatmaps of cell type distribution in the control and diseased conditions.

(c) Number of differentially expressed genes and differentially accessible regions between diseased and control samples

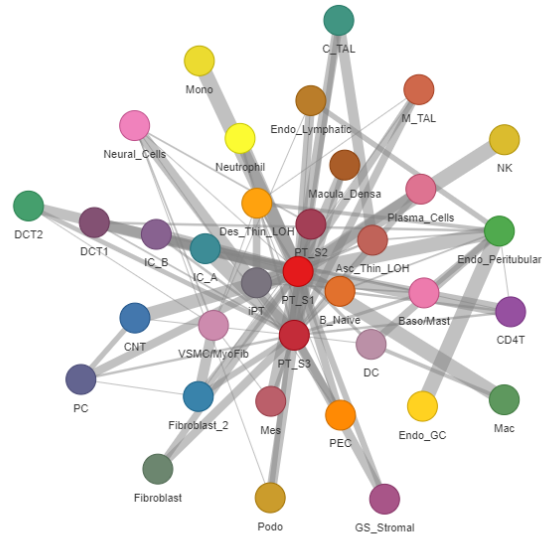
Endo; endothelial cells Mes; mesangial cells, Stroma; stromal cells, PEC; parietal epithelial cells, Podo; podocyte, PT; proximal tubule cells, Injured_PT; injured proximal tubule cells, LOH; loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells.

a

Correlations between Cell Types
Frequency in Spots

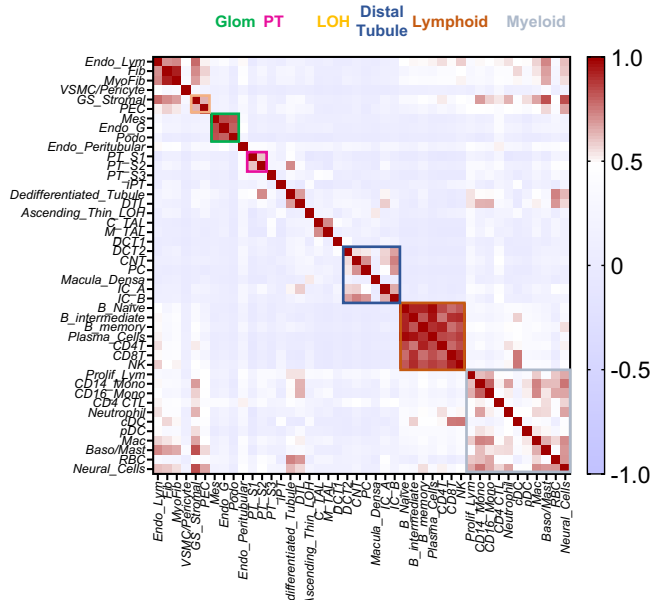
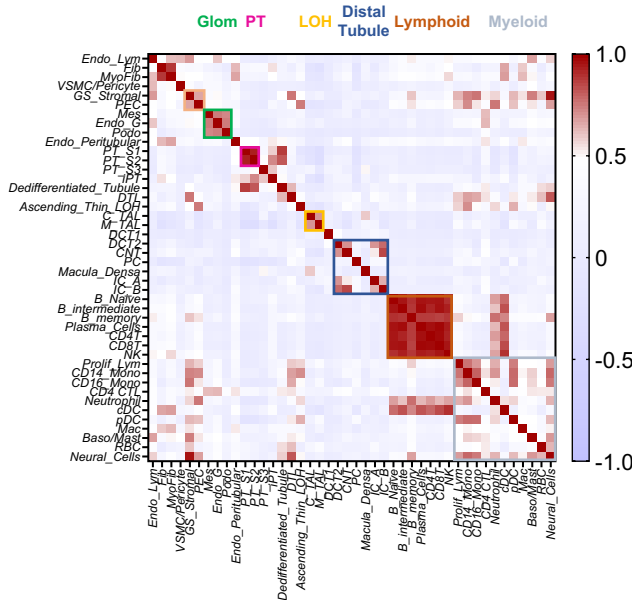
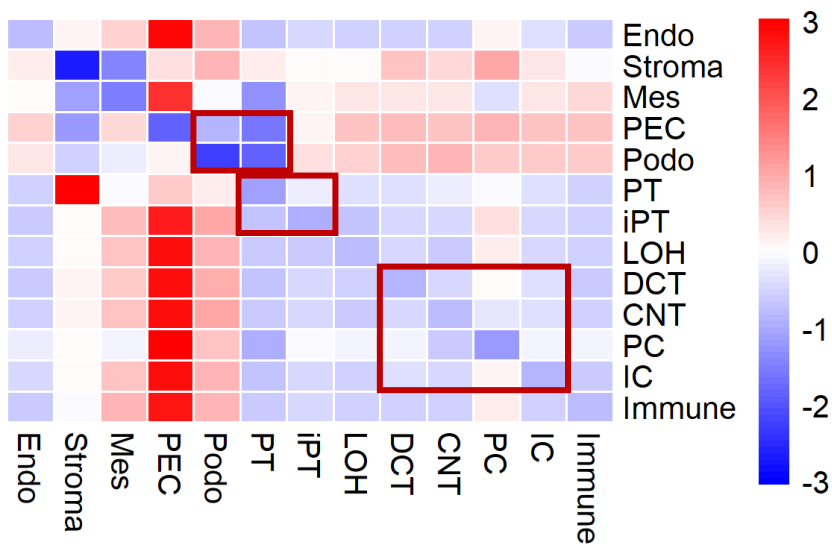
**b**

Spatial cell proximity index

**c**

Control

Disease

**d**

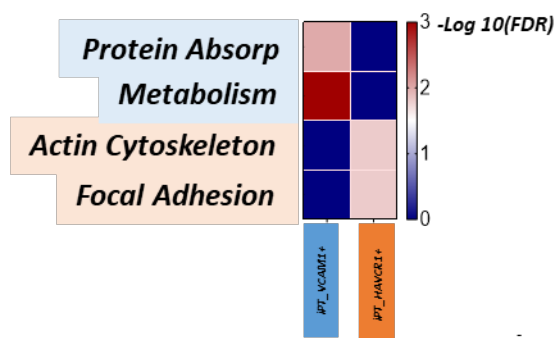
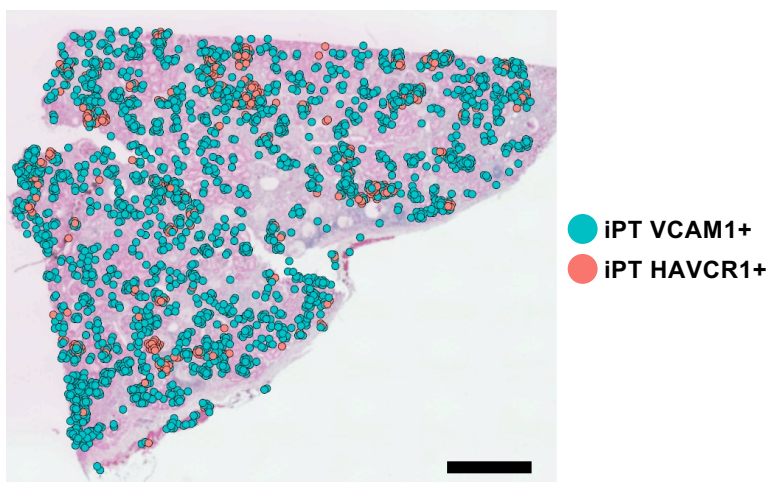
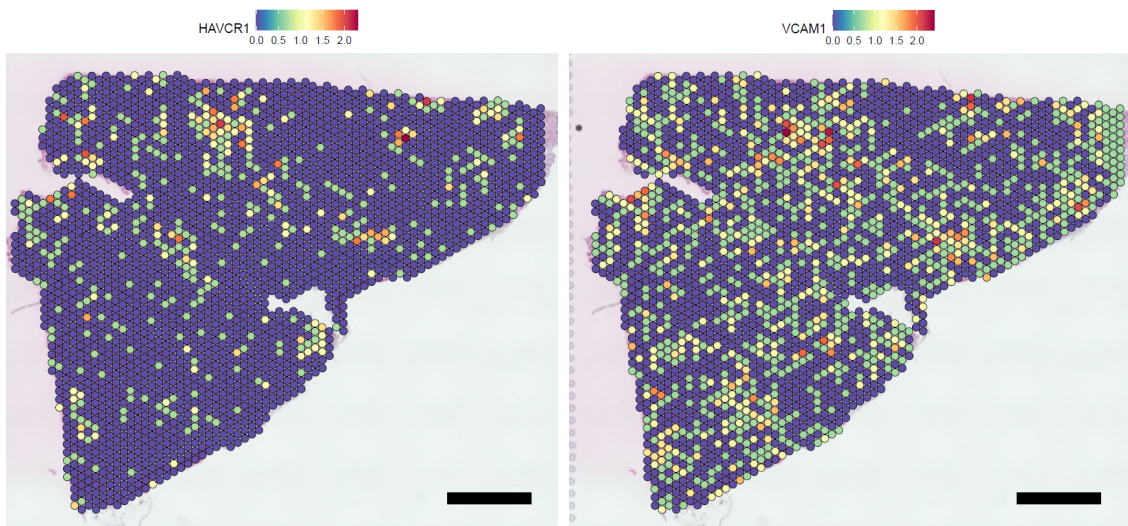
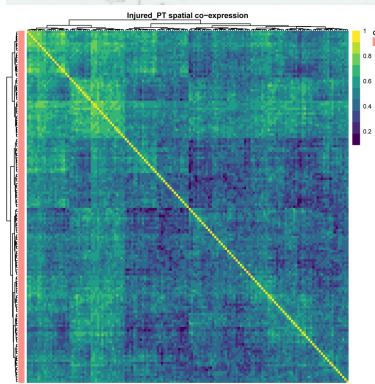
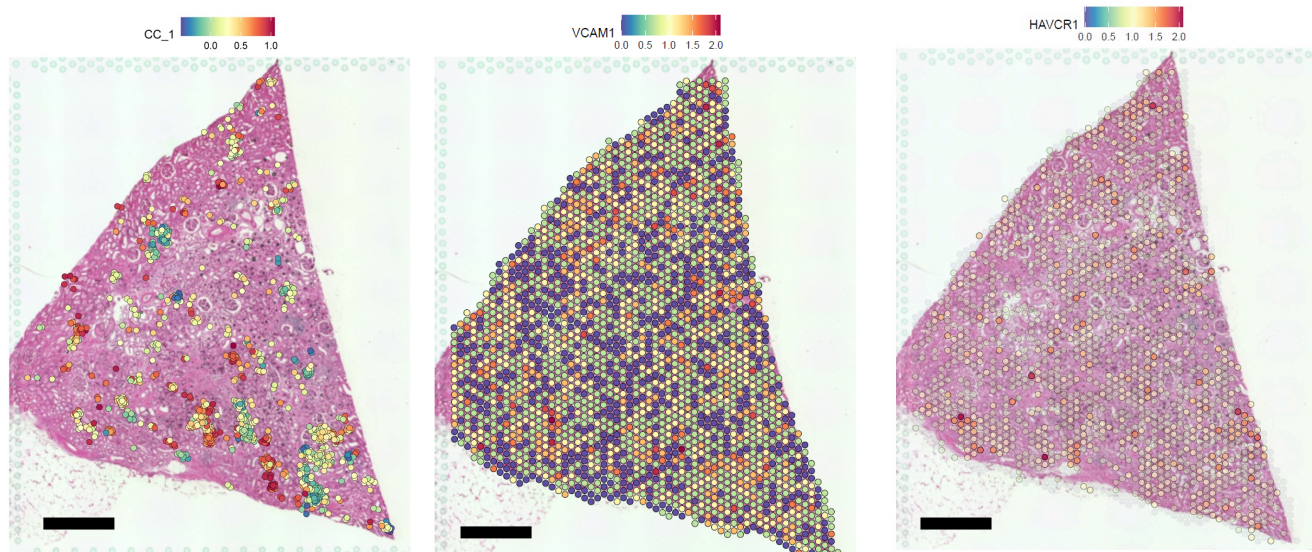
Supplementary Fig 24. Cell type correlations and distances of Cell Trek data.

(a) Identification of neighborhoods based on spatial location of the CellTrek imputations.

(b) Spatial cell proximity index of CellTrek cell types.

(c) Neighborhoods parsed by Control and Disease sample types.

(d) Cell type distances between different cell types. Glomerular, proximal tubular and distal tubular neighborhoods demonstrated by red boxes.

a**b****c****d****e**

Supplementary Fig. 25. Injured PT cells in the control and diseased human kidney samples.

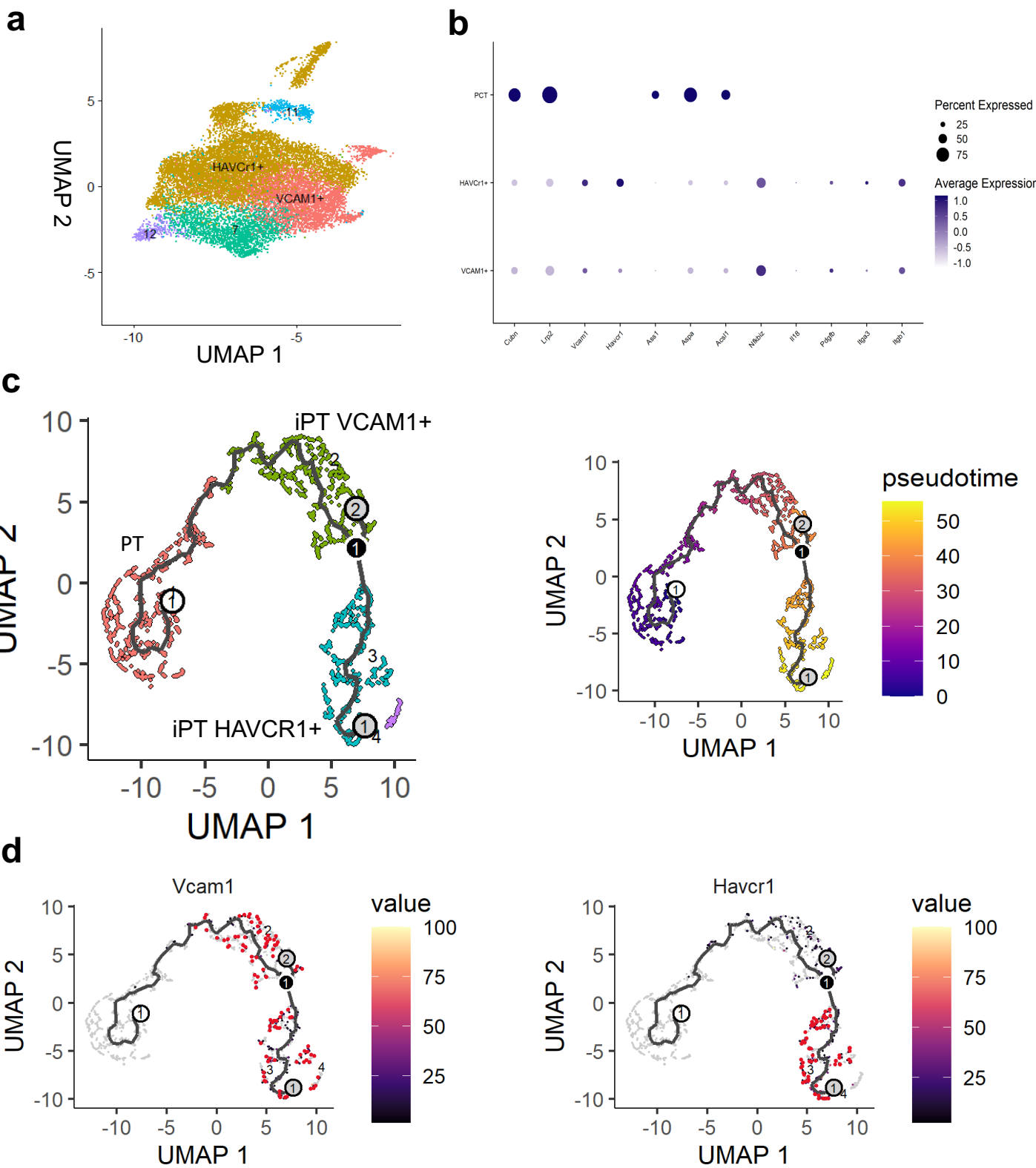
(a) Gene ontology enrichment for top pathways in iPT-VCAM1+ and iPT-HAVCR1+ cells.

(b) The spatial location of iPT_VCAM1+ (blue) and iPT_HAVCR1+ (red) cells in a diseased kidney sample based on CellTrek imputation.

(c) Expression of *HAVCR1* and *VCAM1* (color indicates level of expression).

(d) The gene co-expression network indicates one type of injured_PT in control human kidneys, upper panel shows the heatmap of co-expressed genes.

(e) Expression of *CC1* in imputed iPT from CellTrek, and raw *VCAM1*, and *HAVCR1* expression in a healthy human sample.



Supplementary Fig. 26

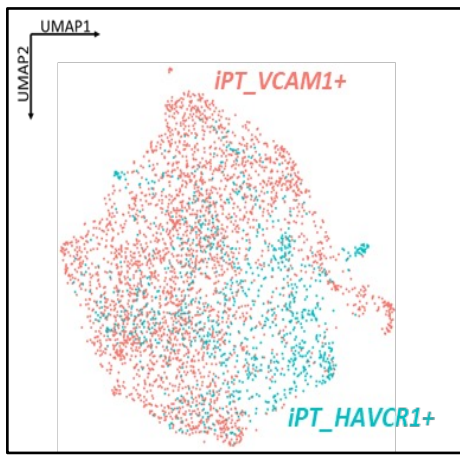
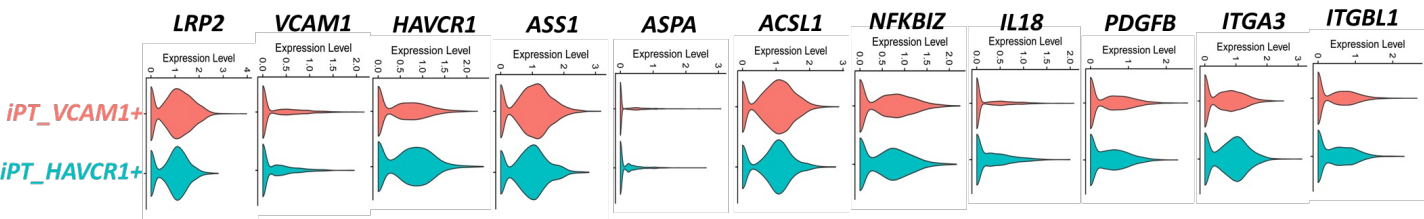
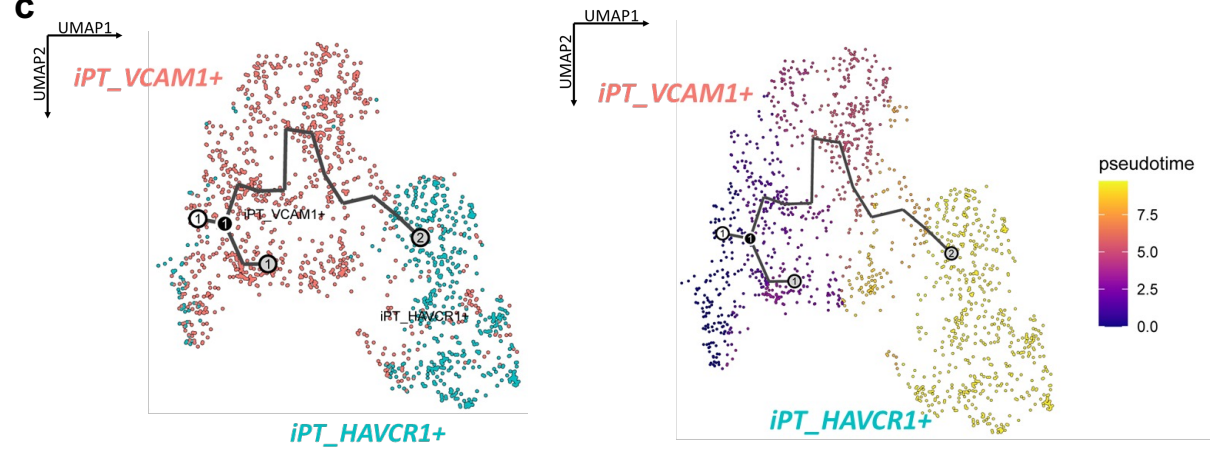
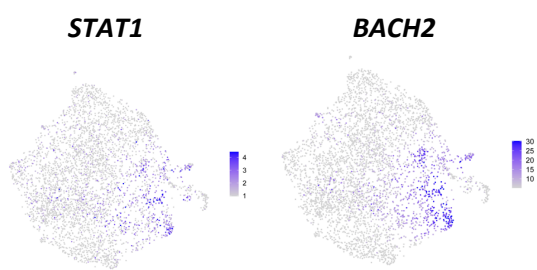
Supplementary Fig. 26. Injured PT cells in mouse diabetic kidney disease samples.

(a) UMAP of (n=28,579) of proximal convoluted tubule and two types of injured PT cell nuclei (left panel).

(b) The dot plots of marker genes used for the annotation of PCT, iPT_VCAM1+, and iPT_HAVCR1+ in the integrated dataset. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

(c) Monocle cell trajectory of PT and iPT cells sub-clustering trajectory from PT to iPT-VCAM1+ and iPT-HAVCR1+ in mouse snRNAseq (left panel). Cells are colored by pseudotime (right panel) .

(d) Expression of the *Vcam1* and *Havcr1* along the trajectory.

a**b****c****d**

Supplementary Fig. 27. Injured PT cells in the snATACseq data.

(a) UMAP of 24,256 injured PT sub-cluster in human snATACseq dataset.

(b) Violin plot of expressions of *VCAM1*, *HAVCR1* and other genes in in two types of iPTs in snATACseq dataset.

(c) Monocle trajectory analysis of PT and injured_PT cells in the snATACseq dataset. The heatmap shows trajectory analysis based on chromatin accessibility.

(d) Feature plots of transcription factor enriched motif in iPT_HAVCR1+ cell type.