Title: Spatially resolved human kidney multi-omics single cell atlas highlights the key role of the fibrotic microenvironment in kidney disease progression

Name of Authors:
Amin Abedini[1,2,3,4,*], Jonathan Levinsohn[1,2,3,4,*], Konstantin A Klötzer[1,2,3,4], Bernhard Dumoulin[1,2,3,4], Ziyuan Ma[1,2,3], Julia Frederick[1,2,3], Poonam Dhillon[1,2,3], Michael S Balzer[1,2,3], Rojesh Shrestha[1,2,3], Hongbo Liu[1,2,3], Steven Vitale[1,2,3], Kishor Devalaraja-Narashimha[6], Paola Grandi[6], Tanmoy Bhattacharyya[7], Erding Hu[7], Steven S. Pullen[8], Carine M Boustany-Kari[8], Paolo Guarnieri [8], Anil Karihaloo[9], Daniel Traum[2,4], Hanying Yan[10], Kyle Coleman[10], Matthew Palmer[11], Lea Sarov-Blat[7], Lori Morton[5], Christopher A. Hunter[12], Klaus H Kaestner[2,4] Mingyao Li[10], Katalin Susztak[1,2,3,4]

[1]Renal, Electrolyte, and Hypertension Division, Department of Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA
[2]Institute for Diabetes, Obesity, and Metabolism, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA
[3]Penn/CHOP Kidney Innovation Center, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA
[4]Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA
[5]Cardiovascular, Renal and Fibrosis Research, Regeneron Pharmaceuticals Inc., Tarrytown, NY 10591, USA
[6]Research and Development, GSK, Cellzome GmbH, Genomic Sciences, GSK, Heidelberg, Germany
[7]Research and Development, GSK, Crescent Drive, Philadelphia, Pennsylvania, USA
[8]Department of Cardiometabolic Diseases Research, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA.
[9]Novo Nordisk Research Center Seattle, Inc., Seattle, USA
[10]Department of Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[11]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, USA
[12]Department of Pathobiology, University of Pennsylvania School of Veterinary Medicine, Philadelphia, PA, USA

Correspondence:
Katalin Susztak, MD, PhD
Professor of Medicine
University of Pennsylvania, Perelman School of Medicine
3400 Civic Center Blvd,
Smilow Translational building 12-123,
Philadelphia, PA 19104
Phone: (215) 898-2009
ksusztak@pennmedicine.upenn.edu

Abstract word count: 179

Total word count: 4766

Abstract

Kidneys possess one of the most intricate three-dimensional cellular structures in the body, yet the spatial and molecular principles of kidney health and disease remain inadequately understood. Here, we have generated high-quality datasets for 81 samples, including single cell (sc), single nuclear (sn), spot level (Visium) and single cell resolution (CosMx) spatial (sp)-RNA expression, and sn open chromatin, capturing cells from healthy, diabetic, and hypertensive diseased human kidneys. By combining the snRNA, snATAC and scRNA sequencing we identify cell types and map these cell types to their locations within the tissue. Unbiased deconvolution of the spatial data identifies 4 distinct spatial microenvironments: glomerular, immune, tubule and fibrotic. We describe the complex, heterogenous cellular and spatial organization of human microenvironments in health and disease. Further, we find that the fibrotic microenvironment spatial gene signature is not only able to molecularly classify human kidneys, but it also offers an improved prognosis prediction compared to traditional histopathological analysis. We provide a comprehensive spatially resolved molecular roadmap of the human kidney and the fibrotic process, demonstrating the clinical utility of spatial transcriptomics.

59 **Introduction**

60 Human kidneys filter over 140 liters of plasma daily, reabsorb vital nutrients, excrete water and electrolytes,
61 and eliminate toxins to maintain the internal milieu[1,2]. Kidney disease is defined by a decline in glomerular
62 filtration. Chronic kidney disease (CKD) is the 9th leading cause of death[3,4] in the United States, affecting 14% of
63 the population. Diabetes and hypertension are responsible for more than 75% of all CKD[5].

64 More than 30 specialized cell types including epithelial, endothelial, interstitial, and immune cells, have been
65 identified in the kidney[6,7]. The development of single cell and single nuclear RNA sequencing (scRNAseq,
66 snRNAseq, respectively) as well as single nuclei Assay for Transposase-Accessible Chromatin sequencing
67 (snATACseq) has provided an unprecedented insight into the molecular and cellular composition of healthy
68 mouse and human kidneys, including changes during development and disease[8-12]. These methods use
69 dissociated cells or nuclei isolated from kidney tissue samples. Despite the significant cellular diversity of the
70 kidney, cell types could be accurately identified in scRNAseq and snRNAseq data, as their specialized cellular
71 function matches with the cellular gene expression signatures[13].

72 The lack of spatial information impedes the accurate mapping of known cell types only described by their
73 anatomical location. This limitation hampers the interrogation of local gene expression changes and cell-cell
74 communication, both of which play a vital role in maintaining cellular health and can be dysregulated in disease.
75 The field of spatial omics is rapidly evolving, but currently available spatial datasets either lack single cell
76 resolution information, unable to provide genome-scale gene expression data, or only evaluate a small spatial
77 area. Clearly, there is a pressing need for large-scale spatial omics datasets to better understand kidney health
78 and disease.

79 Chronic kidney disease (CKD) is associated with a complex change in the kidney's cellular structure[14]. Some
80 histologic changes are specific for disease subtypes; for instance, diffuse thickening of glomerular basement
81 membrane is observed in diabetic kidney disease (DKD)[15]. However, fibrosis is a common manifestation across
82 all forms of progressive CKD. The histologic hallmark of fibrosis is the accumulation of extracellular matrix[16,17],
83 which has been the primary focus of most prior studies. Matrix accumulation can cause organ stiffness, which
84 is implicated in organ failure in cases of pulmonary and heart fibrosis[18-20]. The role of tissue elasticity in the
85 regulation of kidney function is not clear[21], and thus, the mechanism by which matrix accumulation (or fibrosis)
86 influences kidney function remains a subject of controversy[22,23].

87 One of the most significant clinical problems is identifying patients who will develop end-stage renal failure,
88 requiring either a transplant or dialysis. Clinical parameters, including eGFR and albuminuria, can predict kidney
89 function decline with modest accuracy for patients with advanced CKD stages [24-26]. However, for patients in the
90 earlier stages of disease, accurate prognosis remains a challenge[25]. Although some studies suggest that the
91 severity of fibrosis can serve as a predictor, this finding has not been consistently reproduced[27-29].

92 Here, we generate both spot (Visium) and single cell level (CosMx) spatial transcriptomic data for both healthy
93 and diseased human kidneys in conjunction with sn/scRNAseq and snATACseq. By combining spatial gene
94 expression with high quality single cell expression and open chromatin information, we resolve the identity of
95 cells previously only known by their spatial localization. Moreover, we perform a detailed spatial and cellular
96 characterization of tissue fibrosis. We demonstrate the cellular heterogeneity of the fibrotic stroma, which
97 includes not only fibroblasts and myofibroblasts but also endothelial cells and immune cells that follow the
98 organization of a lymphoid organ and are anatomically close to injured proximal tubule cells. We define various
99 tissue microenvironments, including the fibrotic microenvironment (FME). Ultimately, our data show that the
100 FME gene signature can classify kidney samples and predict future kidney function decline. These findings
101 demonstrate the potential clinical utility of spatial transcriptomics.

102

103 **Results**

104 **Spatially resolved multi-omics single cell survey of healthy and diseased human kidneys defines expression,**
105 **gene regulation and spatial location of >100 cell types and states.**

106

107 We generated a comprehensive human kidney single cell atlas with spatial resolution by analyzing 81 human
108 kidney tissue samples from 58 subjects (62% male, and age of 64.1 ± 14.0 years). Samples were divided into two
109 groups: (i) healthy control (N = 36) determined by estimated glomerular filtration rate (eGFR) > 60 ml/min/1.73
110 m^2 and fibrosis score of < 10 % (ii) chronic kidney disease (CKD) (N = 45) determined by (eGFR) < 60 ml/min/1.73
111 m^2 or kidney fibrosis score of > 10%, including 20 with the clinical histopathological diagnosis of diabetic kidney
112 disease (DKD) and 25 with the clinical diagnosis of hypertensive-attributed chronic kidney disease.
113 **Supplementary Table 1** shows the demographic, clinical, and histological characteristics of the included
114 samples.
115 We performed droplet-based single cell analysis using 10X Chromium Next GEM for sc/snRNAseq (N = 47),
116 snATACseq (N = 18). After standard processing and meticulous quality control (**Supplementary Fig. 1** and

117 **Supplementary Table 2**), during which we removed low-quality cells, we included 338,565 cells/nuclei in our
118 final atlas. We used SCVI tools to generate a single unified comprehensive integrated human kidney atlas
119 including all different modalities and disease states[30]. Overall, we identified six cell super families, including
120 endothelial cells, stromal cells, tubule epithelial cells, immune cell types, glomerular cells, and neural/Schwann
121 cells. Within these large families we identified 44 main and 114 distinct cell sub-types or states in healthy and
122 diseased human kidneys (**Fig. 1a**). Key cluster-specific gene markers are shown in **Supplementary Fig. 2** and
123 **Supplementary Tables 3 to 6**. Our sc and sn human kidney atlas captured kidney cell types in healthy and
124 diseased states in all anatomical regions of the kidney. The main identified cell types were: podocytes, different
125 types of proximal tubule segments 1-3 (PT_S1, S2, S3, and injured), descending thin loop of Henle, ascending
126 thin loop of Henle, cortical and medullary thick ascending loop of Henle (C_TAL and M_TAL), distal convoluted
127 tubule (DCT), connecting tubule (CNT), principal cells of collecting duct (PC), intercalated cells type alpha and
128 beta (IC_A and IC_B), stromal, endothelial, and different types of immune cells.
129
130 The combination of single cell and single nuclear methods, the large number of analyzed cells, the high-quality
131 dataset, and the inclusion of samples with different degrees of kidney disease severity in our kidney atlas
132 enabled the capture of rare and novel cell types. We captured 11 different types of endothelial cells, including
133 afferent, efferent arterioles (*FBLN5*+) and vasa recta (*MCTP1*+) cells (**Fig. 1b, c, and Supplementary Fig. 2,3**).
134 We could also identify rare erythropoietin producing cells within the stromal cell population (**Supplementary**
135 **Fig. 3**). We captured proximal tubule (PT) cells expressing high levels of *SLC47A2,* a gene specific for toxin
136 excretion (**Supplementary Fig. 2, 3 and Supplementary Table 6)** and tubule epithelial subtypes mostly seen in
137 diseased kidneys that were positive for *IL18*, *WFDC2*, *SPP1*, and *ITGB6*. Our atlas provides a comprehensive
138 reference for human kidney immune cells. We identified lymphoid (CD4T, CD8T, natural killer cells, double-
139 negative T cells, Th17, B_Naive, B_intermediate B_memory, plasma_cells) and myeloid cells (neutrophil,
140 basophil/mast cells, CD14_monocyte, CD16_monocyte, macrophage, classical and plasmacytoid dendritic cells).
141 We provide comprehensive hierarchical clustering, including information on sex, disease status, sample identity,
142 and analytical methods (**Fig. 1d).** Furthermore, we created an easy-to-use web interface to examine gene
143 expression enabling users to identify cell type-specific and disease state-specific changes in the human kidney
144 (www.susztaklab.com/hk_genemap/)
145

146 To further improve and validate our cell type and state identification, we integrated our human kidney atlas
147 with the human kidney precision medicine project (KPMP) dataset[31] (**Fig. 1c, Supplementary Fig. 4,5**). This
148 combined fully integrated atlas contains ~600K cells or nuclei using three different analytical modalities
149 validating the consistency of the cell type annotations between datasets (**Supplementary Fig. 4, 5).**
150
151 In addition to the gene expression data, the snATACseq of 57,847 human kidney nuclei provided us
152 opportunities to identify transcription factors (TF) and enriched TF motifs in each cell type. Cell gene-expression
153 markers, a comprehensive list of cell-type differentially accessible regions and transcription factors can be found
154 in **Supplementary Fig. 6** and **Supplementary Table 7** and include *WT1* for podocyte and parietal epithelial cells
155 (PEC), *HNF4A* for PT cell types, *FOSL2* for injured_PT (iPT), and *TFAP2A* for C_TAL (**Supplementary Fig. 6)**.
156
157 **High-resolution spatially resolved human kidney atlas**
158 A key limitation to cell type identification has been the lack of high-resolution, spatially resolved cell
159 transcriptomics information. To overcome this limitation, we used the new Visium FFPE platform and generated
160 14 spRNAseq data sets, including 3 control (healthy) and 11 diseased samples (7 DKD, 4 HKD) (**Supplementary**
161 **Table 2, Supplementary Fig. 7-9**).
162
163 We then leveraged our dissociated sn/sc data to identify cell types within our spatial dataset. Firstly, for the
164 Visium data we used the Cell2location[32] package, which is a machine learning method that estimates the
165 contribution of each cell type to the observed gene expression profiles. We then employed CellTrek[33], which
166 uses the dissociated data to impute the spatial location of cell types to near single cell level (149,717 datapoints
167 after imputation). The high-resolution data enabled the projection of all identified cell types from the
168 dissociated datasets to their spatial location (**Fig. 2a, b**). We could also identify markers for cell types previously
169 only known by their anatomical location; for instance, PEC cells expressing *CFH*, and *WT1,* as well as mesangial
170 cells expressing *ITGA8* and *POSTN*. The gene expression-based spatial map was consistent with the Human
171 Protein Atlas[34] data (**Supplementary Fig. 10**) . Careful examination of the Cell2Location and CellTrek-based cell
172 type prediction showed the expected anatomical localization (i.e. glomerular cells colocalizing at morphologic
173 glomeruli, immune cells localizing with histologic immune infiltrate) (**Supplementary Fig. 11 to 14)**.
174

175 Finally, we then used the CosMx (NanoString) platform to assay human kidney samples (1 healthy, 1 DKD), at
176 true single cell resolution analyzing the expression of 1,000 genes (**Supplementary Fig. 7, 15) and**
177 **Supplementary Table 2**). The method generated high quality datasets and distinct UMAP cell clusters that
178 corresponded to specific kidney cell types (**Supplementary Fig. 16**). The CosMx dataset seamlessly integrated
179 with our single nuclei RNA dissociated cell data, further confirming the proper annotation of cell types
180 (**Supplementary Fig. 17**). Furthermore, the gene expression-based cell type prediction fully matched with
181 anatomical and histological annotation of the slide (**Fig. 2b**). The processed spatial transcriptomics (Visium and
182 CosMx) dataset is fully available for the community and on our interactive website
183 (www.susztaklab.com/hk_genemap/).https://susztaklab.com/samui
184
185 Overall, we constructed an integrated high-quality spatially resolved human kidney multi-omics atlas, which
186 allowed the spatial mapping of cellular gene expression and gene regulatory information in healthy and
187 diseased states.
188
189 **Stromal cells heterogeneity and matrix production in human kidney disease**
190
191 We next analyzed kidney fibrosis associated changes, as fibrosis is the common manifestation of all progressive
192 chronic kidney disease. We created an extracellular matrix (ECM) score by combining the expression of
193 collagens, glycoproteins, and proteoglycans[35,36]. **Fig. 3a** shows that myofibroblasts, followed by fibroblasts, and
194 then vascular smooth muscle cells (VSMC)/pericytes had the highest ECM score. We noted that disease samples
195 have a higher frequency of myofibroblasts compared to healthy samples **(Fig. 3b**) and myofibroblasts were
196 located within high ECM score area (**Fig. 3c**).
197
198 Our combined dataset enabled the identification of stromal cell subtypes that were consistent with prior
199 publications and datasets (**Supplementary Fig. 17,18**) and with correct spatial location (**Supplementary Fig. 19-**
200 **20)**. As previously known, stromal cells were positive for *PDGFRB* expression. We were able to identify markers
201 for mesangial cells (*ITGA8*, *POSTN*), VSMC/pericytes (*MYH11*, *NOTCH3*, *NTRK3*), fibroblast (*KCNK2*, *FAP*), and
202 myofibroblast (*COL1A1*, *SYNPO2*) (**Fig. 3d, e, Supplementary Fig. 21**). We validated the consistency of stromal
203 subtypes with prior publications using the MetaNeighbor tool[37] (**Supplementary Fig. 22**) and we verified their

204 spatial location (**Fig. 3f**). The CosMx dataset separated 2 main stromal clusters, one of which comprised
205 VSMC/pericyte/mesangial cells and the other fibroblasts/myofibroblasts. (**Supplementary Fig. 21**).
206
207 Further sub-clustering analysis of stromal cells distinguished 12 different cell subtypes (**Supplementary Fig. 22)**.
208 We captured medullary fibroblasts expressing *SYT1* and *NCAM1* and four different myofibroblasts marked by
209 *COL1A1*+, *CLMP*+, *FGF7*+, or *ITGBL1*+ expression (**Supplementary Fig. 22)**. We could discriminate VSMC from
210 myofibroblasts based on the expression of *MYH11*, *NOTCH3, CNN1*, and *NTRK3* (**Fig 3f, Supplementary Fig. 21).**
211 These annotations were also consistent with protein expression in the Human Protein Atlas (**Supplementary**
212 **Fig. 23**) and snATACseq analysis (**Supplementary Fig. 24**). Gene ontology analysis highlighted important
213 differences between the different stromal cells, such as myofibroblasts showed enrichment for focal adhesion,
214 and TGFB-signaling pathway genes (**Fig. 3g)**. Weighted gene co-expression network analysis (WGCNA) [38,39]
215 indicated enrichment for *GUCY1A2*, *LAMA2*, and *MYO1D* in the fibroblast module and *COL1A1*, *ZEB2*, *SVEP1*,
216 and *PTEN* in the myofibroblast module in disease states (**Supplementary Fig. 21, Supplementary Table 8)**. In
217 the snATACseq data, we could identify enrichment for *NPAS2*, *TEAD2*, and *TCF7* motifs in open chromatin areas
218 in myofibroblasts and *USF2* transcription factor motifs in VSMC (**Supplementary Fig. 24**).
219
220 In summary, we have generated a high-resolution, spatially resolved stromal cell atlas and identified ECM
221 producing cells in diseased human kidneys.
222
223 **The kidney fibrotic microenvironment is established by the complex interaction of injured epithelial cells,**
224 **immune cells, and endothelial cells.**
225
226 Our newly generated spRNAseq dataset is uniquely suited to defining microenvironments (ME) in the human
227 kidney. We ran nonnegative matrix factorization (NMF) on the Visium datasets using the STutility[40] package. We
228 found four major MEs in the human kidney, that visually corresponded to glomerular, tubule, fibrotic, and
229 immune regions. Gene ontology enrichment analysis of genes detected in each microenvironment was
230 consistent with their anatomical annotation (**Supplementary Fig. 25).** The computationally defined fibrotic
231 microenvironment (FME) strongly correlated with kidney ECM gene expression (**Supplementary Fig. 26**) and our
232 pathologist's assessment of fibrosis (r=0.75, *p*=0.003) (**Fig. 4a, Supplementary Fig. 26)**.

233 We also identified a specific immune ME. The immune ME regions were located within the FME, but with patchy
234 distribution. The immune ME consisted of dendritic cells, plasma cells, and B- and T- lymphocytes (**Fig. 4b, c**,
235 **Supplementary Fig. 26, 27**). The immune ME organizations resembled early tertiary lymphoid structures[41].
236 Immunostaining studies with cell type specific antibodies validated the presence of these specific immune cells
237 and immune cell aggregates (**Supplementary Fig. 28**). Furthermore, we have performed *in situ* mass
238 spectrometry (IMC) to understand protein expression levels. Consistent with the Visium data, we confirmed the
239 protein expression of HAVCR1+-positive iPT cells, B-cells, CD4T-cells, plasma, myeloid and endothelial cells by
240 IMC (**Fig. 4d**, **Supplementary Fig. 29).**
241 To further understand cell interactions in FMEs, we implemented CellChat[42] on the integrated sc/snRNAseq and
242 snATACseq datasets as well as communication analysis by Optimal Transport (COMMOT)[43] on the spRNAseq
243 datasets. We found enrichment for *SPP1*, *CXCL12*, *CCL19*, *CCL21, PDGFB,* and *TGFB1* and their receptors in FME
244 regions (**Fig. 4e, Supplementary Fig. 30, Supplementary Table 10**). The iPT cells in the FMEs expressed *SPP1* and
245 *PDGFB* and showed a strong interaction with stromal cells. Also, we found endothelial cells expressing *CD34* and
246 *CDH5* in the FME regions (**Supplementary Fig. 30)**. The stromal cells in FME were enriched for chemotactic
247 factors including *CXCL12*, *CCL19*, *CCL21*, while their receptors were expressed in different immune cells,
248 suggesting that stromal cells might signal to immune cells. We observed *PDGFB* and *TGFB1*, known mediators
249 of fibrosis, in FME associated immune aggregates (**Supplementary Fig. 30**)[43] **.**
250 Overall, using unbiased computational tools, we were able to identify spatial regions in the kidney, including
251 glomerular, tubular regions in healthy samples and fibrotic and immune regions in diseased kidneys. Most
252 importantly, FMEs were not only characterized by matrix-producing fibroblasts, but also associated iPT-stromal
253 ligand-receptor interactions.
254
255 **Injured tubule cells in human kidney fibrosis.**
256
257 Cell type enrichment analysis of the fibrotic microenvironment indicated fibroblast, myofibroblast, immune
258 cells, endothelial cells and injured tubule cells including injured proximal tubule cells (iPT) in FMEs (**Fig. 4e**). iPT
259 cells have not been well characterized, but they have been identified in diseased mouse and human kidney
260 samples [10,11,44,45]. iPT cells were physically close to proximal tubule cells **(Fig. 5a, 4c, Supplementary Fig. 33).**
261 Further, we note that PT cells number correlated with eGFR, fibrosis and disease state within our cohort (**Fig 5b,**
262 **c**) and showed the most drastic decline in CKD **(Supplementary Fig. 32).** In contrast, we observed a clear

263 enrichment of iPT cells in the fibrotic microenvironment and increased number in diseased samples **(Fig. 5d).**
264 Upon careful sub-clustering, of our CosMx data, which we verified by examining our annotations in space, We
265 note that *VCAM1* expression in both PEC and iPT, while *HAVCR1* was specific for iPTs **(Fig. 5e, Supplementary**
266 **Fig. 20a)**.
267 We then sought to examine iPT heterogeneity, starting with our Visium data. We used the single cell co-
268 expression (SCoexp) module of CellTrek[33]. We identified two different iPT modules, corresponding to two iPT
269 subtypes in diseased samples (**Fig. 5f**) and only one iPT type in healthy samples (**Supplementary Fig. 34**). Moving
270 back to the rich integrated atlas, we found that one of the iPT cluster expressed *VCAM1*, *ACSL1*, *ASS1*, and *ASPA*;
271 genes playing role in cellular metabolism. We called this cluster iPT_VCAM1+. This cluster was present in healthy
272 samples and this cluster was close to other injured PT cells in the integrated dataset expressing mitochondrial
273 genes such as *MT1* (**Fig. 5f**). The second iPT cluster expressed *HAVCR1* (or *KIM1*), *NFKBIZ*, *IL18*, *ITGA3*, *PDGFB*,
274 and *ITGB1* and was enriched for the expression of genes associated with cell adhesion and matrix (iPT-HAVCR1+)
275 (**Supplementary Fig. 34**). Both in the Visium spatial and dissociated datasets, most iPT-HAVCR1+ cells were in
276 CKD samples (**Fig. 5g**). Pseudo-time analysis of PT and iPT cells using Monocle, indicated a continuous
277 differentiation path of PT cells to iPT_VCAM1+ and iPT_HAVCR1+ (**Fig. 5h, Supplementary Table 11)**. Studies
278 going back to a decade ago have identified HAVCR1+ as an injured tubule marker[10]. These two different iPT cell
279 subtypes were also present in a recently published mouse DKD snRNAseq[45] dataset (**Supplementary Fig. 35**) as
280 well as in our snATACseq data (**Supplementary Fig. 36)**. We identified *HNF1A* and *BACH2* as enriched TFs for
281 iPT_VCAM1+ and iPT_HAVCR1+, respectively (**Supplementary Fig. 36**).
282 We then subclustered iPT cells using our positionally verified CosMx data (Fig. 5i, Supplementary Fig. 31,
283 34). We noted upon sub-clustering iPT heterogeneity that corresponded with neighboring cell identity
284 (Supplementary Fig. 31), supporting our Visium data that iPT subpopulations co-localize and may be related to
285 the underlying microenvironment. In the CosMx dataset, APOE, SPP1 and KRT7 (**Supplemental Figure 31b**)
286 marked these subclusters. Our identification of these specific markers, as opposed to other iPT markers, may
287 be related to the relative paucity of transcripts used by the CosMx platform compared to whole transcriptome
288 approaches.
289 In summary, the different types of single cell omics (snRNA, scRNA, snATAC) data indicated highly plastic PT cell
290 population, including iPT cells. Our spatial transcriptomics analysis highlighted different types of iPT cells
291 (*VCAM1*+ and *HAVCR1*+). We observed *VCAM1*+ iPT and PEC cells in non-CKD adult kidney samples, while
292 HAVCR+ cells were enriched in the fibrotic microenvironment and diseased kidneys.

293

294 **Fibrotic microenvironment gene signature successfully predicts disease prognosis in a large cohort of human**
295 **kidney samples.**

296 At present, kidney disease classification is performed by an expert renal pathologist. We wanted to understand
297 whether our spRNAseq can aid disease classification and prognosis evaluation. We first generated an FME gene
298 signature (FME-GS) (**Supplementary Table 12**) and used this signature to analyze gene expression data from a
299 large external kidney cohort containing 292 human kidneys (**Fig. 6a**), including healthy samples and samples
300 with varying severity of CKD (attributed to diabetes and hypertension). Our FME-GS successfully discriminated
301 samples with more severe disease (higher fibrosis and lower eGFR) from healthier samples (**Fig. 6b,**
302 **Supplementary Table 13**) despite the fact, that these parameters were not included in the clustering algorithm.
303 Next, we wanted to understand whether FME signature can predict kidney disease prognosis. We found that
304 cluster 1 subgroup, samples with the highest FME-GS load, had 4x higher chance of reaching endpoint, defined
305 as 40% change in eGFR or reaching end stage renal failure (eGFR< 15 ml/min) **(Fig. 6c)**. Subgroup analysis
306 however, indicated that samples with the highest FME-GS has statistically higher degree of fibrosis 12.1% vs
307 6.51%. While this difference is not clinically meaningful the differences was statistically significant.

308 Finally, we wanted to understand whether the FME signature could also predict renal disease progression in
309 patients with preserved eGFR (mean eGFR 70 ml/min/1.73m^2) and no structural damage (interstitial fibrosis
310 less than 10%). These parameters are used to define the lack of structural and functional kidney damage
311 (CKD/fibrosis)[46] **(Fig. 6c)**. The pathologist reported fibrosis score could not predict the outcome for these
312 samples (**HR = 1.17, 95%CI: 0.49 – 2.83**) **(Fig. 6d)**. Kaplan-Meier analysis indicated that cluster 1, with the highest
313 FME-GS load, had significantly higher hazard ratio to reach the endpoint (**HR = 3.95, 95%CI: 1.71 – 9.12**) **(Fig.**
314 **6d)**. While the entire FME gene set better predicted outcome, employing only the top 10 genes from LASSO
315 regression still showed a statistically significant enrichment for renal failure (**Supplementary Fig. 37, p=0.021,**
316 **Supplementary Table 14**). Clustering samples using a random gene set did not correlate with outcome
317 (**Supplementary Fig. 37**). Analyzing the relationship between cell types and kidney disease progression, we
318 found that genes that correlated with eGFR slope were enriched in iPT and stromal cells, potentially implying
319 their causal role (**Supplementary Fig. 37**).

320 In summary, our spatially derived FME-GS can identify subjects with progressive kidney function decline in a
321 large cohort.

322

323 **Discussion**

324 Here we present the spatial molecular principles of kidney health and disease by generating a comprehensive
325 and spatially resolved human kidney atlas and combining single cell omics data (scRNAseq, snRNAseq, and
326 snATACseq) from a large number of human kidney tissue samples with varying degree of disease severity. Our
327 work fills a critical knowledge gap by characterizing the cellular gene expression program of cells previously only
328 defined by their spatial location, showing the anatomical location of cells only observed in dissociated single cell
329 expression data, and defining cell-type specific gene expression changes in diseased states. We define the
330 cellular complexity of the fibrotic microenvironment as the intricate interaction of many different cell types. We
331 furthermore demonstrate the clinical prognostic value of spatial transcriptomics even for samples where the
332 current pathological analysis loses sensitivity.

333

334 Previous single cell analyses, focusing on dissociated human and mouse kidney datasets, have generated gene
335 expression and regulatory matrices for a variety of kidney cell types[8-12,47]. As kidney cell types have been
336 functionally well characterized, most identified cell types have been matched back to a more than half-century
337 old functional cell type definition[6]. A key limitation of these analyses has been the identification and molecular
338 characterization of anatomically defined cell types, such as mesangial cells, PECs, and fibroblasts. Here, we
339 demonstrate that a joint approach, which combines single cell and single nuclear expression, open chromatin,
340 and single cell and spot level spRNAseq data from many diverse samples and a large number of cells, is critical
341 to achieving this goal. The orthogonal analytical tools provide unique opportunities for validation, as each
342 method suffers from specific technological biases. Here, we have not only been able to resolve and validate
343 previously anatomically-known cell types but also identify novel cell types such as specific stromal cells for
344 glomerulosclerosis (expressing *CDH13)*[48]. Furthermore, we clarify genes expressed by cells previously only
345 identified by their anatomical location, such as PECs.

346

347 Fibrotic diseases are responsible for close to 40% of all deaths[49]. Kidney fibrosis is the final common pathway to
348 end stage kidney failure[50]. Combining snRNA and spatial information and multiple datasets we characterize
349 stromal cell subtypes, and we validate their consistency between different datasets. We could conclusively
350 discriminate VSMC and mesangial cells from myofibroblasts that are anatomically distinct but share gene
351 expression signatures in dissociated sc/snRNAseq data[51-53]. We could identify novel markers and, ultimately,
352 new stromal cell types and determine their spatial location. This information could be important in the field of

353 finding therapeutic candidates for renal fibrosis. We noted a cluster of FAP-positive fibroblasts in diseased
354 human kidneys[54-56]. FAP targeted cellular and RNA therapies have been developed and shown to have efficacy
355 in animal models of cardiac fibrosis[54-56]. Our data suggests that these therapeutics may be helpful for treating
356 kidney fibrosis.
357
358 Here, we demonstrate the cellular and architectural complexity of kidney fibrosis. We propose the use of the
359 fibrotic microenvironment to characterize these lesions, to not only focus on matrix accumulation but on the
360 elaborate cellular complexity of these lesions. We show that FME signature contains tubule cells including
361 injured PT. While iPT cells have been observed in dissociated single cell data, here we show that these cells can
362 be consistently identified in all datasets and contain consistent markers. We furthermore identify spatially
363 defined iPT subtypes [10]. We show that in humans *VCAM1*+ cells can be observed in subjects with preserved
364 kidney function. This might be consistent with observations that we show that some parietal epithelial cells
365 express VCAM1 or that low degree of fibrosis is observed in some healthy subjects or during aging. We identify
366 HAVCR1 positive iPT cells that almost exclusively present in CKD samples. *HACVR1* has been identified as marker
367 of acute and chronic kidney disease.[57]
368
369 Immune cell clusters have long been observed in fibrotic kidney samples, even in patients with non-immune-
370 mediated kidney disease, such as diabetes and hypertension[50,58]. Here, we resolve these regions both spatially
371 and at a cell type level. The interaction of a large number of cell types, including iPT, immune, stromal, and
372 endothelial cells, establishes the FME. In patient samples, there is a strong interaction between stromal and
373 immune cells and also signaling by immune and stromal cells to iPT, which might play a role in maintaining their
374 injured PT state. Our kidney scRNAseq data was enriched for immune cells and enabled us to spatially resolve
375 immune cell types and determine the distributions of immune cells in the kidney. We show that immune cell
376 clusters (the immune microenvironment) are localized mostly within some FMEs and to some degree they show
377 a resemblance to the tertiary lymphoid structures (TLS). TLS are organized aggregates of immune cells that form
378 postnatally in nonlymphoid tissues, usually as a persistent antigen production[59], and generate autoreactive
379 effector cells. TLS have been earlier described in mouse kidney tissue samples[60-63].
380
381 One of the most devastating complications of CKD is its progression to ESRD, which requires life-sustaining
382 dialysis or transplantation [64]. Predicting which patients will have progressive disease is of major clinical

383 importance. This has been classically done using eGFR and the extent of fibrosis noted on biopsy for subjects
384 with more advanced disease (eGFR<60 ml/min/1.73m2). While some studies indicated that including the degree
385 of tissue fibrosis can improve kidney prognosis estimates, this has not been uniformly observed. By using the
386 FME-GS score, we were able to improve upon this. We identified subjects at risk of ESRD in a large external
387 dataset of human kidney tissue samples even for samples where traditional fibrosis score had limited predictive
388 value. These results establish FME-GS as a key biomarker and potentially as a causal mechanism of progression.
389 Future studies shall work on optimizing and validating the predictive gene signature, but our preliminary data
390 indicates that a relatively small number of genes might be able to discriminate those patients who progress to
391 renal failure.
392
393 Our study has several limitations. Here, we present the initial interpretation of one of the largest dissociated
394 and spatially resolved human kidney datasets. Here, we used the clinical and histological annotation for disease
395 definition, however, hypertensive kidney disease might not be a single disease entity, where hypertension might
396 not even be the single causal injury. While we applied a variety of tools and methods, future studies will be
397 essential for a more comprehensive analysis of the presented information. Our datasets indicate the presence
398 and key role of injured PT cells and immune cells in disease progression; however, their functional role should
399 be studied in mechanistic experiments using cells and animal models. Furthermore, the FME gene signature will
400 need to be validated in other cohorts.

401 In summary, we develop a spatially defined molecular human kidney cellular atlas, characterize the fibrotic
402 microenvironment, and indicate its role as a clinically meaningful prognostic disease biomarker, demonstrating
403 the utility of spRNAseq for the investigation of complex diseases.

404
405

406 **Methods:**

407 **Single nuclei RNA sequencing**

408 Nuclei were isolated using lysis buffer (Tris-HCl, NaCl, MgCl2, NP40 10%, and RNAse inhibitor (40 U/ul))**.** 10-30
409 mg of frozen kidney tissue was minced with a razor blade into 1-2 mm pieces in 1 ml of lysis buffer. The chopped
410 tissue was transferred into a gentleMACS C tube, and homogenized in 2 ml of lysis buffer using a gentleMACS
411 homogenizer with programs of Multi_E_01 and Multi_E_02 for 45 seconds. The homogenized tissue was filtered
412 through a 40 µm strainer (08-771-1, Fisher Scientific) and the strainer was washed with 4 ml wash buffer. Nuclei
413 were centrifuged at 500xg for 5 minutes at 4°C. The pellet was resuspended in wash buffer (PBS 1X + BSA 10%
414 (50 mg/ml), + RNAse inhibitor (40 U/ul)), filtered through a 40 µm Flowmi cell strainer (BAH136800040-50EA,
415 Sigma Aldrich). Nuclear quality was checked, and nuclei were counted. In accordance with the manufacturer's
416 instructions, 30,000 cells were loaded into the Chromium Controller (10X Genomics, PN-120223) on a Chromium
417 Next GEM chip G Single Cell Kit (10X Genomics, PN-1000120) generate single cell gel beads in the emulsion (10X
418 Genomics, PN-1000121). The Chromium Next GEM Single Cell 3' GEM Kit v3.1 (10X Genomics, PN-1000121) and
419 Single Index Kit T Set A (10X Genomics, PN-120262) were used in accordance with the manufacturer's
420 instructions to create the cDNA and library. Libraries were subjected to quality control using the Agilent
421 Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626). Libraries were sequenced using the
422 Illumina Novaseq 6000 system with 2 × 150 paired-end kits. Demultiplexing was as follows: 28 bp Read1 for cell
423 barcode and UMI, 8 bp I7 index for sample index, and 91 bp Read2 for transcript.

424 **Single nuclei ATAC sequencing**

425 The procedure described above for single nuclei RNA sequencing was used to isolate the nuclei for ATAC
426 sequencing. The resuspension was performed in diluted Nuclei Buffer (10X GEM). Nuclei quality and
427 concentration were measured in Countess AutoCounter (Invitrogen, C10227). Diluted nuclei were loaded and
428 incubated in chromium single cell ATAC library & gel bead kit's transposition mix (10X Genomics, PN-1000110).
429 Chromium Chip E (10X Genomics, PN-1000082) in the Chromium Controller was utilized to capture the GEMs.
430 The Chromium Single Cell ATAC Library & Gel Bead Kit and Chromium i7 Multiplex Kit N Set A (10X Genomics,
431 PN-1000084) were then used to create snATAC libraries in accordance with the manufacturer's instructions.
432 Library quality was examined using an Agilent Bioanalyzer High Sensitivity DNA kit. After sequencing on an
433 Illumina Novaseq system using two 50 bp paired-end kits, libraries were demultiplexed as follows: 50 bp Read1

434 for DNA fragments, 8 bp i7 index for sample index, 16 bp i5 index for cell barcodes, and 50 bp Read2 for DNA
435 fragments.

436 **Single Cell RNAseq**

437 Fresh human kidneys (up to 0.5 gr) collected in RPMI were minced into approximately 2-4 mm cubes using a
438 razor blade. The minced tissue was then transferred to a gentleMACS C tube containing Multi Tissue dissociation
439 kit 1 (Miltenyi, #130-110-201). The tissue was homogenized using the Multi_B program of the gentleMACS
440 dissociator. The tube, containing 100ul of Enzyme D, 50ul of Enzyme R, and 12.5ul of Enzyme A in 2.35 ml of
441 RPMI, was incubated for 30 mins at 37 ^{o}C. Second homogenization was performed using Multi_B program on
442 gentleMACS dissociator. The solution was then passed through a 70 um cell strainer. After centrifugation at
443 1,200 RPM for 7 mins, cell pellet was incubated with 1ml of RBC lysis buffer on ice for 3mins. The reaction was
444 stopped by adding 10 ml PBS. Next, the solution centrifuged at 1,000 RPM for 5 minutes. Finally, after removing
445 the supernatant, the pellet was resuspended in PBS. Cell number and viability were analyzed using Countess
446 AutoCounter (Invitrogen, C10227). This method generated a single cell suspension with greater than 80%
447 viability. Next, 30,000 cells were loaded into the Chromium Controller (10X Genomics, PN-120223) on a
448 Chromium Next GEM chip G Single Cell Kit (10X Genomics, PN-1000120) to generate single cell gel beads in the
449 emulsion (GEM) according to the manufacturer's protocol (10X Genomics, PN-1000121). The cDNA and library
450 were made using the Chromium Next GEM Single Cell 3' GEM Kit v3.1 (10X Genomics, PN-1000121) and Single
451 Index Kit T Set A (10X Genomics, PN-120262) according to the manufacturer's protocol. Quality control for the
452 libraries was performed using Agilent Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626).
453 Libraries were sequenced on Illumina Novaseq 6000 system with 2 × 150 paired-end kits using the following
454 demultiplexing: 28 bp Read1 for cell barcode and UMI, 8 bp I7 index for sample index and 91 bp Read2 for
455 transcript.

456 **Visium FFPE sample preparation**

457 The RNA quality of human kidney FFPE sample was checked by RNeasy FFPE kit (Qiagen-Cat #73504) according
458 to the manufacturer's protocol. RNA quality was examined using Agilent bioanalyzer, and samples with
459 DV200>50% were selected. Then a 5 µm tissue sample was cut onto the 10x Visium Spatial Gene Expression
460 Slide. After deparaffinization, H & E staining was performed. We used Keyence 1266 BZ-X810 microscope for
461 whole slide imaging. After scanning, de-crosslinking, probe hybridization, probe release, and extension, library

462 preparation was performed using a Single Index Kit TS Set A (10X Genomics, PN-3000511) according to the
463 manufacturer's protocol. Quality control for the libraries was performed using the Agilent Bioanalyzer High
464 Sensitivity DNA kit (Agilent Technologies, 5067-4626). Libraries were sequenced using the Illumina Novaseq
465 6000 system with 2 × 150 paired-end kits. Demultiplexing was as follows: 28 bp Read1 for cell barcode and UMI,
466 10 bp I7 index, 10bp i5 index, and 50 bp Read2 for transcript.

467 **CosMx Sample preparation**

468 Samples were prepared according to manufacturer specifications. Tissue sections were cut at 5 um thickness
469 and prepared according to the manufacturer specifications (NanoString Technologies). We used the human
470 universal cell characterization RNA probes, plus 50 additional custom probes which were designed to the
471 following gene transcripts: *ESRRB, SLC12A1, UMOD, CD247, SLC8A1, SNTG1, SLC12A3, TRPM6, ACSL4, SCN2A,*
472 *SATB2, STOX2, EMCN, MEIS2, SEMA3A, PLVAP, NEGR1, SERPINE1, CSMD1, SLC26A7,SLC22A7, SLC4A9, SLC26A4,*
473 *CREB5, HAVCR1, REN, AP1S3, LAMA3, NOS1, PAPPA2, SYNPO2, RET, LHX1, SIX2, CITED1, WNT9B, AQP2, SCNN1G,*
474 *ALDH1A2, CFH, NTRK3, WT1, NPHS2, PTPRQ, CUBN, LRP2, SLC13A3, ACSM2B, SLC4A4, PARD3, XIST,UTY*. Samples
475 were imaged with configuration A. After imaging, the flowcell was kept in xylene overnight, the coverslip was
476 removed and the slide was stained with hematoxylin and eosin.

477 **Human Sample Acquisition**

478 Left-over kidney samples were irreversibly deidentified, and no personal identifiers gathered. Therefore, they
479 were exempt from IRB review (category 4). We engaged an external honest broker who was responsible for
480 clinical data collection without disclosing personal identifiable information. The University of Pennsylvania
481 institutional review board (IRB) gave its approval for the collection of human kidney tissue.

482 Cortical and outer medullary area of the human kidney was used for scRNAseq, snRNAseq, and snATACseq. In
483 addition, a portion of the tissue was formalin-fixed, paraffin-embedded, and stored for spatial trascriptomics,
484 as well as stained with periodic acid-Schiff for pathology scoring. A local renal pathologist performed objective
485 pathological scoring.

486

487 ***Immunostaining***

488 Paraffin blocks were sectioned and deparaffinized. For blocking, 1% bovine serum albumin was used. Slides were
489 then incubated overnight with diluted primary antibodies (CD4 CST [Catalogue #48374], IGKC: Biolegend
490 [Catalogue #392702], and CD79A Abcam [Catalogue #ab79414]). After washing the sections with PBS three
491 times, secondary antibodies were applied for 1 hour at room temperature. After DAPI staining, slides were
492 imaged with an OLYMPUS BX43 Microscope. Positive cells in ten randomly selected fields were counted on each
493 slide.

494

495 ***Imaging Mass Spectrometry***

496 Formalin-fixed paraffin-embedded (FFPE) tissue sections were deparaffinized and rehydrated using standard
497 protocols. Antigen retrieval was performed using citrate buffer (pH 6.0), and sections were blocked with a buffer
498 containing 5% bovine serum albumin (BSA). Primary antibodies against target proteins were conjugated with
499 specific metal isotopes (mention which isotopes) using the Maxpar Antibody Labeling Kit (Fluidigm). Tissue
500 sections were incubated with the antibody-metal conjugates overnight at 4°C. After washing, secondary
501 antibodies, conjugated with a different metal isotope (mention which isotopes), were added to amplify the
502 signal. Finally, the tissue sections were stained with an iridium intercalator for nuclear staining. IMC imaging
503 was performed using a Hyperion imaging system (Fluidigm) equipped with a 20X objective lens. The tissue
504 sections were ablated using a laser at a 1µm pixel size, and the resulting metal signals were detected by a time-
505 of-flight mass spectrometer.

506

507 **Bioinformatic analysis**

508 ***Primary single nuclei and cell RNAseq data processing***

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

511 ***Data Processing and Computational Analyses***

512 Ambient RNA was corrected using SoupX[65] and doublets were removed by DoubletFinder[66] with default
513 parameters. Seurat objects were created from the aligned outputs of multiple samples, retaining genes

514 expressed in more than three cells and cells with at least 300 genes. Further, a merged Seurat object was
515 obtained using the merge function in Seurat v4.0.366. The QC filters applied were: (a) cells with n_feature counts
516 more than 3000 or less than 200, (b) cells with more than 15% mitochondrial counts (for snRNAseq data) and
517 more than 50% mitochondrial counts (for scRNAseq data), and (c) cells with more than 15% ribosomal gene
518 counts.

519 ***Data Normalization and Cell Population Identification***

520 Initially, highly variable genes were identified using the vst method. The data was then natural log-transformed
521 and scaled. The scaled values underwent principle component analysis (PCA) for linear dimension reduction.
522 We used the RunHarmony function in the harmony[67] package for batch effect correction. A shared nearest
523 neighbor network was created based on Euclidean distances between cells in a multidimensional PC space (the
524 first 50 PCs were used for both snRNAseq and scRNAseq). This network was used to generate a 2-dimensional
525 Uniform Manifold Approximation and Projection (UMAP) for visualization.

526 To identify cell-type markers, we used the FindAllMarkers function in Seurat. This method calculates log fold
527 changes, percentages of expression within and outside a group, and p-values of the Wilcoxon-Rank Sum test
528 comparing a group to all cells outside that specific group, including adjustment for multiple testing. A log-fold-
529 change threshold of 0.25 and FDR<0.05 was considered significant. These steps were performed on the
530 snRNAseq and scRNAseq datasets separately. Clusters expressing multiple cell types of specific marker genes
531 were excluded as potential doublets. The dataset was further refined: first, clusters were annotated based on
532 well-known marker genes, then the dataset was projected to a published human kidney single cell[44] reference
533 using the CCA method. The FindTransferAnchors function was used to discover matching genes between the
534 two datasets by using shared correlation patterns in the gene activity matrix and snRNAseq dataset. Predicted
535 labels within the two datasets were identified using the TransferData method. Cells were considered doublets
536 and eliminated from the dataset if the original annotation was incompatible with the predicted cell type.

537

538 ***Single nuclei ATACseq analysis***

539 Raw FASTQ files were aligned to the GRCh38 reference genome and quantified using Cell Ranger ATAC (v. 1.1.0).
540 Outputs from 19 snATACseq datasets were embedded using Signac (v.1.3.0)[68] to generate Signac object. Low-
541 quality cells were removed from each snATAC object using the following criteria: peak_region_fragments < 3000

542 & peak_region_fragments > 20000 & pct_reads_in_peaks < 15 & nucleosome_signal > 4 & TSS.enrichment < 2).
543 The filtered cells in 19 objects were merged in Seurat. Dimension reduction was performed by singular value
544 decomposition (SVD) of the TFIDF matrix and UMAP. Batch effect was corrected using Harmony[67] via the
545 RunHarmony function in Seurat. A KNN graph was created to cluster cells using the Louvain algorithm. Peaks
546 observed in at least 20% of cells were evaluated for differentially accessible chromatin regions (DARs) between
547 different cell types using a likelihood ratio test, a log-fold-change threshold of 0.25, and an FDR of 0.05 with the
548 Signac FindMarkers function. To annotate the genomic regions harboring snATACseq peaks, ChIPSeeker
549 (v1.24.0)70 was used[69].
550

551 ***Removing the doublets from snATACseq dataset***

552 The GeneActivity tool in Signac was used to create a gene activity matrix after clustering the twenty integrated
553 snATACseq datasets using protein-coding genes annotated in the Ensembl database (this technique counts the
554 ATAC peaks inside the gene body and 2 kb upstream of TSS). Log normalization was applied to the gene activity
555 matrix.
556 A published snRNAseq[44] dataset was utilized as a reference, and the FindTransferAnchors function was used to
557 discover matching genes between the snRNAseq and snATACseq datasets by using shared correlation patterns
558 in the gene activity matrix and snRNAseq dataset. Subsequently, the predicted labels within two datasets were
559 identified using the TransferData method. If the original annotation was incompatible with the predicted cell
560 type, the cell was considered a doublet and was eliminated from the dataset.

561

562 ***Motif Enrichment Analysis and Motif Activities***

563 First, we used the AddMotifs function of Signac to perform motif enrichment analysis. This was done by creating
564 a matrix of positional weights for motif candidates from JASPAR2020. To identify potentially important cell-type-
565 specific regulatory sequences and motifs, we evaluated the differentially accessible peaks between each cell
566 type versus other cell types using the 'LR' test and min.pct = 0.05. Subsequently, a hypergeometric test was used
567 to determine the likelihood of observing a specific motif at a particular frequency by chance. We compared this
568 observation with a group of peaks that have the same GC content as the motif, which we considered the

569 background set. This analysis was performed using the Find Motifs function. We used the RunChromVAR
570 function in conjunction with chromVAR (v.1.6.0) [70] to ascertain transcription factor activity.

571 ***DARs between groups***

572 We used the FindMarkers function after selecting DefaultAssay as peaks to identify DARs in each cell type and
573 diseased and healthy conditions, with a log-fold-change threshold of 0.25 and FDR<0.05. Peaks translated to
574 related genes using ChIPSeeker (v1.21.1)[69].

575

576 ***Integration of snRNAseq, scRNAseq, and snATACseq datasets***

577 We used SCVI[30,71] to integrate scRNAseq, snRNAseq, and snATACseq data from multiple samples. Briefly, we
578 used the preprocessed and clean datasets previously described and then combined them into a single AnnData
579 object using the Scanpy v1.4.4 package. After normalization and identification of the 3000 highly variable genes
580 in the integrated dataset, we ran SCVI on the combined dataset to learn a joint latent representation of the
581 data, using the sample, nFeature_RNA, nCount_RNA, and tech factors as covariates. We utilized the scanorama
582 setting of SCVI to account for batch effects across samples. We incorporated four layers in the model and
583 reduced the dimensionality of the data to 45 dimensions using the SCVI's dimensionality reduction method.
584 Finally, we used the Scanpy package to perform downstream analysis on the integrated dataset, including
585 clustering, differential expression analysis, and visualization.

586 ***Ligand–receptor interactions in integrated sc/snRNAseq and snATACseq***

587 The CellChat[42] repository (v1.4.00 was used to assess cellular interactions between different cell types and to
588 infer cell–cell communication networks from integrated sc/snRNAseq and snATACseq data. Only receptors and
589 ligands expressed in more than 10 cells in each cluster were considered. Probability and *P* values were measured
590 for each interaction.

591 ***DEGs between diseased and healthy groups***

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

595 ***Single nuclei RNAseq trajectory analysis***

596 Proximal tubules (PT) and Injured PT cells were subclustered for the trajectory analysis. The trajectory analysis
597 was done in two steps. Different sub-types of iPT cells with equal numbers were randomly subsampled and cell
598 dataset object (CDS) was generated using Monocle3[72,73]. After preprocessing and batch effects correction, the
599 dataset was embedded for dimension reduction and pseudotemporal ordering. We used the order_cell function
600 and indicated the PT as start point for pseudotime analysis. The track genes algorithm was used to identify the
601 DEGs along the trajectories, and genes with q values of 0.05 or higher were considered significant.

602 **MetaNeighbor analysis**

603 MetaNeighbor[37] is a versatile computational tool designed to quantify the degree of cell type replicability
604 between datasets by assessing cell type-specific gene expression patterns. The method involves constructing a
605 cell-cell similarity network using a shared, reduced gene space and subsequently calculating the area under the
606 receiver operating characteristic curve (AUROC) to estimate cell type-specific gene expression conservation.
607 High AUROC values indicate a high degree of cell type replicability, whereas low values suggest limited
608 conservation. We followed the original protocol and used default settings for parameter selection, including the
609 50 nearest neighbors for network construction and 1,000 most variable genes for gene space reduction.

610 ***Single nuclei WGCNA***

611 We used the R packages hdWGCNA and WGCNA to perform WGCNA on the fibroblast and myofibroblast. We
612 used the hdWGCNA function to construct metacells to aggregate transcriptionally neighboring single cells into
613 pseudo-bulk metacells. Then, the metacells were processed by regular Seurat functions including
614 NormalizeMetacells , ScaleMetacells, RunPCAMetacells, RunHarmonyMetacells, and RunUMAPMetacells
615 were performed . Then, the expression matrix was set for fibroblasts and myofibroblasts separately. Next,
616 after selecting the appropriate soft power by running TestSoftPowers, the co-expression network was created.
617 In the next step harmonized module eigengenes were calculated to summarize the gene expression profile for
618 each gene and in each sample. Then, the module connectivity was calculated, genes assigned to each module
619 were extracted, and the results were shown as the average gene expression in all clusters. Also, the module
620 eigengenes for each group were calculated and shown by heatmap.
621

622 ***Visium data analysis***

623 Data was aligned using Space Ranger (v1.0.0) with the reference genome GRCh38 and the human probe dataset
624 (Visium_Human_Transcriptome_Probe_Set_v1.0_GRCh38). To correct for technical noise and improve the
625 quality of our spatial transcriptomics dataset, we used the SpotClean method[74]. SpotClean is a computational
626 method that leverages the negative control spots on the spatial transcriptomics array to remove technical noise
627 and improve the accuracy of the expression measurements. First, we preprocessed the spatial transcriptomics
628 data by filtering out lowly expressed genes and normalizing the expression values using the SCTransform
629 method. We then used the SpotClean package to clean the data using the negative control spots. Finally, we
630 performed downstream analysis on the cleaned expression matrix. This step was done for all 14 samples, which
631 we merged, using the merge function of Seurat. The data were subjected to principal component analysis (PCA)
632 for linear dimension reduction and Harmony was used to integrate the datasets. A shared nearest neighbor
633 network was created based on Euclidean distances between cells in a multidimensional PC space (30 PCs were
634 used) and a fixed number of neighbors per cell. This network was used to generate a 2-dimensional Uniform
635 Manifold Approximation and Projection (UMAP) for visualization.

636 To identify spot specific markers, we used Seurat's FindAllMarkers function. This method calculates log-fold
637 changes, percentages of expression within and outside a group, and p-values of the Wilcoxon-Rank Sum test
638 comparing a group to all cells outside that specific group. This also includes adjustments for multiple testing. A
639 log-fold-change threshold of 0.25 and FDR < 0.05 was considered significant. The basic functions of Seurat were
640 used for visualization.

641

642 ***Cell2location analysis***

643 To identify cell types and their spatial locations within a tissue, we used the Cell2location tool[32]. Cell2location is
644 a Bayesian probabilistic model that integrates transcriptomic and spatial information to estimate cell type-
645 specific gene expression patterns and their spatial distribution. We first preprocessed the spatial transcriptomics
646 data by filtering out lowly expressed genes and normalizing the expression values. We then mapped the spatial
647 locations of the cells onto a two-dimensional coordinate space using the cell coordinates provided by the
648 dataset. We subsequently trained the Cell2location model on the preprocessed data using our integrated
649 sc/snRNAseq and snATACseq data and prior information on cell type co-localization. This prior information,
650 obtained from imaging data, was used to constrain the spatial distribution of each cell type.

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

654 ***Deconvolution of Visium Dataset***

655 To deconvolve the spots in our spatial transcriptomics dataset, we utilized the Cell2location method. We began
656 by preprocessing the raw data using standard quality control and filtering steps to remove low-quality spots and
657 genes with low expression. We then normalized the data using the total-count normalization method
658 implemented in the Cell2location package. Next, we trained the Cell2location model using our integrated
659 sc/snRNAseq and snATACseq atlas and our spatial transcriptomics data. The single-cell RNA sequencing data
660 were used to define the cell types present in the tissue and to estimate their gene expression signatures, while
661 the spatial transcriptomics data were used to estimate the spatial distribution of the cell types within the tissue.
662 We estimated the model parameters using the variational Bayesian inference algorithm implemented in the
663 Cell2location package, running the algorithm for 5,000 iterations with a convergence threshold of 0.01 for the
664 change in the lower bound on the marginal likelihood of the observed data.

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

670

671 ***Mapping sn/scRNAseq to spatial location***

672 To map back the cell types identified in the dissociated data (sn/scRNAseq datasets), we used the Celltrek[33]
673 package. We first downsampled the integrated sc/snRNAseq and snATACseq datasets to 20,000 cells. Then,
674 using the traint function, sn/scRNAseq datasets were co-embedded with spRNAseq datasets. Subsequently,
675 using the random forest model, single cells were mapped to their spatial locations. This analysis was performed
676 by merging snRNAseq and immune cell types to enrich the dataset for immune cells. Regarding colocalization,

677 the sColoc function of CellTrek was used. To identify the different cell type modules in the spRNAseq, we
678 performed a spatial-weighted gene co-expression analysis.

679

680 **Ligand–Receptor Interactions in Spatial Transcriptomics Data**

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

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

691

692 **Finding microenvironments in spRNAseq**

693 To identify microenvironments on the merged spatial dataset after SC transformation (SCT) we performed an
694 NMF reduction using the STUtility package[40]. The dataset was then clustered using the first 20 factors from the
695 NMF reduction. To identify MEs specific markers, we used Seurat's FindAllMarkers function. Identified MEs were
696 then annotated based on marker genes. To find correlations between the number of spots of each MEs and
697 clinical characteristics, we extracted the number of spots for each MEs per sample and then performed a
698 Pearson correlation analysis with clinical findings.

699

700 ***ECM production score***

701 To calculate the extracellular matrix production (ECM) score, we calculated the proportion of the expressions
702 of collagen, proteoglycan, and glycoprotein[36] genes in each cell types.

703

704 **CosMx Data Analysis**

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

713 ***Bulk RNAseq Analysis***

714 The human bulk kidney RNAseq data were downloaded from GEO (GSE115098 and GSE173343). We used
715 previously published pipeline for bulk RNAseq analysis [75]. Briefly, the FASTQC was used to check the QC of the
716 sequencing results. The adapters and low-quality bases were trimmed using TrimGalore (v0.4.5). The trimmed
717 FASTQ files were aligned to the human genome (hg19/GRCh37) using STAR (v2.7.3a)[76,77] based on GENCODE
718 v19 annotations[76,77]. Gene expression was quantified using RSEM by calculating uniquely mapped reads as
719 transcripts per million (TPM). To ensure accurate microdissection, we projected the TPM matrix of the tubule
720 and glomerular compartments onto a PCA plot and samples that did not separate clearly were removed.

721

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

723 Hierarchical clustering was performed on the scaled TPM matrix of microdissected human tubules datasets
724 using the FME-GS list as the base. Ward's method with Euclidean distances was used to cluster the datasets.
725 The optimal number of clusters was determined by average silhouette method.

726

727 ***Kaplan Meier and Cox-regression analysis***

728 To compare the probability of reaching the outcome; end stage kidney disease (eGFR < 15 ml/min/1.73 m^2) or
729 40% eGFR decline, Kaplan-Meier life table survival analysis was performed. The log-rank test was used to

730 compare the survival probability in different clusters. Cox proportional hazard ratio analysis was applied to
731 estimate the hazard ratio of outcome of the subjects in different groups.
732
733 ***Linear regression analysis***
734 To perform the linear regression analysis between fibrosis, eGFR or eGFR decline with gene expression levels in
735 each cell type, cells were subset based on the frequency of each cell type with a maximum number of 1000 cells
736 proportional to its original fraction. A regression model adjusted for age, sex, race, diabetes mellitus,
737 hypertension, and QC parameters (nFeature_Count, nCount_RNA, and percent_mt) was employed. FDR < 0.05
738 was considered as significant threshold.
739

740 **Statistics**

741 Data were expressed as means ± SEM. Independent sample t-test was used to compare the continuous variable
742 in two groups and one-way ANOVA was used to compare continuous parameters between more than two
743 groups followed by Bonferroni post hoc test for subgroup comparisons. A $p < 0.05$ was considered as a
744 significant. LASSO regression was performed with scikit-learn's LASSO CV function, 10 n_splits with 3 repeats
745 were used for parameter tuning.
746

747 **Data Availability**

748 Raw data, processed data, and metadata from the snRNAseq, scRNAseq, snATACseq, and spRNAseq have been
749 deposited in Gene Expression Omnibus (GEO) with the accession code of GSE211785 (**reviewer token:**
750 **srahoicgfzqbfkj**) The human bulk kidney RNAseq data are available under following accession numbers:
751 GSE115098 and GSE173343**.** The single cell and nuclear expression and open chromatin and spatial data is also
752 available at www.susztaklab.com (https://susztaklab.com/hk_genemap/).
753

754 **Code Availability**

All the codes used for the analysis were deposited on GitHub (https://github.com/amin69upenn/Human_Kidney_Multiomics_and_Spatial_Atlas_and https://github.com/jlevins2010/FME_atlas)

Acknowledgement

Work in the Susztak lab is supported by the NIH P50DK114786, DK076077, DK087635, DK132630 and DK105821. The study is supported by GSK, Regeneron, Boehringer Ingelheim, and Novo Nordisk. The funders have no influence on the reported results. The authors thank the Molecular Pathology and Imaging Core (P30-DK050306) and Diabetes Research Center (P30-DK19525) at University of Pennsylvania for their services.

Competing interests

KD and LM are employees of Regeneron Pharmaceuticals. GP, TB, EH, and LSB are employees of GSK. SP, CMB, and PG are employees of Boehringer Ingelheim. AK is an employee of Novo Nordisk.

Author Contributions

AA, JL, ZM, JF, RS, PD, GP, DT, and TB performed experiments. AA, JL, KK, MSB, HL, SV, MSB, HY, and KC performed computational analysis. KD, BD, LM, EH, LSB, CAH, AK, PG, CMB, GP, KHK and ML offered experimental and analytical suggestions. KS was responsible for overall design and oversight of the experiments. MP performed pathological scorings. KS supervised the experiment. AA and KS wrote the original draft. All authors contributed to and approved the final version of the manuscript.

References

1. Gueutin, V., Deray, G. & Isnard-Bagnis, C. Renal physiology. *Bulletin du cancer* **99**, 237-249 (2012).
2. Smith, H.W. The fate of sodium and water in the renal tubules. *Bulletin of the New York Academy of Medicine* **35**, 293 (1959).
3. Vart, P. *et al.* National trends in the prevalence of chronic kidney disease among racial/ethnic and socioeconomic status groups, 1988-2016. *JAMA network open* **3**, e207932-e207932 (2020).
4. Yinusa, A., Faezipour, M. & Faezipour, M. A Study on CKD Progression and Health Disparities Using System Dynamics Modeling. in *Healthcare* Vol. 10 1628 (MDPI, 2022).

5. Kakitapalli, Y., Ampolu, J., Madasu, S.D. & Kumar, M.S. Detailed review of chronic kidney disease. *Kidney Diseases* **6**, 85-91 (2020).
6. Balzer, M.S., Rohacs, T. & Susztak, K. How Many Cell Types Are in the Kidney and What Do They Do? *Annual review of physiology* **84**, 507 (2022).
7. Schreibing, F. & Kramann, R. Mapping the human kidney using single-cell genomics. *Nature Reviews Nephrology* **18**, 347-360 (2022).
8. Balzer, M.S. *et al.* Single-cell analysis highlights differences in druggable pathways underlying adaptive or fibrotic kidney regeneration. *Nature Communications* **13**, 1-18 (2022).
9. Miao, Z. *et al.* Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nature communications* **12**, 1-17 (2021).
10. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature communications* **12**, 1-17 (2021).
11. Wilson, P.C. *et al.* The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proceedings of the National Academy of Sciences* **116**, 19619-19625 (2019).
12. Young, M.D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *science* **361**, 594-599 (2018).
13. Dixon, E.E., Wu, H., Muto, Y., Wilson, P.C. & Humphreys, B.D. Spatially resolved transcriptomic analysis of acute kidney injury in a female murine model. *Journal of the American Society of Nephrology* **33**, 279-289 (2022).
14. Chen, T.K., Knicely, D.H. & Grams, M.E. Chronic kidney disease diagnosis and management: a review. *Jama* **322**, 1294-1304 (2019).
15. Palmer, M.B. *et al.* The role of glomerular epithelial injury in kidney function decline in patients with diabetic kidney disease in the TRIDENT cohort. *Kidney international reports* **6**, 1066-1080 (2021).
16. Armutcu, F. Organ crosstalk: the potent roles of inflammation and fibrotic changes in the course of organ interactions. *Inflammation Research* **68**, 825-839 (2019).
17. Majo, J., Klinkhammer, B.M., Boor, P. & Tiniakos, D. Pathology and natural history of organ fibrosis. *Current opinion in pharmacology* **49**, 82-89 (2019).
18. Deng, Z. *et al.* The extracellular matrix and mechanotransduction in pulmonary fibrosis. *The International Journal of Biochemistry & Cell Biology* **126**, 105802 (2020).
19. Humphreys, B.D. Mechanisms of renal fibrosis. *Annual review of physiology* **80**, 309-326 (2018).
20. Parimon, T., Hohmann, M.S. & Yao, C. Cellular senescence: pathogenic mechanisms in lung fibrosis. *International Journal of Molecular Sciences* **22**, 6214 (2021).
21. Makita, A. *et al.* The association between renal elasticity evaluated by real-time tissue elastography and renal fibrosis. *Clinical and Experimental Nephrology* **25**, 981-987 (2021).
22. Lausecker, F., Lennon, R. & Randles, M.J. The kidney matrisome in health, aging and disease. *Kidney International* (2022).
23. Zhou, D. & Liu, Y. Understanding the mechanisms of kidney fibrosis. *Nature Reviews Nephrology* **12**, 68-70 (2016).
24. Grams, M.E. *et al.* Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. *Kidney international* **93**, 1442-1451 (2018).
25. Hallan, S.I. *et al.* Combining GFR and albuminuria to classify CKD improves prediction of ESRD. *Journal of the American Society of Nephrology* **20**, 1069-1077 (2009).
26. Neuen, B.L. *et al.* Changes in GFR and albuminuria in routine clinical practice and the risk of kidney disease progression. *American Journal of Kidney Diseases* **78**, 350-360. e1 (2021).
27. An, Y. *et al.* Renal histologic changes and the outcome in patients with diabetic nephropathy. *Nephrology Dialysis Transplantation* **30**, 257-266 (2015).
28. Rodríguez-Iturbe, B., Johnson, R.R. & Herrera-Acosta, J. Tubulointerstitial damage and progression of renal failure. *Kidney international* **68**, S82-S86 (2005).
29. Srivastava, A. *et al.* The prognostic value of histopathologic lesions in native kidney biopsy specimens: results from the Boston Kidney Biopsy Cohort Study. *Journal of the American Society of Nephrology* **29**, 2213-2224 (2018).

30. Luecken, M.D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41-50 (2022).
31. Lake, B.B. *et al.* An atlas of healthy and injured cell states and niches in the human kidney. *bioRxiv*, 2021.07.28.454201 (2021).
32. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology* **40**, 661-671 (2022).
33. Wei, R. *et al.* Spatial charting of single-cell transcriptomes in tissues. *Nature biotechnology*, 1-10 (2022).
34. Thul, P.J. & Lindskog, C. The human protein atlas: a spatial map of the human proteome. *Protein Science* **27**, 233-244 (2018).
35. Kuppe, C. *et al.* Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281-286 (2021).
36. Naba, A. *et al.* The extracellular matrix: Tools and insights for the "omics" era. *Matrix Biology* **49**, 10-24 (2016).
37. Crow, M., Paul, A., Ballouz, S., Huang, Z.J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature communications* **9**, 884 (2018).
38. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1-13 (2008).
39. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature genetics* **53**, 1143-1155 (2021).
40. Nirmal, A.J. *et al.* The spatial landscape of progression and immunoediting in primary melanoma at single cell resolution. *Cancer discovery* **12**, 1518 (2022).
41. Sato, Y., Tamura, M. & Yanagita, M. Tertiary lymphoid tissues: A regional hub for kidney inflammation. *Nephrology Dialysis Transplantation* (2021).
42. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nature communications* **12**, 1-20 (2021).
43. Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods* **20**, 218-228 (2023).
44. Wilson, P.C. *et al.* Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature communications* **13**, 5253 (2022).
45. Wu, H. *et al.* Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. *Cell metabolism* **34**, 1064-1078. e6 (2022).
46. Kluger, A.Y. *et al.* Class effects of SGLT2 inhibitors on cardiorenal outcomes. *Cardiovascular diabetology* **18**, 1-13 (2019).
47. Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758-763 (2018).
48. Zhang, L. *et al.* Genome-wide analysis of histone H3 lysine 4 trimethylation in peripheral blood mononuclear cells of minimal change nephrotic syndrome patients. *American journal of nephrology* **30**, 505-513 (2009).
49. Wynn, T.A. & Ramalingam, T.R. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nature medicine* **18**, 1028-1040 (2012).
50. Djudjaj, S. & Boor, P. Cellular and molecular mechanisms of kidney fibrosis. *Molecular aspects of medicine* **65**, 16-36 (2019).
51. Liu, M. & Gomez, D. Smooth muscle cell phenotypic diversity: At the crossroads of lineage tracing and single-cell transcriptomics. *Arteriosclerosis, thrombosis, and vascular biology* **39**, 1715-1723 (2019).
52. Pedroza, A.J. *et al.* Single-cell transcriptomic profiling of vascular smooth muscle cell phenotype modulation in Marfan syndrome aortic aneurysm. *Arteriosclerosis, thrombosis, and vascular biology* **40**, 2195-2211 (2020).
53. Zhang, F., Guo, X., Xia, Y. & Mao, L. An update on the phenotypic switching of vascular smooth muscle cells in the pathogenesis of atherosclerosis. *Cellular and Molecular Life Sciences* **79**, 1-19 (2022).
54. Aghajanian, H. *et al.* Targeting cardiac fibrosis with engineered T cells. *Nature* **573**, 430-433 (2019).
55. Aghajanian, H., Rurik, J.G. & Epstein, J.A. CAR-based therapies: opportunities for immuno-medicine beyond cancer. *Nature Metabolism* **4**, 163-169 (2022).
56. Rurik, J.G. *et al.* CAR T cells produced in vivo to treat cardiac injury. *Science* **375**, 91-96 (2022).

881 57. Song, J. *et al.* Understanding kidney injury molecule 1: a novel immune factor in kidney pathophysiology. *Am J*
882 *Transl Res* **11**, 1219-1229 (2019).
883 58. Huang, E. *et al.* The roles of immune cells in the pathogenesis of fibrosis. *International journal of molecular sciences*
884 **21**, 5203 (2020).
885 59. Barone, F. *et al.* Stromal fibroblasts in tertiary lymphoid structures: a novel target in chronic inflammation.
886 *Frontiers in immunology* **7**, 477 (2016).
887 60. Huang, Y. *et al.* Identification of novel genes associated with renal tertiary lymphoid organ formation in aging
888 mice. *PloS one* **9**, e91850 (2014).
889 61. Kitching, A.R. Immunoaging within the kidney via injury-associated tertiary lymphoid tissue. *Kidney Int* **102**, 9-11
890 (2022).
891 62. Luo, R. *et al.* Tertiary lymphoid organs are associated with the progression of kidney damage and regulated by
892 interleukin-17A. *Theranostics* **11**, 117 (2021).
893 63. Sato, Y. *et al.* CD153/CD30 signaling promotes age-dependent tertiary lymphoid tissue expansion and kidney
894 injury. *The Journal of clinical investigation* **132**(2022).
895 64. Zhao, J., Gu, S. & McDermaid, A. Predicting outcomes of chronic kidney disease from EMR data based on Random
896 Forest Regression. *Math Biosci* **310**, 24-30 (2019).
897 65. Young, M.D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA
898 sequencing data. *Gigascience* **9**(2020).
899 66. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data
900 Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4 (2019).
901 67. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**,
902 1289-1296 (2019).
903 68. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *BioRxiv*
904 (2020).
905 69. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and
906 visualization. *Bioinformatics* **31**, 2382-2383 (2015).
907 70. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated
908 accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).
909 71. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics.
910 *Nature methods* **15**, 1053-1058 (2018).
911 72. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
912 73. Trapnell, C. *et al.* Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate
913 decisions. *Nature biotechnology* **32**, 381 (2014).
914 74. Ni, Z. *et al.* SpotClean adjusts for spot swapping in spatial transcriptomics data. *Nature Communications* **13**, 2971
915 (2022).
916 75. Sheng, X. *et al.* Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms
917 of disease and potential treatments. *Nature genetics* **53**, 1322-1333 (2021).
918 76. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
919 77. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*
920 **22**, 1760-1774 (2012).

921

922 **Figure Legends**

923 **Fig 1. Comprehensive integrated multimodal human kidney single cell atlas**

924 **(a).** Overview of the multimodal analysis. Basic clinical characteristics of the samples.

925 **(b)** UMAP of 338,565 integrated human kidney sn/scRNAseq and snATACseq data generated in this study.
926 Annotated cell types are indicated on the plot.

927 **(c)** UMAP of 588,425 integrated human kidney sn/scRNAseq and snATACseq data of the present study and
928 kidney precision medicine project (KPMP).

929 (**d**) Hierarchical sub-clustering identified 114 distinct cell types or cell states. The bar charts depict the relative
930 abundance of each group (method, sex, group, and samples) contributing to the cluster.

931 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
932 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-
933 specific stromal cells, VSMC/Pericyte; vascular smooth muscle cells/pericyte, PEC; parietal epithelial cells, Podo;
934 podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule
935 segment 3, iPT; injured proximal tubule cells, DTL; descending thin loop of Henle, C_TAL; cortical thick ascending
936 loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting
937 tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated
938 cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, dnT; double negative T cells,
939 Prolif_Lym; proliferative lymphocyte, Th17; T helper 17 lymphocyte, gDT; gamma delta T cells, TCM; T Cell
940 memory, TEM; T effector memory, Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red blood cells,
941 Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac;
942 macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+, CTL; control, HKD; hypertensive
943 kidney disease, DKD; diabetic kidney disease.

944

945 **Fig 2. Spatially resolved human kidneys.**

946 **(a)** Spatial transcriptomics data was generated from human formalin fixed paraffin embedded kidney samples
947 using two platforms. Visium (top) employs a spot-based approach with each spot of 55 um, with the ability to
948 detect >18,000 genes able for detection and requires deconvolution to identify presence of individual cell types.
949 CosMx (bottom) imaging generates single cell level data and identifies 1000 genes, and which permits
950 annotation of cell types based on the expression patterns of these genes.

951 **(b)** Spatial location and marker gene expression of identified cell types using both Visium and CosMx. H&E
952 sections shown on the left of the figure show individual tissue histology of our Visium sections. Using our

953 calculated Cell2Location scores for each of these tissues, we imputed the presence of cell types on each of these
954 sections with cell type scores shown in blue overlaying the H&E. We identify glomerular cell types, proximal
955 tubular cell types, immune and fibrotic cell types, distal tubular cell types and loop of Henle cell types. To the
956 right of these images, we validated the location of these cell types using the CosMx assay, with allows for
957 individual cell type annotations, which are overlayed on the H&E image of the assayed tissue.

958 Endo_GC; endothelial cells of glomerular capillary tuft, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-
959 specific stromal cells, Myo_VSMC; myofibroblast/vascular smooth muscle cells PEC; parietal epithelial cells,
960 Podocyte; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal
961 tubule segment 3, iPT; injured proximal tubule cells, ATL: ascending thin limb of loop of Henle, C_TAL; cortical
962 thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule,
963 CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type
964 beta intercalated cells, , CD4T; T lymphocytes CD4+, Plasma cell; plasma cell.

965

966 **Fig 3. Human kidney stromal cell atlas.**

967 **(a)** Extracellular matrix (ECM) gene expression score in different kidney cell types in the integrated sn/scRNAseq
968 and snATACseq. ECM score was calculated based on the expression of collagen, proteoglycan, and glycoprotein
969 genes.

970 **(b)** The fraction of myofibroblasts in control and CKD samples. The bars indicate SEM. Independent t-test was
971 used to compare the fractions between two groups.

972 **(c)** The ECM score in the spatial transcriptomics datasets of healthy and CKD samples (upper panel). Red
973 indicates higher ECM gene expression (upper panel). The lower panel indicates the location of mesangial cells,
974 fibroblasts, myofibroblasts and VSCMC/pericytes in healthy and CKD samples based on their Cell2Location score.

975 **(d)** UMAP representation of sub-clustering of 32,706 stomal cells in the integrated dataset -- present study and
976 Kidney Precision Medicine Project (KPMP).

977 (**e**) The dot plots of marker genes used for stromal cell type annotation in the integrated dataset. The size of the
978 dot indicates the percent positive cells, and the darkness of the color indicates average expression.

979 **(f)** The spatial location and stromal cell subtypes and specific marker genes. The left panel shows the relative
980 abundance of each cell type using Cell2location. The right panel shows relative gene expression using CellTrek
981 (red higher).

982 (**g**) The heatmap indicates the -Log_{10} (FDR) enrichment of the top KEGG pathways in each stromal cell type.

983 Mes; mesangial cells, GS_Stromal; glomerulosclerois-specific stromal cells, Fib; fibroblast, VSMC; vascular
984 smooth muscle cells, MyoFib; myofibroblast.

985

986 **Fig 4. Human kidney fibrotic microenvironment.**

987 **(a)** Human kidney microenvironments defined by non-negative matrix factorization (NMF) of the spRNAseq.
988 Glomerular (green), tubule (brown), fibrotic (red), immune (blue). The spatial distribution of the calculated ECM
989 score (right panel). Red indicates higher ECM gene expression.

990 **(b)** Correlation of microenvironment with ECM score (left) and cell type correlation with each microenvironment
991 (right).

992 (**c**) Cell2location (left), Celltrek (right) cell type imputations of the Visium spatial data and CosMx (below)
993 annotations (below) showing the location of different cell types in the fibrotic microenvironment in a diseased
994 kidney sample. CosMx annotations of iPT, PT, fibroblasts and immune cells are shown at 2 different
995 magnifications.

996 (**d**) *In situ* mass spec image of CKD kidney (from panel E) labeled with LRP2 (PT) KIM1 (iPT), CD4 (T-cell) CD20 (B-
997 cell), CD16 (myeloid) and CD31 (endo) in a fibrotic human kidney sample.

998 (**e**) The spatial location of the identified cell-cell interaction pathway (*SPP1* and *CXCL*). The arrows indicate the
999 source and targets of the identified pathways, and the color indicates the received signals (red higher). The
1000 results were obtained using COMMOT package.

1001 ME; microenvironment, GME; glomerular microenvironment, TME; tubular microenvironment, FME; fibrotic
1002 microenvironment, IME; immune microenvironment, Endo; endothelial cells, MyoFib; myofibroblast, Fib;
1003 fibroblast, VSMC/Pericyte; vascular smooth muscle cells and pericyte, Podo; podocyte, Mes; mesangial cells, PT;
1004 proximal tubule cells, iPT; injured proximal tubule cells, LOH; loop of Henle, DCT; distal convoluted tubule, CNT;
1005 connecting tubule, PC; principal cells of the collecting duct.

1006

1007 **Fig 5. Injured proximal tubule cells in diseased human kidneys.**

1008 **(a)** Localization of iPT and PT cells within control and diseased samples in the human kidney. iPT is scattered in
1009 a patchy pattern, and colocalize amongst themselves.

1010 (**b**) Heatmap of number of genes correlated with eGFR and renal fibrosis in each cell type. The red shows higher
1011 number of genes and blue indicates lower number of genes.

1012 **(c)** Heatmap of number of genes comparing prevalence in healthy vs. diseased samples in each cell type. PT had
1013 the highest changes between control and diseased samples, with red indicating more genes.

1014 **(d)** The fractions of PT and iPT in control and diseased samples type in the integrated sn/scRNAseq and
1015 snATACseq dataset. Bar indicated the SEM. For the comparison between two groups, t-test was used.

1016 **(e)** CosMx spatial transcriptomics shows that annotation of iPT and PECs localize to tubules and glomeruli,
1017 respectively. Dot plot shows *VCAM1* is expressed in both populations, while *HAVCR1* and *CFH* appear to be more
1018 specific for iPT and PECs, respectively.

1019 **(f)** Gene co-expression network analysis of human kidney spRNAseq data indicates two modules expressing
1020 *VCAM1* or *HAVCR1*. The right panels indicate the spatial location of *VCAM1* or *HAVCR1* (KIM1) positive injured
1021 PT cells. The color indicates gene expression of iPT modules (red higher expression).

1022 **(g)** Sub-clustering of iPT cells from integrated atlas. The fraction of VCAM1 or HAVCR1 positive iPT cells in control
1023 and diseased kidneys, shown on UMAP and quantified in frequency. The dot plots show the expression marker
1024 genes in iPT and PT cells. The size of the dot indicates the percent positive cells, and the darkness of the color
1025 indicates average expression.

1026 **(h)** Cell trajectory analysis (Monocle) representation of PT and iPT cells (Upper panels). The heatmap shows the
1027 differentially expressed genes along the trajectory, cells ordered by pseudotime. Red indicates higher
1028 expression.

1029 **(i)** Sub-clustering of iPT within the CosMx data shows 3 subtypes of iPT. These iPT subtypes are located near PT
1030 cells and are visualized in specific fields of view—one in each sample. These iPT subtypes have different

1031 neighboring cells (within a 50 micron radius) with iPT_KRT7 having most frequent fibroblast and immune
1032 neighbors within our diseased sample (*p*-value = 1e-177 , 1e-17 for fibroblasts and immune cells respectively
1033 within our disease sample by Wilcoxon rank sum test). iPT_APOE had fewer immune (*p*-value = 0, *p*-value = 0,
1034 for HK3039 and HK2844) and fibroblast neighbors than the iPT_SPP1 subtype. (*p*-value = 5e-52, *p*-value = 1e-52,
1035 for HK3039 and HK2844).

1036 Endo; endothelial cells Mes; mesangial cells, Stroma; stromal cells, PEC; parietal epithelial cells, Podo; podocyte,
1037 PT; proximal tubule cells, iPT; injured proximal tubule cells, LOH; loop of Henle, DCT; distal convoluted tubule,
1038 CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type
1039 beta intercalated cells.

1040

1041 **Fig 6. Fibrotic microenvironment gene signature successfully predicts disease prognosis in a large cohort of**
1042 **human kidney samples.**

1043 **(a)** Clinical characteristics of 292 human kidney tubule RNA samples.

1044 **(b)** Unbiased cluster dendrogram of 292 human kidney tubule bulk RNAseq samples based on FME gene
1045 signature. Clinical characteristics of each cluster. Chi-square test for categorical variables and one-way ANOVA
1046 for continuous variables were used to compare groups.

1047 (**c**) Unbiased cluster dendrogram of 245 human kidney tubule bulk RNAseq samples with fibrosis <10% based
1048 on expression of FME genes. Clinical characteristics of each cluster are shown in the table. Chi-square test for
1049 categorical variables and one-way ANOVA for continuous variables were used to compare groups.

1050 (**d**) Kaplan-Meier analysis of 292 kidney samples based on FME gene signature. Renal survival was defined as
1051 cases reaching end stage renal disease or greater than 40% eGFR decline (above). Kaplan-Meier analysis of 245
1052 kidney samples based on pathologist defined kidney fibrosis degree (<3%, 3-6% or 7-10% as defined by an expert
1053 pathologist) (below). Renal survival was defined as cases reaching end stage renal disease (eGFR of 15
1054 ml/min/1.73m2) or greater than 40% eGFR decline.

1055 FME-GS; fibrotic microenvironment gene signature, HR; hazard ratio, HTN: hypertension, DM: diabetes, AA:
1056 African American,

1057

1058 **Supplementary material**

1059 **Supplementary Tables:**

1060 **Supplementary Table 1. Demographic and clinical characteristics of the subjects in the study.**

1061 **Supplementary Table 2. QC metrics of the snRNAseq, scRNAseq, snATACseq, and spRNAseq data.**

1062 **Supplementary Table 3. Cell type marker genes in the snRNAseq data.**

1063 **Supplementary Table 4. Cell type marker genes in the scRNAseq data.**

1064 **Supplementary Table 5. Cell type specific accessible regions and genes in the snATACseq data.**

1065 **Supplementary Table 6. Cell type specific markers genes in the integrated snRNAseq, snATACseq, and**
1066 **scRNAseq.**

1067 **Supplementary Table 7. Cell type specific transcription factor motif enrichment.**

1068 **Supplementary Table 8. Module specific genes and scores in fibroblast and myofibroblast (WGCNA)**

1069 **Supplementary Table 9. Differentially expressed genes between diseased and control samples in each cell**
1070 **type in the integrated snRNAseq, snATACseq, and scRNAseq.**

1071 **Supplementary Table 10. CellChat interaction scores.**

1072 **Supplementary Table 11. Differentially expressed genes along injured PT trajectory in the integrated**
1073 **snRNAseq, snATACseq, and scRNAseq.**
1074

1075 **Supplementary Table 12. Fibrotic microenvironment specific genes.**

1076 **Supplementary Table 13. Clinical characteristics of the 298 and 245 tubule bulk RNAseq cohorts.**

1077 **Supplementary Table 14. Lasso coefficients for each gene of the FME gene signature against eGFR.**

1078

1079 Supplementary Figures.

1080 Supplementary Fig. 1. Quality metrics of the sn/scRNAseq, snATACseq and spatial transcriptomics data.

1081 (a) QC parameters of the scRNAseq. The plots depict the distribution of nFeature_Count (number of genes per cell),
1082 nCount_RNA (number of reads per cell), and mitochondrial percentage (fraction of mitochondrial genes per cell) in each
1083 sample. X axis indicates the individual sample and y-axis shows the values.

1084 (b) QC parameters of the snRNAseq. The plots depict the distribution of nFeature_Count (number of genes per cell),
1085 nCount_RNA (umber of reads per cell), and mitochondrial percentage (fraction of mitochondrial genes per cell) in each
1086 sample. X axis indicates the individual sample and y-axis shows the values.

1087 (c) QC parameters of the snATACseq. The plots depict the distribution of pct_reads in peaks, nFeature_ATAC per sample.
1088 X-axis indicates the individual sample and y-axis shows the values.

1089

1090 Supplementary Fig. 2. Multimodal single cell atlas

1091 (a) UMAP of snRNAseq, scRNAseq, and snATACseq datasets before integration.

1092 (b) UMAP of integrated snRNAseq, scRNAseq, and snATACseq datasets of 338,565 cells and nuclei using the SCVI tool.

1093 (c) Annotations of cell types on integrated UMAP.

1094 (d) The dot plots of marker genes used to annotation of 44 main cell types in the integrated dataset. The size of the dot
1095 indicates the percent positive cells, and the darkness of the color indicates average expression.

1096 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
1097 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; meseangial cells, GS_Stromal; glomerulosclerosis-specific
1098 stromal cells, VSMC/Pericyte; vascular smooth muscle cells/pericyte, PEC; parietal epithelial cells, Podo; podocyte,
1099 PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT;
1100 injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle,
1101 M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC;
1102 principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer
1103 cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, dnT; double negative T cells, Prolif_Lym; proliferative
1104 lymphocyte, Th17; T helper 17 lymphocyte, B_Naive; Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red
1105 blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac;
1106 macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

1107

1108 Supplementary Fig. 3. High resolution clustering of the integrated data.

1109 (a)UMAP of integrated snRNAseq, scRNAseq, and snATACseq datasets of 338,565 cells and nuclei and identification of
1110 114 clusters. Right panel shows the original source data.

1111 (b)The dot plots of marker genes used for the annotation of 144 fine clusters in the integrated dataset. The size of the
1112 dot indicates the percent positive cells, and the darkness of the color indicates average expression.

1113 (c)The dot plot of erythropoietin (EPO) expression in the integrated snRNAseq, scRNAseq, and snATACseq datasets. The
1114 size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression. Stromal
1115 cells are the main source of EPO in the kidney.

1116 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
1117 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific
1118 stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte,
1119 PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT;
1120 injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle,
1121 M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC;
1122 principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer
1123 cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naive; Naive B lymphocyte, B_memory; memory B
1124 lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical
1125 dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

1126

1127 Supplementary Fig. 4. Integrations of snRNAseq, scRNAseq, and snATACseq datasets from multiple sources

1128 (a) UMAP of integrated snRNAseq, scRNAseq, and snATACseq datasets (n=588,425 cells and nuclei) from Susztaklab and
1129 KPMP using the SCVI tool. Left panel shows the new annotation after integration the dataset and right panel indicates
1130 the original annotation used by KPMP.

1131 (b) Bar charts showing cell abundance in each cell clusters (method, lab, present study annotation and KPMP
1132 annotation).

1133 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
1134 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific
1135 stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte,
1136 PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT;
1137 injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle,
1138 M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC;
1139 principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer
1140 cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naive; Naive B lymphocyte, B_memory; memory B
1141 lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical
1142 dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

1143

1144 Supplementary Fig. 5. Integrations of snRNAseq, scRNAseq, and snATACseq datasets from multiple sources

1145 (a) Dot plots of marker genes used for the annotation of 39 main cell types in the integrated dataset. The size of the dot
1146 indicates the percent positive cells, and the darkness of the color indicates average expression.

1147 (b) Cell type consistency of the different datasets using MetaNeighbor. The heatmap displays the Area Under the
1148 Receiver Operating Characteristic (AUROC) scores among different cell subtypes in two different data sources (this study
1149 vs KPMP). Dendrograms were created via hierarchical clustering, based on Euclidean distances, and used average linkage
1150 for this purpose. Each row and each column are a cell type in one of the source data.

1151 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
1152 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific
1153 stromal cells, VSMC/Pericyte; vascular smooth muscle cells/pericyte, PEC; parietal epithelial cells, Podo; podocyte,
1154 PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT;
1155 injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle,
1156 M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC;
1157 principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer
1158 cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, dnT; double negative T cells, Prolif_Lym; proliferative
1159 lymphocyte, Th17; T helper 17 lymphocyte, B_Naive; Naive B lymphocyte, B_memory; memory B lymphocyte, RBC; red
1160 blood cells, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac;
1161 macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.

1162

1163 Supplementary Fig. 6. Human kidney single nuclear open chromatin (snATACseq) data.

1164 (a) UMAP of 58,155 nuclei after QC filtering in snATACseq dataset. 26 clusters were identified.

1165 (b) Heatmap of top 10 differentially accessible regions per cells in the snATACseq dataset. Rows indicate the top 10
1166 differentially accessible regions and their degree of accessibility pre cluster and columns show the cell types.

1167 (c) Accessibility of marker genes across annotations. Reads at each locus are shown as a with a higher peak indicating
1168 more open chromatin.

1169 (d) Enriched motif and their activity per cluster in snATACseq dataset. The table shows the enriched motif in each cell
1170 type.

1171 (e) The dot plot indicates the motifs activity in main clusters of snATACseq dataset. The size of the dot indicates the
1172 percent positive cells, and the darkness of the color indicates average activity.

1173

1174 Supplementary Fig. 7. Histological images of the Visium data.

1175 H&E imaging of the 3 control and 11 diseased samples. Each sample is H&E stained and the number of Visium spots with
1176 significant RNA quantity is indicated below. Scale bar is 1 mm.

1177

1178 Supplementary Fig 8. Quality Control metrics of spatial transcriptomic data.

1179 (a) QC parameters of the Visium spatial RNA assay. We show number of features and number of RNA counts for each
1180 individual sample.

1181 (b) QC parameters for the CosMx analysis represented as histograms, which are parsed by sample. False codes indicates
1182 reads that do not correspond to a known gene barcode, and negative probe counts corresponds to reads to a probe that
1183 is designed not to anneal to a human RNA. Counts were aggregated by field of view (FOV) to inspect if a particular region
1184 of tissue performed poorly. These metrics are then shown in space on the slide. Lastly, we also show the location of each
1185 sample on the CosMx slide (below).

1186

1187 Supplementary Fig. 9. Human kidney Visium spatial transcriptomics data

1188 (a) UMAP of the 37,436 spots, across all 14 samples. Spots were annotated based on markers shown in B, though
1189 multiple cell types likely exist within each spot.

1190 (b) The dot plots of marker genes used for spots annotation in the spRNAseq. The size of the dot indicates the percent
1191 positive cells, and the darkness of the color indicates average expression.

1192 (c) The heatmap of the top 10 differentially expressed genes in the identified spots in spRNAseq. Rows indicate the top
1193 10 differentially expressed genes and their degree of accessibility pre cluster and columns show the cell types.

1194 Endo_G; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels,
1195 Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific
1196 stromal cells, Fib; Fibroblast, MyoFib; Myofibroblast, VSMC/Pericyte; vascular smooth muscle cells/Pericyte, PEC;
1197 parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3;
1198 proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL;
1199 cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule,
1200 CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta
1201 intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naive; Naive B

1202 lymphocyte, B_memory; memory B lymphocyte, RBC; red blood cells, Baso/Mast; basophil or mast cells, pDC;
1203 plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+,
1204 CD16_Mono; monocyte CD16+.

1205

1206 Supplementary Fig. 10. Protein expression of human kidney cell type markers in the Human Protein Atlas Images were
1207 taken https://www.proteinatlas.org/.

1208

1209 Supplementary Fig. 11. Glomerular, proximal and distal tubule cells in each kidney sample (Cell2Location was used to
1210 map cells).

1211

1212 Supplementary Fig. 12. Immune and stromal cells in each kidney sample (Cell2Location was used to map cells).

1213

1214 Supplementary Fig. 13. Proximal tubule, glomerular, stromal and LOH cells in each kidney sample (CellTrek was used to
1215 map cells).

1216

1217 Supplementary Fig. 14. Distal tubule and immune cells in each kidney sample (CellTrek was used to map cells).

1218

1219 Supplementary Fig 15. Tissue staining of the CosMx slides.

1220 (a) Immunostained CosMx slide. CosMx slides were stained with DAPI, Pan-Cytokeratin, CD-45, and membrane stain.

1221 (b) After imaging the slide, the flowcell (sample) was removed and the slides were then H&E stained and re-imaged.

1222

1223 Supplementary Fig. 16. CosMx Cell populations.

1224 (a) Original UMAP of all cells that passed QC along with their annotations. The “PT?” population appeared to be similar
1225 to PTs but lacked classic markers.

1226 (b) Cell Populations with annotatable markers of the CosMx data on UMAP.

1227 (c) UMAP of all CosMx cells by sample.

1228 (d) CosMx cell annotations across the UMAP with a single population being shown for a given UMAP demonstrating that
1229 these clusters are indeed relatively localized within the UMAP.

1230 (e) Dot plot showing markers for each annotated CosMx population

1231 (f) Annotation Frequency for each cell type.

1232

1233 Supplementary Fig. 17. CosMx SCVI integration with snRNAseq data.

1234 After integrating with the snRNAseq data, we compared annotations from our CosMx analysis and the separately
1235 annotated snRNAseq.

1236 (a) UMAP of integrated data demonstrating technology type of each cell within the UMAP.

1237 (b) Comparison of CosMx annotations and snRNAseq annotations, demonstrating concordance of location within the
1238 integrated UMAP.

1239

1240 Supplementary Fig. 18 Location of CosMx annotated cell types within the slide.

1241 Location of annotated cell types within the two tissue sections.

1242 Location of glomerular cell subtypes

1243 Location of injured PT, fibroblasts and immune cells

1244

1245 Supplementary Fig. 19 Microanatomy of the CosMx slide

1246 Location of glomerular cell types within a subsection of tissue, and in a single field of view (right)

1247 Location of injured thick ascending limb, healthy injured thick ascending limb, principal cells and immune cell types
1248 within a subsection of tissue, and in a single field of view (right).

1249 Location of distal nephron cell subtypes within a subsection of tissue and in a single field of view (right)

1250 Single field of view showing many cell types.

1251

1252 Supplementary Fig. 20 Neighborhood characteristics of CosMx slide

1253 Relative type cell frequency between each sample. Orange indicates HK3039 (healthy) and blue indicates HK2844
1254 (diseased), and frequency of neighbor annotations for each cell type for a 20-micron neighborhood (right).

1255 Neighborhood enrichment by permuting annotations for the 20-micron neighborhood size. Lighter color indicates higher
1256 enrichment and co-localization of a given population.

1257 Dot plots for iPT, PEC, Podocytes and PT cells expression of iPT and PEC markers across genomics modalities and protein
1258 staining of VCAM1 in PECs from the human protein atlas: https://www.proteinatlas.org/.

1259

1260 Supplementary Fig. 21. Sub-clustering of stromal cells in the integrated dataset

1261 (a) UMAP of 41,422 stromal cells in the integrated dataset.

1262 (b) The dot plots of marker genes used for cell type annotation. The size of the dot indicates the percent positive cells,
1263 and the darkness of the color indicates average expression.

1264 (c) The single cell Weighted correlation network analysis (WGCNA) on fibroblast and myofibroblast identified specific
1265 modules in control and disease samples. The heatmaps shows the identified modules in each cell types and their scores.
1266 The lower panel shows the representative genes with high scores in each cell type.

1267 Mes; Mesangial cells, GS_Stromal; glomerulosclerosis associated stromal cells, JGC; juxta-glomerular cells, VSMC;
1268 vascular smooth muscle cell, Fib; fibroblast, MyoFib; myofibroblast.

1269

1270 Supplementary Fig. 22. The integration of stromal cells with prior studies

1271 (a) UMAP of integration of 86,481 stromal cells of current study prior papers by KPMP and Kramann et al. The color
1272 indicates the data source.

1273 (b) Heatmap display of MetaNeighbor analysis of the similarities of the identified stromal cell types. The heatmap
1274 displays the Area Under the Receiver Operating Characteristic (AUROC) scores among different cell subtypes in three
1275 different sources of datasets, which were determined using a set of highly variable genes (HVG). Dendrograms were
1276 created via hierarchical clustering, based on Euclidean distances, and used average linkage for this purpose. Each row
1277 and column is a cell type in the source data.

1278

1279 Supplementary Fig. 23. Expression of stromal markers in the Human Protein Atlas kidney samples. Images were taken
1280 https://www.proteinatlas.org/.

1281

1282 Supplementary Fig. 24. Stromal sub-clustering in human snATACseq data.

1283 (a) UMAP of 5,822 stromal cells in human snATACseq dataset.

1284 (b) The dot plot of gene markers used to annotate the different stromal cell types in snATACseq dataset. The dot plot
1285 shows the gene activity in different cell types. The size of the dot indicates the percent positive cells, and the darkness of
1286 the color indicates average expression.

1287 (c) The dot plot of motif activity of enriched motifs in each identified stromal cluster. The size of the dot indicates the
1288 percent positive cells, and the darkness of the color indicates average expression.

1289 Mes; Mesangial cells, GS_Stromal; glomerulosclerosis associated stromal cells, JGC; juxta-glomerular cells, VSMC;
1290 vascular smooth muscle cell, Fib; fibroblast, MyoFib; myofibroblast.

1291

1292 Supplementary Fig. 25. Gene ontology analysis of microenvironment gene signature

1293 (a) Gene ontology analysis (DAVID) of the microenvironment gene signature. The X axis shows the -Log10 (FDR) of the
1294 identified pathways and Y axis lists the pathways.

1295 (b) Person correlation of the number of spots assigned to each microenvironment and eGFR and fibrosis.

1296

1297 Supplementary Fig. 26. The spatial microenvironments in the human kidney. Four spatial microenvironments were
1298 identified in human kidney samples using matrix factorization: glomerular (green), tubule (brown), fibrotic (red) and
1299 immune (blue). The ECM score (calculated by aggregate expression collagen and other glycoprotein, and proteoglycan)
1300 in Visium data shown next to the tissue microenvironment information.

1301

1302 Supplementary Fig. 27. Different cell types in kidney microenvironment.

1303 Left panel kidney microenvironment in tissue samples, right panel shows cell abundance mapped using Cell2location
1304 (endothelial cell, injured_PT cells, stromal cells and immune cells)

1305

1306 Supplementary Fig. 28. Immunostaining of the lymphocyte and plasma cells in regions of fibrosis-associated immune
1307 cells aggregates.

1308 Immunofluorescent staining of T-cell markers (CD4) and plasma cells (IGKC) in diseased human kidney samples.

1309 Spatial RNAseq (10x Visium with CellTrek imputation) and immunohistochemistry staining of T-cell, B-cell, and plasma
1310 cells in diseased human kidney samples in the regions of fibrosis-associated immune cells aggregates.

1311

1312 Supplementary Fig. 29. *In situ* mass spectrometry of the control, and diseased samples to verify cell types in fibrotic
1313 microenvironments.

1314 The left panel shows the Visium 10x data cellular microenvironment. The right panel shows the IMC staining of the area,
1315 markers and colors are listed on the right.

1316 Supplementary Fig. 30. Cell-cell interaction analysis in the spRNAseq dataset in fibrotic microenvironment.

1317 (a) Weighted total interaction strength of the CXCL, SPP1, TGFb, and PDGF pathways in control and diseased samples in
1318 the fibrotic microenvironment (left panel). The spatial location of the identified cell-cell communications pathways
1319 (CXCL, SPP1, TGFb, and PDGF) in control and diseased sample in the fibrotic microenvironment were shown in the right
1320 panel. The arrows indicate the source and targets of the identified pathways.

1321 (b) Expression of CD34 and CDH5 as the markers of high endothelial venules (HEVs) in the fibrotic microenvironments.

1322 (c) Volcano plot of differentially expressed genes from CosMx data. Cells with an immune neighbor within 20 microns
1323 were compared against cells without an immune neighbor for both the PT and iPT population. Logfold change > 0
1324 indicates increased expression in cell with an immune neighbor, while logfoldchange < 0 indicates increased expression
1325 in cells without an immune neighbor. Log10(pvalue) is indicated by the y-axis. Genes with an adjusted p-value < 0.01 are
1326 marked in orange.

1327 (d) The dot plot of expression of ligands and receptors in regions of FME in integrated sn/scRNAseq and snATACseq data.
1328 The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression (right
1329 panel). The gray indicates control and red indicates diseased group.

1330 Endo; endothelial cells, MyoFib; myofibroblast, Fib; fibroblast, VSMC/Pericyte; vascular smooth muscle cells and
1331 pericyte, Podo; podocyte, Mes; mesangial cells, PT; proximal tubule cells, iPT; injured proximal tubule cells, LOH; loop of
1332 Henle, DCT; distal convoluted tubule, CNT; connecting tubule, PC; principal cells of the collecting duct.

1333

1334 **Supplementary Fig. 31 Spatial characteristics of injured PT subclusters on CosMx.**

1335 **(a) The CosMx iPT population for each sample was subclustered using a leiden algorithm with resolution of 0.3.**

1336 **(b) The top 10 differentially expressed genes for each subcluster.**

(c) iPT subcluster localization within the entire slide. Views of specific regions indicated by inset boxes are shown in Supplementary Fig. 31d.

(d) iPT subclusters visualized on H&E. Subset images showing populations on H&E. Blue cells correspond with cluster 0 (iPT_APOE), orange with cluster 1 (iPT_SPP1) and green with cluster 2 (iPT_KRT7).

(e) Frequency of immune and fibroblast neighbors for each iPT subtype within each sample within 20 microns is shown below. We performed testing using a Wilcoxon rank sum test between each population within a sample. These subtypes had significantly different immune neighbors and fibroblast neighbors with each sample. HK3039 fibroblasts: iPT_APOE vs iPT_SPP1, p-value = 7e-39. HK3039 immune cells: iPT_APOE vs iPT_SPP1, p-value = 4e-67. HK2844 fibroblasts iPT_APOE vs iPT_SPP1, p-value = 9e-9, iPT_KRT7 vs iPT_SPP1, p-value = 9e-46. HK2844 immune cells iPT_APOE vs iPT_SPP1, p-value = 3e-16, iPT_KRT7 vs iPT_SPP1, p-value = 9e-11.

(f) iPT subcluster neighborhood enrichment within a 20-micron neighborhood size. Lighter color indicates higher enrichment and co-localization of a given population.

Supplementary Fig. 32. 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) Heat of gene expression changes correlating with eGFR and fibrosis in each cell type. Red indicates larger number of correlating cells.

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.

Supplementary Fig 33. 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.

Supplementary Fig. 34. 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. 35. 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.

Supplementary Fig. 36. 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.

Supplementary Fig. 37. FME gene expression predicts kidney outcomes

Hierarchical clustering of 245 human kidney tubule samples based on the expression of 1,100 randomly picked genes.

Kaplan Meier analysis with Log-rank test was used to compare the survival of the 3 different clusters. Renal survival was defined as cases reaching end stage renal disease or greater than 40% eGFR decline.

Single cell expression enrichment of genes associated with eGFR decline. The heatmap shows the cell type enrichment of genes associated with eGFR decline (red indicates more genes with cell type expression). Endo; endothelial cells, Stroma; stromal cells, PEC; parietal epithelial cells, Podo; podocyte, PT; proximal tubule cells, LOH; loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells. Spatial expression and microenvironment enrichment of genes associated with eGFR decline. GME: glomerular, TME: tubule, FME: fibrosis, IME: immune microenvironment.

Using the LASSO regression of all FME genes against eGFR, Kaplan Meier Analysis was re-performed using clustering of subsets of gene-- those with a Lasso coefficient that is non-zero and the genes with the most negative coefficients (Supplementary Table 14).

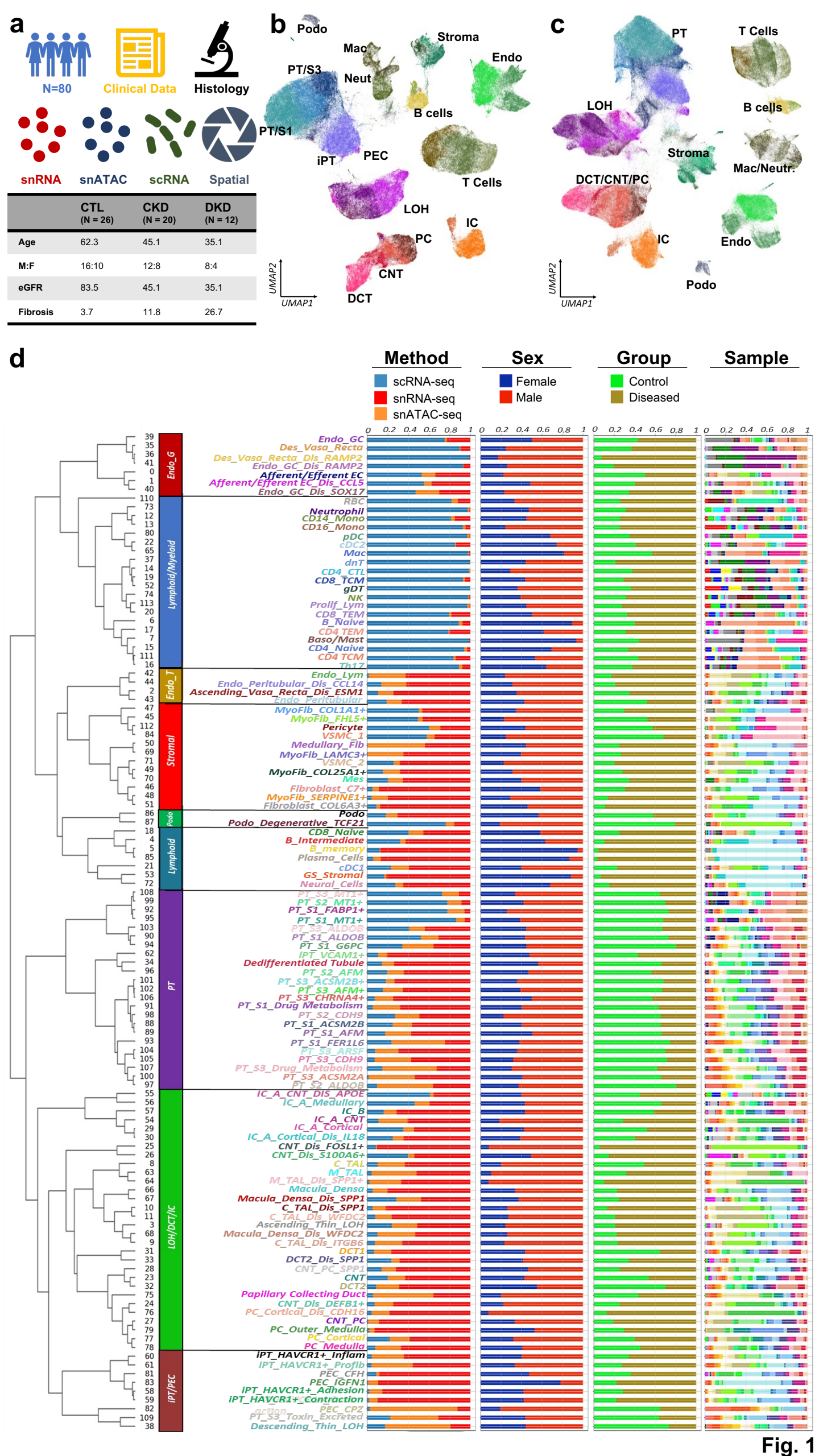

a
N=80
Clinical Data
Histology
snRNA
snATAC
scRNA
Spatial
b
Podo
Mac
Stroma
Endo
PT/S3
Neut
B cells
PT/S1
iPT
PEC
T Cells
LOH
IC
PC
CNT
DCT
UMAP2
UMAP1
c
PT
T Cells
LOH
B cells
Stroma
Mac/Neutr.
DCT/CNT/PC
IC
Endo
Podo
UMAP2
UMAP1
d
Method
scRNA-seq
snRNA-seq
snATAC-seq
Sex
Female
Male
Group
Control
Diseased
Sample
Endo_G
Lymphoid/Myeloid
Endo_T
Stromal
Podo
Lymphoid
PT
LOH/DCT/IC
iPT/PEC
Endo_GC
Des_Vasa_Recta
Des_Vasa_Recta_Dis_RAMP2
Endo_GC_Dis_RAMP2
Afferent/Efferent EC
Afferent/Efferent EC_Dis_CCL5
Endo_GC_Dis_SOX17
RBC
Neutrophil
CD14_Mono
CD16_Mono
pDC
cDC2
Mac
dnT
CD4_CTL
CD8_TCM
gDT
NK
Prolif_Lym
CD8_TEM
B_Naive
CD4 TEM
Baso/Mast
CD4_Naive
CD4 TCM
Th17
Endo_Lym
Endo_Peritubular_Dis_CCL14
Ascending_Vasa_Recta_Dis_ESM1
Endo_Peritubular
MyoFib_COL1A1+
MyoFib_FHL5+
Pericyte
VSMC_1
Medullary_Fib
MyoFib_LAMC3+
VSMC_2
MyoFib_COL25A1+
Mes
Fibroblast_C7+
MyoFib_SERPINE1+
Fibroblast_COL6A3+
Podo
Podo_Degenerative_TCF21
CD8_Naive
B_Intermediate
B_memory
Plasma_Cells
cDC1
GS_Stromal
Neural_Cells
PT_S3_MT1+
PT_S2_MT1+
PT_S1_FABP1+
PT_S1_MT1+
PT_S3_ALDOB
PT_S1_ALDOB
PT_S1_G6PC
iPT_VCAM1+
Dedifferentiated Tubule
PT_S2_AFM
PT_S3_ACSM2B+
PT_S3_AFM+
PT_S3_CHRNA4+
PT_S1_Drug Metabolism
PT_S2_CDH9
PT_S1_ACSM2B
PT_S1_AFM
PT_S1_FER1L6
PT_S3_ARSF
PT_S3_CDH9
PT_S3_Drug_Metabolism
PT_S3_ACSM2A
PT_S2_ALDOB
IC_A_CNT_DIS_APOE
IC_A_Medullary
IC_B
IC_A_CNT
IC_A_Cortical
IC_A_Cortical_Dis_IL18
CNT_Dis_FOSL1+
CNT_Dis_S100A6+
C_TAL
M_TAL
M_TAL_Dis_SPP1+
Macula_Densa
Macula_Densa_Dis_SPP1
C_TAL_Dis_SPP1
C_TAL_Dis_WFDC2
Ascending_Thin_LOH
Macula_Densa_Dis_WFDC2
C_TAL_Dis_ITGB6
DCT1
DCT2_Dis_SPP1
CNT_PC_SPP1
CNT
DCT2
Papillary Collecting Duct
CNT_Dis_DEFB1+
PC_Cortical_Dis_CDH16
CNT_PC
PC_Outer_Medulla
PC_Cortical
PC_Medulla
iPT_HAVCR1+_Inflam
iPT_HAVCR1+_Profib
PEC_CFH
PEC_IGFN1
iPT_HAVCR1+_Adhesion
iPT_HAVCR1+_Contraction
PEC_CPZ
PT_S3_Toxin_Excreted
Descending_Thin_LOH

Fig. 1

Fig 1. Comprehensive integrated multimodal human kidney single cell atlas

(a). Overview of the multimodal analysis. Basic clinical characteristics of the samples.

(b) UMAP of 338,565 integrated human kidney sn/scRNAseq and snATACseq data generated in this study. Annotated cell types are indicated on the plot.

(c) UMAP of 588,425 integrated human kidney sn/scRNAseq and snATACseq data of the present study and kidney precision medicine project (KPMP).

(**d**) Hierarchical sub-clustering identified 114 distinct cell types or cell states. The bar charts depict the relative abundance of each group (method, sex, group, and samples) contributing to the cluster.

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, iPT; injured proximal tubule cells, DTL; descending thin 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, gDT; gamma delta T cells, TCM; T Cell memory, TEM; T effector memory, 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+, CTL; control, HKD; hypertensive kidney disease, DKD; diabetic kidney disease.

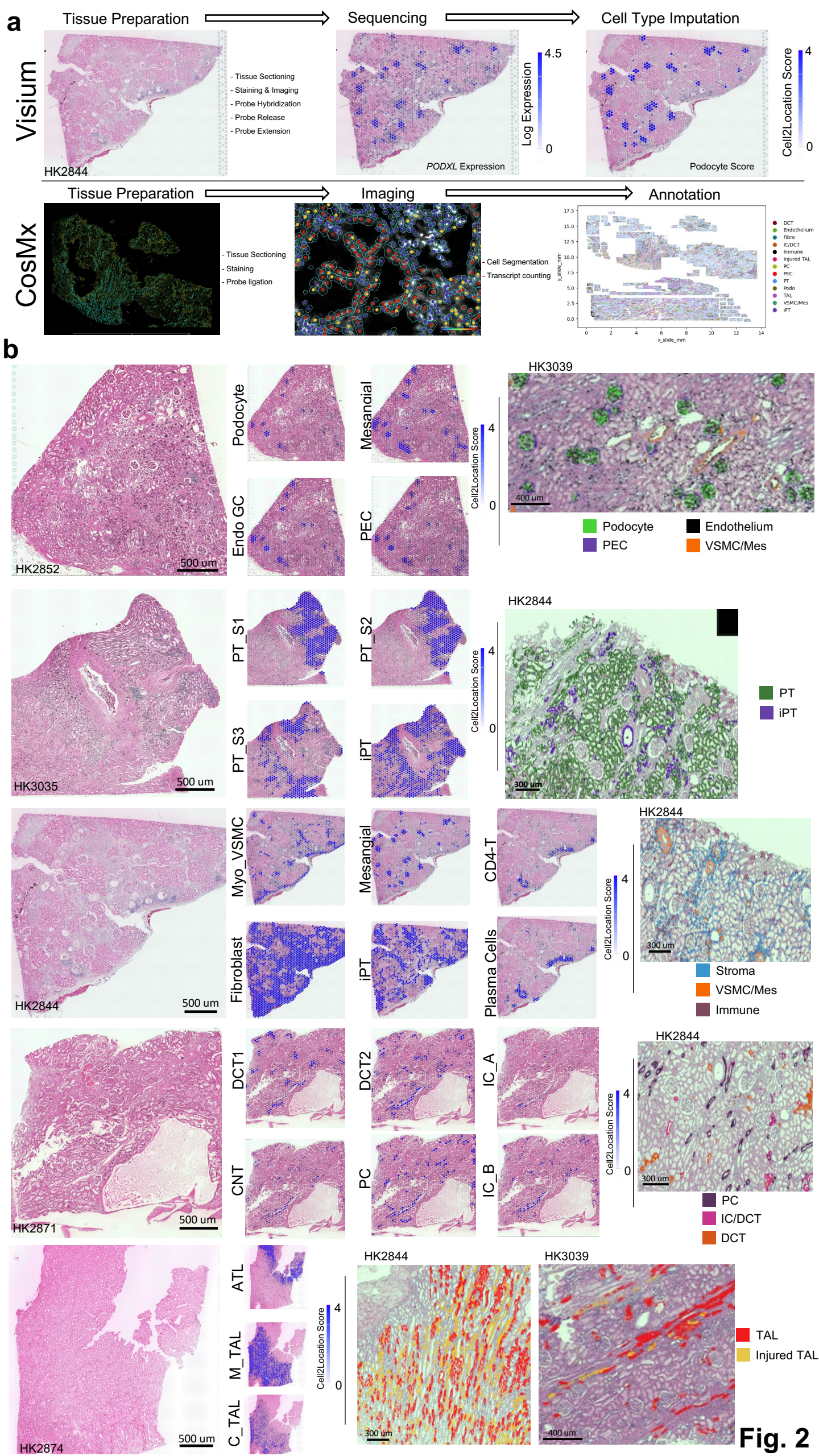

a
Tissue Preparation
Sequencing
Cell Type Imputation
Visium
- Tissue Sectioning
- Staining & Imaging
- Probe Hybridization
- Probe Release
- Probe Extension
HK2844
PODXL Expression
Log Expression
4.5
0
Podocyte Score
Cell2Location Score
4
0
Tissue Preparation
Imaging
Annotation
CosMx
- Tissue Sectioning
- Staining
- Probe ligation
- Cell Segmentation
- Transcript counting
DCT
Endothelium
Fibro
IC/DCT
Immune
Injured TAL
PC
PEC
PT
Podo
TAL
VSMC/Mes
iPT
b
HK2852
500 um
Podocyte
Mesangial
Endo GC
PEC
HK3039
400 um
Podocyte
Endothelium
PEC
VSMC/Mes
HK3035
500 um
PT_S1
PT_S2
PT_S3
iPT
HK2844
300 um
PT
iPT
HK2844
500 um
Myo_VSMC
Mesangial
CD4-T
Fibroblast
iPT
Plasma Cells
HK2844
300 um
Stroma
VSMC/Mes
Immune
HK2871
500 um
DCT1
DCT2
IC_A
CNT
PC
IC_B
HK2844
300 um
PC
IC/DCT
DCT
HK2874
500 um
ATL
M_TAL
C_TAL
HK2844
300 um
HK3039
400 um
TAL
Injured TAL

Fig. 2

Fig 2. Spatially resolved human kidneys.

(a) Spatial transcriptomics data was generated from human formalin fixed paraffin embedded kidney samples using two platforms. Visium (top) employs a spot-based approach with each spot of 55 um, with the ability to detect >18,000 genes able for detection and requires deconvolution to identify presence of individual cell types. CosMx (bottom) imaging generates single cell level data and identifies 1000 genes, and which permits annotation of cell types based on the expression patterns of these genes.

(b) Spatial location and marker gene expression of identified cell types using both Visium and CosMx. H&E sections shown on the left of the figure show individual tissue histology of our Visium sections. Using our calculated Cell2Location scores for each of these tissues, we imputed the presence of cell types on each of these sections with cell type scores shown in blue overlaying the H&E. We identify glomerular cell types, proximal tubular cell types, immune and fibrotic cell types, distal tubular cell types and loop of Henle cell types. To the right of these images, we validated the location of these cell types using the CosMx assay, with allows for individual cell type annotations, which are overlayed on the H&E image of the assayed tissue.

Endo_GC; endothelial cells of glomerular capillary tuft, Mes; mesangial cells, GS_Stromal; glomerulosclerosis-specific stromal cells, Myo_VSMC; myofibroblast/vascular smooth muscle cells PEC; parietal epithelial cells, Podocyte; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, iPT; injured proximal tubule cells, ATL: ascending thin limb of 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, CD4T; T lymphocytes CD4+, Plasma cell; plasma cell.

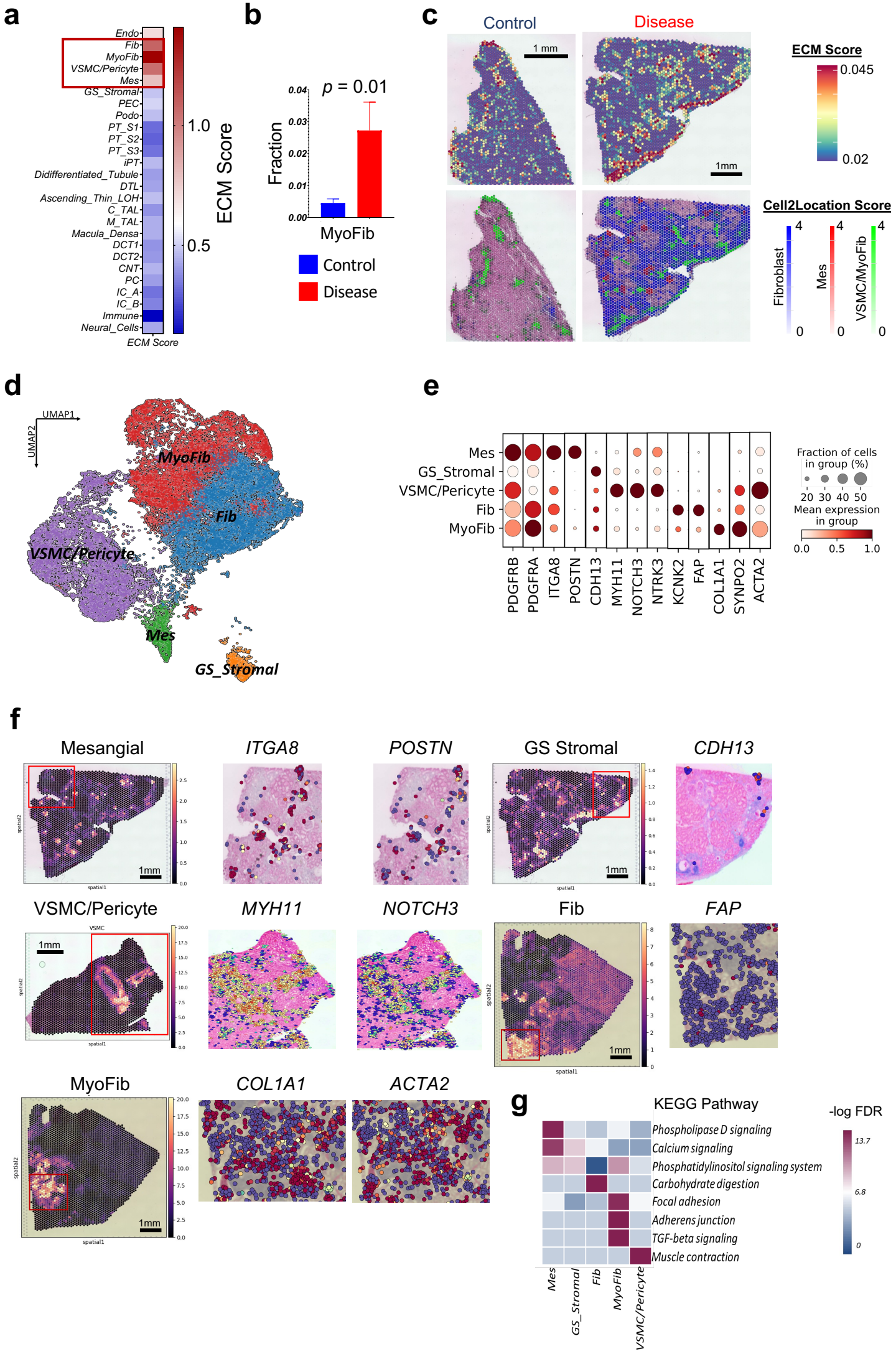

a
Endo
Fib
MyoFib
VSMC/Pericyte
Mes
GS_Stromal
PEC
Podo
PT_S1
PT_S2
PT_S3
iPT
Didifferentiated_Tubule
DTL
Ascending_Thin_LOH
C_TAL
M_TAL
Macula_Densa
DCT1
DCT2
CNT
PC
IC_A
IC_B
Immune
Neural_Cells
ECM Score
1.0
0.5
ECM Score
b
p = 0.01
Fraction
0.04
0.03
0.02
0.01
0.00
MyoFib
Control
Disease
c
Control
Disease
1 mm
1mm
ECM Score
0.045
0.02
Cell2Location Score
Fibroblast
Mes
VSMC/MyoFib
4
0
d
UMAP1
UMAP2
MyoFib
Fib
VSMC/Pericyte
Mes
GS_Stromal
e
Mes
GS_Stromal
VSMC/Pericyte
Fib
MyoFib
PDGFRB
PDGFRA
ITGA8
POSTN
CDH13
MYH11
NOTCH3
NTRK3
KCNK2
FAP
COL1A1
SYNPO2
ACTA2
Fraction of cells in group (%)
20 30 40 50
Mean expression in group
0.0 0.5 1.0
f
Mesangial
ITGA8
POSTN
GS Stromal
CDH13
VSMC/Pericyte
MYH11
NOTCH3
Fib
FAP
MyoFib
COL1A1
ACTA2
1mm
g
KEGG Pathway
-log FDR
13.7
6.8
0
Phospholipase D signaling
Calcium signaling
Phosphatidylinositol signaling system
Carbohydrate digestion
Focal adhesion
Adherens junction
TGF-beta signaling
Muscle contraction
Mes
GS_Stromal
Fib
MyoFib
VSMC/Pericyte

Fig. 3

Fig 3. Human kidney stromal cell atlas.

(a) Extracellular matrix (ECM) gene expression score in different kidney cell types in the integrated sn/scRNAseq and snATACseq. ECM score was calculated based on the expression of collagen, proteoglycan, and glycoprotein genes.

(b) The fraction of myofibroblasts in control and CKD samples. The bars indicate SEM. Independent t-test was used to compare the fractions between two groups.

(c) The ECM score in the spatial transcriptomics datasets of healthy and CKD samples (upper panel). Red indicates higher ECM gene expression (upper panel). The lower panel indicates the location of mesangial cells, fibroblasts, myofibroblasts and VSCMC/pericytes in healthy and CKD samples based on their Cell2Location score.

(d) UMAP representation of sub-clustering of 32,706 stomal cells in the integrated dataset -- present study and Kidney Precision Medicine Project (KPMP).

(**e**) The dot plots of marker genes used for stromal cell type annotation in the integrated dataset. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

(f) The spatial location and stromal cell subtypes and specific marker genes. The left panel shows the relative abundance of each cell type using Cell2location. The right panel shows relative gene expression using CellTrek (red higher).

(**g**) The heatmap indicates the $-Log_{10}$ (FDR) enrichment of the top KEGG pathways in each stromal cell type.

Mes; mesangial cells, GS_Stromal; glomerulosclerois-specific stromal cells, Fib; fibroblast, VSMC; vascular smooth muscle cells, MyoFib; myofibroblast.

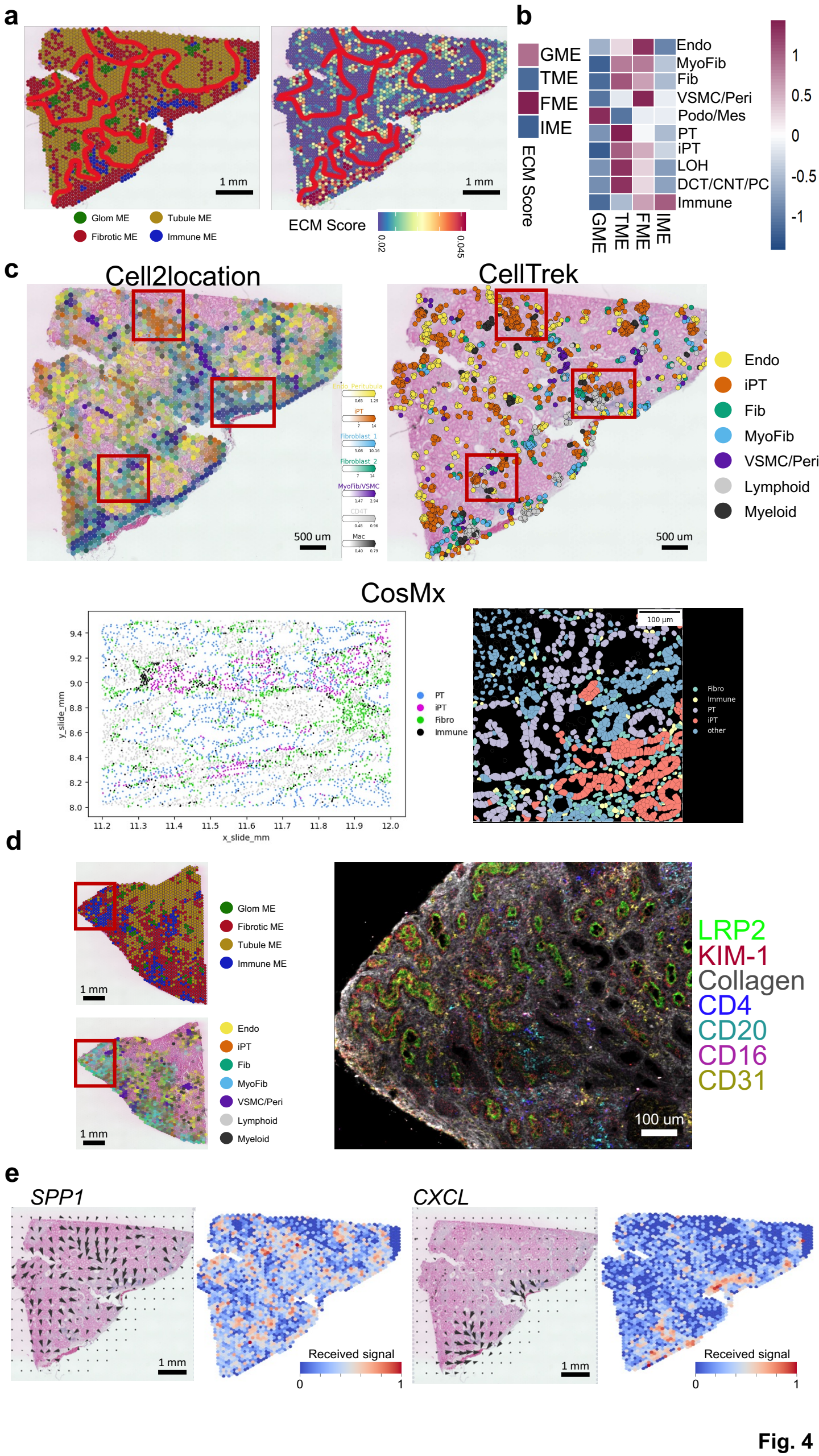

a
1 mm
1 mm
Glom ME
Tubule ME
Fibrotic ME
Immune ME
ECM Score
0.02
0.045
b
GME
TME
FME
IME
ECM Score
Endo
MyoFib
Fib
VSMC/Peri
Podo/Mes
PT
iPT
LOH
DCT/CNT/PC
Immune
GME
TME
FME
IME
1
0.5
0
-0.5
-1
c
Cell2location
CellTrek
500 um
500 um
Endo
iPT
Fib
MyoFib
VSMC/Peri
Lymphoid
Myeloid
CosMx
y_slide_mm
x_slide_mm
PT
iPT
Fibro
Immune
100 µm
Fibro
Immune
PT
iPT
other
d
Glom ME
Fibrotic ME
Tubule ME
Immune ME
1 mm
Endo
iPT
Fib
MyoFib
VSMC/Peri
Lymphoid
Myeloid
1 mm
LRP2
KIM-1
Collagen
CD4
CD20
CD16
CD31
100 um
e
SPP1
CXCL
1 mm
Received signal
0
1
1 mm
Received signal
0
1

Fig. 4

Fig 4. Human kidney fibrotic microenvironment.

(a) Human kidney microenvironments defined by non-negative matrix factorization (NMF) of the spRNAseq. Glomerular (green), tubule (brown), fibrotic (red), immune (blue). The spatial distribution of the calculated ECM score (right panel). Red indicates higher ECM gene expression.

(b) Correlation of microenvironment with ECM score (left) and cell type correlation with each microenvironment (right).

(**c**) Cell2location (left), Celltrek (right) cell type imputations of the Visium spatial data and CosMx (below) annotations (below) showing the location of different cell types in the fibrotic microenvironment in a diseased kidney sample. CosMx annotations of iPT, PT, fibroblasts and immune cells are shown at 2 different magnifications.

(**d**) *In situ* mass spec image of CKD kidney (from panel E) labeled with LRP2 (PT) KIM1 (iPT), CD4 (T-cell) CD20 (B-cell), CD16 (myeloid) and CD31 (endo) in a fibrotic human kidney sample.

(**e**) The spatial location of the identified cell-cell interaction pathway (*SPP1* and *CXCL*). The arrows indicate the source and targets of the identified pathways, and the color indicates the received signals (red higher). The results were obtained using COMMOT package.

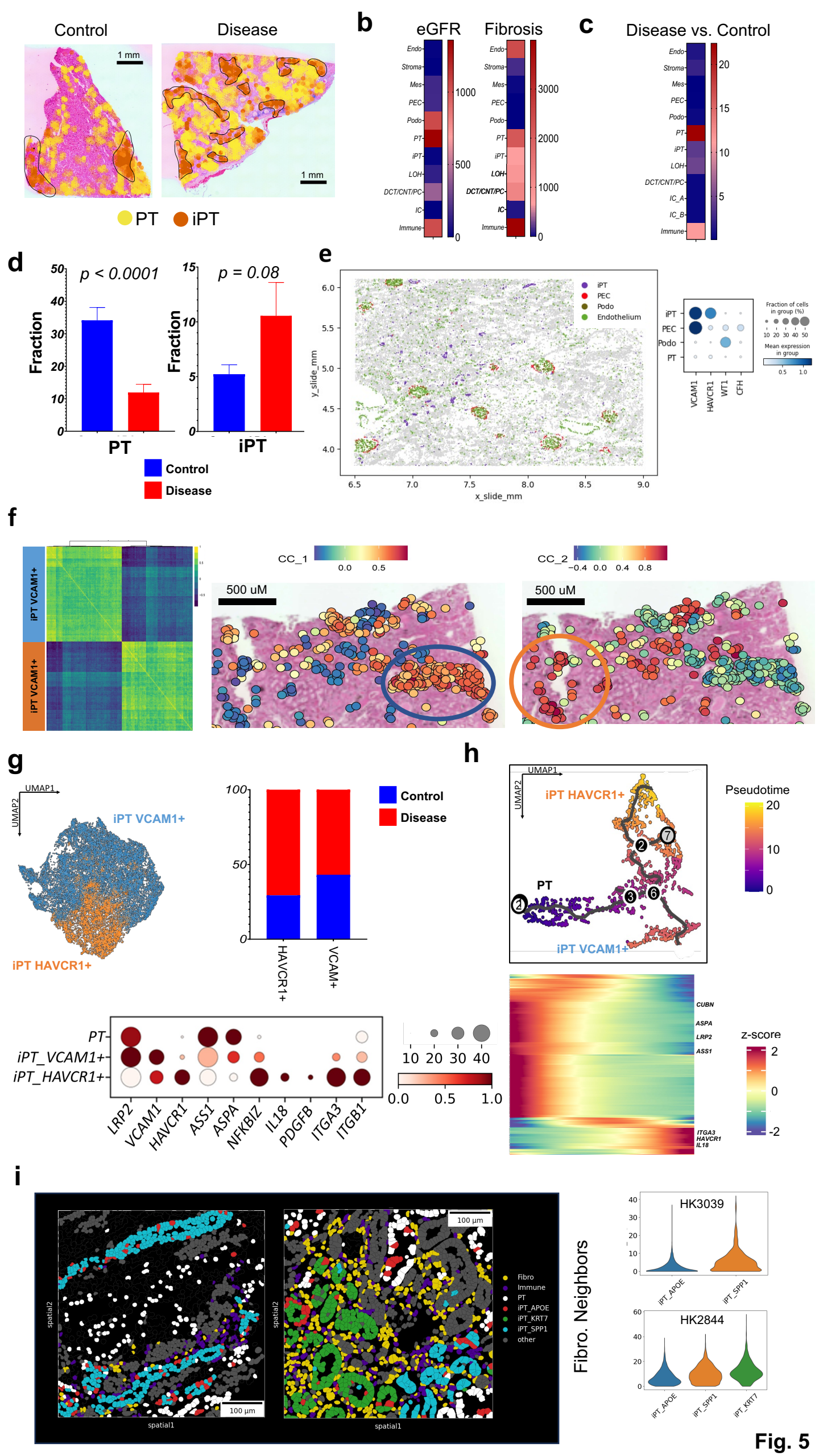

Control
Disease
1 mm
PT
iPT
b
eGFR
Fibrosis
Endo
Stroma
Mes
PEC
Podo
PT
iPT
LOH
DCT/CNT/PC
IC
Immune
c
Disease vs. Control
IC_A
IC_B
d
p < 0.0001
p = 0.08
Fraction
PT
iPT
Control
Disease
e
iPT
PEC
Podo
Endothelium
y_slide_mm
x_slide_mm
VCAM1
HAVCR1
WT1
CFH
Fraction of cells in group (%)
Mean expression in group
f
iPT VCAM1+
CC_1
CC_2
500 uM
g
UMAP1
UMAP2
iPT VCAM1+
iPT HAVCR1+
HAVCR1+
VCAM+
PT
iPT_VCAM1+
iPT_HAVCR1+
LRP2
VCAM1
HAVCR1
ASS1
ASPA
NFKBIZ
IL18
PDGFB
ITGA3
ITGB1
h
iPT HAVCR1+
Pseudotime
z-score
CUBN
ASPA
LRP2
ASS1
ITGA3
HAVCR1
IL18
i
100 µm
spatial1
spatial2
Fibro
Immune
PT
iPT_APOE
iPT_KRT7
iPT_SPP1
other
Fibro. Neighbors
HK3039
HK2844

Fig. 5

Fig 5. Injured proximal tubule cells in diseased human kidneys.

(a) Localization of iPT and PT cells within control and diseased samples in the human kidney. iPT is scattered in a patchy pattern, and colocalize amongst themselves.

(**b**) Heatmap of number of genes correlated with eGFR and renal fibrosis in each cell type. The red shows higher number of genes and blue indicates lower number of genes.

(c) Heatmap of number of genes comparing prevalence in healthy vs. diseased samples in each cell type. PT had the highest changes between control and diseased samples, with red indicating more genes.

(d) The fractions of PT and iPT in control and diseased samples type in the integrated sn/scRNAseq and snATACseq dataset. Bar indicated the SEM. For the comparison between two groups, t-test was used.

(e) CosMx spatial transcriptomics shows that annotation of iPT and PECs localize to tubules and glomeruli, respectively. Dot plot shows *VCAM1* is expressed in both populations, while *HAVCR1* and *CFH* appear to be more specific for iPT and PECs, respectively.

(f) Gene co-expression network analysis of human kidney spRNAseq data indicates two modules expressing *VCAM1* or *HAVCR1*. The right panels indicate the spatial location of *VCAM1* or *HAVCR1* (KIM1) positive injured PT cells. The color indicates gene expression of iPT modules (red higher expression).

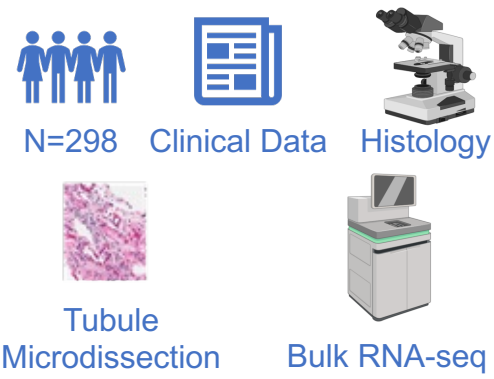
(g) Sub-clustering of iPT cells from integrated atlas. The fraction of VCAM1 or HAVCR1 positive iPT cells in control and diseased kidneys, shown on UMAP and quantified in frequency. The dot plots show the expression marker genes in iPT and PT cells. The size of the dot indicates the percent positive cells, and the darkness of the color indicates average expression.

(h) Cell trajectory analysis (Monocle) representation of PT and iPT cells (Upper panels). The heatmap shows the differentially expressed genes along the trajectory, cells ordered by pseudotime. Red indicates higher expression.

(i) Sub-clustering of iPT within the CosMx data shows 3 subtypes of iPT. These iPT subtypes are located near PT cells and are visualized in specific fields of view—one in each sample. These iPT subtypes have different neighboring cells (within a 50 micron radius) with iPT_KRT7 having most frequent fibroblast and immune neighbors within our diseased sample (*p*-value = 1e-177 , 1e-17 for fibroblasts and immune cells respectively within our disease sample by Wilcoxon rank sum test). iPT_APOE had fewer immune (*p*-value = 0, *p*-value = 0, for HK3039 and HK2844) and fibroblast neighbors than the iPT_SPP1 subtype. (*p*-value = 5e-52, *p*-value = 1e-52, for HK3039 and HK2844).

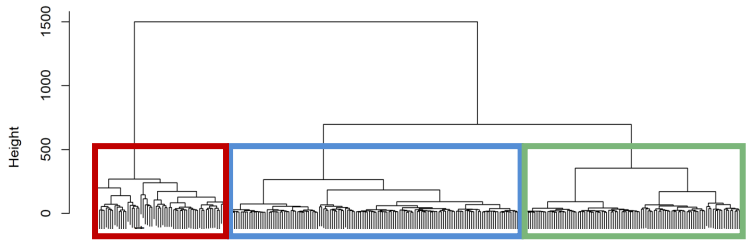
Endo; endothelial cells Mes; mesangial cells, Stroma; stromal cells, PEC; parietal epithelial cells, Podo; podocyte, PT; proximal tubule cells, iPT; injured proximal tubule cells, LOH; loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct,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

a



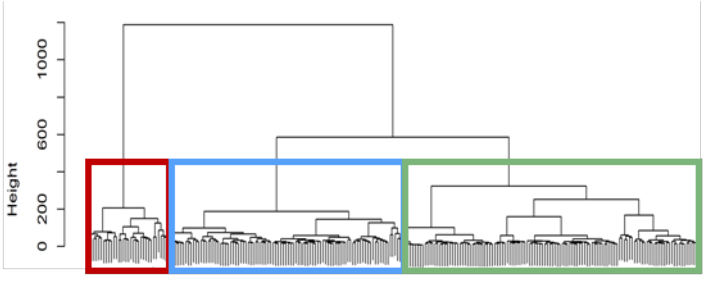
Subjects (n)	N=298
eGFR (ml/min/1.73 m2)	66.41 (27.29)
Gender (%Male)	180 (60.4%)
Age	61.16 (13.61)
Race	
Asian, n (%)	1 (0.3%)
Caucasian, n (%)	140 (46.97%)
African American, n (%)	99 (33.2%)
Hispanic, n (%)	10 (3.35%)
Multi-racial/Unknown, n (%)	28 (9.4%)
Diabetes Mellitus, n (%)	87 (29.2%)
Hypertension, n (%)	196 (65.77%)
BMI (kg/m2)	30.7 (7.68)
Serum glucose (mg/dl)	127.45 (51.64)
SBP (mmHg)	136.81 (19.15)
DBP (mmHg)	77.15 (12.17)
Interstitial Fibrosis (%)	15.82 (25.44)

b



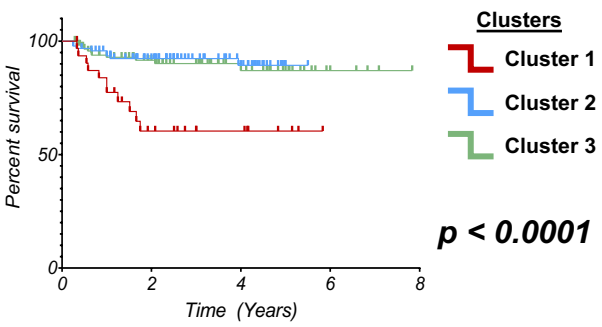
Variables	N = 62	N = 136	N = 103	*p*
Age	60.35 ± 15.48	60.51 ± 12.23	61.81 ± 12.41	0.86
Male (%)	48.3 %	55.9 %	68.9 %	**0.02**
AA (%)	46.8 %	39.7 %	15.53 %	**0.0001**
HTN (%)	71%	63.2%	64.1%	0.18
DM (%)	27.4 %	28.7 %	29.1 %	0.23
eGFR (ml/min/1.73 m^2)	54.22 ± 33.15	67.59 ± 23.16	72.05 ± 22.57	**0.001**
Fibrosis (%)	45.47 ± 41.31	8.45 ± 10.98	6.67 ± 9.255	**0.0001**

c

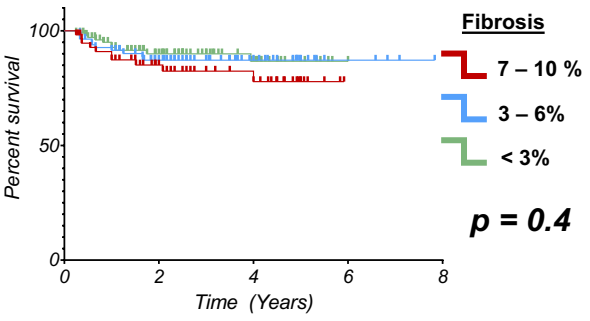


Variables	N = 33	N = 94	N = 118	*P*
Age	63.33 ± 11.59	62.49 ± 11.26	60.29 ± 12.4	0.26
Male (%)	66.7 %	70.21 %	47.45 %	**0.002**
AA (%)	6.1 %	11.7 %	21.2 %	**0.03**
HTN (%)	66.7 %	70.21 %	65.2 %	0.78
DM (%)	42.4 %	33 %	33.9 %	0.59
eGFR Baseline (ml/min/1.73m^2)	70.57 ± 18.28	73.18 ± 19.95	71.45 ± 21.65	0.75
eGFR Follow Up (ml/min/1.73m^2)	54.42 ± 25.12	58.06 ± 21.63	61.1 ± 24.71	0.31
eGFR Slope (ml/min/1.73m^2)	-23.5 ± 40.61	-9.4 ± 17.35	-6.6 ± 13.38	**0.01**
Fibrosis (%)	4.82 ± 3.41	4.64 ± 3.44	4.19 ± 3.08	0.47

d



Cluster 1 → HR = 3.95 (P = 0.001) Renal Function Decline



Fibrosis 7 – 10% → HR = 1.17 (P = 0.71) Renal Function Decline

Fig. 6

Fig 6. Fibrotic microenvironment gene signature successfully predicts disease prognosis in a large cohort of human kidney samples.

(a) Clinical characteristics of 292 human kidney tubule RNA samples.

(b) Unbiased cluster dendrogram of 292 human kidney tubule bulk RNAseq samples based on FME gene signature. Clinical characteristics of each cluster. Chi-square test for categorical variables and one-way ANOVA for continuous variables were used to compare groups.

(**c**) Unbiased cluster dendrogram of 245 human kidney tubule bulk RNAseq samples with fibrosis <10% based on expression of FME genes. Clinical characteristics of each cluster are shown in the table. Chi-square test for categorical variables and one-way ANOVA for continuous variables were used to compare groups.

(**d**) Kaplan-Meier analysis of 292 kidney samples based on FME gene signature. Renal survival was defined as cases reaching end stage renal disease or greater than 40% eGFR decline (above). Kaplan-Meier analysis of 245 kidney samples based on pathologist defined kidney fibrosis degree (<3%, 3-6% or 7-10% as defined by an expert pathologist) (below). Renal survival was defined as cases reaching end stage renal disease (eGFR of 15 ml/min/1.73m2) or greater than 40% eGFR decline.

FME-GS; fibrotic microenvironment gene signature, HR; hazard ratio, HTN: hypertension, DM: diabetes, AA: African American.