bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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:
Authors: Amin Abedini[1,2,3], 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[4], Paola Grandi[5], Tanmoy Bhattacharyya[6], Erding Hu[6], Steven S. Pullen[7], Carine M Boustany-Kari[7], Paolo Guarnieri[7], Anil Karihaloo[8], Hanying Yan[9], Kyle Coleman[9], Matthew Palmer[10], Lea Sarov-Blat[6], Lori Morton[4], Christopher A. Hunter[11], Mingyao Li[9], Katalin Susztak[1,2,3*]

Affiliations:

[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]Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA

[4]Cardiovascular, Renal and Fibrosis Research, Regeneron Pharmaceuticals Inc., Tarrytown, NY 10591, USA

[5] Research and Development, GSK, Cellzome GmbH, Genomic Sciences, GSK, Heidelberg, Germany

[6] Research and Development, GSK, Crescent Drive, Philadelphia, Pennsylvania, USA

[7]Department of Cardiometabolic Diseases Research, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA.

[8]Novo Nordisk Research Center Seattle, Inc., Seattle, USA

[9]Department of Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[10]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[11]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

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Kidneys have one of the most complex three-dimensional cellular organizations in the body, but the spatial molecular principles of kidney health and disease are poorly understood. Here we generate high-quality single cell (sc), single nuclear (sn), spatial (sp) RNA expression and sn open chromatin datasets for 73 samples, capturing half a million cells from healthy, diabetic, and hypertensive diseased human kidneys. Combining the sn/sc and sp RNA information, we identify > 100 cell types and states and successfully map them back to their spatial locations. Computational deconvolution of spRNA-seq identifies glomerular/vascular, tubular, immune, and fibrotic spatial microenvironments (FMEs). Although injured proximal tubule cells appear to be the nidus of fibrosis, we reveal the complex, heterogenous cellular and spatial organization of human FMEs, including the highly intricate and organized immune environment. We demonstrate the clinical utility of the FME spatial gene signature for the classification of a large number of human kidneys for disease severity and prognosis. We provide a comprehensive spatially-resolved molecular roadmap for the human kidney and the fibrotic process and demonstrate the clinical utility of spatial transcriptomics.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

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

More than 30 specialized cell types including epithelial, endothelial, interstitial and immune cells have been identified in the kidney(*6, 7*). The development of novel single cell and single nuclear RNA-sequencing (scRNA-seq, snRNA-seq, respectively) as well as single nuclei Assay for Transposase-Accessible Chromatin sequencing (snATAC-seq) have provided an unprecedented insight into the molecular and cellular composition of healthy mouse and human kidneys as well as changes during development and disease(*8-12*). These methods use dissociated cells or nuclei isolated from kidney tissue samples. Despite the significant cellular diversity of the kidney, cell types could be identified even after cell dissociation as specialized cellular function matches with gene expression signatures, allowing investigators to estimate the location of cells (*13*).

The kidney is an architectural masterpiece. A critical limitation of dissociated single cell datasets has been the lack of information on the spatial cellular context(*14*). Without spatial information, it has been difficult to map known cell types that are only described by their anatomical location, for example, cells that mostly provide structural support. The spatial context is also critical for mapping cell types and cell states identified by novel single cell tools. We observe important regional differences in disease severity, the dissociated single cell data is unable to interrogate local gene expression changes and cell-cell communication, which plays a critical role in maintaining cellular health and dysregulated in disease. Spatial omics analysis is a rapidly evolving

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

field. Currently available spatial datasets either lack single cell resolution information, are unable to provide genome-scale gene expression data, or only evaluate a small spatial area (*13, 15, 16*). There is a clear need for large-scale spatial omics datasets to better understand kidney health and disease.

Chronic kidney disease (CKD), regardless of disease etiology, is associated with a complex change in the kidney's cellular architecture(*17*). Some of the histological changes are specific for disease type, such as thickening of glomerular basement membrane in diabetic kidney disease (DKD)(*18*). Architectural changes, collectively called fibrosis, are present in all kidneys with advanced CKD. The narrow definition of fibrosis focuses on accumulation extracellular matrix (*19, 20*). Most prior studies, therefore, concentrated on understanding matrix accumulation in diseased organs. Matrix accumulation can cause organ stiffness, which is likely responsible for organ failure in pulmonary and heart fibrosis(*21-23*). As the role of tissue elasticity in kidney function regulation is not immediately obvious(*24*), the mechanism by which matrix accumulation (or fibrosis) affects kidney function has been controversial(*25, 26*). Kidney function only modestly correlate with fibrosis (r = 0.4)(*18, 27*).

Here, we generated spRNA-seq data for healthy and diseased human kidneys in conjunction with sn/scRNA-seq, snATAC-seq. By combining spatial gene expression with high quality single cell expression and open chromatin information, we resolve the identity of cells previously only known by their spatial localization and perform a detailed two-dimensional characterization of tissue fibrosis. We demonstrate the cellular heterogeneity of the fibrotic stroma, which includes not only immune and matrix-producing fibroblasts but also endothelial cells and immune cells that follow the organization of a lymphoid organ that are anatomically close to injured proximal tubule cells. We define tissue microenvironments, including the fibrotic microenvironment (FME) and show

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that the FME gene signature can classify kidney samples and predict future kidney function decline.

Results

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

We generated a comprehensive human kidney single cell and spatial resolution atlas by analyzing 73 human kidney tissue samples from 49 subjects (59.2% male and age of 63.75 ± 12.44 years). Samples were divided into two groups: (i) healthy control (N=35) determined by estimated glomerular filtration rate (eGFR) > 60 ml/min/1.73 m^2 and fibrosis < 5 % (ii) chronic kidney disease (CKD) (N=38) determined by (eGFR) < 60 ml/min/1.73 m^2 or kidney fibrosis > 10% including 18 with diabetic kidney disease (DKD) and 20 with hypertensive kidney disease (HKD). **Supplementary Table 1** shows the detailed demographic, clinical, and histological characteristics of the included samples.

We performed droplet-based single cell analysis using 10X Chromium Next GEM (sc/snRNA-seq (*N=46*) and snATAC-seq (*N=20*)) and used the Visium formalin-fixed paraffin embedded (FFPE) tissue (*N=7*) platform for spRNA-seq. After standard processing and meticulous quality control (QC), removing low-quality cells, we included 453,782 cells/nuclei into our final atlas. **Supplementary Fig. 1** and **Supplementary Table 2** contains QC metrices of the included samples. Overall, we could identify six cell super families, including endothelial cells, stromal cells, tubule epithelial cells, immune cell types, glomerular cells, and neural cells. Clustering identified 37 main and 111 distinct cell sub-types or states in healthy and diseased human kidneys

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Fig. 1,** and **Supplementary Fig. 2,3**). Key cluster-specific gene markers are shown in **Fig. 1, Supplementary Fig. 3** and **Supplementary tables 3 to 6**. Our sc and sn human kidney atlas captured all kidney cell types in healthy and disease status in all anatomical regions. The main identified cell types were: podocytes, different types of proximal tubules segments 1-3 (PT_S1, S2, S3, and injured), descending loop of Henle (DLOH), cortical and medullary thick ascending loop of Henle (C_TAL and M_TAL), distal convoluted tubule (DCT), connecting tubule (CNT), principal cells of collecting duct (PC), intercalated cells type alpha and beta (IC_A and IC_B), stromal, and different types of immune cells.

The combination of single cell and single nuclear methods, the large number of analyzed cells, the high-quality dataset, and inclusion of samples with different degrees of kidney disease severity in our kidney atlas enabled the capture of rare and novel cell types. We could identify different stromal cell types we called fibroblast_1 (*COL1A1*+, *COL1A2*+), fibroblast_2 (*VIM*+, *IGFBP7*+, *B2M*+), and cells specifically present in sclerosed glomeruli (*CDH12*+, *CDH13*+) we called GS_stromal cells (**Fig. 1C, D, and Supplementary Fig 2,3**). We could capture 19 different types of endothelial cells and erythropoietin producing cells (Endo_peritubular_RAMP3+) (**Supplementary Fig. 3**). We captured proximal tubule (PT) cells expressing high levels of *SLC47A2,* specific for toxin excretion (**Supplementary Fig. 2, 3 and Supplementary Table 6)** and tubule epithelial subtypes mostly seen in diseased kidneys that were positive for *CTSD*, *CALB1*, *SPP1*, *CXCL14*.

Our atlas provides a comprehensive reference for human kidney immune cells. We could capture all lymphoid (CD4T, CD8T, natural killer cells, T_regulatory, B_Naiive, B_memory, plasma_cells) and myeloid cells (neutrophil, basophil/mast cells, CD14_monocyte, CD16_monocyte, macrophage, classical and plasmacytoid dendritic cells). In summary, we were

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

not only able to generate the most comprehensive analysis of human kidney cells, including multiple novel cell types, but these cell types were present in multiple analyzed samples and captured by multiple analytical methods (sn/scRNA and snATAC analysis) (**Supplementary Fig. 4**).

In addition to the gene expression data, the snATAC-seq of 80,845 human kidney nuclei provided us opportunities to identify transcription factors (TF) and enriched TF motifs in each cell type. Cell gene-expression markers and a comprehensive list of cell types' differentially accessible regions and transcription factors can be found in **Supplementary Fig 5, Supplementary Table 5, 7** and include *WT1* for podocyte and parietal epithelial cells (PEC), *HNF4A* for PT cell types, *FOSL2* for injured_PT (iPT), and *TFAP2A* for C_TAL.

A key limitation to cell type identification has been the lack of high-resolution spatially resolved cell transcriptomics information. To overcome this limitation, we used the new Visium FFPE platform and generated seven spRNA-seq data sets, including two control (healthy) and five diseased samples (3 DKD, 2 HKD) (**Supplementary Fig. 6**). We captured 2,043 ± 374 spots per sample and detected 3,471 ± 1,390 genes per spot, providing an extremely rich dataset and information (**Supplementary Fig. 6 and Supplementary Table 2**); enabling the identification of all key kidney cell types (24 clusters) now at spatial level (**Supplementary Fig. 7).**

As a next step, we co-embedded the dissociated sc/snRNA-seq and snATAC-seq with the spRNA-seq data, and generated an augmented high-resolution spatial dataset (94,696 datapoints) using CellTrek(*28*). The high-resolution data enabled the projection of all identified cell types from the dissociated datasets to its spatial location. Given differences in efficiencies of the cell capturing of the scRNA and snRNA datasets, we generated three cellular resolution spatially resolved atlases using our snRNA-seq (**Fig. 2**), scRNA-seq (**Supplementary Fig. 8**), and snATAC-seq

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Supplementary Fig. 9**). Via this method, we could successfully match the dissociated cell type expression data to their anatomical, cellular locations including all types of tubules, different interstitial cell types and endothelial cells. Furthermore, we could verify and highlight cell types, such as iPT, previously observed in dissociated datasets without anatomical location. We could identify markers for cell types previously only known by their anatomical location for instance, PEC cells express *CFH*, *VCAN*, and *VCAM1* as well as mesangial cells express *ITGA8* and *POSTN*. The different types of omics information (scRNA/snRNA/snATAC) provided a critical validation for our datasets. Our computational kidney spatial map was consistent with the reading of our renal pathologist as well as the Human Protein Atlas data (**Supplementary Fig. 10**).

Overall, we constructed a high-quality spatially resolved human kidney multiome atlas, which allowed the spatial mapping of high-resolution cellular and gene expression, gene regulatory information in health and disease states. The entire dataset is now available for the community on our easy-to-search website www.susztaklab.com.

The presence and spatial proximity of injured proximal tubule cells to stromal cells indicates their critical role in human kidney fibrosis

To identify key cell types and mechanisms of fibrosis in DKD and HKD, we applied a variety of unbiased computational tools. Differential gene expression (DEG) and accessible region (DAR) analysis between healthy and CKD samples highlighted PT, stroma, and immune cell types with the highest numbers of DEGs and DARs (**Supplementary Fig. 4**). As fibrosis is patchy, it has been difficult to understand driver pathways purely based on dissociated scRNAseq information(*29*). To understand the proximity of cells, we performed an in silico cellular

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

deconvolution of the analyzed spots using our snRNA-seq dataset as a reference. We determined the frequency when cells were captured together in the spatial data by running a correlation analysis. We found that the coexistence correlation of cell types frequency follows the anatomical regions in the kidney for example glomerular cells; glomerular endothelial cells, podocyte, PEC, mesangial were mostly captured together. We observed a similar pattern for PT, iPT, LOH, and distal tubes (**Fig. 3A, Supplementary Fig. 11A, 12, 13**). We found a strong correlation between stromal, immune cells, and iPT cells, indicating their co-existence/proximity in the measured spots (**Fig. 3A, Supplementary Fig. 11 A**). Healthy and diseased samples showed similar patterns. However, the colocalization of stromal, immune, and iPT cell types was more robust in diseased samples (**Supplementary Fig. 11B**).

Next, we generated an unbiased cell-cell distance matrix (measuring physical cell-cell distance) in the Cell-Trek imputed spRNA-seq dataset (**Fig. 3B**). Similarly, to the spot deconvolution method, we observed the proximity of glomerular cells and also the different types of fibroblast clusters **(Fig. 3C, Supplementary Fig. 11C)**. In this analysis, we found that PT cells, specifically injured proximal tubules (iPT), were the most common scattered cells in the kidney, indicating that iPT cells had the most diverse set of neighboring cells. We found that almost every kidney cell type; especially stromal and immune cells, colocalized with iPT cells. In summary, differential expression analysis indicated the high plasticity of PT cells and the close proximity of injured PT cells to other cell types (**Supplementary Fig. 11-13)**.

The spatial proximity and plasticity of PT cells made us focus on these cells. We found that the fraction of iPT cells was markedly higher in diseased kidneys (**Fig. 3C**). However, we also observed iPT cells in healthy kidneys. Using the single cell co-expression (SCoexp) module of CellTrek(*28*) we identified two different iPT modules, corresponding to two iPT subtypes in

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

diseased samples (**Fig. 3D**) and one iPT type in healthy samples (**Supplementary Fig. 14**). Moving back to the rich snRNA-seq data, we found that one iPT cluster was enriched for the expression of *VCAM1*, *ACSL1*, *ASS1*, and *ASPA*, genes playing roles in cellular metabolism. We called them iPT_VCAM1+. This cluster was more frequent in healthy samples. The second iPT cluster expressed *HAVCR1 (or KIM1)*, *NFKBIZ*, *IL18*, *SPP1*, *ITGA3*, and *ITGB1* and was enriched for genes associated with cell adhesion and matrix (iPT-HAVCR1+) (**Fig. 3E, Supplementary Fig. 15**). Most iPT-HAVCR1+ cells were in the fibrotic samples. Trajectory analysis indicated that iPT_HAVCR1+ were located at the end of pseudotime, suggesting that they have accrued greater damage (**Fig. 3F)**. Gene expression changes along the trajectory are listed in **Supplementary Table 8**. Our snATAC-seq recapitulated our results (**Supplementary Fig. 16**). We identified *TFEC* and *BACH2* as specific TFs for iPT_VCAM1+ and iPT_HAVCR1+, respectively (**Fig. 3G**). Our results are consistent with prior snRNAseq results identifying VCAM1+ cells and prior mechanistic studies recognizing HAVCR1+ as an injured PT marker (*10*).

iPTs were often captured together with stromal cells and were the closest to stromal fibroblasts (**Supplementary Fig. 17**). Our trajectory analysis indicated a continuous transition between iPT and fibroblasts similar to the previously described epithelial-mesenchymal transition (EMT)(*30, 31*) including the expression of *ZEB1*, *ZEB2*, *SNAI2*, and *ACTA2* (**Supplementary Fig. 17, 18)**. Module analysis of the spRNA-seq dataset highlighted fibroblast_1 and fibroblast_2 subtypes with different characteristics; fibroblast_1 was enriched for matrix protein expression and fibroblast_2 for inflammatory genes (**Supplementary Fig. 17, 19**).

In summary, differential expression analysis indicated highly plastic PT cells and the close proximity of injured PT cells to the fibrotic stroma. Using spatial profiling, we could identify two

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

types of injured PT cells (VCAM1+ and HAVCR1+) in healthy and diseased samples and show their close proximity to fibroblasts.

Fibroblast heterogeneity in human kidney disease

To further examine fibroblast heterogeneity and its relationship to the development of fibrosis, we created an extracellular matrix (ECM) score by calculating the expression of collagen, glycoprotein, and proteoglycan specific genes in different cell types(*32, 33*). **Fig. 4A** shows that fibroblast_1, 2, MyoFib/VSMC, and mesangial cells had the highest ECM score. Consistently, fibroblast_1, and VSMC/myofibroblast fractions were higher in diseased samples (**Fig. 4B**). The ECM score was consistent with the presence of fibroblasts in the spRNA-seq data, which was compatible with the presence of these cells (**Fig. 4C**).

Sub-clustering analysis of stromal cells identified 10 different cell types, including 6 different fibroblasts; SERPINE1+, FAP+, COL1A1+, CR2+, B2M+, and CXCL14+ fibroblasts. The sub-clustering also indicated *REN*-expressing juxtaglomerular cells and *ITGA8* and *POSTN*-expressing mesangial cells. We could discriminate VSMC expressing *MYH11*, *RSG6*, and myofibroblast expressing *ACTA2* and *SYNPO2* (**Fig 4. D**). While several snRNA-seq studies proposed stromal cell subtypes, our spRNA-seq dataset provides an unbiased verification and spatial localization for these cells (**Fig 4. E**). Our spRNA-seq data was consistent with protein expression in the Human Protein Atlas (**Supplementary Fig. 20**) and by snATAC-seq analysis (**Supplementary Fig. 21**).

Within the stromal cells, SEPRINE1+, COL1A1+, FAP+ cells, and myofibroblast had the highest ECM score. Consistently, this cell type was enriched in diseased kidneys compared to controls (**Fig. 4F**). Cell trajectory analysis indicated that myofibroblasts are located at the end of pseudo time originating from pericytes, as previously shown(*32*) (**Supplementary Fig. 22,**

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Table 9). Using the snATAC-seq data, we could identify *TCF12* for SERPINE1+ and *E2F1* transcription factor motifs in myofibroblast (**Fig. 4G, Supplementary Fig. 21**).

The interaction of stromal, immune, endothelial and injured epithelial establishes the kidney fibrotic microenvironment

Our newly generated spRNA-seq dataset is uniquely suited to defining microenvironments (ME) in the human kidney. We ran nonnegative matrix factorization (NMF) on the spRNA-seq datasets. We found four major MEs in the human kidney, including glomerular/vascular MEs, tubule MEs, fibrotic MEs (FMEs), and immune MEs. The gene ontology enrichment analysis of genes detected in each microenvironment was consistent with their anatomical annotation (**Supplementary Fig. 23).** It is important to note that the method identified patchy areas in the kidneys that were labelled as fibrotic microenvironments. The computationally defined FME strongly correlated with kidney ECM scores (**Fig. 5A, Supplementary Fig. 24**) and our pathologist's assessment of fibrosis. Cell type enrichment analysis indicated iPT, fibroblast_1, fibroblast_2, and different immune cell types around the endothelial cells in FMEs (**Fig. 5B, Supplementary Fig. 24, 25**).

We also identified a specific immune ME. These immune MEs were located within the FME, but again with patchy distribution. The immune ME consisted of follicular dendritic cells, plasma cells, B-cell and T lymphocytes (**Supplementary Fig. 26**). The immune ME organizations resembled early tertiary lymphoid structures(*34*). Immunostaining studies with cell type specific antibodies validated the presence of these specific immune cells and immune cell aggregates (**Supplementary Fig. 27**).

To further understand cell interactions in FMEs, we implemented CellChat(*35*) on sn/scRNA-seq and spRNA-seq datasets. We found enrichment for *C3*, *IL7*, *SPP1*, *IL17A*, *CXCL12*, *CXCL13*,

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CCL19, *CCL21, PDGFB, TGFB1* and their receptors in FME regions (**Fig. 5C, D, Supplementary Fig. 28**). We observed that iPT_HAVCR1+ expressed *IL7*, *C3*, and *SPP1* while their receptors were present on CD4T, CD8T, macrophages, and stromal cells, respectively, indicating that iPT cells might be responsible for the influx of these cells (**Supplementary Fig. 28, 29)**. The stromal cells in FME were enriched for chemotactic factors including *CXCL12*, *CXCL13*, *CCL19*, *CCL21* and while their receptors where expressed in different immune cell, suggesting that stromal cells might signal to immune cell. We observed expression of *PDGFB* and *TGFB1*, known mediators of fibrosis, in FME associated immune aggregates (**Fig. 5 C, D**). CellChat analysis of sn/scRNA-seq and spRNA-seq indicated FME stromal cells with the highest secretory score (**Supplementary Fig. 28, 29).**

Overall, using unbiased NMF we identified spatial kidney regions, including well established glomerular and tubular regions, but also fibrotic and immune regions. Most importantly, FMEs were not only characterized by matrix-producing fibroblasts but we identified an intricate cell-cell interaction, indicating a complex cellular architecture (**Fig. 5E**).

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

Next, to understand whether our spatially resolved human kidney atlas information can be used for disease classification and prognosis evaluation, we analyzed a large cohort of human kidney samples. We first generated an FME gene signature (FME-GS) (**Supplementary Table 10**) and analyzed our large external kidney cohorts' gene expression data from 298 human kidney samples (**Fig. 6A**), including healthy samples and samples with varying severity of DKD and HKD.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our FME-GS was able to successfully cluster 298 human kidney samples into 3 separate groups (**Fig. 6B**). These 3 groups corresponded to samples with varying degrees of disease severity as indicated by differences in clinical parameters such as eGFR and fibrosis (**Fig. 6B**) (despite the fact that these parameters were not included in the clustering algorithm).

Next, we wanted to know whether FME-GS could be used as a disease prognostic marker. Here we used a different set of large external human kidney gene expression datasets (N = 218), with a mean follow-up time of 2.49 (SD: 1.96) years. Our FME-GS successfully clustered samples based on disease severity (**Fig. 6C**). The top FME genes showing the greatest difference between clusters were mostly stromal and immune cell specific genes, including *PDGFB*, *MYH9*, *NFKB1*, and *STAT3* (**Fig. 6D**). Next, we analyzed the relationship between cell types and kidney disease progression. We found that genes correlated with eGFR slope were enriched in PT, stromal and immune cells (**Fig. 6E**). Finally, we performed a Kaplan-Meier analysis to predict the probability of reaching to end stage kidney disease (eGFR < 15 ml/min/1.73 m^2) or 40% eGFR decline/year. These are hard outcomes identified by the FDA for drug effectiveness(*36*). Our data indicated that cluster 1, with the highest FME-GS score, had the highest hazard ratio to reach the end-point (**HR = 3.61, 95%CI: 1.25 – 10.4**). We found that FME-GS has the strongest predictive value when compared to other microenvironments (**Supplementary Fig. 30).**

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

Discussion

Here we present the spatial molecular principles of kidney health and disease via generating a comprehensive and spatially resolved human kidney atlas by combining single cell omics data and

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a large number of human kidney tissue samples with varying degrees of disease severity. Our work fills a critical knowledge gap by characterizing cell types previously only defined by their spatial location, showing the anatomical location of cells only observed in dissociated single cell expression data and defining cell-type specific gene expression changes in diseased areas. We define the cellular complexity of the fibrotic microenvironment as the intricate interaction of a large number of cell types. We demonstrate the clinical prognostic value of spatial transcriptomics.

Previous single cell analyses, focusing on dissociated human and mouse kidney datasets, have generated gene expression and regulatory matrices for a variety of kidney cell types(*8-12, 37*). As kidney cell types have been functionally well characterized, most identified cell types have been matched back to a more than half-century old functional cell type definition(*6*). A key limitation of these analyses has been the identification and molecular characterization of anatomically defined cell types, such as mesangial cells, PEC cells, and fibroblasts. Here we demonstrate that a joint approach that includes large single cells, single nuclear expression, open chromatin, and spRNA-seq combined with large and diverse samples and large cell numbers is critical to achieve this goal. The orthogonal analytical tools provide unique opportunities for validation, as each method suffers from specific technological biases. Here, we have not only been able to resolve and validate previously anatomically-known cell types but also identify novel cell types such as specific stromal cells for glomerulosclerosis (expressing *CDH13)(38)*.

Fibrotic diseases are responsible for close to 40% of all deaths(*39*). Kidney fibrosis is the final common pathway to end stage kidney failure(*40*). Fibrosis, however, is an anatomically defined lesion and most emphasis has been placed on matrix accumulation and characterization of matrix

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

producing cells. Here, we demonstrate the cellular and architectural complexity of kidney fibrosis. We propose the use of the fibrotic microenvironment to characterize these lesions, to not only focus on matrix accumulation but on the elaborate cellular complexity of these lesions. We show that they are anatomically localized close to injured PT, indicating that iPT is likely to be an important nidus of fibrosis. We identify spatially defined iPT subtypes. These iPT subtypes are consistent with previous mechanistic studies and animal model single cell data (*10*). Furthermore, our data suggest that some iPT cells can directly convert into fibroblasts, consistent with the previously proposed EMT hypothesis(*30, 31*).

Combining snRNA and spatial information, we not only define the stromal cell subtypes but also the cellular and architectural heterogeneity of fibrosis. We could conclusively discriminate VSMC and mesangial cells from myofibroblasts that are anatomically distinct but share gene expression signatures in sc/snRNAseq data(*41-43*). We identify two key fibroblast modules; matrix secreting and immune fibroblast and show 10 different stromal cell types. We identify the key cell types that contribute to ECM production. Our data indicates that fibroblasts are the precursors of myofibroblasts in the kidney, but tubule cells could also become fibroblasts(*32*). We could identify novel markers and, ultimately, new fibroblast types and determine their spatial location. This information could be important in the field of finding therapeutic candidates for renal fibrosis. We noted a large cluster of FAP-positive fibroblasts in diseased human kidneys(*44-46*). FAP targeted cellular and RNA therapies have been developed and shown to have efficacy animal models of cardiac fibrosis(*44-46*). Our data suggests that these therapeutics may be helpful for treating kidney fibrosis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Most importantly, we demonstrated that human kidney fibrosis is an established microenvironment, not just a simple collagen accumulation problem. The interaction of a large number of cell types, including iPT, immune, stromal, and endothelial cells, establishes the FME. While we did not perform side-by-side comparison, the cell heterogeneity and cell interaction network of human kidney fibrosis appear far more complex than what has been published for mouse models(*47, 48*). For example, in mice, we identified a large number of secreted cytokines from iPT cells responsible for the influx of immune and stromal cells(*48*). In patient samples, there is a strong interaction between stromal and immune cells and also signaling by immune and stromal cells to iPT, which might play a role maintaining their injured PT state.

Immune cell clusters have long been observed in fibrotic kidney samples, even in patients with non-immune-mediated kidney disease, such as diabetes and hypertension(*40, 49*). Here we resolve these regions both spatially and at a cell type level. Our kidney scRNA-seq data was enriched for immune cells and enabled us to spatially resolve immune cell types and determine the distributions of immune cells in the kidney. We show that immune cell clusters (the immune microenvironment) are localized mostly within some FMEs. While we did not perform a systematic comparison of human and mouse kidney fibrosis, our data indicate lymphocyte prominence compared to myeloid cells in human fibrosis, while mouse fibrosis models are strongly enriched for macrophages(*48*). The fibrosis-associated immune aggregates show a resemblance to the tertiary lymphoid structures (TLS). TLS are organized aggregates of immune cells that form postnatally in nonlymphoid tissues, usually as a persistent antigen production(*50*) and generate autoreactive effector cells. TLS have been earlier described in mouse kidney tissue samples(*51-54*). Future studies will be needed

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to define TLSs in CKD and kidney fibrosis; however, they could have tremendous therapeutic potential.

One of the most devastating complications of CKD is its progression to ESRD, which requires life-sustaining dialysis or transplantation (*55*). At present, we cannot predict which patients will progress to ESRD, representing an important clinical problem. Our data indicate that FME-GS can identify subjects at risk of ESRD in a large external dataset of human kidney tissue samples. These results establish FME-GS as a key biomarker and potentially as a causal mechanism of progression.

In summary, we develop a spatially defined molecular human kidney cellular atlas, characterize the fibrotic microenvironment, and indicate their role as a clinically meaningful prognostic disease biomarker, demonstrating the utility of spRNA-seq for the investigation complex diseases.

.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods:

Single nuclei RNA sequencing

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Single nuclei ATAC sequencing

The procedure described above was used to isolate the nuclei. The resuspension was performed in diluted Nuclei Buffer (10X GEM). Nuclei quality and concentration were measured with Countess AutoCounter (Invitrogen, C10227). The diluted nuclei were loaded and incubated in chromium single cell ATAC library & gel bead kit's transposition mix (10X Genomics, PN-1000110). Chromium Chip E (10X Genomics, PN-1000082) in the Chromium Controller was utilized to capture the GEMs. The Chromium Single Cell ATAC Library & Gel Bead Kit and Chromium i7 Multiplex Kit N Set A (10X Genomics, PN-1000084) were then used to create snATAC libraries in accordance with the manufacturer's instructions. Library quality was examined using an Agilent Bioanalyzer High Sensitivity DNA kit. Libraries were demultiplexed, as follows, after sequencing on an Illumina Novaseq system using two 50-paired-end kits: 50 bp Read1 for DNA fragments, 8 bp i7 index for sample index, 16 bp i5 index for cell barcodes, and 50 bp Read2 for DNA fragments.

Single Cell RNA-seq

Fresh human Kidneys (up to 0.5 gr) collected in RPMI were minced into approximately 2-4 mm cubes using a razor blade and then transferred to a gentlMACS C tube contains Multi Tissue dissociation kit 1 (Miltenyi, #130-110-201). The tissue was homogenized using Multi-B program of gentleMACS dissociator with Multi_B program in the tube contains 100ul of Enzyme D, 50ul of Enzyme R and 12.5ul of Enzyme A in 2.35 ml of RPMI and incubated for 30mins at 37 degrees. Second homogenization were performed using Multi_B program on gentleMACS dissociator. The solution was then passed through a 70um cell strainer. After centrifugation at 1,200 RPM for 7mins, cell pellet was incubated with 1ml of RBC lysis buffer on ice for 3mins. The reaction was

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stopped by adding 10 ml PBS. Next the solution centrifuged at 1,000 RPM for 5 minutes. Finally, after removing the supernatant, the pellet was resuspended in PBS. Cell number and viability were analyzed using Countess AutoCounter (Invitrogen, C10227). This method generated single cell suspension with greater than 80% viability. Next, 30,000 cells were loaded into the Chromium Controller (10X Genomics, PN-120223) on a Chromium Next GEM chip G Single Cell Kit (10X Genomics, PN-1000120) to generate single cell gel beads in the emulsion (GEM) according to the manufacturer's protocol (10X Genomics, PN-1000121). The cDNA and library were made using the Chromium Next GEM Single Cell 3′ GEM Kit v3.1 (10X Genomics, PN-1000121) and Single Index Kit T Set A (10X Genomics, PN-120262) according to the manufacturer's protocol. Quality control for the libraries were performed using Agilent Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626). Libraries were sequenced on Illumina Novaseq 6000 system with 2 × 150 paired-end kits using the following demultiplexing: 28 bp Read1 for cell barcode and UMI, 8 bp I7 index for sample index and 91 bp Read2 for transcript.

Visium FFPE for SpRNA-seq

RNA quality of human kidney FFPE sample was checked by extracting RNA using RNeasy FFPE kit (Qiagen-Cat #73504) according to the manufacturer's protocol. RNA quality was examined using Agilent bioanalyzer and samples with DV200>50% were selected. Then a 5 μm tissue samples was cut onto the Visium Spatial gene Expression Slide. After deparaffinization, H & E staining was performed. We used Keyence 1266 BZ-X810 microscope for whole slide imaging. After scanning, de-crosslinking, probe hybridization, probe release and extension, library preparation was performed by single Index Kit TS Set A (10X Genomics, PN-3000511) according to manufacturer's protocol. Quality control for the libraries were performed using Agilent

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626). Libraries were sequenced on Illumina Novaseq 6000 system with 2 × 150 paired-end kits using the following demultiplexing: 28 bp Read1 for cell barcode and UMI, 10 bp I7 index, 10bp i5 index and 50 bp Read2 for transcript.

Microdissection and Bulk RNA sequencing

Under a dissecting microscope, human kidney tissues were microdissected in RNA-later solution using a microdissection forceps. After removing glomeruli, the remaining tissue was treated as a tubule. Total RNA was extracted using the Qiagen RNeasy kit (catalog #74106). Agilent Bioanalyzer RNA 6000 Pico kit (Agilent Technologies, 5067-1513) was used to assess the quality of the RNA. All samples with an RNA integrity number (RIN) of at least 6 were utilized. Following the manufacturer's instructions, strand-specific RNA-seq libraries were created using the NEBNext® UltraTM RNA Library Prep Kit for Illumina (catalog #E7530L). RNA-seq libraries were then sequenced to a depth of 20 million 2x150 pair end reads.

Human Sample Acquisition

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A portion of the tissue were formalin-fixed, paraffin-embedded, and stained with periodic acid-Schiff. A local renal pathologist performed objective pathological scoring of the abnormal parameters.

Immunostaining

Paraffin blocks were sectioned. After deparaffinization, 1% bovine serum albumin was used for blocking. Diluted primary antibodies on slides were incubated overnight (CD4 CST (Catalogue #48374), IGKC: Biolegend (Catalogue #392702), and CD79A Abcam (Catalogue #ab79414). After washing the sections with PBS, three times, secondary antibodies were used for 1h at room temperature. The stains were imaged with OLYMPUS BX43 Microscope. Positive cells in ten randomly selected fields were counted on each slide.

Bioinformatic analysis

Primary single nuclei and cell RNA-seq data processing

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

Data Processing and Computational Analyses

After ambient RNA correction using "SoupX"(*56*) and doublet removal by "DoubletFinder"(*57*) using default parameters, Seurat objects from the aligned outputs (from multiple samples) were created where genes expressed in more than 3 cells and cells with at least 300 genes were retained. Further, a merged Seurat object was obtained using "merge" function of Seurat v (4.0.3)(*58*). The following QC filtering were used: (a) cells having n_feature counts of more than 3000 and less

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

than 200 as well as (b) more than 15% mitochondrial counts (for snRNA-seq data) and more than 50% mitochondrial counts (for scRNA-seq data) were filtered.

Data Normalization and Cell Population Identification

First, highly variable genes were identified using the method "vst". The data was natural log transformed and scaled. The scaled values were then subjected to principle component analysis (PCA) for linear dimension reduction. We used the "harmony"(*59*) package by "RunHarmony" function for batch effect correction. A shared nearest neighbor network was created based on Euclidean distances between cells in a multidimensional PC space (the first 50 and 30 PCs were used for snRNA-seq and scRNA-seq, respectively) and a fixed number of neighbors per cell, which was used to generate a 2-dimensional Uniform Manifold Approximation and Projection (UMAP) for visualization.

In order to identify cell-type markers, we used Seurat's "FindAllMarkers" function of "Seurat". This method calculates log fold changes, percentages of expression within and outside a group, and p-values of Wilcoxon-Rank Sum test comparing a group to all cells outside that specific group including adjustment for multiple testing. A log-fold-change threshold of 0.25 and FDR<0.05 was considered significant. These steps were performed on the snRNA-seq and scRNA-seq datasets, separately. Clusters expressing multiple cell types specific marker genes were excluded as potential doublets.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DEGs between diseased and healthy groups

To identify DGEs between experimental groups, we utilized the "FindMarkers" tool for each cell type and condition, a log-fold-change threshold of 0.25, and an FDR 0.05.

Single nuclei RNA-seq trajectory analysis

PT, Injured PT cells and different types of fibroblasts were subclustered for the trajectory analysis. The trajectory analysis was done in two steps. Different sub-types of iPT and stromal cells with equal numbers were randomly subsampled and cell dataset object (CDS) was generated using Monocle3(*60, 61*). After preprocessing, batch effects correction, the dataset was embedded for dimension reduction and pseudotemporal ordering. We used the “order_cell” function and indicated the PT as start point for “pseudotime” analysis. The "track genes" algorithm was used to identify the DGEs along the trajectories, and genes with q values of 0.05 or higher were considered significant.

Ligand–receptor interactions

CellChat(*35*) repository was used to assess cellular interactions between different cell types and to infer cell–cell communication networks from snRNA-seq data. Package CellChat v1.4.0 was used to predict cell type-specific ligand–receptor interactions (1939 interactions). Only receptors and ligands expressed in more than 10 cells in each cluster were considered. Probability and *P* values were measured for each interaction.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Single nuclei ATAC-seq analysis

Raw FASTQ files were aligned to the GRCh38 reference genome and quantified using Cell Ranger ATAC (v. 1.1.0). The cell ranger outputs of four snATAC-seq datasets were embedded using Signac (v.1.3.0)(*62*) to generate Signac object. Low-quality cells were removed from each snATAC object using the following criteria: peak_region_fragments < 3000 & peak_region_fragments > 20000 & pct_reads_in_peaks < 15 & nucleosome_signal > 4 & TSS.enrichment < 2). The filtered cells in twenty objects were merged together using "merge" function in Seurat. Dimensional reduction was done by singular value decomposition (SVD) of the TFIDF matrix and UMAP. Batch effect was corrected using Harmony(*59*) via the "RunHarmony" function in Seurat. A KNN graph was made to cluster cells using the Louvain algorithm.

Cluster annotation of snATAC-seq

With the Signac "FindMarkers" function, peaks observed in at least 20% of cells were evaluated for differentially accessible chromatin regions (DARs) between different cell types using a likelihood ratio test, a log-fold-change threshold of 0.25, and an FDR of 0.05.

To annotate the genomic regions harboring snATAC-seq peaks, ChIPSeeker (v1.24.0)(*63*) was used.

Motif Enrichment Analysis and Motif Activities

The "AddMotifs" function of Signac was used to run a motif enrichment analysis after creating a matrix of positional weights for motif candidates from JASPAR2020. The related function of "RunChromVAR" and chromVAR (v.1.6.0)(*64*) were used to determine transcription factor activity. The "FindMarkers" program was used to calculate the differences in motif activity

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

between clusters, and an FDR of 0.05 was deemed significant. The "FindMotif" tool was used to carry out motif enrichment analysis on the differentially accessible regions.

DARs between groups

We used the “FindMarkers” function after selecting “DefaultAssay” as “peaks” to identify DARs in each cell type and diseased and healthy conditions, with a log-fold-change threshold of 0.25 and FDR<0.05. Peaks translated to related genes using ChIPSeeker (v1.21.1)(*63*).

Annotation based on snRNA-seq and Integration snATAC-seq and snRNA-seq

The "GeneActivity" tool in Signac was used to create a gene activity matrix following clustering of the twenty integrated snATAC-seq datasets. Using protein-coding genes annotated in the Ensembl database, this technique counts the ATAC peaks inside the gene body and 2 kb upstream of the transcriptional start point. Next, log normalization was applied to the gene activity matrix. The snRNA-seq dataset was utilized as a reference, and the "FindTransferAnchors" function was used to discover matching genes between the snRNA-seq and snATAC-seq datasets by using shared correlation patterns in the gene activity matrix and snRNA-seq dataset. Next, the predicted labels within two datasets were identified using the "TransferData" method.

Integration of snRNA-seq, scRNA-seq and snATAC-seq datasets

In order to create a single snRNA-seq, scRNA-seq, and snATAC-seq dataset we used a step-by-step integration method. First, we used our snRNA-seq dataset as a reference and the snATAC-seq data (which gene activity was already calculated) to project to the snRNA-seq dataset using

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

"FindTransferAnchors", and "TransferData" functions. Then the imputed snATAC-seq dataset was merged with snRNA-seq dataset and after scaling, the data dimensions were reduced using PCA and UMAP. After creating a single data matrix of snRNA-seq and snATAC-seq, the scRNA-seq was projected to this dataset by finding the shared anchors. Then the imputed scRNA-seq dataset was merged with integrated snRNA-seq, snATAC-seq datasets and after scaling, the data dimensions were reduced using PCA and UMAP.

SpRNA-seq data analysis

The data was aligned using Space Ranger (v1.0.0) with reference genome GRCh38 and human probe dataset (Visium_Human_Transcriptome_Probe_Set_v1.0_GRCh38). The data then was loaded to make the Seurat object and normalized using SCT. This step was done for all seven samples. The samples were merged together, using "merge" function of Seurat. Next, the data was subjected to principle component analysis (PCA) for linear dimension reduction and Harmony was used to integrate the datasets. A shared nearest neighbor network was created based on Euclidean distances between cells in a multidimensional PC space (30 PCs were used) and a fixed number of neighbors per cell, which was used to generate a 2-dimensional Uniform Manifold Approximation and Projection (UMAP) for visualization.

In order to identify spot specific markers, Seurat's "FindAllMarkers" function was used. In this method log fold changes, percentages of expression within and outside a group, and p-values of Wilcoxon-Rank Sum test comparing a group to all cells outside that specific group including

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

adjustment for multiple testing was calculated. A log-fold-change threshold of 0.25 and FDR<0.05 was considered as significant. Basic functions of Seurat were used for visualization.

Deconvolution of SpRNA-seq Dataset

Two different methods were used to deconvolute the spRNAseq data; the RCTD(*65*) method using the default parameters and the CCA(*66*) method using Seurat. The "FindAnchors" function in Seurat, the shared genes between two datasets was determined and cell type prediction was performed using "TransferData" function and the prediction score of each cell type in each spot was considered as the frequency of each cell type in the spot. The distribution score was calculated as the number of spots with more than 10% probability of one cell type.

In order to determine the colocalization of the identified cells in each spot, Pearson correlation test was performed which indicate the probability of co-existing of different cell types.

Mapping sn/scRNA-seq to Spatial Location

In order to map back the cell types identified in the dissociated data (sn/scRNA-seq datasets), Celltrek(*28*) package was used. Firstly, the sn and scRNA-seq data were down sampled to 20,000 cells. Then, by using "traint" function, sn/scRNA-seq datasets were co-embedded with spRNAseq datasets. Next, using the random forest model, single cells were mapped to their spatial locations. This analysis was performed by merging snRNA-seq and immune cell types to enrich the dataset for immune cells. Regarding colocalization, the "sColoc" function of the CellTrek was used.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to find the different cell type modules in the spRNAseq, spatial-weighted gene co-expression analysis was performed.

Finding microenvironments in spRNA-seq

In order to identify microenvironments on the merged dataset the NMF reduction was performed then, the clustering by default parameters using NMF reduction was done. In order to identify MEs specific markers, Seurat's "FindAllMarkers" function was used.

ECM production score

In order to calculate the extracellular matrix production (ECM), the proportion of the expressions of the collagen, proteoglycan and glycoprotein(*33*) genes in each cells were calculated.

Bulk RNA-seq Analysis

FASTQC was used to check the QC of the sequencing results. Next, the adapters and low-quality bases were trimmed using TrimGalore (v0.4.5). The trimmed FASTQ files were aligned to the to the human genome (hg19/GRCh37) using STAR (v2.7.3a)**(*67, 68*)** based on GENCODE v19 annotations**(*67, 68*)**. The expression of different genes was measured using RSEM by calculating uniquely mapped reads as transcripts per million (TPM).

Hierarchical clustering analysis

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Hierarchical clustering was performed on the scaled TPM matrix of microdissected human tubules datasets based on the FME-GS list. Ward's method with Euclidean distances was used to cluster the datasets. The optimal number of clusters was determined by average silhouette method. After clustering, the data was presented as a cluster dendrogram.

Statistics

The data were expressed as means ± SEM. Independent sample t test was used to compare the continuous variable in two groups and One-way ANOVA was used to compare the continuous parameters between more than two groups followed by Bonferroni post hoc test for subgroup comparisons. $P < 0.05$ was considered as a significance.

Data Availability

Raw data, processed data, and metadata from the snRNA-seq, scRNA-seq, snATAC-seq, and spRNA-seq have been deposited in Gene Expression Omnibus (GEO) and the accession number will be provided when it will be available. The human bulk kidney RNA-seq data are available under following accession numbers: GSE115098 and GSE173343**.** The single cell and nuclear expression and open chromatin and spatial data is also available at www.susztaklab.com.

Code Availability

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

Acknowledgement

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Work in the Susztak lab is supported by the NIH DK076077, DK087635 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 an employee of GSK. SP, CMB, and PG are employees of Boehringer Ingelheim. AK is the employee of Novo Nordisk.

Author Contributions

AA, ZM, JF, RS, PD, GP, and TB performed experiments. AA, MSB, HL, SV, MSB, HY, and KC performed computational analysis. KD, LM, EH, LSB, CAH, AK, PG, CMB, GP 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

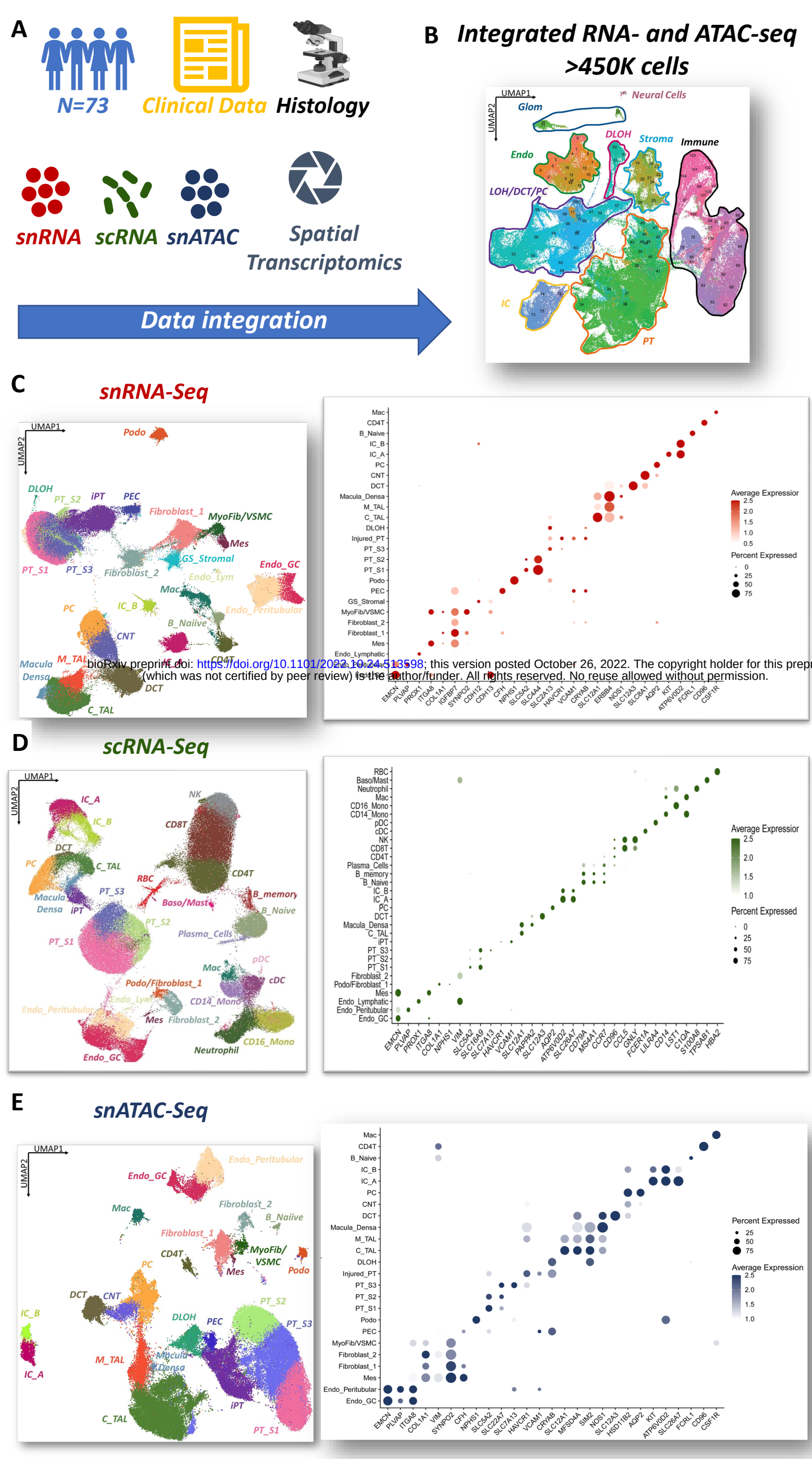
1. V. Gueutin, G. Deray, C. Isnard-Bagnis, Renal physiology. *Bull. Cancer* **99**, 237-249 (2012).
2. H. W. Smith, The fate of sodium and water in the renal tubules. *Bull. N. Y. Acad. Med.* **35**, 293 (1959).
3. P. Vart *et al.*, National trends in the prevalence of chronic kidney disease among racial/ethnic and socioeconomic status groups, 1988-2016. *JAMA network open* **3**, e207932-e207932 (2020).
4. A. Yinusa, M. Faezipour, M. Faezipour, in *Healthcare*. (MDPI, 2022), vol. 10, pp. 1628.
5. Y. Kakitapalli, J. Ampolu, S. D. Madasu, M. S. Kumar, Detailed review of chronic kidney disease. *Kidney Diseases* **6**, 85-91 (2020).
6. M. S. Balzer, T. Rohacs, K. Susztak, How Many Cell Types Are in the Kidney and What Do They Do? *Annu. Rev. Physiol.* **84**, 507 (2022).
7. F. Schreibing, R. Kramann, Mapping the human kidney using single-cell genomics. *Nature Reviews Nephrology* **18**, 347-360 (2022).
8. M. S. Balzer *et al.*, Single-cell analysis highlights differences in druggable pathways underlying adaptive or fibrotic kidney regeneration. *Nature Communications* **13**, 1-18 (2022).
9. Z. Miao *et al.*, Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nature communications* **12**, 1-17 (2021).
10. Y. Muto *et al.*, Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature communications* **12**, 1-17 (2021).
11. P. C. Wilson *et al.*, The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proceedings of the National Academy of Sciences* **116**, 19619-19625 (2019).
12. M. D. Young *et al.*, Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594-599 (2018).
13. E. E. Dixon, H. Wu, Y. Muto, P. C. Wilson, B. D. Humphreys, Spatially resolved transcriptomic analysis of acute kidney injury in a female murine model. *J. Am. Soc. Nephrol.* **33**, 279-289 (2022).
14. E. E. Dixon, H. Wu, E. Sulvarán-Guel, J. Guo, B. D. Humphreys, Spatially resolved transcriptomics and the kidney: many opportunities. *Kidney Int.*, (2022).
15. R. M. Ferreira *et al.*, Integration of spatial and single-cell transcriptomics localizes epithelial cell–immune cross-talk in kidney injury. *JCI insight* **6**, (2021).
16. J. L. Marshall *et al.*, High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *Iscience* **25**, 104097 (2022).
17. T. K. Chen, D. H. Knicely, M. E. Grams, Chronic kidney disease diagnosis and management: a review. *JAMA* **322**, 1294-1304 (2019).
18. M. B. Palmer *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).
19. F. Armutcu, Organ crosstalk: the potent roles of inflammation and fibrotic changes in the course of organ interactions. *Inflamm. Res.* **68**, 825-839 (2019).
20. J. Majo, B. M. Klinkhammer, P. Boor, D. Tiniakos, Pathology and natural history of organ fibrosis. *Curr. Opin. Pharmacol.* **49**, 82-89 (2019).
21. Z. Deng *et al.*, The extracellular matrix and mechanotransduction in pulmonary fibrosis. *The International Journal of Biochemistry & Cell Biology* **126**, 105802 (2020).
22. B. D. Humphreys, Mechanisms of renal fibrosis. *Annu. Rev. Physiol.* **80**, 309-326 (2018).
23. T. Parimon, M. S. Hohmann, C. Yao, Cellular senescence: pathogenic mechanisms in lung fibrosis. *Int. J. Mol. Sci.* **22**, 6214 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

24. A. Makita *et al.*, The association between renal elasticity evaluated by real-time tissue elastography and renal fibrosis. *Clin. Exp. Nephrol.* **25**, 981-987 (2021).
25. F. Lausecker, R. Lennon, M. J. Randles, The kidney matrisome in health, aging and disease. *Kidney Int.*, (2022).
26. D. Zhou, Y. Liu, Understanding the mechanisms of kidney fibrosis. *Nature Reviews Nephrology* **12**, 68-70 (2016).
27. G. Z. Quinn *et al.*, Renal histologic analysis provides complementary information to kidney function measurement for patients with early diabetic or hypertensive disease. *J. Am. Soc. Nephrol.* **32**, 2863-2876 (2021).
28. R. Wei *et al.*, Spatial charting of single-cell transcriptomes in tissues. *Nat. Biotechnol.*, 1-10 (2022).
29. H. Wu, Y. Kirita, E. L. Donnelly, B. D. Humphreys, Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23-32 (2019).
30. M. T. Grande *et al.*, Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat. Med.* **21**, 989-997 (2015).
31. J. Li *et al.*, Renal protective effects of empagliflozin via inhibition of EMT and aberrant glycolysis in proximal tubules. *JCI insight* **5**, (2020).
32. C. Kuppe *et al.*, Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281-286 (2021).
33. A. Naba *et al.*, The extracellular matrix: Tools and insights for the "omics" era. *Matrix Biol.* **49**, 10-24 (2016).
34. Y. Sato, M. Tamura, M. Yanagita, Tertiary lymphoid tissues: A regional hub for kidney inflammation. *Nephrology Dialysis Transplantation*, (2021).
35. S. Jin *et al.*, Inference and analysis of cell-cell communication using CellChat. *Nature communications* **12**, 1-20 (2021).
36. A. Y. Kluger *et al.*, Class effects of SGLT2 inhibitors on cardiorenal outcomes. *Cardiovasc. Diabetol.* **18**, 1-13 (2019).
37. J. Park *et al.*, Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758-763 (2018).
38. L. Zhang *et al.*, Genome-wide analysis of histone H3 lysine 4 trimethylation in peripheral blood mononuclear cells of minimal change nephrotic syndrome patients. *Am. J. Nephrol.* **30**, 505-513 (2009).
39. T. A. Wynn, T. R. Ramalingam, Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat. Med.* **18**, 1028-1040 (2012).
40. S. Djudjaj, P. Boor, Cellular and molecular mechanisms of kidney fibrosis. *Mol. Aspects Med.* **65**, 16-36 (2019).
41. M. Liu, D. Gomez, Smooth muscle cell phenotypic diversity: At the crossroads of lineage tracing and single-cell transcriptomics. *Arterioscler. Thromb. Vasc. Biol.* **39**, 1715-1723 (2019).
42. A. J. Pedroza *et al.*, Single-cell transcriptomic profiling of vascular smooth muscle cell phenotype modulation in Marfan syndrome aortic aneurysm. *Arterioscler. Thromb. Vasc. Biol.* **40**, 2195-2211 (2020).
43. F. Zhang, X. Guo, Y. Xia, L. Mao, An update on the phenotypic switching of vascular smooth muscle cells in the pathogenesis of atherosclerosis. *Cell. Mol. Life Sci.* **79**, 1-19 (2022).
44. H. Aghajanian *et al.*, Targeting cardiac fibrosis with engineered T cells. *Nature* **573**, 430-433 (2019).
45. H. Aghajanian, J. G. Rurik, J. A. Epstein, CAR-based therapies: opportunities for immuno-medicine beyond cancer. *Nature Metabolism* **4**, 163-169 (2022).
46. J. G. Rurik *et al.*, CAR T cells produced in vivo to treat cardiac injury. *Science* **375**, 91-96 (2022).

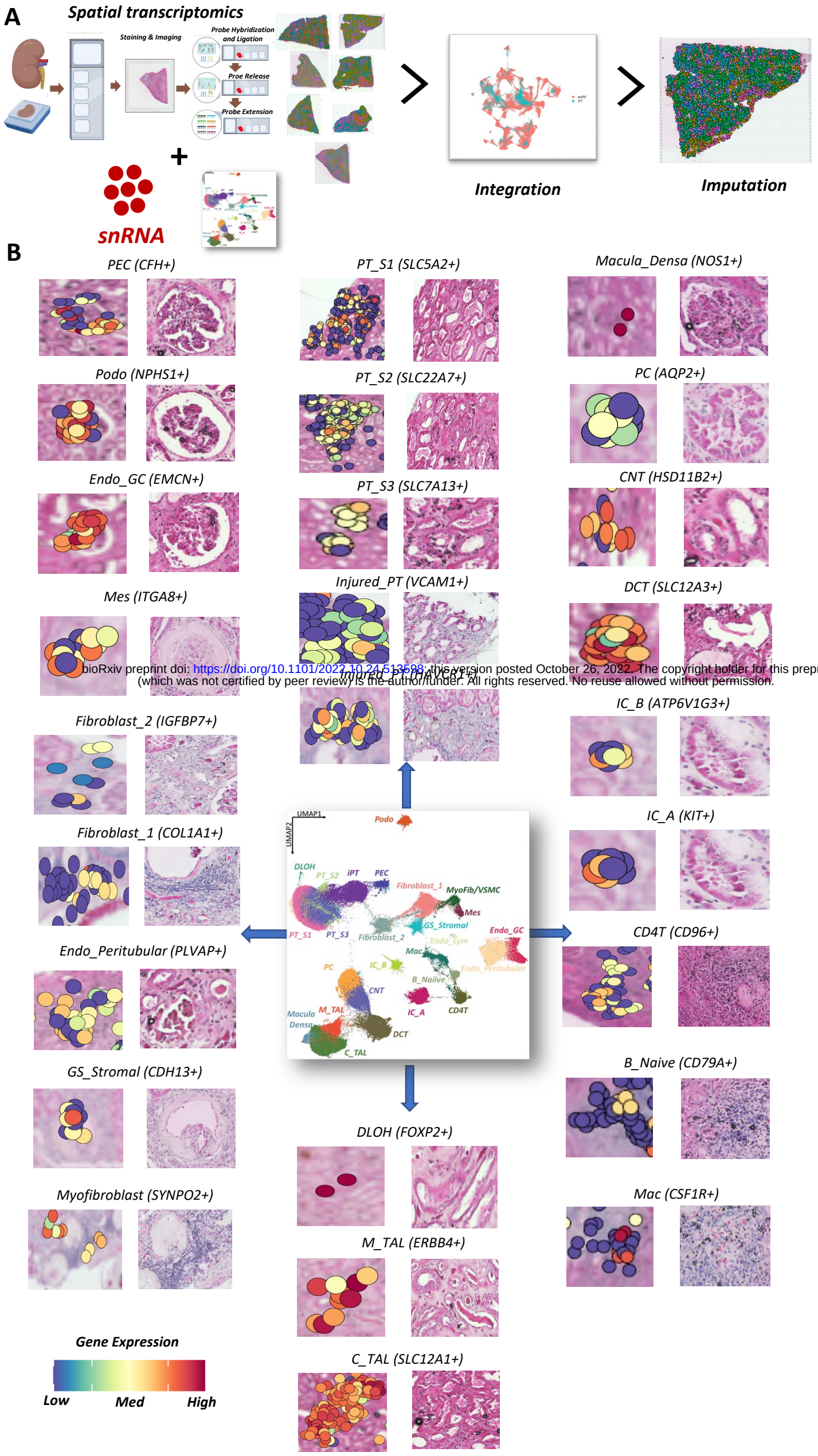
bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

47. P. Dhillon *et al.*, The nuclear receptor ESRRA protects from kidney disease by coupling metabolism and differentiation. *Cell Metab.* **33**, 379-394. e378 (2021).
48. T. Doke *et al.*, Single-cell analysis identifies the interaction of altered renal tubules with basophils orchestrating kidney fibrosis. *Nat. Immunol.*, 1-13 (2022).
49. E. Huang *et al.*, The roles of immune cells in the pathogenesis of fibrosis. *Int. J. Mol. Sci.* **21**, 5203 (2020).
50. F. Barone *et al.*, Stromal fibroblasts in tertiary lymphoid structures: a novel target in chronic inflammation. *Front. Immunol.* **7**, 477 (2016).
51. Y. Huang *et al.*, Identification of novel genes associated with renal tertiary lymphoid organ formation in aging mice. *PLoS One* **9**, e91850 (2014).
52. A. R. Kitching, Immunoaging within the kidney via injury-associated tertiary lymphoid tissue. *Kidney Int.* **102**, 9-11 (2022).
53. R. Luo *et al.*, Tertiary lymphoid organs are associated with the progression of kidney damage and regulated by interleukin-17A. *Theranostics* **11**, 117 (2021).
54. Y. Sato *et al.*, CD153/CD30 signaling promotes age-dependent tertiary lymphoid tissue expansion and kidney injury. *The Journal of clinical investigation* **132**, (2022).
55. J. Zhao, S. Gu, A. McDermaid, Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Math. Biosci.* **310**, 24-30 (2019).
56. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).
57. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e324 (2019).
58. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell*, (2021).
59. I. Korsunsky *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).
60. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
61. C. Trapnell *et al.*, Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* **32**, 381 (2014).
62. T. Stuart, A. Srivastava, C. Lareau, R. Satija, Multimodal single-cell chromatin analysis with Signac. *BioRxiv*, (2020).
63. G. Yu, L.-G. Wang, Q.-Y. He, ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383 (2015).
64. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).
65. D. M. Cable *et al.*, Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517-526 (2022).
66. T. Stuart *et al.*, Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
67. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
68. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760-1774 (2012).



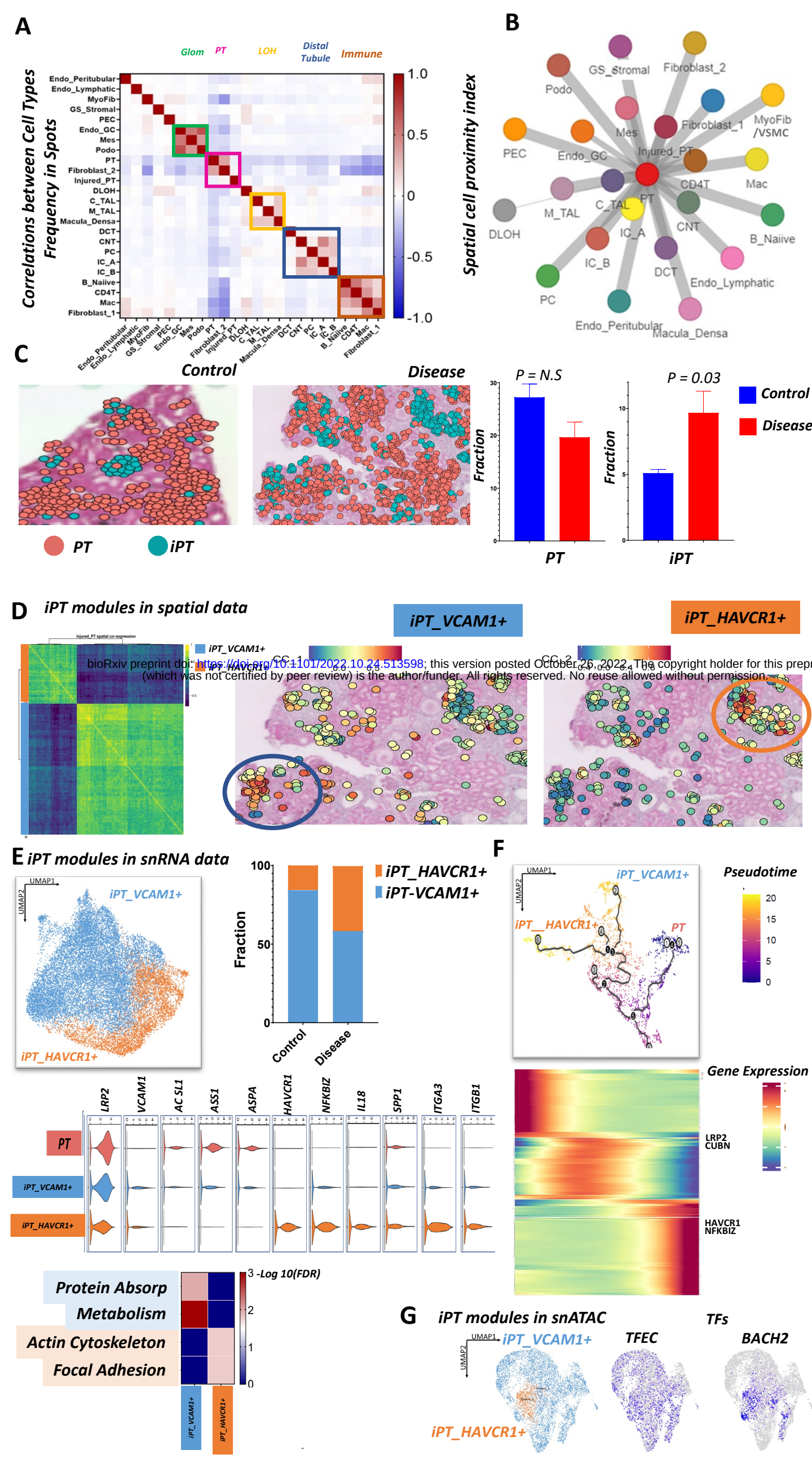
bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig 1. Single cell resolution comprehensive human kidney multi-omics atlas. (A) Study overview. **(B)** Combined UMAP representation of 453,718 integrated human kidney sn/sc-RNA-seq and sn-ATAC-seq data. (**C**) UMAP of 223,438 human kidney snRNA-seq data and bubble dot plots of cluster specific marker genes. The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression. **(D)** UMAP of 149,498 human kidney scRNA-seq data and bubble dot plots of cluster specific marker genes. The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression. **(E)** UMAP of 80,845 human kidney snATAC-seq data and bubble dot plots of cluster specific marker genes using gene activity score. The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression. Endo_GC; endothelial cells of glomerular capillary tuft, Endo_peritubular; endothelial cells of peritubular vessels, Endo_lymphatic; endothelial cells of lymphatic vessels, Mes; meseangial cells, GS_Stromal; glomerulosclerois-specific stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, NK; natural killer cells, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naiive; Naiive 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+.



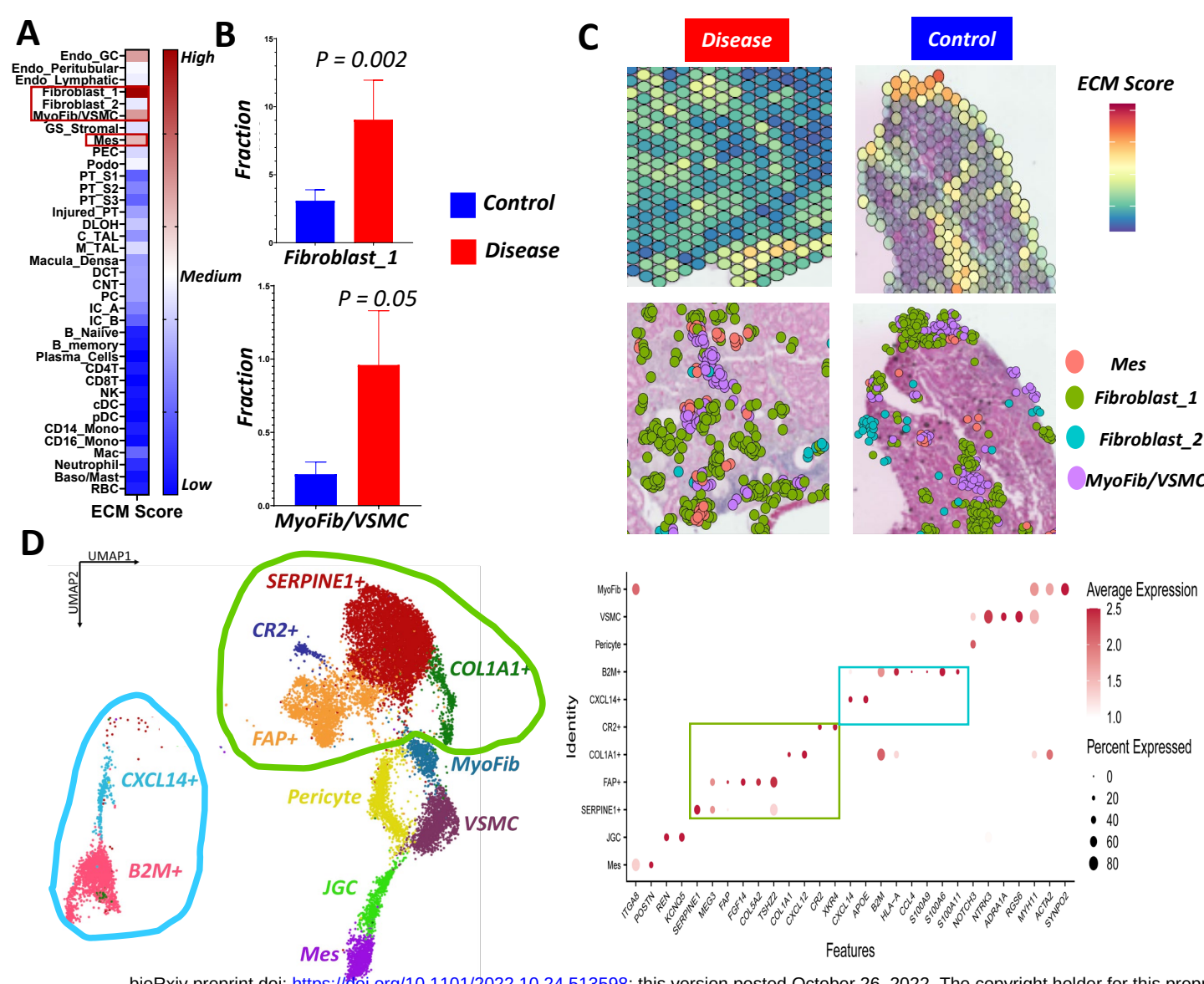
bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig 2. Spatially resolved human kidney gene expression atlas. (A) Overview of the data generation and analysis of spRNA-seq. The spRNASeq data was integrated with snRNA-seq information and spots were imputed to obtain near single cell level information using "CellTrek". **(B)** Spatial location and specific marker genes expression of identified cell types in snRNA-seq. The dots show cells mapped back to their spatial location in the human kidney tissue. For each spatial location the original (H&E) image of the slide shown. The color indicates the gene expression level of specific marker genes, from blue to red indicates higher expression. Endo_GC; 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; glomerulosclerois-specific stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, CD4T; T lymphocytes CD4+, B_Naiive; Naiive B lymphocyte, Mac; macrophage.



bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig 3. Spatial analysis highlights the proximity of the injured PT cells to the fibrotic niche. **(A)** The co-occurrence (Pearson correlation) of different kidney cells in the spatial transcriptome data after spot deconvolution using RCTD method using snRNA-seq as reference. The color indicates the degree of correlation. Higher correlation indicates the higher probability of co-occurrence of the cells. Different kidney compartments were encircled on the heatmap. **(B)** The spatial distance of kidney cells in the spRNAseq data using sColoc of CellTrek. The circles indicate each cell type and the distance from the center indicates the lower frequency and distribution of cell types. The distance of the circles correlates with the distance of those cell types. **(C)**. Bar graphs indicated the mean fractions of proximal tubule and injured proximal tubule cells in healthy control and diseased samples in snRNA-seq data. Bars indicate SEM. P values were calculated using independent *t* test. **(D)** The gene co-expression network indicates two types of injured_PT in diseased human kidneys, left panel shows the heatmap of co-expressed genes. The right panels indicate the spatial location of the identified injured PT cells. The color scheme of the heatmap indicate the expressions of the genes in each iPT modules. **(E)** Two types of injured PT cells in snRNA-seq dataset. The bar graph shows the frequency of the different iPT types in healthy control and diseased samples. Violin plots show the different gene markers in PT and iPT cells. The heatmap indicates the enriched pathways for iPT-VCAM1+ and iPT-HAVCR1+ (lower panel). **(F)** UMAP representation of PT and iPT cell sub-clustering trajectory from PT to iPT-VCAM1+ and iPT- HAVCR1+ in snRNA-seq (Upper panels). Cells are colored by pseudotime and the arrow indicates the direction of the pseudotime. The heatmap shows the differentially expressed genes along the trajectory. The color scheme indicates the z scores of expression along the trajectory. **(G)** Representative feature plot of motif activity of specific transcription factors in iPT-VCAM1+ and iPT-HAVCR1+ using snATAC-seq data. Endo_GC; 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; glomerulosclerois-specific stromal cells, VSMC/Myofib; vascular smooth muscle cells/myofibroblast, PEC; parietal epithelial cells, Podo; podocyte, PT_S1; proximal tubule segment 1, PT_S2; proximal tubule segment 2, PT_S3; proximal tubule segment 3, Injured_PT; injured proximal tubule cells, DLOH; thin descending loop of Henle, C_TAL; cortical thick ascending loop of Henle, M_TAL; medullary thick ascending loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule cells, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells, IC_B; Type beta intercalated cells, CD4T; T lymphocytes CD4+, B_Naiive; Naiive B lymphocyte, Mac; macrophage.



bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this prepr
(which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

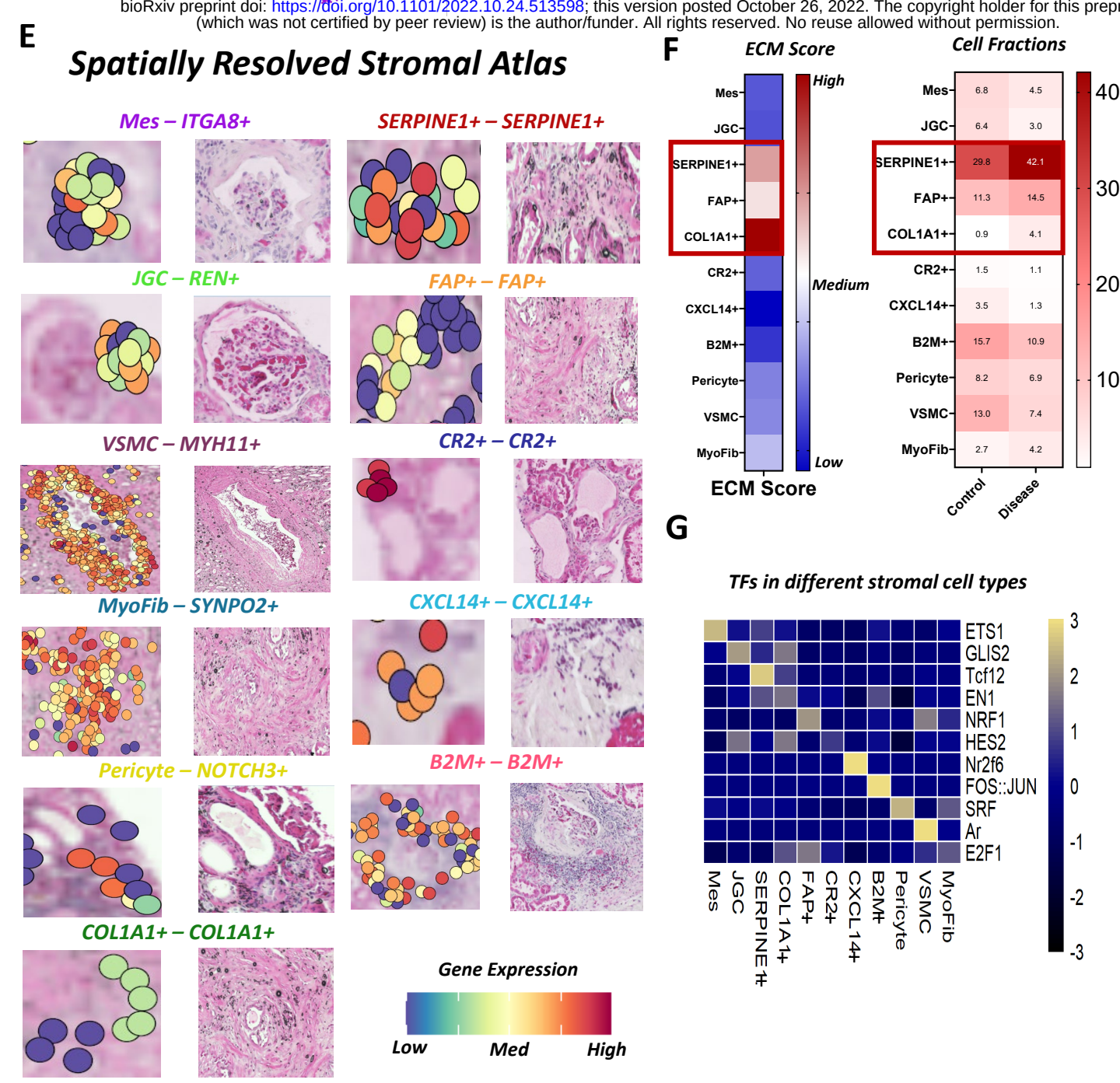
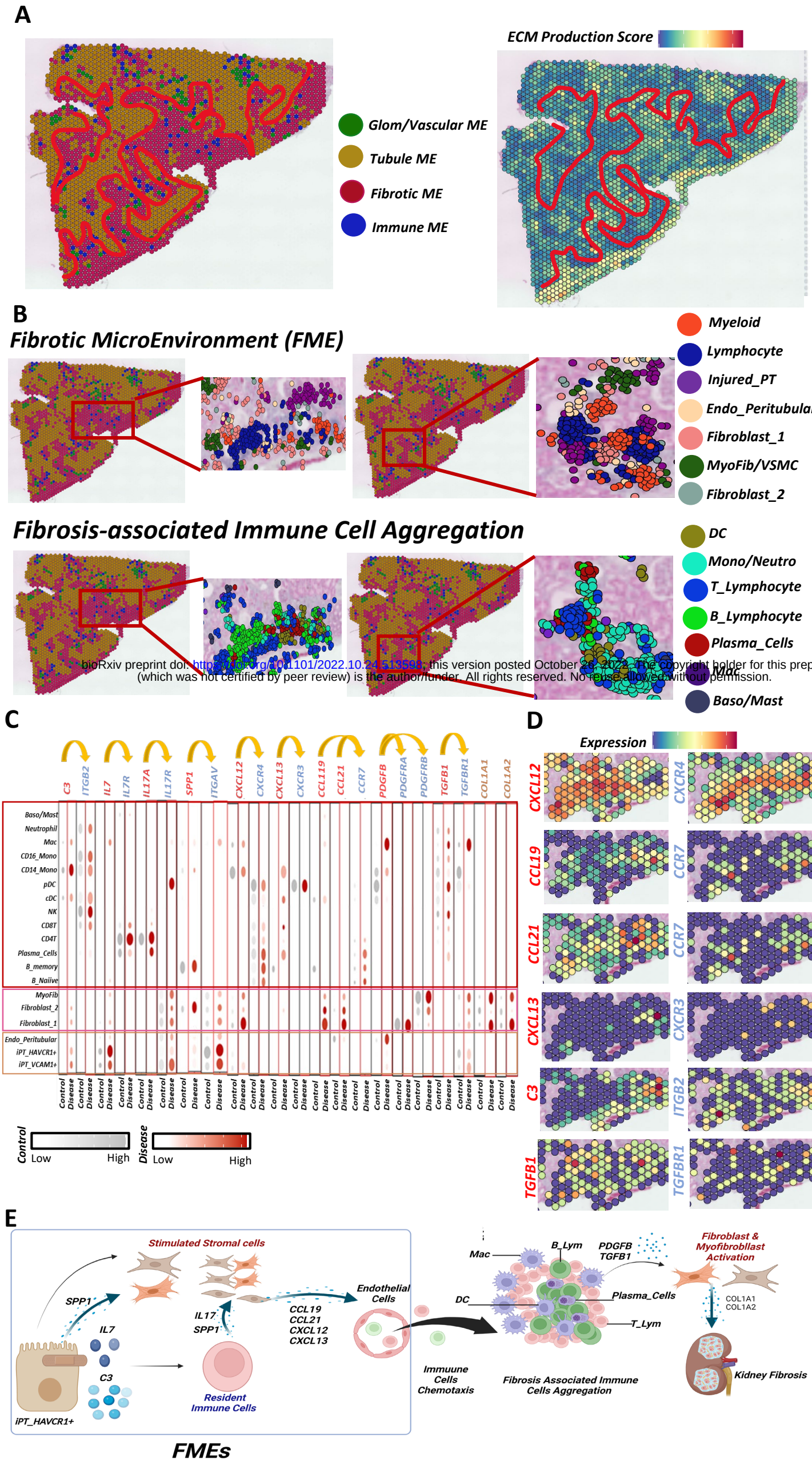
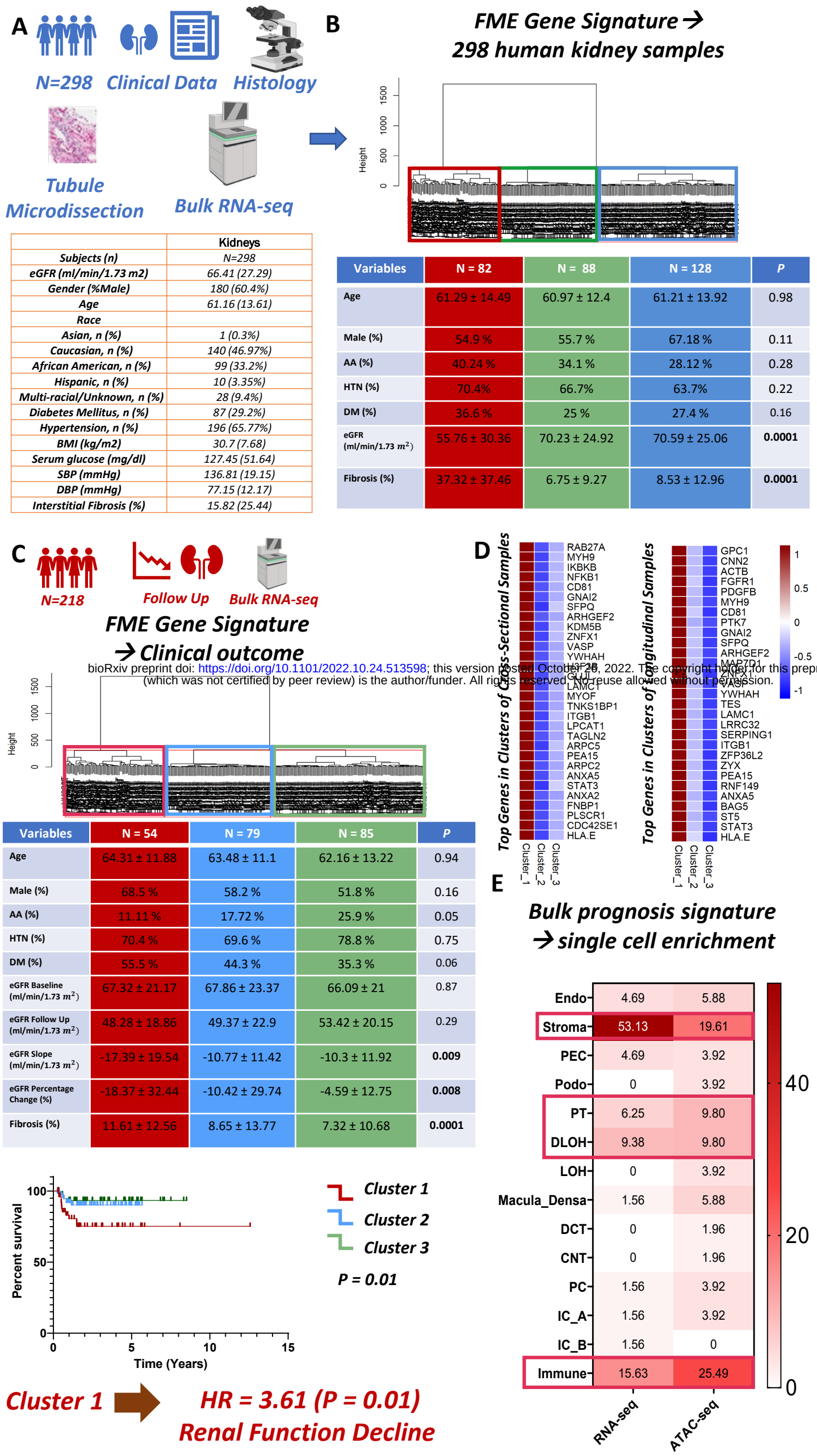


Fig 4. Spatially and transcriptionally resolved fibroblasts heterogeneity in human kidney fibrosis. (A) Extracellular matrix production score in different kidney cells in the sn/sc-RNA-seq. The color scheme indicates the ECM score in each cell type, calculated by the expression of the collagen, proteoglycan and glycoprotein genes. **(B)** The comparison between fractions of fibroblast_1 and VSMC/myofibroblast cells with the highest ECM score in healthy control and diseased samples. The bars indicate SEM. Independent t test was used to compare the fractions between two groups. **(C)** The ECM score in spRNA-seq data of healthy and diseased samples (upper panel). The color scheme indicates the z score. The spatial location of the cells in the regions with high ECM in spRNAseq data (lower panel). **(D)** UMAP representation of sub-clustering of stomal cell in snRNA-seq dataset (left panel). The bubble dot plots of cluster specific marker genes in the snRNA-seq. The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression (right panel). **(E)** The spatial location and specific marker genes expression of identified stromal cell types in snRNA-seq. The dots show the cells mapped back to their spatial location in the human kidney. The original (H&E) image of the slide is shown side by side. The colors indicate the gene expression level of specific marker genes. (**F**) The heatmap of ECM score in the sub-clustered stromal cells (left panel). The heatmap of the fractions of different types of stromal cells in healthy control and diseased samples (right panel). (**G**) Transcription factor enrichment motifs in each stromal cell sub-cluster. The heatmap shows the z score of motif activity in each cell type using chromvar. Mes; mesangial cells, JGC; juxta glomerular cells, SERPINE1+; *SERPINE1* positive fibroblast, FAP+; *FAP* positive fibroblast, COL1A1+; collagen 1 producing fibroblasts, CR2+; *CR2* positive fibroblast, CXCL14+; *CXCL14* positive fibroblast, B2M+: *B2M* positive fibroblast, VSMC; vascular smooth muscle cells, MyoFib; myofibroblast.



bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig 5. The kidney fibrotic microenvironment relies on complex and organized epithelial, stromal, endothelial and immune cell interaction. (A) Different human kidney microenvironments in the (left panel) spRNA-seq dataset and the calculated ECM score (right panel). The color indicates ECM gene expression score in the kidney. **(B)** Key cell types located in the fibrotic microenvironment (upper panel). Fibrosis-associated immune cell aggregation in FMEs showing lymphocytes, plasma cells and macrophages (lower panel). The dots indicate cells mapped back to their spatial location using the merged spRNAseq, snRNA-seq and scRNA-seq datasets. (**C**) The bubble plot of expression of ligands and receptors in regions of FME in integrated sn/scRNA-seq data. The size of the dot indicates the percent positive cells and the darkness of the color indicates average expression (right panel). The gray indicates control and red indicates diseased group. **(D)** Ligands and receptor expression in specific cell types in FME regions (in spRNA-seq data). The color intensity indicates gene expression level and the dot indicates the location of the expressions. (**E**) Summary of the putative mechanism of the human kidney fibrosis. iPT; Injured_PT, MyoFib; Myofibroblast, CD4T; T lymphocytes CD4+, CD8T; T lymphocytes CD8+, B_Naiive; Naiive B lymphocyte, B_memory; memory B lymphocyte, Baso/Mast; basophil or mast cells, pDC; plasmacytoid dendritic cells, cDC; classical dendritic cells, Mac; macrophage, CD14_Mono; monocyte CD14+, CD16_Mono; monocyte CD16+.



	Kidneys
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)*

Variables	N = 82	N = 88	N = 128	*P*
Age	61.29 ± 14.49	60.97 ± 12.4	61.21 ± 13.92	0.98
Male (%)	54.9 %	55.7 %	67.18 %	0.11
AA (%)	40.24 %	34.1 %	28.12 %	0.28
HTN (%)	70.4%	66.7%	63.7%	0.22
DM (%)	36.6 %	25 %	27.4 %	0.16
eGFR (ml/min/1.73 m^2)	55.76 ± 30.36	70.23 ± 24.92	70.59 ± 25.06	**0.0001**
Fibrosis (%)	37.32 ± 37.46	6.75 ± 9.27	8.53 ± 12.96	**0.0001**

Variables	N = 54	N = 79	N = 85	*P*
Age	64.31 ± 11.88	63.48 ± 11.1	62.16 ± 13.22	0.94
Male (%)	68.5 %	58.2 %	51.8 %	0.16
AA (%)	11.11 %	17.72 %	25.9 %	0.05
HTN (%)	70.4 %	69.6 %	78.8 %	0.75
DM (%)	55.5 %	44.3 %	35.3 %	0.06
eGFR Baseline (ml/min/1.73 m^2)	67.32 ± 21.17	67.86 ± 23.37	66.09 ± 21	0.87
eGFR Follow Up (ml/min/1.73 m^2)	48.28 ± 18.86	49.37 ± 22.9	53.42 ± 20.15	0.29
eGFR Slope (ml/min/1.73 m^2)	-17.39 ± 19.54	-10.77 ± 11.42	-10.3 ± 11.92	**0.009**
eGFR Percentage Change (%)	-18.37 ± 32.44	-10.42 ± 29.74	-4.59 ± 12.75	**0.008**
Fibrosis (%)	11.61 ± 12.56	8.65 ± 13.77	7.32 ± 10.68	**0.0001**

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.24.513598; this version posted October 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig 6. Fibrotic microenvironment gene signature successfully predict disease prognosis in a large cohort of human kidney samples. (A) Clinical characteristics of 298 human kidney tubule RNA-seq samples. (**B**) Unbiased cluster dendrogram of 298 human kidney tubule bulk RNA-seq samples based on expression of FME genes. 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 dendrograms of 218 human kidney tubule bulk RNA-seq samples with follow up eGFR based on expression of FME genes. The characteristics of each cluster were shown in the table. Chi-square test was used for categorical variables and one-way ANOVA for continuous variables was used to compare groups. The lower panel shows the Kaplan-Meier analysis of comparison of 3 the groups for renal survival. Renal survival was defined as cases reaching end stage renal disease or greater than 40% eGFR decline. (**D**) Heatmap of mean expression of top 30 genes cluster driving genes in 298 cross-sectional human kidney bulk RNA-seq (left panel) and 218 longitudinal human kidney bulk RNA (right panel). Top genes were defined based on the highest variation between groups using ANOVA. The heatmap shows the z score of mean gene expression in each cluster. (**E**) Enrichments of eGFR decline associated genes in sn/scRNA-seq and snATAC-seq clusters. The heatmap shows the percentage of the eGFR decline associated genes in each cluster with highest expression. HR; hazard ration, Endo; endothelial cells, Stroma; stromal cells, PEC; parietal epithelial cells, Podo; podocyte, PT; proximal tubule cells, DLOH; thin descending loop of Henle, LOH; loop of Henle, DCT; distal convoluted tubule, CNT; connecting tubule, PC; principal cells of collecting duct, IC_A; Type alpha intercalated cells,.