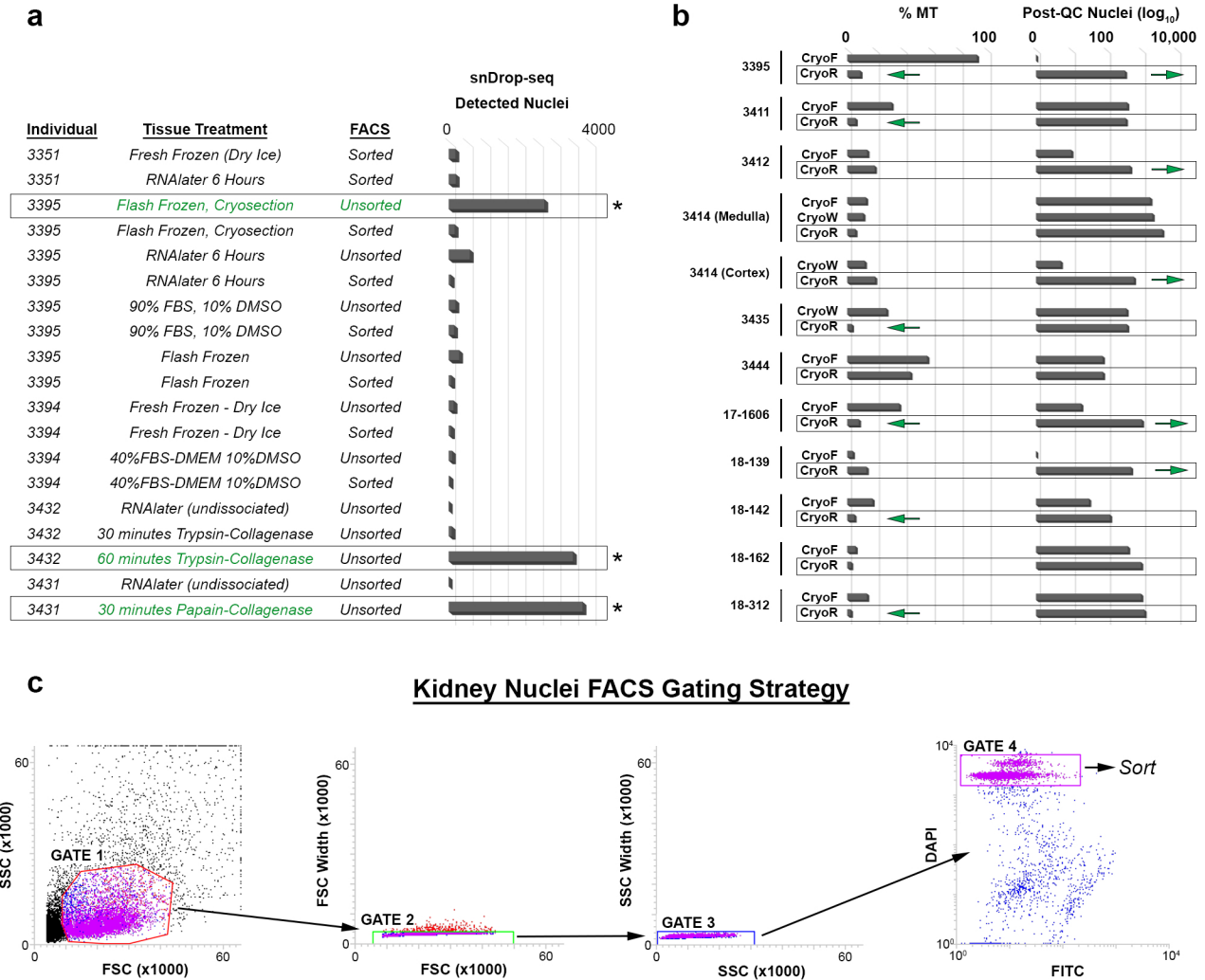


SUPPLEMENTARY INFORMATION

A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys

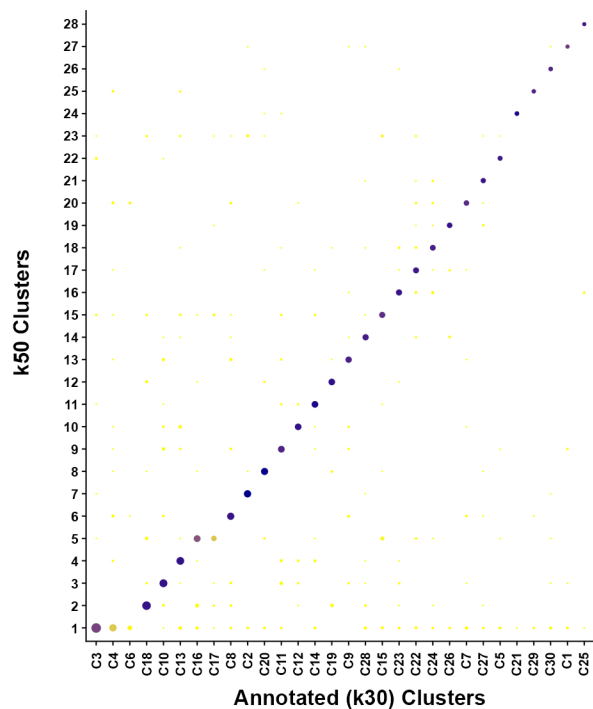
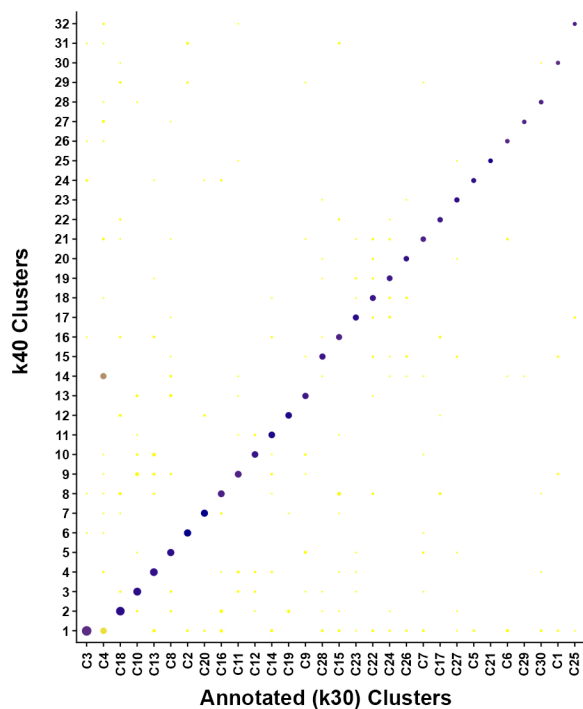
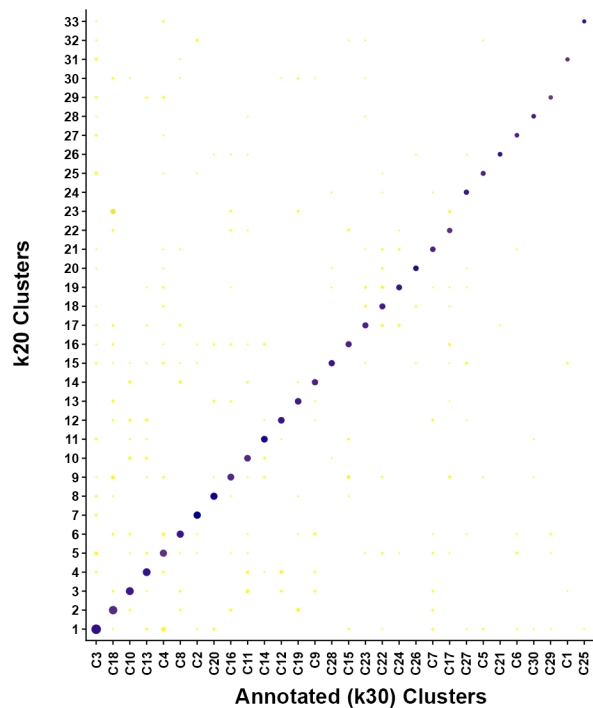
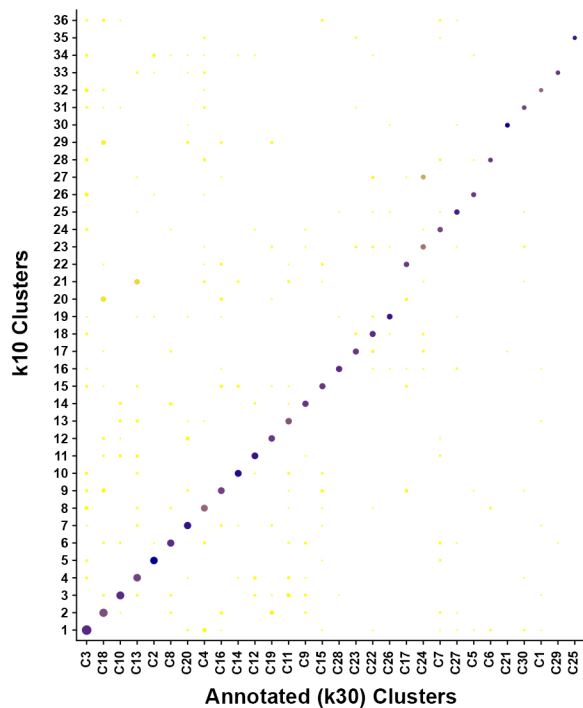
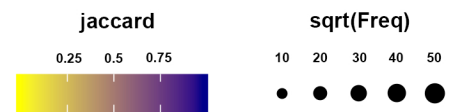
Lake et al.

Tissue Processing Assessments

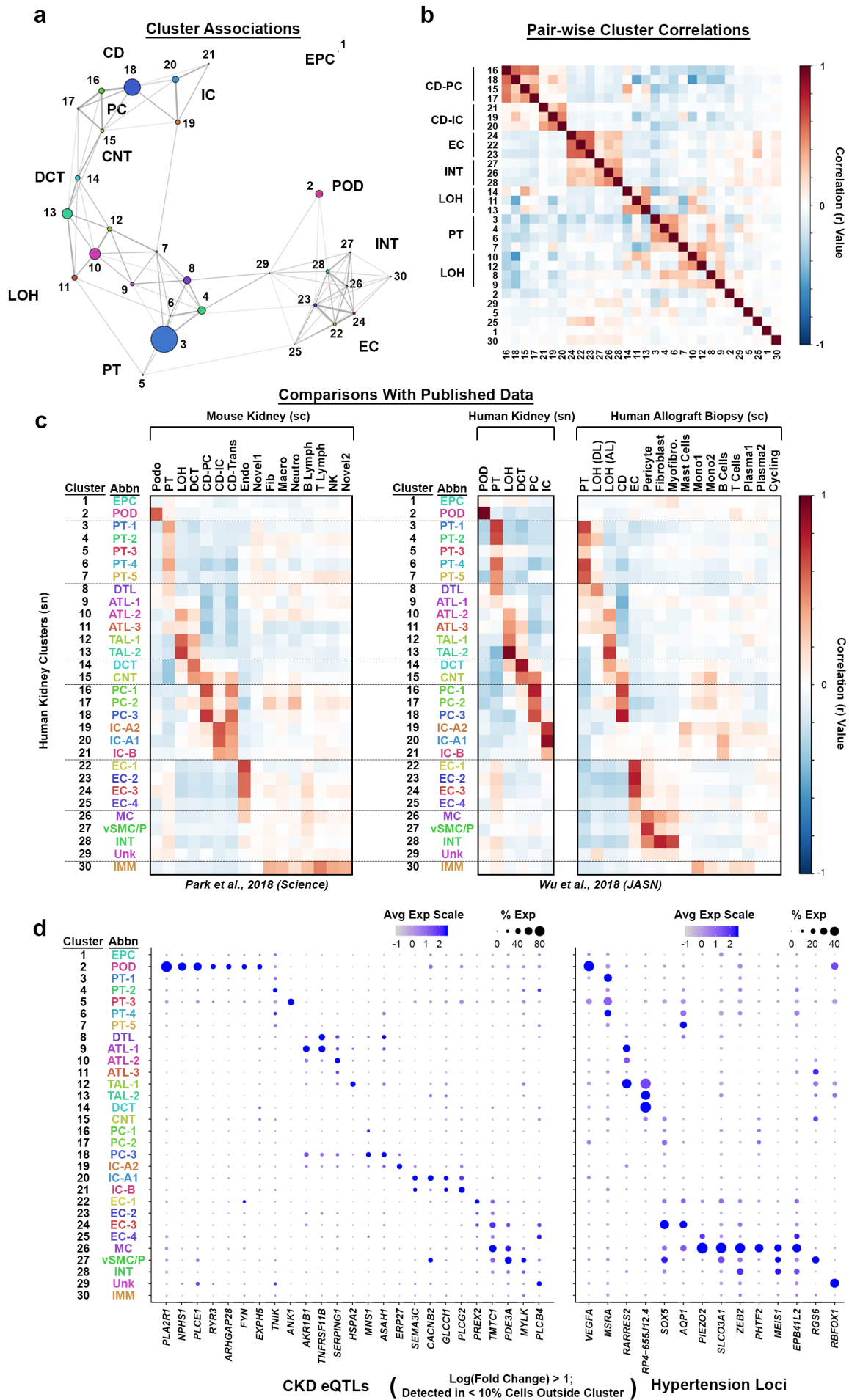


Supplementary Figure 1. Adult human kidney tissue processing assessment. **a.** Nuclei from different tissue preservation and processing methods within and across individuals ($n = 5$) were processed by snDrop-seq and the number of nuclei (> 400 UMI detected) relative to total number of nuclei loaded is shown. Asterisks indicate conditions showing positive outcome. **b.** Comparison of snDrop-seq outcome using nuclei isolated from cryosections from the same samples ($n = 12$) that were treated in different ways (cryoF, cryoW, cryoR). Percent MT (nuclei > 400 UMI) and number of post-QC nuclei (nuclei > 400 nonMT UMI, > 400 and < 5000 genes detected) are shown. Arrows indicate lower percent MT and higher post-QC yields for cryoR samples. **c.** FACS gating strategies for DAPI stained kidney nuclei used in (a). Source data for (a) and (b) are provided as a Source Data File.

Supplementary Figure 2. Single nucleus interrogation of the human kidney. **a.** Data processing pipeline. Count matrices across experiments were combined and only nuclei detecting more than 400 non-MT UMI were included. Mitochondrial genes not expressed in the nucleus were excluded. Nuclei detecting more than 400 and less than 5000 genes were then subjected to a gene/UMI ratio filter (see Methods) to further remove low quality data prior to clustering using Pagoda2 software. Two rounds of clustering were performed to remove nuclei that failed to cluster or that formed clusters of less than 30 nuclei. This allowed further removal of low-quality nuclei or multiplets. Remaining clusters that were derived from the 17,659 post-QC nuclei were visualized by UMAP. The number of nuclei after each stage of processing are shown. **b.** Heatmap of expression of marker genes defining clusters shown in **Fig. 1b** using criteria indicated and for differentially expressed genes found in **Supplementary Data 7**. **c.** Dot plot of marker gene expression values (log scale) and percentage of nuclei expressing these genes within each cluster (**Fig. 1b**) using criteria indicated and for differentially expressed genes found in **Supplementary Data 7**.

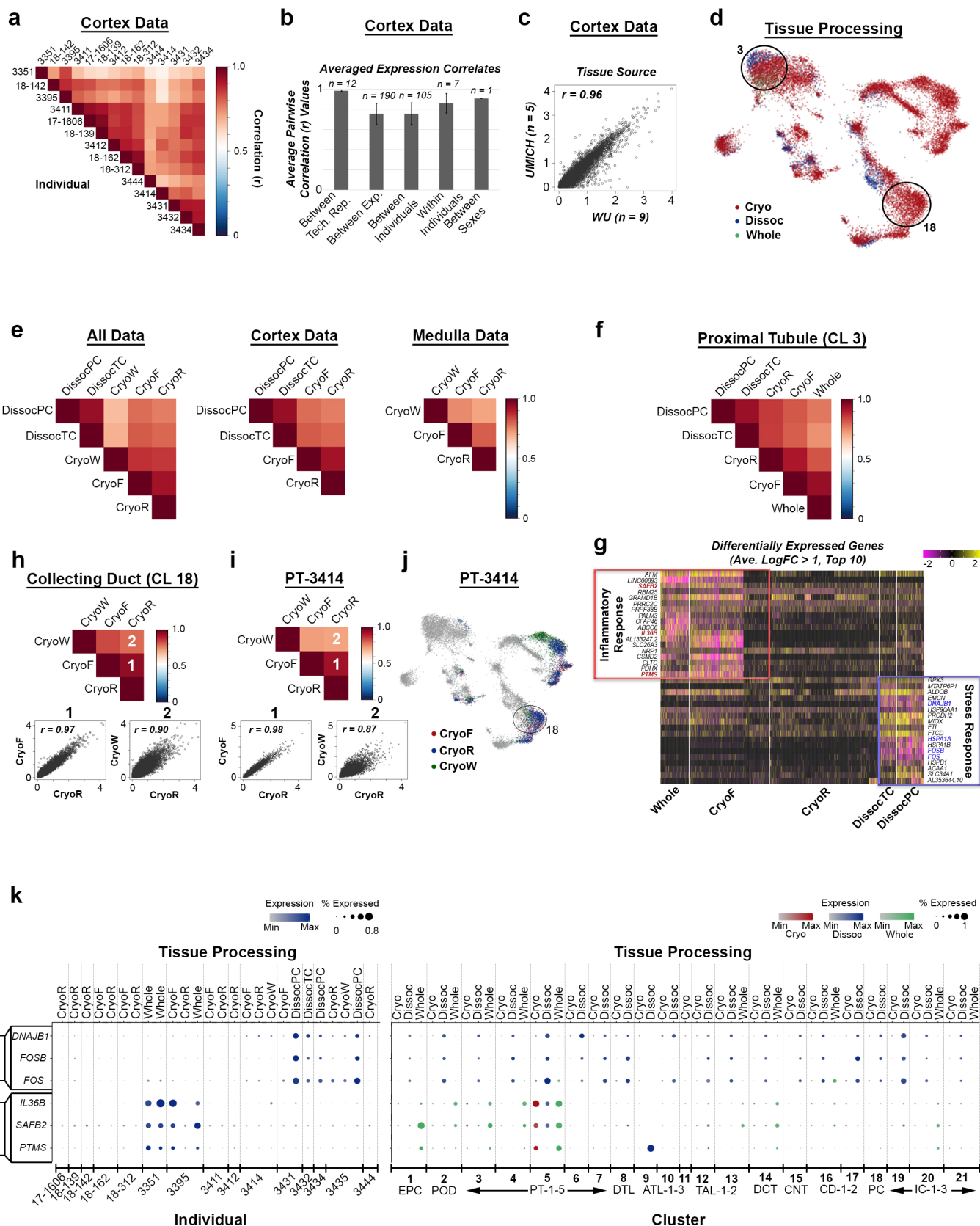


Supplementary Figure 3. Stability of single nucleus clusters. Cluster identities derived using different k values were compared against the final cluster identities shown in **Fig.1b** using jaccard similarity index to indicate extent of overlap for the single nuclei barcodes (see Methods). For the most part nuclei were similarly clustered at differing k values, with a few smaller sub-clusters arising at lower k values (e.g. C18 or PC-3 splitting into two additional small sub-clusters 20 and 29 at k = 10), and a few clusters merging at high k values (e.g. PT clusters C3, C4 and C6 merging into a single PT cluster at k = 50).

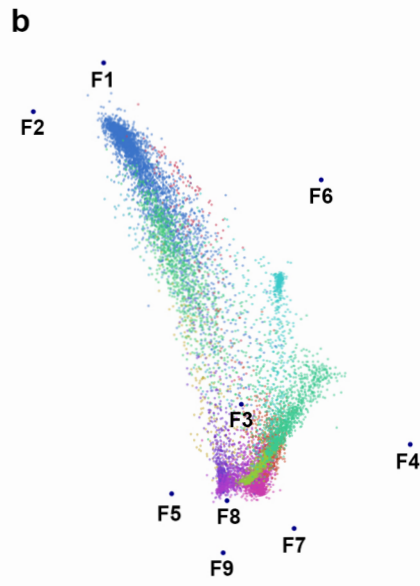
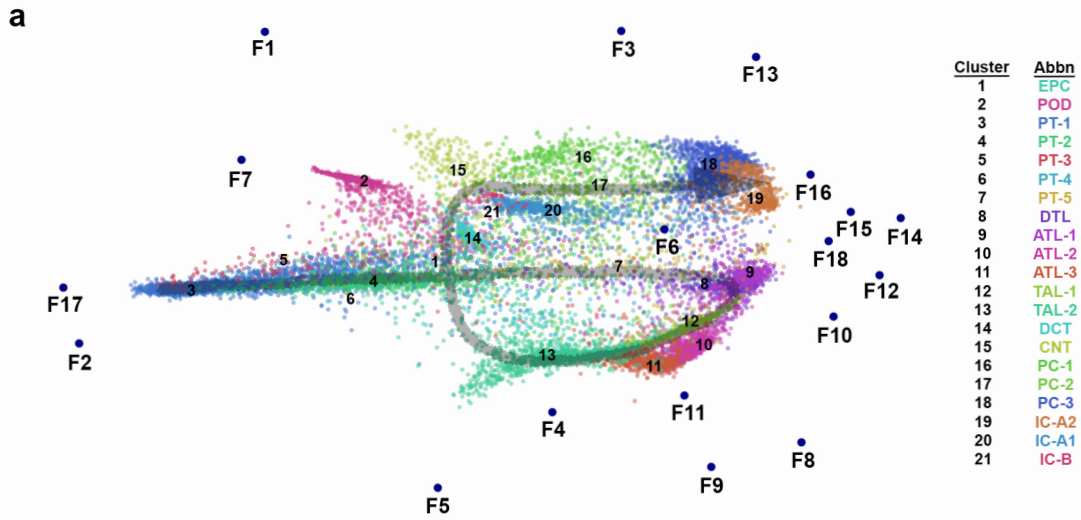


Supplementary Figure 4. Distinct cell types resolved from single nucleus data. **a.** A cluster association plot where thickness of connecting lines indicates relative correlation of averaged scaled expression values for the top 2000 variant genes used to define clusters, and point size indicates number of nuclei within each cluster. **b.** Hierarchical clustering (ward.D method) heatmap for pairwise correlation values that were generated on averaged scaled expression values for the top 2000 variant genes used for PAGODA2 clustering. **c.** Clusters shown in **(a)** were compared with published single-cell or single-nucleus data from the mouse or human kidney. Heatmaps show correlation values for cluster-averaged scaled expression values for shared top variant genes detected in each data set (see Methods). **d.** Dot plot of a subset of distinct CKD-associated eQTLs and hypertension risk loci expression values (log scale), and percentage of nuclei expressing these genes within each cluster (**Fig. 1b**), using the selection criteria indicated and for differentially expressed genes found in **Supplementary Data 8**.

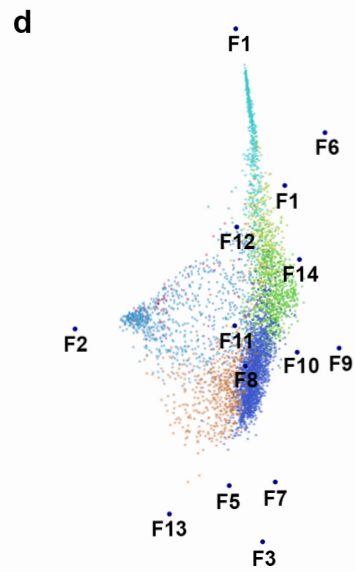
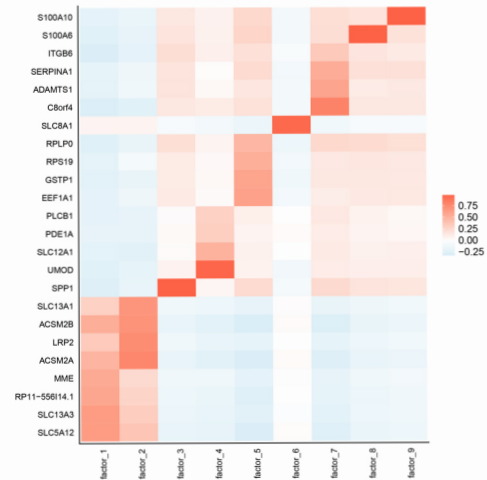
Supplementary Figure 5. QC assessment of snDrop-seq data. **a.** Violin plots indicating number of genes and UMI detected per nucleus and the percent MT that were detected for 17,659 post-QC nuclei (**Supplementary Data 4**) grouped by tissue processing conditions. **b.** Similar metrics as in (**a**) except for nuclei grouped by cluster identity. Boxes indicate S3 PT (PT-5, cluster 7) and TAL (TAL-1, cluster 12) clusters showing elevated %MT. **c.** UMAP plot as shown in **Fig. 1b** showing percent MT across single nuclei. Clusters highlighted in (**b**) are indicated. Violin plots show percent MT for cluster 12 (TAL-1) nuclei only grouped by either tissue processing or tissue procurement method. **d.** UMAP plots as shown in **Fig. 1b** showing different metadata components associated with each nucleus (**Supplementary Data 4**), including: tissue region, experiment, batch (library), individual, tissue processing method used, collection or procurement method, sex and gene depth. **e.** Pie charts for each cluster showing proportions contributed from different individuals, different tissue processing methods and different tissue regions. Predicted origin for each cluster to the cortex or medulla is indicated. *Predicted artifactual cortical PT cluster derived mostly from a single individual. **Predicted artifactual CD cluster derived primarily from tissue processing methods involving enzymatic dissociation (dissocPC and dissocTC). Source data for (**e**) is provided as a Source Data File.



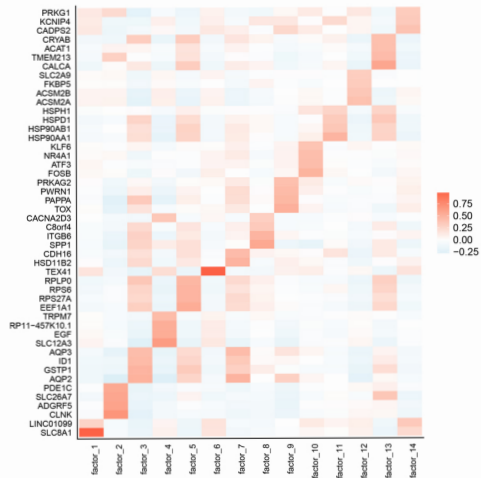
Supplementary Figure 6. QC and reproducibility assessment across individuals and tissue processing methods. **a.** Heatmap of Pearson correlation values (r) that were generated on averaged expression values (all genes) for nuclei grouped by the different individuals they were isolated from (renal cortex data only, 14 individuals). **b.** Bar chart of averaged correlation values generated on nuclei (cortex data only) grouped by the conditions indicated. Correlations were based on average expression values (all genes), n refers to the number of comparisons performed, and error bars indicate standard deviation (sd) between comparisons. Source data is provided as a Source Data File. **c.** Scatter plot for averaged gene expression values (log transformed) for nuclei grouped by tissue source. Associated correlation value and number of individuals (cortex only) the nuclei were derived from are indicated. **d.** UMAP as shown in **Fig. 1b** showing general tissue processing methods used. PT cluster 3 and CD PC cluster 18, used in subsequent analyses, are indicated. **e.** Correlation heatmaps for averaged expression of all genes separately performed on all nuclei, cortex-only nuclei or medulla-only nuclei that were grouped by tissue processing method. **f.** Correlation heatmap for averaged expression of all genes performed on PT cluster 3 nuclei grouped by tissue processing method. **g.** Heatmap of differentially expressed genes identified between tissue processing methods (average log(fold-change) > 1, top 10 genes shown) for PT cluster 3. Gene categories (Inflammatory Response and Stress Response) were based on highlighted genes. **h.** Correlation heatmap for averaged expression of all genes performed on CD cluster 18 nuclei grouped by tissue processing method. Representative scatter plots for comparisons indicated are shown. **i.** Correlation heatmap for averaged expression of all genes performed on nuclei from a single individual (patient 3414) grouped by tissue processing method. Representative scatter plots for comparisons indicated are shown. **j.** UMAP as shown in **Fig. 1b** showing general tissue processing methods used specifically for patient 3414. Cluster 18 is indicated. **k.** Dot plot showing gene expression and associated detection rates for genes highlighted in **(g)** for all 17,659 post-QC nuclei grouped by the conditions indicated. Stress response genes show as specific for dissociated PC/TC and across all clusters, indicating a general tissue processing effect. Inflammatory response genes associate with specific individuals (3351/3395) and were found more in PT clusters, likely reflecting an individual-related artifact rather than a tissue processing effect.



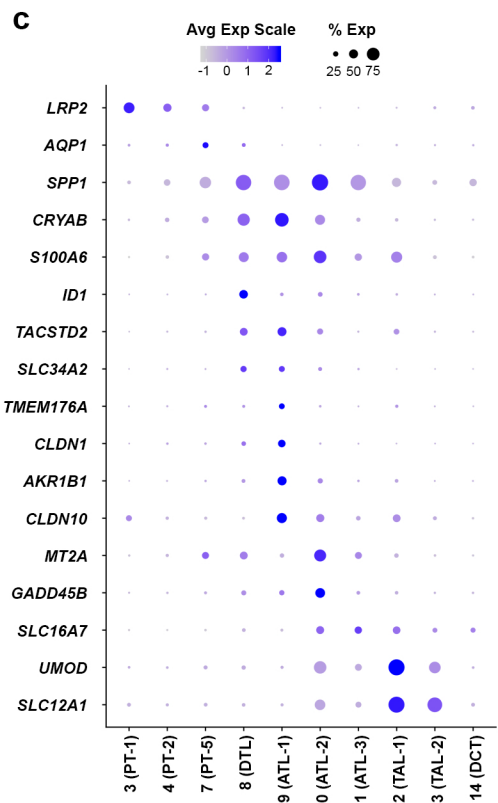
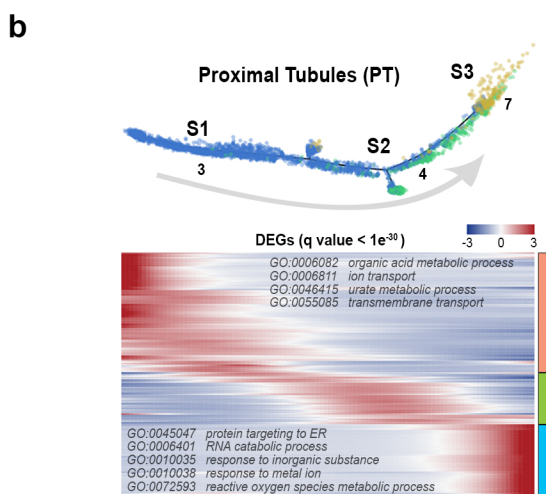
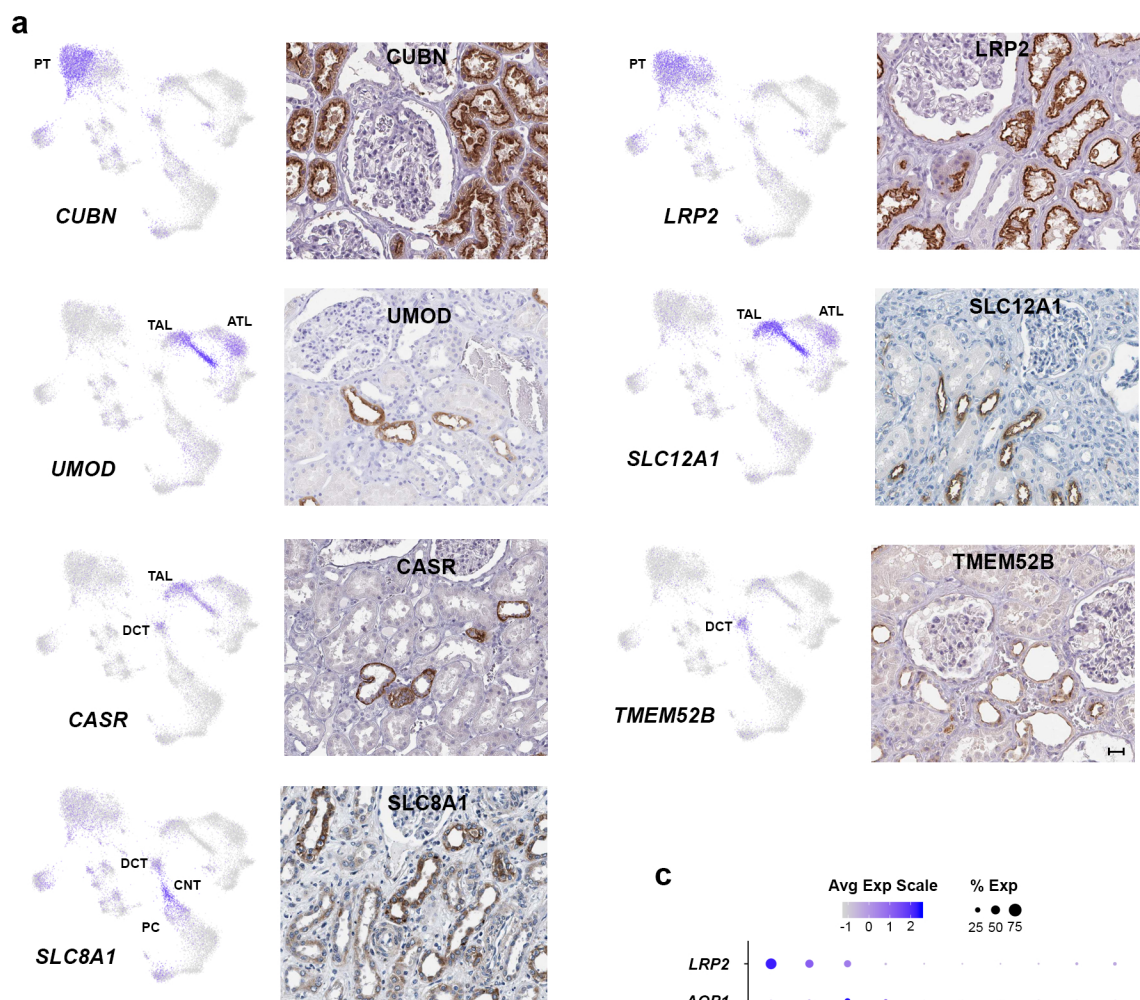
c Top Factor Associated Genes (Score > 0.25)



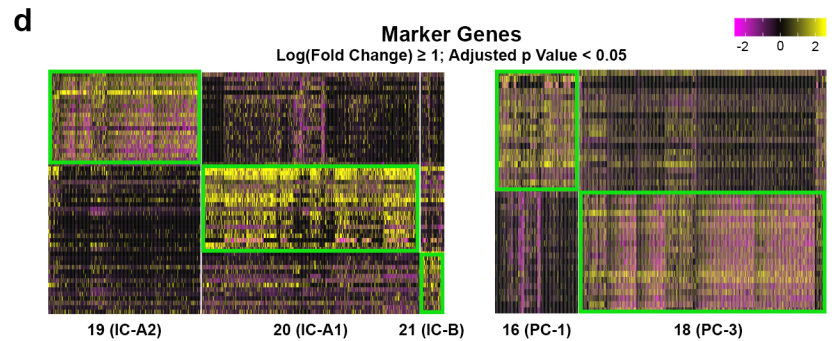
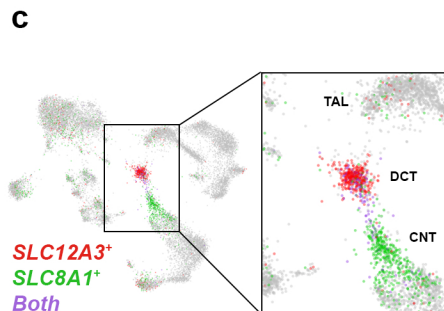
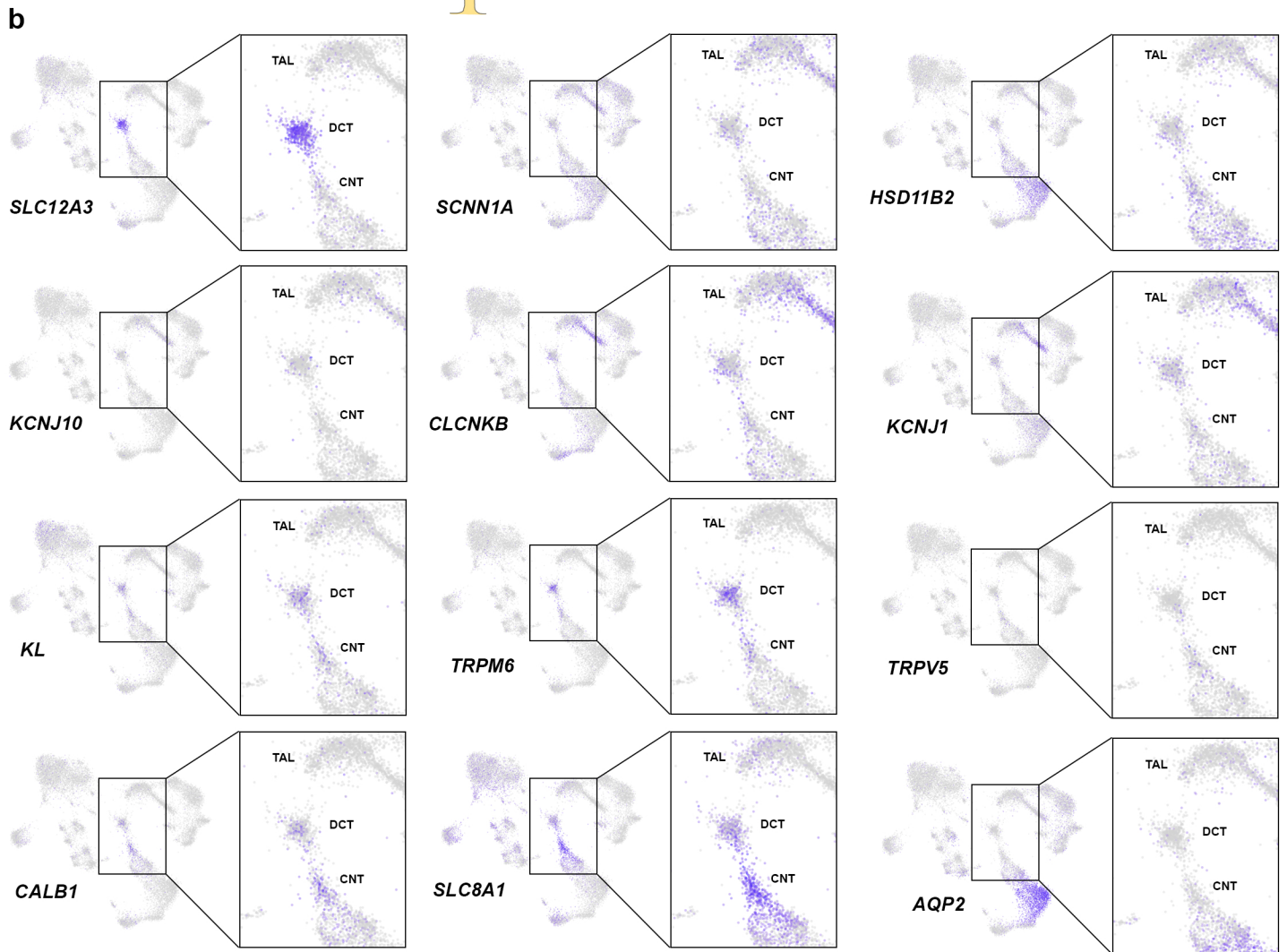
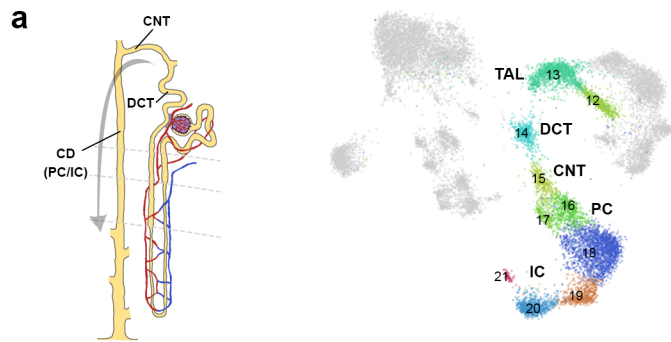
e Top Factor Associated Genes (Score > 0.25)



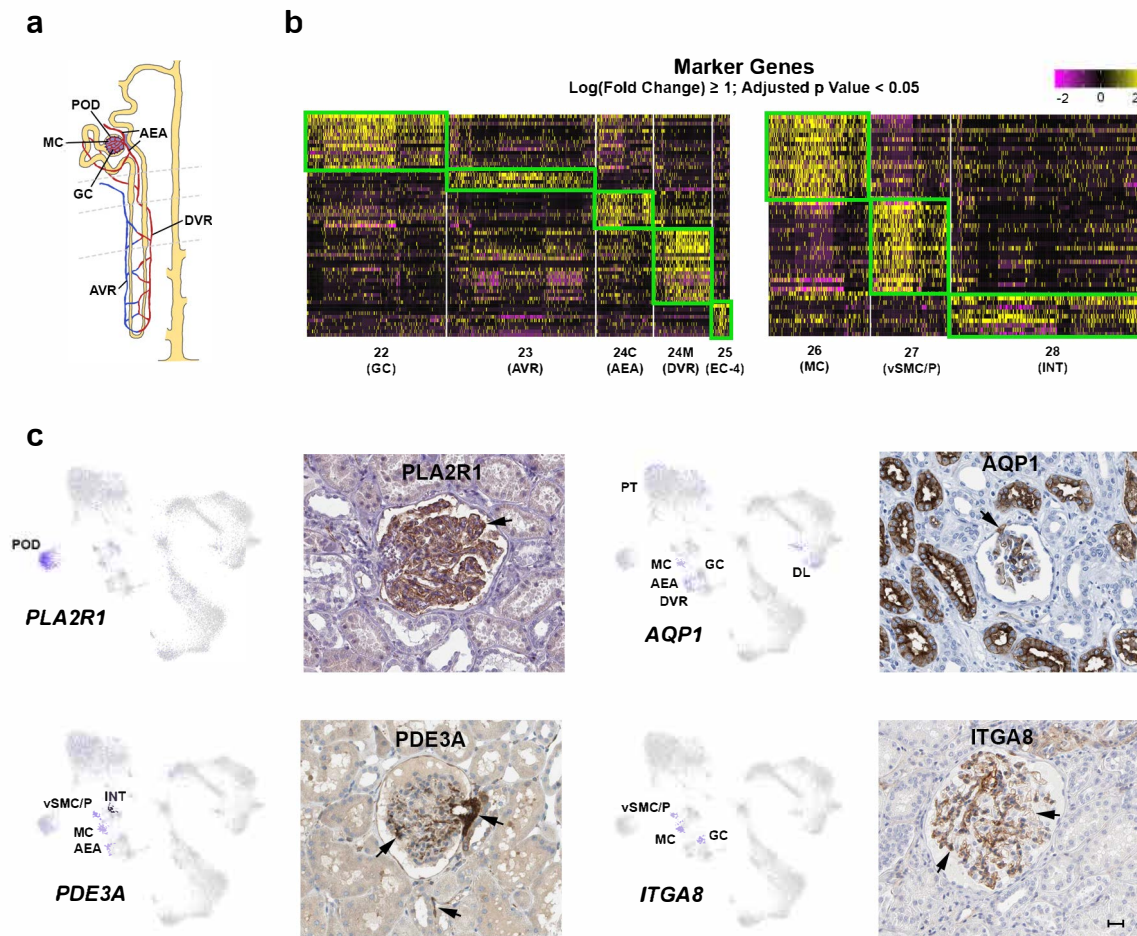
Supplementary Figure 7. SWNE analyses. **a.** Nephron and collecting duct cell populations (clusters 1-21) were visualized using SWNE (see Methods). Arrow indicates expected tubular progression of cell types seen *in vivo*. NMF factors (F1-F18) driving the spatial distribution of the nuclei are indicated. **b.** Spatial distribution of renal tubule cell populations (clusters 2-14) visualized using SWNE. Associated NMF factors (F1-F9) are shown. **c.** Heatmap of the top genes associated with factors shown in **(b)**. **d.** Spatial distribution of distal tubules through collecting duct cell populations (clusters 14-21) visualized using SWNE. Associated NMF factors (F1-F14) are shown. **e.** Heatmap of the top genes associated with factors shown in **(d)**. All factor associated genes can be found in **Supplementary Data 9**.



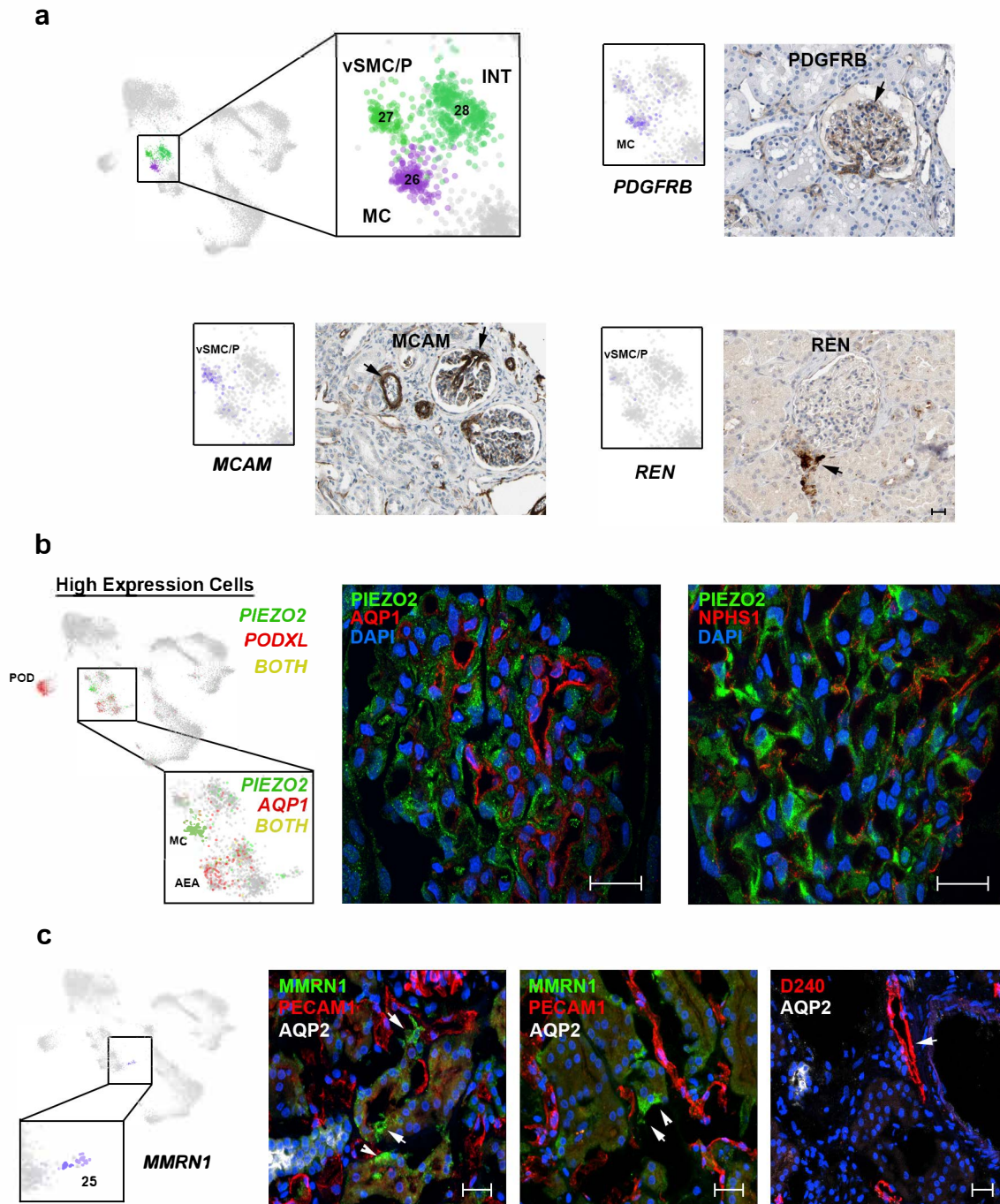
Supplementary Figure 8. Progressive renal tubule cell types. **a.** UMAP plots as in **Fig. 1b** showing relative expression levels (scaled from low - gray to high - blue) of associated markers in indicated cell types and corresponding protein immunostainings (Human Protein Atlas¹, **Supplementary Data 16**). Scale bar indicates 25 μ m. **b.** Trajectory analysis of main PT clusters supporting their S1, S2 and S3 PT segment identities as shown in **Fig. 3b**. Heatmap of expression (row Z-scores) for genes differentially expressed (q value < $1e^{-30}$, see Methods) along the proximal tubule trajectory shown in **(b)** that were grouped into three gene sets by hierarchical clustering (**Supplementary Data 10**, see Methods). Associated gene ontologies are indicated for gene sets 1 and 3. **c.** Dot plot of average marker gene expression values (log scale) and percentage of nuclei expressing these genes for a subset of nephron clusters (**Fig. 1b**). Genes shown are those found to be differentially expressed (**Supplementary Data 7**) and that exhibit distinct segment-specific expression in prior studies (**Supplementary Data 5**).



Supplementary Figure 9. Distinct DCT, CNT and CD expression profiles. **a.** Schematic of the kidney nephron and UMAP as shown in **Fig. 1b** showing relevant cell populations. **b.** UMAP plots as in **Fig. 1b** showing relative expression levels (scaled from low - gray to high - blue) of genes associated with DCT and CNT function. **c.** UMAP plots as in **Fig. 1b** showing high expressing cells for the DCT marker *SLC12A3* (red) or the CNT marker *SLC8A1* (green), or both (purple). **d.** Heatmap of expression of marker genes defining IC and PC clusters using criteria indicated and for differentially expressed genes found in **Supplementary Data 11-12**.

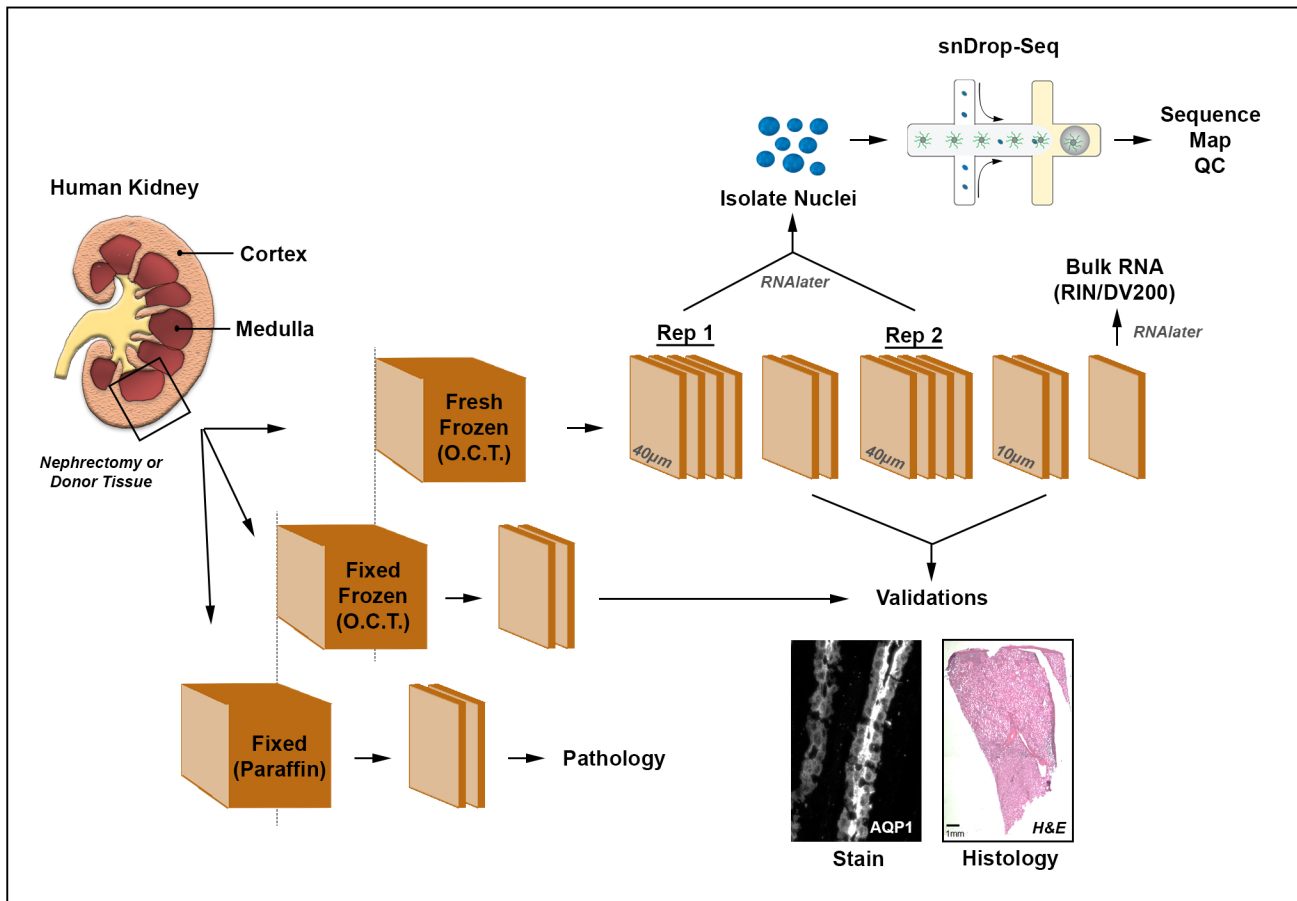


Supplementary Figure 10. Endothelial and interstitial populations. **a.** Schematic of the kidney nephron showing relevant cell populations. **b.** Heatmap of expression of marker genes defining EC and INT populations using criteria indicated and for differentially expressed genes found in **Supplementary Data 13-14**. **c.** UMAP plots as in **Fig. 1b** showing relative expression levels (scaled from low - gray to high - blue) of associated markers in indicated cell types and corresponding protein immunostainings (Human Protein Atlas¹, **Supplementary Data 16**). Arrows indicate representative protein localization. Scale bar indicates 25 μm .



Supplementary Figure 11. Endothelial and interstitial populations. **a.** UMAP plots as in **Fig. 1b** showing relative expression levels (scaled from low - gray to high - blue) of associated markers in indicated cell types and corresponding protein immunostainings (Human Protein Atlas¹, **Supplementary Data 16**). Arrows indicate representative protein localization. Scale bar indicates 25 μ m. **b.** UMAP plots as in **Fig. 1b** indicating high expressing cells for *PIEZO2* (green) and *PODXL* or *AQP1* (red). Cells expressing both are indicated in yellow. Protein fluorescent immunostaining (**Supplementary Data 17**) for *PIEZO2* in MC, *AQP1* in GCs and *NPHS1* in POD. Scale bar indicates 20 μ m. **d.** UMAP plots as in **Fig. 1b** showing relative expression levels (scaled from low - gray to high - blue) of *MMRN1* specific to EC-4 (cluster 25). Protein fluorescent immunostaining (**Supplementary Data 17**) for *MMRN1*, vascular marker *PECAM1* (CD31), PC marker *AQP2* and lymphatic marker *D240*. Closed arrows indicate *MMRN1* extra-tubular staining in interstitial vessels, open arrows indicate *MMRN1* intra-tubular staining. Scale bar indicates 25 μ m.

Tissue Processing Pipeline



Supplementary Figure 12. Optimized tissue processing pipeline. Overview of pipeline applied to adult human kidney. Tissue is segmented into three blocks: paraffin fixed for pathological assessment; fixed and O.C.T. embedded/frozen for protein immunostaining validation/spatial registration assays; unfixed and O.C.T. embedded/frozen for single nucleus assays, bulk RNA assessment and validation/spatial registration assays. Frozen tissues permit parallel sections to be used in multiple complementary assays, specifically: 1) Thick sections are transferred to RNAlater for nuclei isolation and snDrop-Seq; 2) adjacent thin sections are used for histological assessment for tissue integrity and composition; 3) parallel thick sections can be used for technical replicates and increased sampling depth; 4) parallel thick section can be transferred to RNAlater for bulk RNA isolation and quality assessment (RNA integrity number or RIN, DV200) and for bulk RNA-seq.

Supplementary Note 1

Kidney Precision Medicine Project Consortium

Theodore Alexandrov¹, Charles Alpers², Chris Anderton³, Paul Appelbaum⁴, Joseph Ardayfio⁵, Tanim Arora⁶, Tarek Ashkar⁷, Mark Aulizio⁸, Evren Azeloglu⁹, Olivia Balderes⁴, Ulysses Balis¹⁰, Jonathan Barasch⁴, Laura Barisoni¹¹, Daria Barwinska⁷, Jack Bebiak⁷, Kristina Blank², Andrew Bombach⁴, Mary Bray², Keith Brown¹², Will Bush¹³, Taneisha Campbell¹⁰, Catherine Campbell¹⁴, Leslie Cooperman¹³, Dana Crawford¹³, Vivette D'agati⁴, Pierre Dagher⁷, Ian De Boer², Ashveena Dighe², Dejan Dobi¹⁵, Kenneth Dunn⁷, Michael Eadon⁷, Michele Elder¹⁶, Michael Ferkowicz⁷, Malia Fullerton², Yury Goltsev¹⁷, Nir Hacohen¹⁸, Daniel Hall¹⁶, Habib Hamidi¹⁰, Lynda Hayashi¹², Cijang (John) He⁹, Oliver He¹⁰, Susan Hedayati¹⁴, Leal Herlitz¹³, Jonathan Himmelfarb², Jeffrey Hodgins¹⁰, Paul Hoover¹⁸, Ravi Iyengar⁹, Nichole Jefferson², Maria Joanes¹⁵, John Kellum¹⁶, Katherine Kelly⁷, Asra Kermani¹⁴, Krzysztof Kiryluk⁴, Richard Knight¹⁹, Robert Koewler²⁰, Matthias Kretzler¹⁰, Mary Kruth¹⁶, Zoltan Laszik¹⁵, Stewart Lecker²², Simon Lee¹⁴, Chrysta Lienczewski¹⁰, Christopher Lu¹⁴, Randy Luciano⁶, Laura Mariani¹⁰, Robyn McClelland², Gearoid McMahon²², Karla Mehl⁴, Steven Menez²³, Raji Menon¹⁰, Tyler Miller¹⁴, Orson Moe¹⁴, Dennis Moledina⁶, Sean Mooney², Raghav Murugan¹⁶, Garry Nolan¹⁷, George (Holt) Oliver²⁴, John O'toole¹³, Edgar Otto¹⁰, Paul Palevsky¹⁶, Chirag Parikh²³, Samir Parikh²⁵, Christopher Park², Harold Park¹⁴, Ljiljana Pasa-Tolic³, Emilio Poggio¹³, Parmjeet Randhawa¹⁶, Glenda Roberts², Sylvia Rosas⁵, Avi Rosenberg⁶, Matthew Rosengart¹⁶, Kamalanathan Sambandam¹⁴, Francisco Sanchez¹⁴, Minnie Sarwal¹⁵, John Saul², Jennifer Schaub¹⁰, Andrew Schroeder¹⁵, Rachel Sealfon²⁶, John Sedor¹³, Ning (Sunny) Shang⁴, Stuart Shankland², Kumar Sharma²⁷, Anna Shpigel²³, Tara Sigdel¹⁵, Sunny Sims-Lucas¹⁶, Rebecca Steck¹⁰, Mary Stefanick¹⁶, Isaac Stillman²², Stacy Stull¹⁶, Edith Christine Stutzke¹², Swastika Sur¹⁵, Timothy Sutton⁷, Jose Torrealba¹⁴, Robert Toto¹⁴, Olga Troyanskaya²⁶, Mitchell Tublin¹⁶, Katherine Tuttle¹², Ugwuowo Ugochukwu⁶, Miguel Vasquez¹⁴, Dusan Velickovic³, Pam Villalobos²³, Sus Waikar²², Nancy Wang¹⁴, Astrid Weins²², Chenhua Weng⁴, Mark Williams⁵, Kayleen Williams², Francis Perry Wilson⁶, Seth Winfree⁷, Lorenda Wright²⁷, Guanshi Zhang²⁷, Blue B. Lake²⁸, Masato Hoshi²⁰, Diane Salamon²⁰, Amanda Knoten²⁰, Anitha Vijayan²⁰, Joseph Gaut²⁰, Kun Zhang²⁸, Sanjay Jain²⁰

¹European Molecular Biology Laboratory, Germany

²University of Washington, WA

³Pacific Northwest National Laboratories, WA

⁴Columbia University, NY

⁵Joslin Diabetes Center, MA

⁶Yale University, CT

⁷Indiana University, IN

⁸Case Western Reserve, OH

⁹Mount Sinai, NY

¹⁰University of Michigan, MI

¹¹Duke University, NC

¹²Providence Health, WA

¹³Cleveland Clinic, OH

¹⁴UT Southwestern, TX

¹⁵UC San Francisco, CA

¹⁶University of Pittsburgh, PA

¹⁷Stanford University, CA

¹⁸Broad Institute, MA

¹⁹American Association of Kidney Patients, FL

²⁰Washington University St. Louis, MO

²¹Beth Israel Deaconess, MA

²²Brigham & Women's Hospital, MA

²³John Hopkins University, MD

²⁴Parkland Hospital, TX

²⁵Ohio State University

²⁶Princeton University, NJ

²⁷UT Health San Antonio, TX

²⁸UC San Diego, CA

REFERENCES

1. Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).