Supplemental Information

Advantages of single nucleus over single cell RNA-seq in adult kidney

Haojia Wu[1], Yuhei Kirita[1], Erinn L. Donnelly[1] and Benjamin D. Humphreys[1,3]

Supplemental Table of Contents

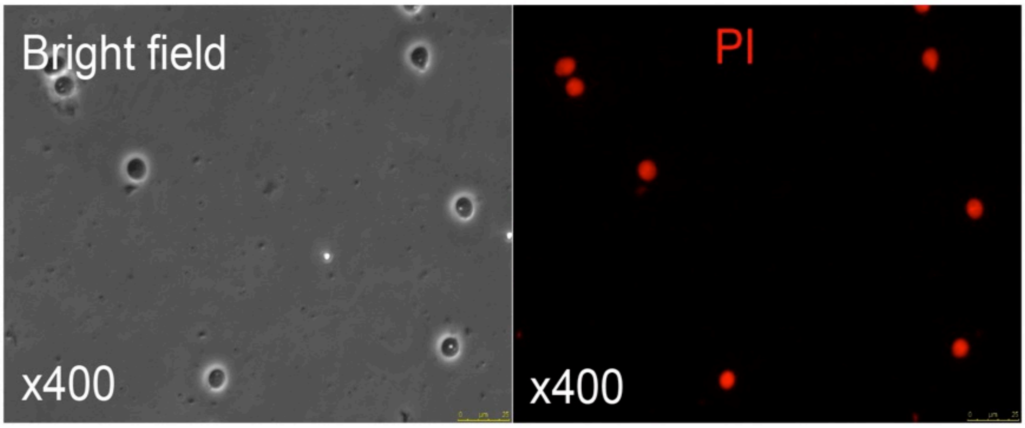

Figure S1. Nuclear preparation from adult mouse kidney. Brightfield (left) and propidium iodide (PI) stained (right) nuclei after preparation from adult mouse kidney. Note that nuclei remain intact.

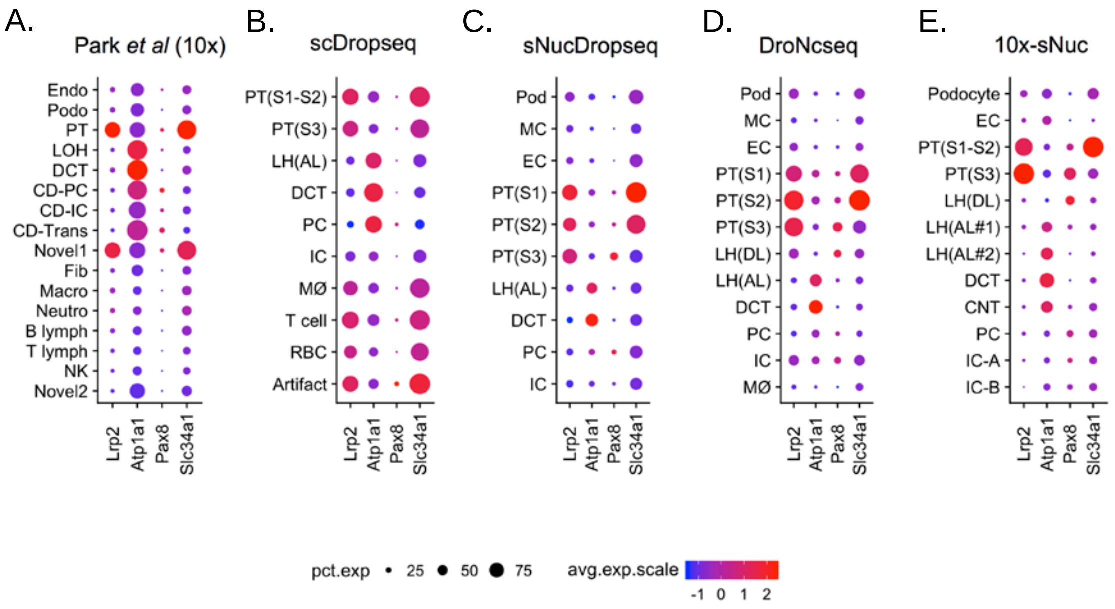

Figure S2. Comparison of tubular mRNA contamination across platforms. A. Reanalysis of Park et al. dataset showing the degree of contamination of four highly expressed tubular genes across all cell types. **B – E.** The same analysis applied to the datasets described in this manuscript. In general, there is somewhat less contamination with snRNA-seq, but some degree of tubular gene expression can be detected in all cell clusters regardless of dissociation or platform.

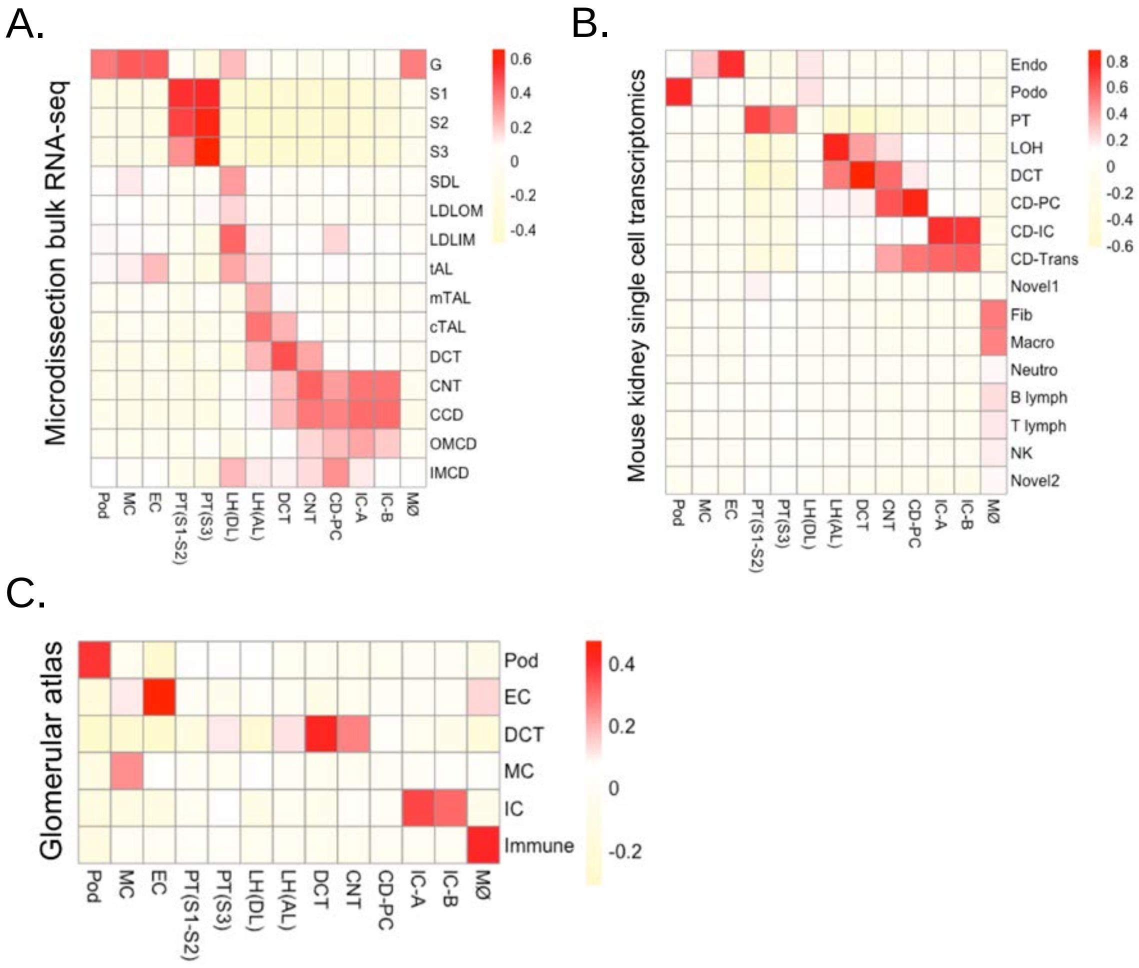

Figure S3. Correlation of combined single cell and single nucleus RNA-seq clusters with microdissected tubule segment RNA-seq, mouse kidney single cell atlas and mouse glomerular single cell atlas. A. The horizontal axis lists the 13 annotated clusters from our combined dataset. The vertical axis shows the Pearson's correlation of these single cell types to bulk RNA-seq of microdissected tubule segments.[1] G, glomerulus; SDL, short descending limb; LDLOM, long descending limb, outer medulla; LDLIM, long descending limb, inner medulla; tAL, thick ascending limb;, mTAL, medullary thick ascending limb; cTAL, cortical thick ascending limb; DCT, distal convoluted tubule, CNT, connecting tubule; CCD, cortical collecting duct; OMCD, outer medullary collecting duct; IMCD, inner medullary collecting duct. **B.** The horizontal axis lists the 13 annotated clusters from our combined dataset. The vertical axis shows the Pearson's correlation of these single cell types to a recently published adult mouse kidney single cell atlas.[2] **C.** The horizontal axis lists the 13 annotated clusters from our combined dataset. The vertical axis shows the Pearson's correlation of these single cell types to a recently published adult mouse glomerulus single cell atlas.[3]

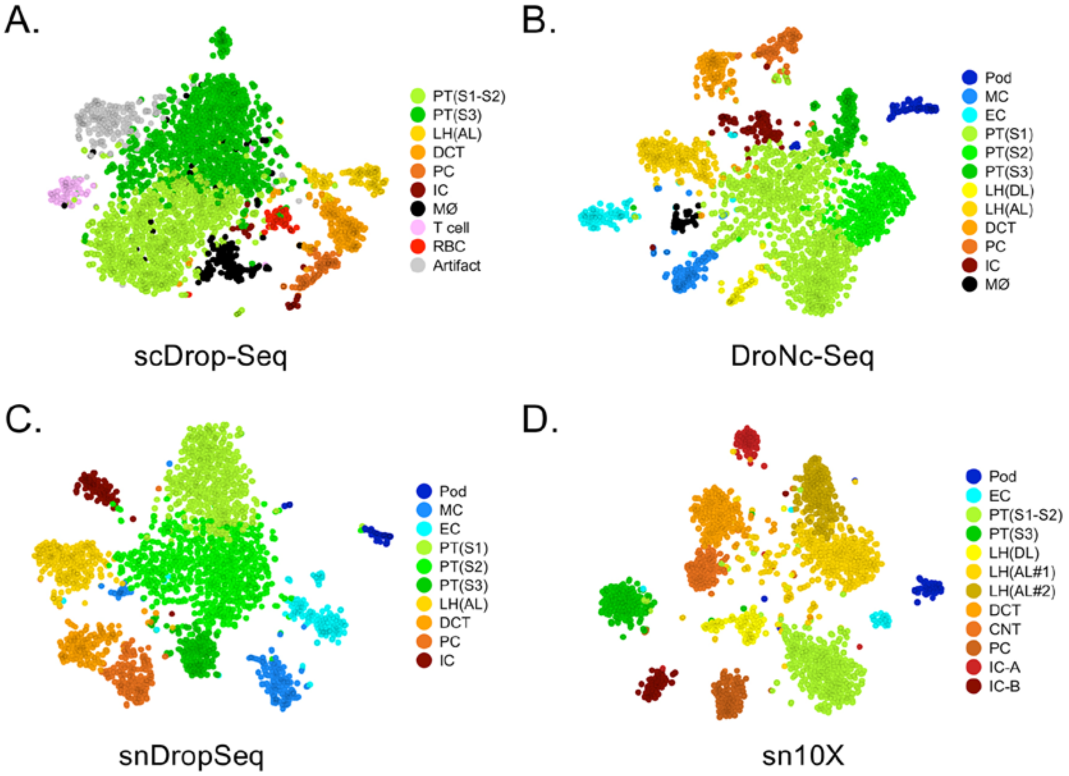

Figure S4. Clustering of all four datasets by tSNE. A. scDropSeq identified ten cell clusters, but one of these is an artifactual cluster arising from cell dissociation, and one of them is a red blood cell cluster. No podocytes or endothelial cells could be detected. **B.** DroNc-Seq identified 12 independent cell clusters, including podocytes, mesangial cells and endothelial cells. **C.** snDropSeq identified ten clusters including podocytes, mesangial cells and endothelial cells. **D.** sn10X identified 12 separate cell clusters including podocytes and endothelial cells and both type A and type B intercalated cells.

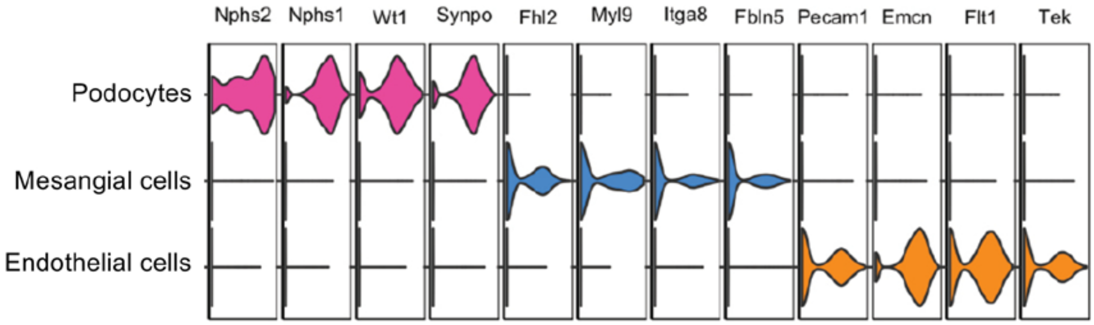

Figure S5. Violin plot showing cell specific markers in glomerular clusters from the combined single cell and single nucleus dataset.

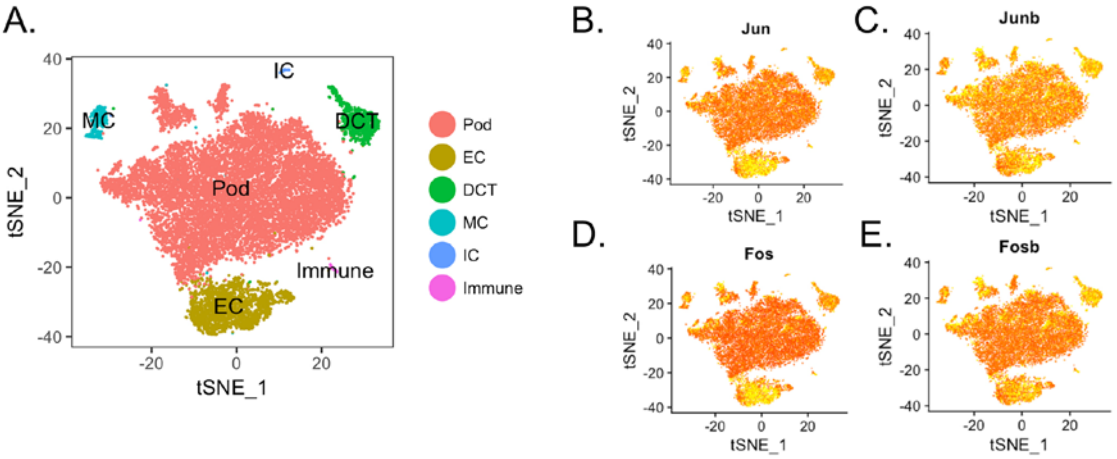

Figure S6. Immediate early gene expression in the mouse glomerular cell atlas generated by scDropSeq. A. 14,382 single glomerulus cell transcriptomes from a recently published glomerular atlas[3] were reclustered using Seurat, reproducing the published cell clusters. **B-D.** Substantial stress response gene expression could be detected in all six clusters. These immediate early genes are known to be induced by proteolytic digestion of tissue at 37 °C.[4]

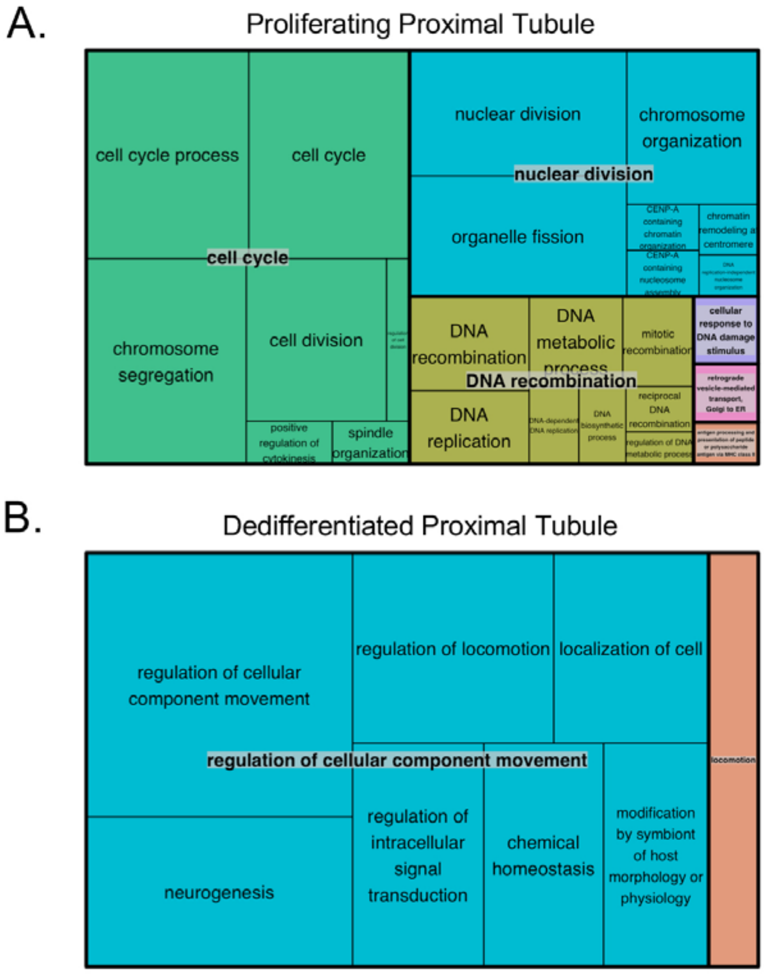

Figure S7. Gene ontology terms for differentially expressed genes between the proliferating and dedifferentiated proximal tubule clusters. A. Cell cycle, cell division and DNA replication terms characterize the proliferating proximal tubule cluster. **B.** Terms related to cell movement and locomotion characterize the dedifferentiated proximal tubule cluster.

Supplementary Methods

Computational data analysis

1. Preprocessing of Dropseq, sNucDropSeq, DroNcSeq and sn10x data

We used a newly developed pipeline, zUMIs[5], to process the single cell and single nucleus sequencing data from mouse kidney. In brief, we first filtered out the low-quality barcodes or UMIs based on sequence with the internal read filtering algorithm built in zUMIs. We then used zUMIs to map the filtered reads to mouse reference genome (mm10) using STAR 2.5.3a (two-pass mapping mode). Next, zUMIs quantified the reads that were uniquely mapped to exonic, intronic or intergenic region of the genome and inferred the true barcodes that mark cells/nuclei by fitting a k-dimensional multivariate normal distribution with mclust package. Finally, a UMI count table utilizing both exonic and intronic reads were generated for downstream analysis. The whole data processing was executed by running the script on a HPC cluster with 96×2.3GHz computing cores (http://brc.wustl.edu/?page_id=12). We summarized the mapping and count statistics from the output files generated by zUMIs and visualized the data in various formats such as bar chart, box plot, dot plot and trend line using *ggplot2* R package.

2. Unsupervised clustering of the single cell/nucleus RNA-seq datasets and cell type annotation

Seurat was used for quality control, dimensionality reduction and cell clustering for the datasets generated from all platforms. In brief, raw UMI count matrix from each platform was loaded separately into the Seurat. For normalization, the UMI count matrix was scaled by total UMI counts, multiplied by 10,000 and transformed to log space. Only genes found to be expressing in >3 cells were retained. Cells with a relatively high percentage of UMIs mapped to mitochondrial genes (>=0.5) were discarded. Moreover, we only kept cells that had more than 300 genes detected to remove the low quality cells or nuclei. Before clustering, variants arising from library size and percentage of mitochondrial and ribosomal genes were regressed out by specifying the *vars.to.regress* argument in Seurat function *ScaleData*. The highly variable genes were identified using the function *FindVariableGenes*. The expression level of highly variable genes in the cells was scaled and centered along each gene, and was conducted to principal component analysis. We then assessed the number of PCs to be included in downstream analysis by (1) plotting the cumulative standard deviations accounted for each PC using the function *PCElbowPlot* in Seurat to identify the ‘knee’ point at a PC number after which successive PCs explain diminishing degrees of variance, and (2) by exploring primary sources of heterogeneity in the datasets using the *PCHeatmap* function in Seurat. Based on these two methods, we selected first top significant PCs for two-dimensional t-distributed stochastic neighbor embedding (tSNE), implemented by the Seurat software with the default parameters. We used *FindCluster* in Seurat to identify cell clusters for each protocol. To identify the marker genes, differential expression analysis was performed by the function *FindAllMarkers* in Seurat with Wilcoxon Rank Sum test. Differentially expressed genes that were expressed at least in 25% cells within the cluster and with a fold change more than 0.25 (log scale) were considered marker genes.

We annotated the cell clusters by two approaches. First, we inspected the top differential genes from each cluster and labeled the cluster with cell type names based on the expression of putative cell type markers (e.g. Nphs1 for podocyte, Emcn for endothelial cells, Slc34a1 for proximal tubular cells, etc). Second, we computed pairwise Pearson correlation between each pair of cell clusters identified by our dataset and the annotated cell types from the published datasets[1-3]. Third, we cross-validated the cell types by computing the Pearson correlation between the cell clusters identified from the single cell and single nucleus datasets used in this study to ensure that the cell type annotations were consistent within this study.

3. Integrated analysis of sCellDropseq, sNucDropseq, DroNcseq and sNuc-10x datasets

To compare the cell types derived from different techniques, we performed comparative analysis on multiple datasets by utilizing a recently developed computational strategy for integrated analysis (implemented in Seurat v2.0)[6]. We first selected the union of the top 3,000 genes with the highest dispersion from all datasets for a canonical correlation analysis (CCA) to identify common sources of variation across the datasets. Then CCA was performed based on the normalized expression value of the highly-dispersed genes. Next, we selected the top dimensions of the CCA by examining a saturation in the relationship between the number of principle components and the percentage of the variance explained using the *MetageneBicorPlot* function in Seurat. We obtained a new dimensional reduction matrix by aligning the CCA subspaces with the top dimensions computed above. With the new dimensional reduction matrix, we performed clustering analysis on the cells or nuclei from different datasets by setting an optimal clustering parameters. We visualized the cells by their original identity or by their cluster identity classified by this integrated analysis. Differential gene analysis was performed on the cells or nuclei from different datasets but grouped in the same cluster after the alignment analysis. Differential genes were visualized using the *FeatureHeatmap* or *DotPlot* function in Seurat.

4. Comparative analysis of tubular single cell and single nucleus transcriptomes

To compare the tubular single cell and single nucleus transcriptomes, we applied a similar approach reported by Bakken TE *et al*[7] to match each nucleus from the sNucDropseq dataset to the most similar cell from the single cell Dropseq dataset based on the maximum correlated expression of all genes weighted for gene dropouts. First, we extracted 1,469 nuclei from the clusters identified as PT (S1, S2 and S3 segments), LH, DCT and PC based on the unsupervised clustering analysis on the sNucDropseq dataset. Our analysis only included the tubular segments as those cell populations can be detected by both single nucleus and single cell datasets. We then estimated the gene dropout probabilities for the selected single nuclei and all single cells from the sCellDropseq dataset (3,531 cells) following the tutorial of the R package *scde* (http://hms-dbmi.github.io/scde/diffexp.html). In brief, expression noise models were fit separately to single nuclei and cells using the *knn.error.models* function with default settings. A mode-relative weighting approach was used to generate a dropout weight matrix where the probability of dropout event was estimated using *scde.failure.probability* and *scde.posteriors* functions. Dropout weighted Pearson correlations between all pairs of nuclei and cells were calculated using the *cov.wt* function from the *stats* R package. A cell was selected if it had highest correlation with any nucleus, and this matched pair of cell and nucleus was removed from the next round of cell selection. This process was repeated until 1,469 best matching cells were selected, and the expression correlations were compared to correlations of the best matching pairs of nuclei.

We next employed two strategies to assess the robustness of cell clustering on the matched cells and nuclei. First, we performed graph-based clustering separately on the single cells and single nuclei (Jaccard- Louvain algorithm) using the Seurat package. To assess the clustering outcome with UMI count matrices generated by exonic read counts and by exonic + intronic read counts, we used the same parameters in both datasets to select highly variable genes for PCA, same number of significant PCA and resolution for clustering and same parameter for tSNE dimensional reduction. Clusters were annotated with examining the expression of cell type specific markers. While this approach provides an intuitive way to visually inspect the clustering outcome (e.g. based on the number of clusters produced in each condition, or how well each cluster was separated with or without including the intronic reads, etc), this graph-based approach can hardly be used for quantitative assessment of the clustering. We therefore followed the methods established by Bakken TE *et al*[7] to compute the cluster cohesion and separation and used these parameters to quantify the clustering. Briefly, we used an iterative clustering pipeline *scrattch.hicat* developed by Allen Institute (https://github.com/AllenInstitute/scrattch.hicat) to perform clustering separately on the matched single cell and single nucleus datasets. The pipeline consists of five steps: 1) selection of HVGs, 2) dimensionality reduction, 3) dimension filtering, 4) hierarchical clustering and 5) cluster merging based on differential genes. This process was iteratively repeated within each resulting cluster until one of the following criteria was met: 1) no more clusters met the differential gene expression and 2) the minimum cluster size threshold was met. The robustness of the clustering was assessed by repeating the clustering procedure 100 times (i.e. n_iter=100) on 80% of randomly subsampled cells. A co-clustering matrix was generated that represented the proportion of clustering iterations that each pair of samples were assigned to the same cluster. This bootstrapped interactive process was wrapped into the function *run_consensus_clust* from the *scrattch.hicat* package. We computed the cluster cohesion (average within cluster co-clustering) and separation (difference between within cluster co-clustering and maximum between cluster co-clustering) for all clusters based on the consensus clustering matrix generated by the above clustering pipeline. Data was visualized by *ggplot2* R package.

To compare the tubular cell transcriptomes profiled by sc- and sn-RNA-seq, we estimated the proportion of cells and nuclei expressing each detected gene. We followed the similar randomly-splitting approach developed by Bakken TE *et al*[7] to estimate the expected variability of gene detection as a result of population sampling. Data were summarized with a hexagonal binned scatter plot and a color code representing the number of genes using the R package *ggplot2*. We performed differential expression between nuclei and cells using the function *WilcoxDETest* from *Seurat* R package. Data was visualized as volcano plot using *ggplot2*. Single cell or single nucleus enriched genes were manually selected from the top DE genes list and were presented as violin plot using *ggplot2* package. GO analysis was performed on the single cell-enriched and single nucleus-enriched genes using the ToppGene Suite (https://toppgene.cchmc.org). Significant enriched GO terms (defined by Benjamini-Hochberg corrected P-value <0.05) were summarized by REVIGO[8] and visualized by *treemap* R package.

5. Comparative analysis of glomerular single cell and single nucleus transcriptomes

As our single cell Dropseq dataset failed to resolve glomerular cell populations, we reanalyzed a recently published dataset from single cell profiling of mouse glomerulus

(GSE111107, Karaiskos N *et al*[3]). Using the same clustering approach described by the authors, we reproduced all the cell types reported in the original paper, including podocytes, mesangial cells, endothelial cells, immune cells and two tubular cell types. To compare the glomerular cell types identified by our single nucleus RNA-seq techniques (sNucDropseq and DroNcSeq) to those reported by the glomerulus single cell study, we trained a multiclass random forest classifier [9, 10] on the clusters from our single nucleus data and used it to map the single cell data. First, we composed a 'training set' by sampling 60% of the cells from 3 glomerular cell types representing podocyte, mesangial cells and endothelial cells. We next trained a random forest using 1,000 trees on the training set using the R package *randomForest*. We then used the remaining 40% of the cells from each cluster from our single nuclei dataset to validate the performance of the trained classifier. We used this model to assign a class label (one of the 3 glomerular cell types) to each cell from glomerular cell atlas dataset. Finally, we selected the same number of cells based on the number of nuclei in each glomerular subtype that have highest number of votes to each class label. We performed CCA integrated analysis on the selected glomerular cells and nuclei with similar parameters described above. Cell identities were annotated based on the cell type-specific marker expression in each cluster. Cells were color coded based on the unsupervised clustering or the origin of dataset in the tSNE graph. Differential gene analysis was performed on nuclei and cells to identify the enriched genes for each technique using Wilcoxon Rank Sum test. Expression of enriched genesets were visualized by violin plot and heatmap using ggplot2 and pheatmap R packages, respectively.

We used MetaNeighbor[11] to assess the replicability of the glomerular cell types identified from our single nuclei datasets. We ran MetaNeighbor using an unsupervised mode (https://github.com/maggiecrow/MetaNeighbor). We first identified the variable genes using the function *variableGenes* from *MetaNeighbor* R package. We then used the cell type information annotated from the clustering analysis mentioned above to label the cells and computed area under the receiver operator characteristic curve (AUROC) using the function *run_MetaNeighbor_US.* AUROC was plotted as heatmap using *heatmap.2* function in *gplots* R package.

6. Cell cycle analysis

We assigned a cell cycle score (from -1 to 1) on each cell according to its gene expression of G2/M and S phase markers[12] using the *CellCycleScoring* function in *Seurat*. We assigned each cell with a cell cycle phase based on the following criteria: 1) If S_{score} (S phase score) > 0.15 and S_{score} > $G2M_{score}$ (G2M phase score), S phase; 2) $G2M_{score}$ > 0.15 and $G2M_{score}$ > S_{score}, G2/M phase; 3) If S_{score} <0.15 and $G2M_{score}$ < 0.15, G1 phase. The cells at different cell cycle classifications were visualized in the tSNE map.

7. Ligand-receptor interaction analysis

To study ligand-receptor interactions across the cell types identified from the UUO kidney single nucleus dataset, we used a human ligand–receptor list comprising 2,557 ligand–receptor pairs curated by the Database of Ligand−Receptor Partners (DLRP), IUPHAR and Human Plasma Membrane Receptome (HPMR)[13, 14]. We selected the receptors that were only differentially expressed in each cell type from the UUO dataset. To determine the ligand-receptor pairs to plot on the heatmap, we required (i) the ligands and receptors are uniquely expressed in each cell type (q-val<0.05 and logFC>0.5); (ii) Each receptor should have at least one corresponding ligand to pair with. We used heatmap.2 function from gplots package to visualize the ligand- receptor pairs in each cell type.

Supplemental References

1. Lee, JW, Chou, CL, Knepper, MA: Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment-Specific Transcriptomes. *J Am Soc Nephrol,* 26**:** 2669-2677, 2015.
2. Park, J, Shrestha, R, Qiu, C, Kondo, A, Huang, S, Werth, M, Li, M, Barasch, J, Susztak, K: Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science,* 360**:** 758-763, 2018.
3. Karaiskos, N, Rahmatollahi, M, Boltengagen, A, Liu, H, Hoehne, M, Rinschen, M, Schermer, B, Benzing, T, Rajewsky, N, Kocks, C, Kann, M, Muller, RU: A Single-Cell Transcriptome Atlas of the Mouse Glomerulus. *J Am Soc Nephrol,* 29**:** 2060-2068, 2018.
4. Adam, M, Potter, AS, Potter, SS: Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. *Development,* 144**:** 3625-3632, 2017.
5. Parekh, S, Ziegenhain, C, Vieth, B, Enard, W, Hellmann, I: zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience,* 7, 2018.
6. Butler, A, Hoffman, P, Smibert, P, Papalexi, E, Satija, R: Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol,* 36**:** 411-420, 2018.
7. Bakken, TE, Hodge, RD, Miller, JM, Yao, Z, Nguyen, TN, Aevermann, B, Barkan, E, Bertagnolli, D, Casper, T, Dee, N, Garren, E, Goldy, J, Gray, LT, Kroll, M, Lasken, RS, Lathia, K, Parry, S, Rimorin, C, Scheuermann, RH, Schork, NJ, Shehata, SI, Tieu, M, Phillips, JW, Bernard, A, Smith, KA, Zeng, H, Lein, ES, Tasic, B: Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA-sequencing. *bioRxiv*, 2018.
8. Supek, F, Bosnjak, M, Skunca, N, Smuc, T: REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One,* 6**:** e21800, 2011.
9. Habib, N, Avraham-Davidi, I, Basu, A, Burks, T, Shekhar, K, Hofree, M, Choudhury, SR, Aguet, F, Gelfand, E, Ardlie, K, Weitz, DA, Rozenblatt-Rosen, O, Zhang, F, Regev, A: Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods,* 14**:** 955-958, 2017.
10. Shekhar, K, Lapan, SW, Whitney, IE, Tran, NM, Macosko, EZ, Kowalczyk, M, Adiconis, X, Levin, JZ, Nemesh, J, Goldman, M, McCarroll, SA, Cepko, CL, Regev, A, Sanes, JR: Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell,* 166**:** 1308-1323 e1330, 2016.
11. Crow, M, Paul, A, Ballouz, S, Huang, ZJ, Gillis, J: Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun,* 9**:** 884, 2018.
12. Tirosh, I, Izar, B, Prakadan, SM, Wadsworth, MH, 2nd, Treacy, D, Trombetta, JJ, Rotem, A, Rodman, C, Lian, C, Murphy, G, Fallahi-Sichani, M, Dutton-Regester, K, Lin, JR, Cohen, O, Shah, P, Lu, D, Genshaft, AS, Hughes, TK, Ziegler, CG, Kazer, SW, Gaillard, A, Kolb, KE, Villani, AC, Johannessen, CM, Andreev, AY, Van Allen, EM, Bertagnolli, M, Sorger, PK, Sullivan, RJ, Flaherty, KT, Frederick, DT, Jane-Valbuena, J, Yoon, CH, Rozenblatt-Rosen, O, Shalek, AK, Regev, A, Garraway, LA: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science,* 352**:** 189-196, 2016.
13. Ramilowski, JA, Goldberg, T, Harshbarger, J, Kloppmann, E, Lizio, M, Satagopam, VP, Itoh, M, Kawaji, H, Carninci, P, Rost, B, Forrest, AR: A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun,* 6**:** 7866, 2015.
14. Wu, H, Malone, AF, Donnelly, EL, Kirita, Y, Uchimura, K, Ramakrishnan, SM, Gaut, JP, Humphreys, BD: Single-Cell Transcriptomics of a Human Kidney Allograft Biopsy Specimen Defines a Diverse Inflammatory Response. *J Am Soc Nephrol,* 29**:** 2069-2080, 2018.