Supplemental Information

Single Cell Transcriptomics of a Human Kidney Allograft Biopsy Defines a Diverse Inflammatory Response

Haojia Wu*, Andrew F. Malone*, Erinn Donnelly, Yuhei Kirita, Kohei Uchimura, Sai M. Ramakrishnan, Joseph Gaut and Benjamin D. Humphreys

* These authors contributed equally

Inventory of Supplemental Information

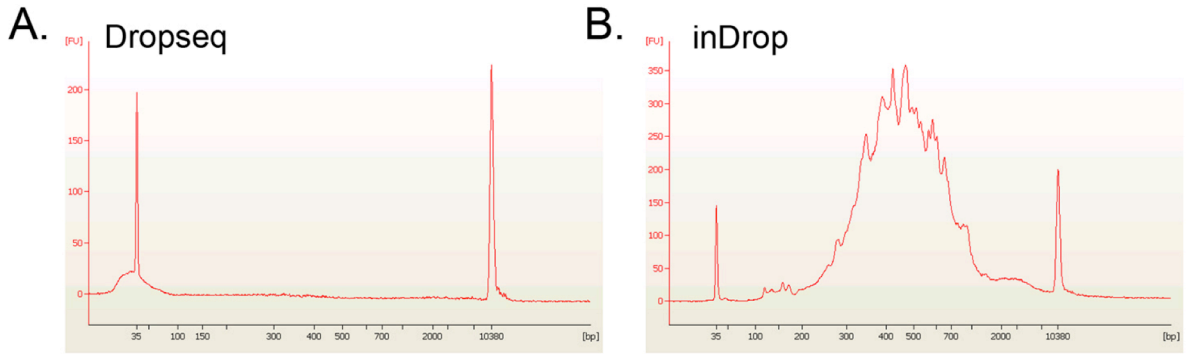


Figure S1. Bioanalyzer traces of library preparations resulting from the two microfluidic-based single cell techniques, Dropseq and inDrops. (A) cDNA trace for a library prepared using the Dropseq method. This method was unsuccessful in making quality libraries from a human biopsy tissue core sample. **(B)** cDNA trace for a library prepared using the inDrops method. Using this method we were able to create a quality cDNA library from a human biopsy tissue core.

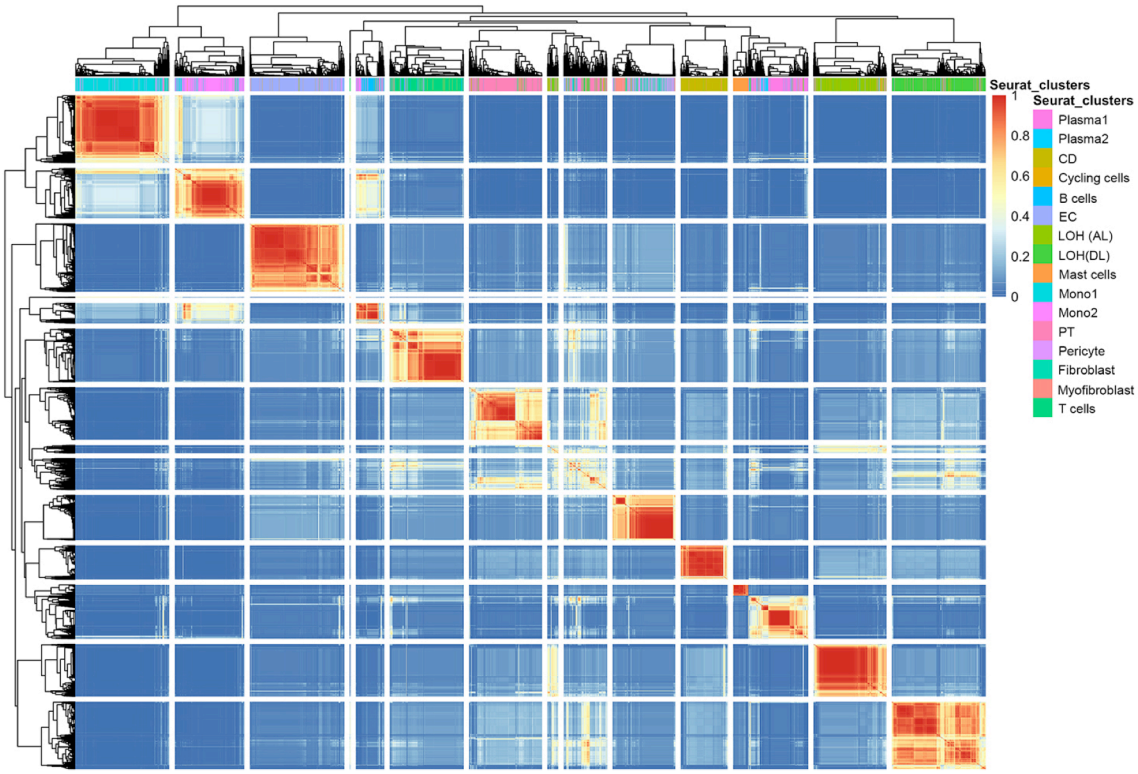


Figure S2. Consensus clustering of our single cell RNA-seq data using SC3. The SC3 method is another unsupervised method used to cluster cells. This method is highly accurate and robust and combines multiple clustering approaches into one. Using SC3 we found that our initial clustering using Seurat was replicated using SC3.

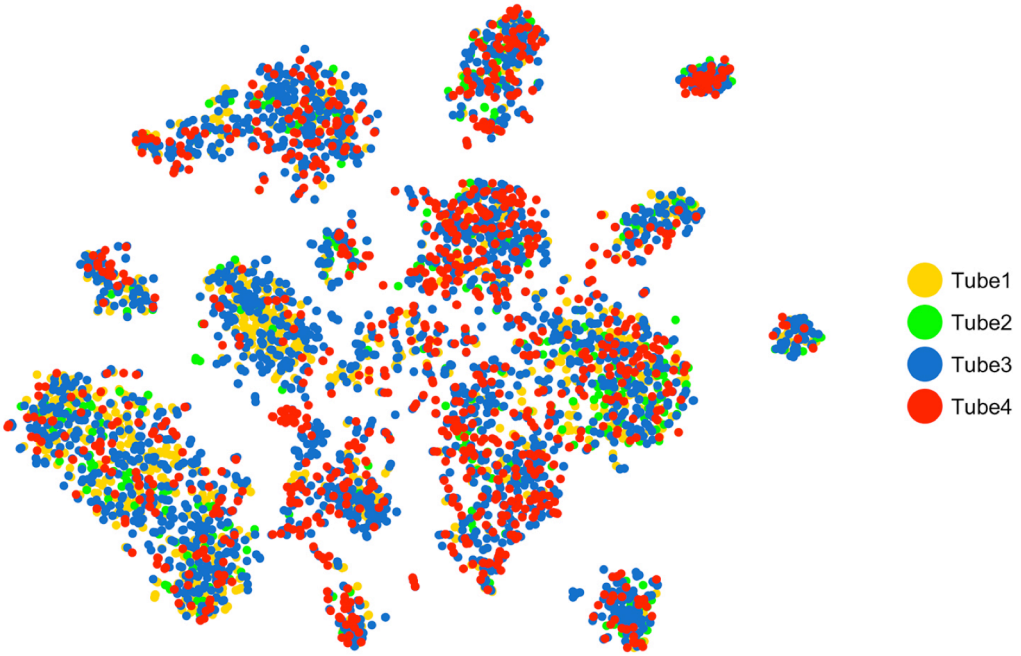


Figure S3. Minimal batch effects observed between sequencing platforms and library preparations. Cells processed using the inDrops method on one biopsy sample were collected sequentially into 4 tubes. Libraries prepared from tubes 1 to 3 were sequenced using the Illumina HiSeq 2500 platform and the library prepared from tube 4 was sequenced using the Illumina NextSeq platform. Projecting cells into one tSNE map shows that all clusters from each tube overlap.

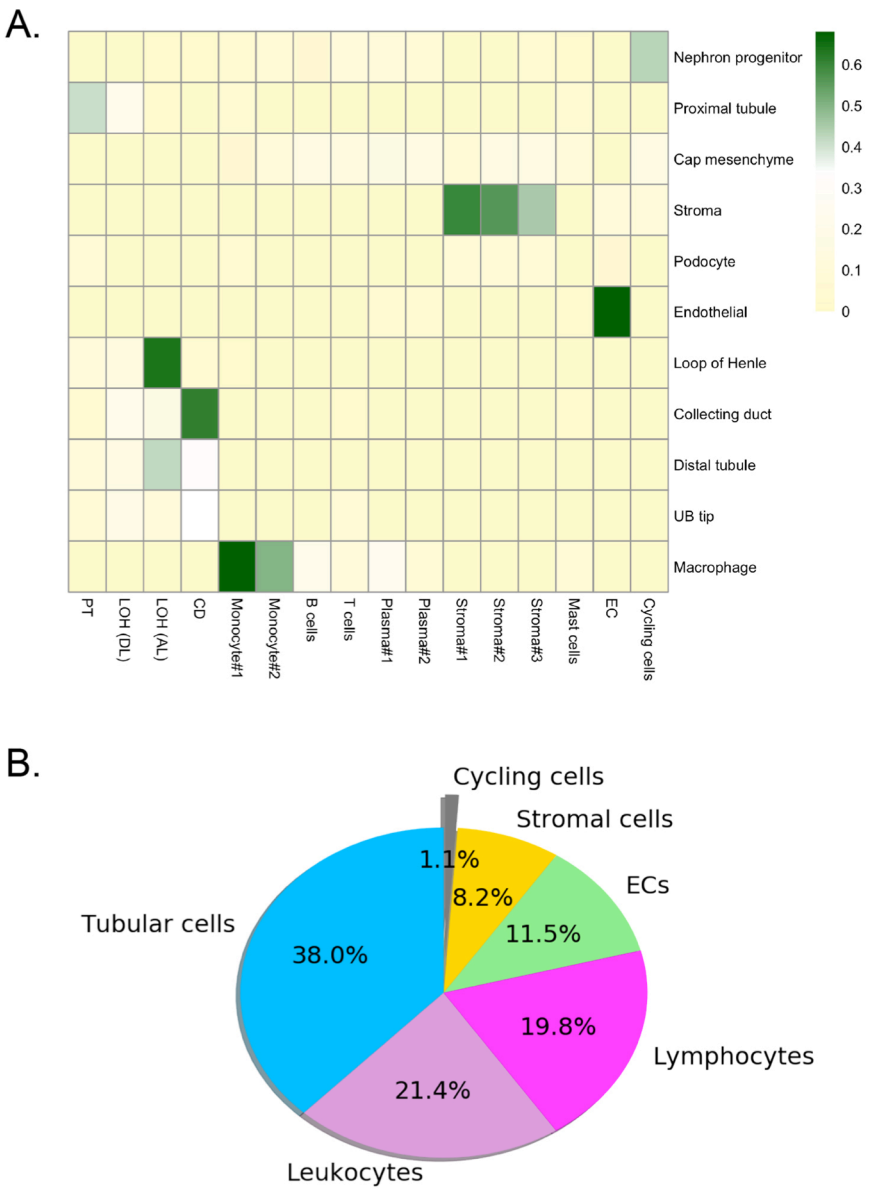


Figure S4. Comparison of cell-type specific gene expression from a P1 mouse kidney and a human PBMC RNAseq dataset. (A) Heatmap comparing cell-type specific gene expression from a P1 mouse kidney RNAseq dataset (y-axis) with our cell-type specific gene expression (x-axis). The results confirm our cluster annotations. Color key denotes the Pearson correlation score. **(B)** Proportion of cell types within our biopsy sample.

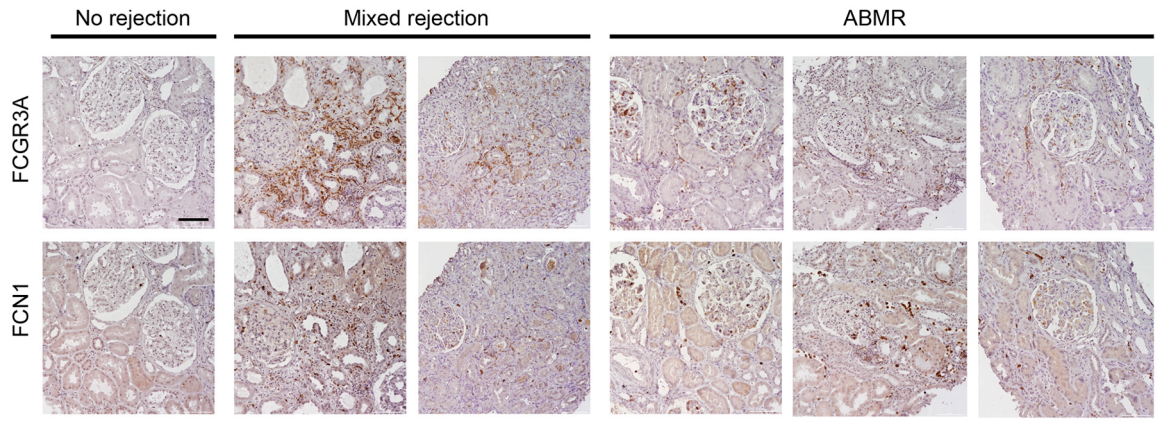


Figure S5. Low power views of independent transplant biopsies stained for monocyte#1 and monocyte#2 subsets. FCGR3A identifies the monocyte#1 subset, and staining was absent from transplant biopsies with a histologic diagnosis of no rejection, but strong staining was present in biopsies with a histologic diagnosis of mixed rejection, and intermediate staining in ABMR biopsies. The monocyte#2 marker FCN1 exhibited a very similar pattern. The top and bottom slides were from serial sections. Scale bar 100 μm.

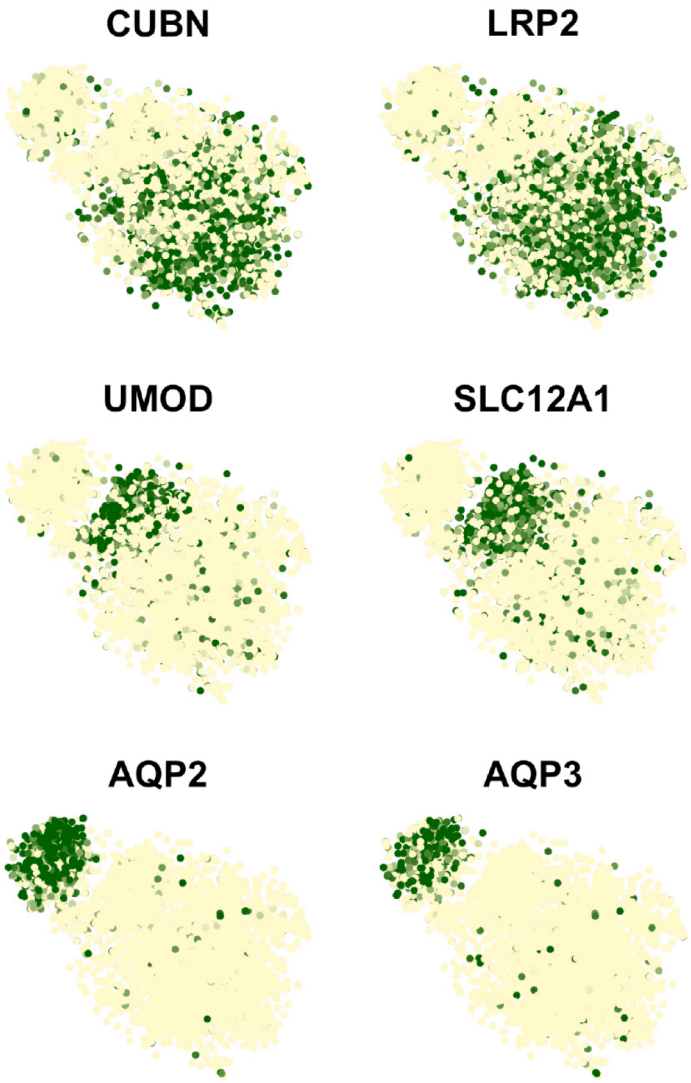


Figure S6. Confirmation of proximal tubule, loop of Henle and collecting duct cell clusters. Clustering analysis of epithelial cell types from the combined biopsy and normal human samples show proximal tubular cell markers CUBN and LRP2 in the proximal tubule cluster, UMOD and SLC12A1 in the loop of Henle cluster and AQP2 and AQP3 in the collecting duct cluster.

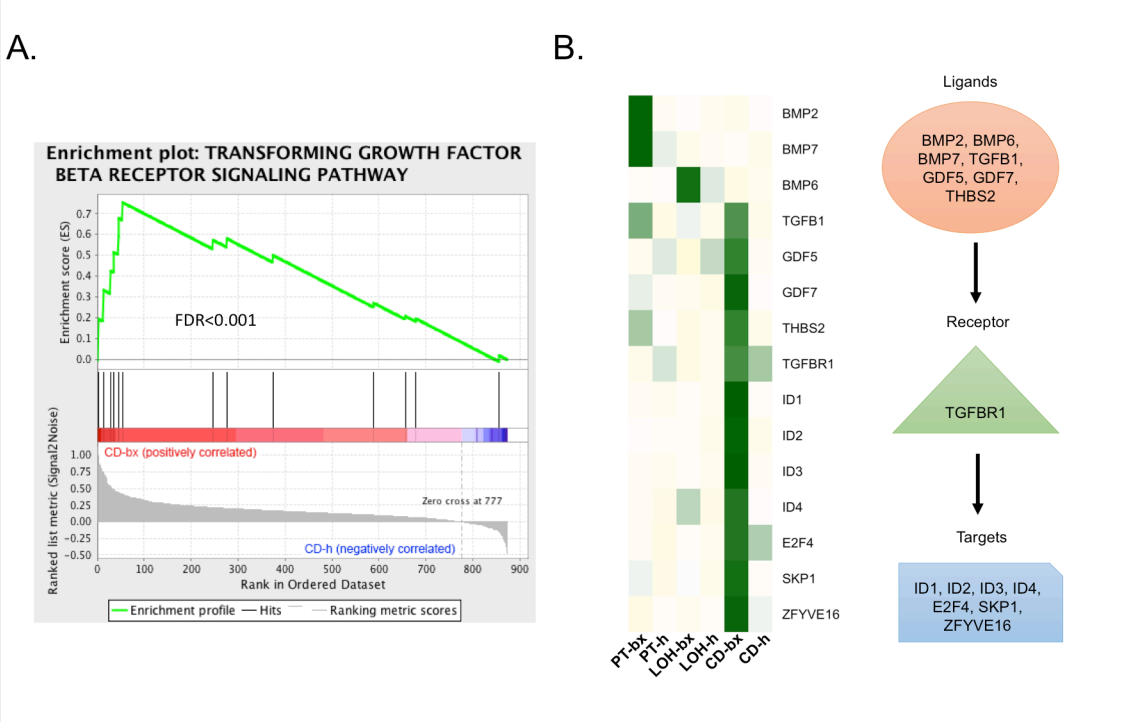


Figure S7. Upregulated TGF-β/BMP signaling signature in collecting duct cells from biopsy. (**A**) GSEA enrichment plot of the TGF-β receptor signaling pathway, one of the top pathways upregulated in biopsy CD when compared with the healthy CD. (**B**) Expression of selected TGF-β/BMP pathway genes were mapped to healthy vs. biopsy epithelial cell types. Unexpectedly, a very strong TGF-β/BMP signature was observed in collecting duct epithelia from the biopsy kidney sample (CD-bx) compared to healthy collecting duct (CD-h). Receptor-ligand pairs were identified from human the Database of Ligand−Receptor Partners (DLRP), IUPHAR and Human Plasma Membrane Receptome (HPMR).

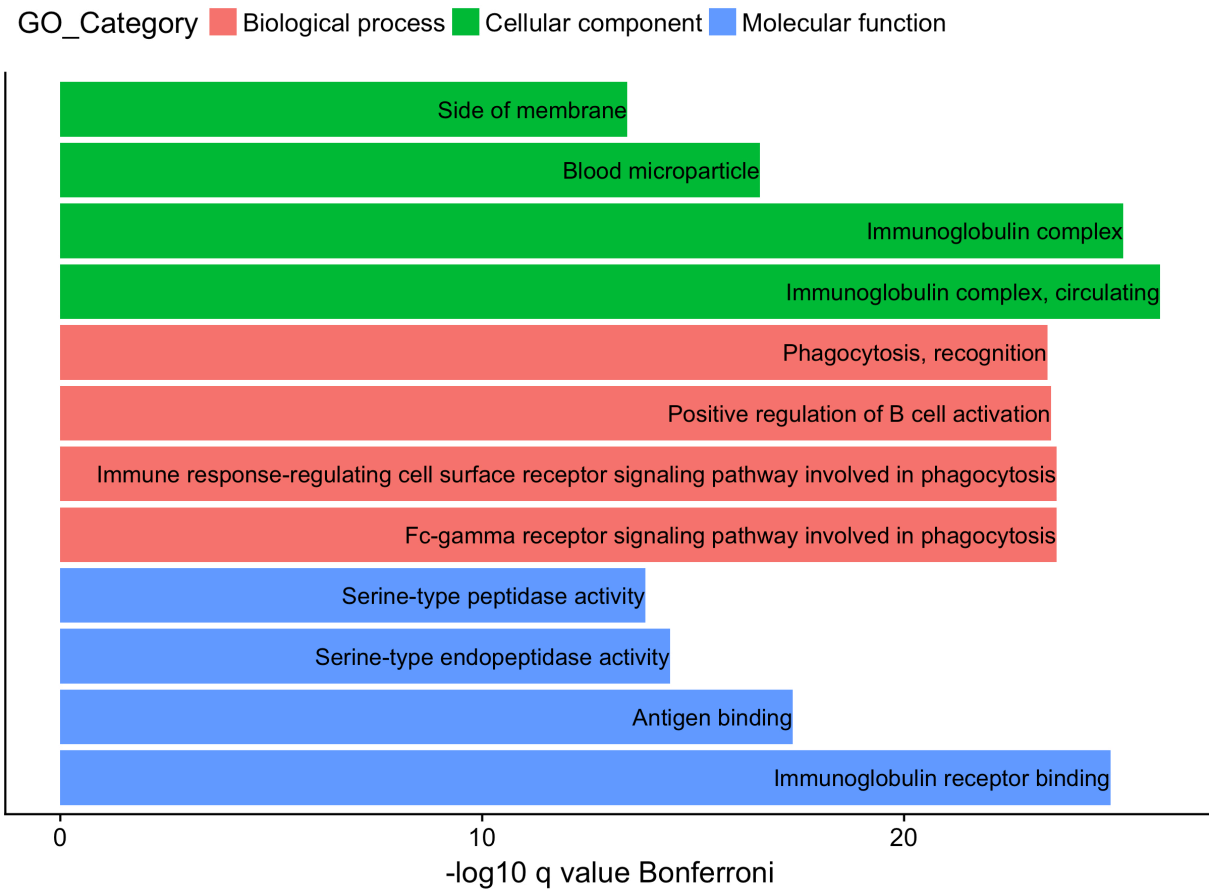
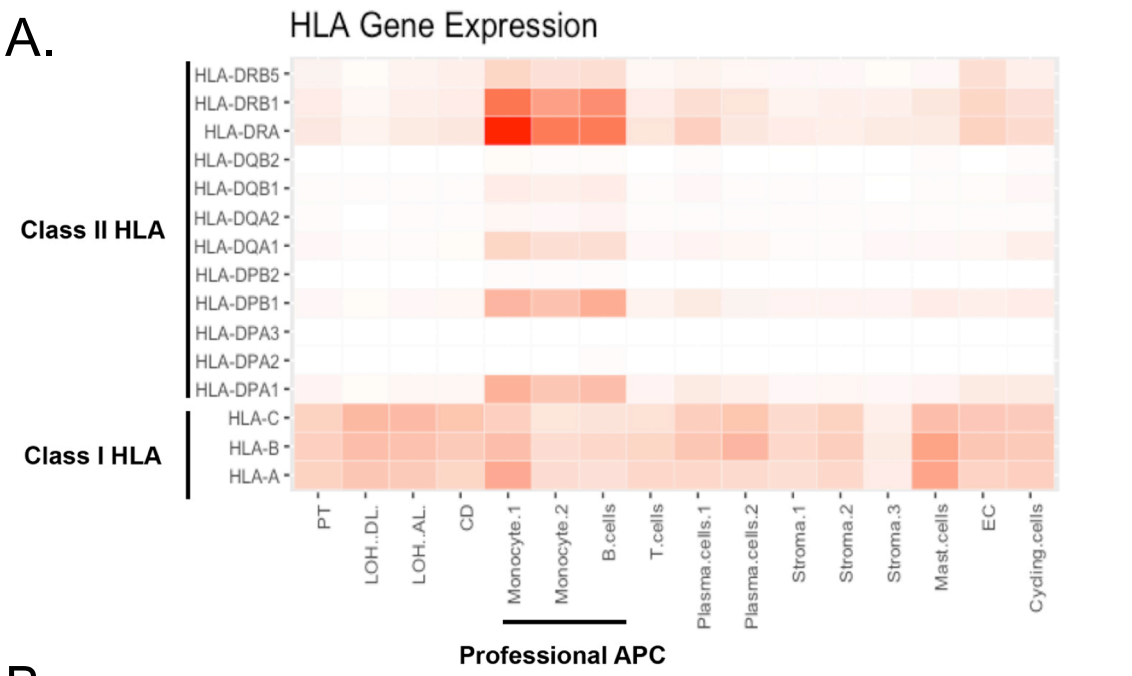


Figure S8. GO analysis on the EC subgroup that expressed immunoglobulins.
The results were obtained from the ToppGene suite using differentially expressed genes in the activated EC subcluster (p value<0.01) as the input gene list. Note the enrichment for immunoglobulin binding and phagocytosis, suggesting antibody-mediated activation of endothelial cells in this biopsy with antibody-mediated rejection.



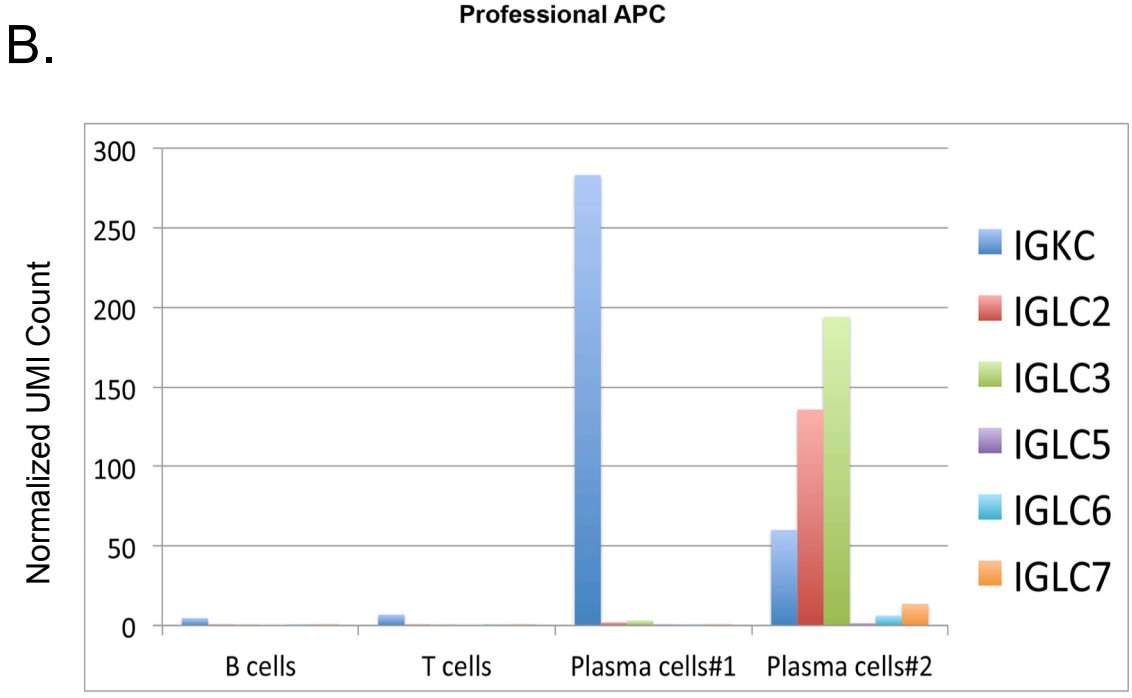


Figure S9. Distribution of HLA and light chain transcript expression across cell types. **(A)** The class I HLA transcripts are expressed equally across all cell types. The Class II HLA transcripts are predominantly expressed in the professional antigen presenting cells (APCs), monocyte cluster 1 and 2 and B cells. The class II transcripts corresponding to the donor specific antibodies circulating at the highest concentration (anti-HLA-DRB1/DRA and anti-HLA-DRB5/DRA) are expressed highly in the APCs**. (B)** Light chain transcript expression in lymphocytes and plasma cells. Plasma cell cluster number 1 is expressing only kappa light chains (IGKC) and is thus light chain restricted. This suggests that this cluster represents a single clone. Plasma cell cluster number 2 expresses kappa and lambda light chain constant regions (IGKC and IGLC) and is therefore not light chain restricted is thus expresses polyclonal.

Supplementary Methods

InDrops single cell RNA-seq

InDrops was performed as described[1]. In brief, cells were diluted into 60,000 cells/mL in 9% Optiprep buffer. Single cell encapsulation was carried out using an inDrops instrument and microfluidic chip manufactured by 1CellBio. In total, four tubes of cells with about 1,000 cells/tube were collected. Library preparation was performed according to the protocol provided by the manufacturer. Of note, we used low PCR cycles (8-9 cycles) to amplify the libraries from this sample as our previous trials with higher PCR cycles (12-13 cycles) resulted in an overamplification of the libraries where averaged transcripts/cell detected was less than 300. Libraries were sequenced by HiSeq 2500 and NextSeq with a sequencing depth of 50K mapped reads/cell.

InDrops single nucleus RNA-seq

Nuclei were isolated with Nuclei EZ Lysis buffer (Sigma #NUC-101) supplemented with protease inhibitor (Roche #5892791001) and RNase inhibitor (Promega #N2615, Life Technologies #AM2696). Samples were cut into <2 mm pieces and homogenized using a Dounce homogenizer (Kimble Chase #885302-0002) in 2ml of ice-cold Nuclei EZ Lysis buffer and incubated on ice for 5 min with an additional 2ml of lysis buffer. The homogenate was filtered through a 40-µm cell strainer (pluriSelect #43-50040-51) and then centrifuged at 500 x for 5 min at 4 ºC. The pellet was resuspended and washed with 4 ml of the buffer and incubated on ice for 5 min. After another centrifugation, the pellet was resuspended with Nuclei Suspension Buffer (1x PBS, 0.07% BSA, 0.1% RNase inhibitor), filtered through a 20-µm cell strainer (pluriSelect 43-50020-50) and counted. RNA from single nucleus was encapsulated, barcoded and reversed transcribed using an InDrop microfluidics system (1CellBio). The library was sequenced in HiSeq2500 with custom primers.

InDrops data preprocessing

We used a recently developed inDrops computational pipeline, dropEst,[1] to process the single cell and single nuclei InDrops data. In brief, cell barcodes and UMIs from the library were extracted from read1 by the dropTag program and added to the names of the transcript reads, resulting in a new fastq file for read alignment. We used STAR (version 2.5.3a) to map the high quality reads to the human genome (GRCh38). Only reads that were uniquely mapped to the genome (~70% of the total reads) were used for UMIs counts. We next ran the dropEst program to estimate the accurate molecular counts, which generated a UMI count matrix for each gene in each cell.

Unsupervised clustering and cell type identification

UMI count matrices from four tubes were combined and loaded into the R package Seurat. For normalization, the DGE matrix was scaled by total UMI counts, multiplied by 10,000 and transformed to log space. Only genes found to be expressed in >10 cells were retained. Cells with a relatively high percentage of UMIs mapped to mitochondrial genes (>=0.3) were discarded. Moreover, cells with fewer than 300 or more than 4,000 detected genes were omitted, resulting in 4,487 cells. We also regressed out the variants arising from library size and percentage of mitochondrial genes using the function RegressOut in R package Seurat. The highly variable genes were identified using the function MeanVarPlot with the parameters: x.low.cutoff = 0.0125, x.high.cutoff = 6 and y.cutoff = 1, resulting in an output of 2,404 highly variable genes. 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 20 PCs for two-dimensional t-distributed stochastic neighbor embedding (tSNE), implemented by the Seurat software with the default parameters. Based on the tSNE map, sixteen clusters were identified using the function FindCluster in Seurat with the resolution

parameter set to 0.6. Alternatively, we used SC3[2] to validate the clusters identified by Seurat with cluster number set to 14 (ks=14). We found that the cells clustered by Seurat as a cell type were also grouped together by SC3. 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 to be marker genes. In total, 2,837 marker genes were identified for the clusters in the biopsy dataset. A heatmap of selected marker gene expression across clusters was plotted using a Python plotting library Matplotlib. We applied the same unsupervised clustering analysis on the single nucleus dataset. First, we generated the digital expression matrix using the same dropEst pipeline that utilizes both exonic and intronic reads. After filtering low quality nuclei, 4,259 nuclei with > 400 genes expressed were imported into Seurat for clustering analysis. In total, we identified six kidney cell types in the single nucleus dataset, including PT, LOH, intercalated cells (IC), principal cells (PC), DCT and podocytes.

Comparison of immune cell types to a published PBMC single cell dataset
Cell-type-specific expression patterns of the cell clusters identified in our dataset were compared to signatures previously defined in a PBMC dataset by calculating the pairwise Pearson correlations coefficients between each pair of cell types for the same set of genes. First, a precomputed Seurat object containing cell cluster information for 33K human PBMCs was downloaded from the Satija lab (http://satijalab.org/seurat/get_started.html). Only genes detected in both our dataset and the PBMC dataset were used for downstream correlation analysis. Second, pearson correlation was computed between the cell clusters in our dataset and the cell clusters identified in the PBMC dataset, using the previously defined cell-type annotations and normalized average gene expression values for each cell type. Data was shown by pheatmap R package.

To compare monocyte transcriptomes between the biopsy and PBMC datasets, we extracted the expression profiles for the cluster we annotated as monocyte#1 and the corresponding monocyte cluster (mono_CD16+_C1qa) in the PBMC dataset. We then clustered the cells after removing cell-cell variations driven by the number of detected molecules and mitochondrial gene contents. Cells in tSNE were colored by the original dataset where they are extracted, or the cell cluster where they are assigned based on unsupervised clustering analysis. To compare the marker gene expression in the monocyte from different sources, we extracted the scaled UMI expression values for the selected genes that are known as monocyte markers, receptors and DC differentiation markers. Average expression level of these selected genes was visualized as the violin plot using ggplot2 package in R.

Correlation analysis of the kidney cell types from allograft biopsy and P1 mouse kidney
To assess the similarity between biopsy cell types and embryonic kidney, we re-analyzed a recently published Dropseq dataset from P1 kidney (GSE94333).[3] We used the Seurat clustering parameters described by the authors and reproduced the same cell types from the datasets. We calculated the Pearson correlation based on the expression patterns of highly variable genes between cell populations within the mouse embryonic kidney dataset against the cell types identified in our biopsy dataset. Correlation matrix were visualized by R package pheatmap. Color keys represent the range of the coefficients of determination (r2) in this analysis.

Ligand-receptor interaction analysis
To study ligand-receptor interactions across the cell types identified from the transplant biopsy, 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).[4, 5] We selected the receptors that were only differentially expressed in each cell type from the biopsy dataset. To determine the ligand-receptor pairs to plot on the heatmap, we required (i) the receptors are uniquely expressed in each cell type (q-val<0.05 and logFC>0.6); (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.

Comparison of healthy kidney and allograft kidney
To compare the transcriptional difference across each tubular cell type in healthy and disease state, we extracted the expression profiles for PT, LOH, and PC (CD) from the healthy kidney single nucleus dataset and the biopsy single cell dataset. We then performed integrated analysis on both datasets using a recently developed computational strategy[6] (implemented in Seurat v2.0). First, we selected the union of the top 2,000 genes with the highest dispersion from both datasets for a canonical correlation analysis (CCA) to identify common sources of variation between the two datasets. Then we aligned the CCA subspaces using the first 15 dimensions of the CCA. After CCA alignment, we performed clustering analysis on the nuclei and cells with the resolution parameter set to 0.6. 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 nuclei and cells from different datasets but grouped in the same cluster by the alignment analysis. Differential genes were visualized using DotPlot function in Seurat.

Pseudotemporal ordering of PT single cells and TF analysis
We used Monocle2[7] to draw a minimal spanning tree connecting the PT cells collected from healthy and diseased kidneys. As input into Monocle2, we selected the highly variable genes for cell ordering as described in the Monocle2 tutorial (http://cole-trapnell-lab.github.io/monocle-release/docs_mobile/). We then reduced the data space to two dimensions using the reduceDimension function with 'DDRTree' method and ordered the cells using the orderCells function in Monocle2. Individual cells were color-coded based on the kidneys where they were collected. To identify the transcription factors whose expression are dynamically changing across the trajectory from healthy PT to diseased PT, we first identified the genes that were differentially expressed across pseudotime using the differentialGeneTest function in Monocle2 with the fullModelFormulaStr parameter set to 'Psudeotime'. To identify the TFs, we crossed the differential genes to the human TF list downloaded from AnimalTFDB (http://bioinfo.life.hust.edu.cn/AnimalTFDB/). We selected some of the important TFs that have been reported by the literature as key regulators of proximal tubule injury and repair and visualized them by R package ggplot2.

Analysis of heterogeneity in the EC population
We performed unsupervised clustering analysis on the endothelial cells extracted from the biopsy dataset using Seurat with similar parameters as described above. Top markers in each subcluster were visualized with the DoHeatmap function in Seurat. We used a stacked violin plot to show the expression of endothelial markers, angiogenic markers, and immunoglobulin genes across the three EC subgroups. Marker genes were selected based on literature search, and were differentially expressed among the EC subclusters. GO analysis was performed by uploading the differential genes from the EC subclusters to ToppGene Suite (http://toppgene.cchmc.org). Top 4 GO terms from categories of biological process, molecular function and cellular component were visualized using ggplot2 package.

Re-assessing disease-associated gene lists from the public datasets
To map the disease-associated genes to single cell, we selected the top genes from three public datasets where the authors identified EC enriched transcripts[8], TCMR enriched genes,[9] and ABMR enriched markers[10] by comparing the microarray data from the biopsies with different rejection patterns. Expression values of these genes in our single cell dataset were normalized, z-scored and visualized the heatmap.2 in gplots R package.

DropSeq single-cell RNA sequencing
Dropseq was performed on the dissociated biopsy cells as previously described.[11] Briefly, biopsy was dissociated into single cell using the same methods as being used in inDrops. Single cell suspension was visually inspected under a microscope, counted by hemocytometer (INCYTO C-chip) and resuspended in PBS + 0.01% BSA. Single cells were coencapsulated in droplets with barcoded beads purchased from Chemgenes (MACOSKO-

2011-10). cDNA libraries were constructed according to an online Dropseq protocol (McCarroll's lab: http://mccarrolllab.com/dropseq/). cDNA quality was determined by BioAnalyzer (Agilent) with a high sensitivity DNA chip.

GSEA pathway analysis
GSEA (http://software.broadinstitute.org/gsea/index.jsp) was used to estimate the enriched pathways in the collecting duct subgroups from biopsy and healthy kidney. We used the normalized UMI count matrix generated by Seurat (NormalizeData and ScaleData) as an expression dataset, and a gene set containing all human pathways downloaded from Bader Lab (http://download.baderlab.org/EM_Genesets/current_release/Human/Entrezgene/) as gene set file input to GSEA. We used 1,000 gene label permutations and identified significantly enriched pathways defined by adjusted p value < 0.05 between the collecting duct cells from biopsy and healthy kidney.

Supplementary Tables

Table S1

HLA DSAs at biopsy	Mean fluorescence intensity
DRB5*01:01	mfi 12143 (Immunodominant DSA)
DRB1*15:01	mfi 2354
DQA1*03:02/DQB1*03:03	mfi 1750
DQA1*03:01/DQB1*03:03	mfi 1660

Table S2

Banff criteria	t	i	v	ptc	g	cg	ci	ct	mm	ah	cv	C4d
score	3	3	0	3	2	0	0	0	0	1	0	0

Table S3

HLA	A	B	C	Bw4	Bw6	DRB1	DRB3/4/5	DQB1	DPB1
Recipient	24/11	40(61)/51	01/03(10)	+	+	07/14	B4*01:03N B3*02	03(9)/05	
Donor	03:01/ 24:02	07:02/ 07:61		-	+	15:01/ 07:01	DRB4*01:03N DRB5*01	03:03/ 06:02	04:01

Tables S1-3: Clinical Features of Study Biopsy.
(1) Haplotype matching of donor a recipient at class I and class II loci. (2) Full Banff criteria scoring for the study biopsy showed histologic features consistent with Banff 1B T-cell mediated rejection and acute C4d-negative antibody mediated rejection. (3) Donor specific antibody (DSA) found in the patients serum at the time of biopsy. mfi-mean fluorescent intensity, t = tubulitis, i = interstitial lymphocyte infiltration, v = intimal arteritis, ptc = peritubular capillaritis, g = glomerulitis, cg = transplant glomerulopathy, ci = interstitial fibrosis, ct = tubular atrophy, mm = mesangial matrix expansion, ah = arteriolar hyalinosis, cv = vascular intimal fibrosis.

Supplementary Table S4.
Link to data set .xls

Supplementary References

1. Klein, AM, Mazutis, L, Akartuna, I, Tallapragada, N, Veres, A, Li, V, Peshkin, L, Weitz, DA, Kirschner, MW: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell,* 161**:** 1187-1201, 2015.
2. Kiselev, VY, Kirschner, K, Schaub, MT, Andrews, T, Yiu, A, Chandra, T, Natarajan, KN, Reik, W, Barahona, M, Green, AR, Hemberg, M: SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods,* 14**:** 483-486, 2017.
3. 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.
4. 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.
5. Hrvatin, S, Hochbaum, DR, Nagy, MA, Cicconet, M, Robertson, K, Cheadle, L, Zilionis, R, Ratner, A, Borges-Monroy, R, Klein, AM, Sabatini, BL, Greenberg, ME: Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci*, 2017.
6. Butler, A, Satija, R: Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*, 2017.
7. Qiu, X, Mao, Q, Tang, Y, Wang, L, Chawla, R, Pliner, HA, Trapnell, C: Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods,* 14**:** 979-982, 2017.
8. Einecke, G, Sis, B, Reeve, J, Mengel, M, Campbell, PM, Hidalgo, LG, Kaplan, B, Halloran, PF: Antibody-mediated microcirculation injury is the major cause of late kidney transplant failure. *Am J Transplant,* 9**:** 2520-2531, 2009.
9. Venner, JM, Famulski, KS, Badr, D, Hidalgo, LG, Chang, J, Halloran, PF: Molecular landscape of T cell-mediated rejection in human kidney transplants: prominence of CTLA4 and PD ligands. *Am J Transplant,* 14**:** 2565-2576, 2014.
10. Sellares, J, Reeve, J, Loupy, A, Mengel, M, Sis, B, Skene, A, de Freitas, DG, Kreepala, C, Hidalgo, LG, Famulski, KS, Halloran, PF: Molecular diagnosis of antibody-mediated rejection in human kidney transplants. *Am J Transplant,* 13**:** 971-983, 2013.
11. Macosko, EZ, Basu, A, Satija, R, Nemesh, J, Shekhar, K, Goldman, M, Tirosh, I, Bialas, AR, Kamitaki, N, Martersteck, EM, Trombetta, JJ, Weitz, DA, Sanes, JR, Shalek, AK, Regev, A, McCarroll, SA: Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell,* 161**:** 1202-1214, 2015.