Report – Single cell transcriptomes from human kidneys reveal the cellular identity of renal tumors§

Matthew D Young[†1], Thomas J Mitchell[†1,2,3], Felipe A Vieira Braga[†1], Maxine GB Tran[4,5], Benjamin J Stewart[6], John R Ferdinand[6], Grace Collord[1,2,7], Rachel A Botting[8], Dorin-Mirel Popescu[8], Kevin W Loudon[6], Roser Vento-Tormo[1], Emily Stephenson[8], Alex Cagan[1], Sarah Farndon[1,9,10], Martin Del Castillo Velasco-Herrera[1], Charlotte Guzzo[1], Nathan Richoz[6], Lira Mamanova[1], Tevita Aho[2], James N Armitage[3], Antony CP Riddick[3], Imran Mushtaq[9], Stephen Farrell[2], Dyanne Rampling[9], James Nicholson[2,7], Andrew Filby[8], Johanna Burge[2], Steven Lisgo[11], Patrick H Maxwell[12], Susan Lindsay[11], Anne Y Warren[2], Grant D Stewart[2,3], Neil Sebire[9,10], Nicholas Coleman[2,13], Muzlifah Haniffa[8,14*], Sarah A Teichmann[1*], Menna Clatworthy[2,6*], Sam Behjati[1,2,7*]

Affiliations:

[1]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.

[2]Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK.

[3]Department of Surgery, University of Cambridge, Cambridge, CB2 0QQ, UK.

[4]UCL Division of Surgery and Interventional Science, Royal Free Hospital, London NW3 2PS, UK.

[5]Specialist Centre for Kidney Cancer, Royal Free Hospital, London, NW3 2PS, UK.

[6]Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK.

[7]Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK.

[8]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK.

[9]Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, UK.

[10]UCL Great Ormond Street Hospital Institute of Child Health, London WC1N 1E, UK.

[11]Human Developmental Biology Resource, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, NE1 3BZ, UK.

[12]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, CB2 0XY, UK

[13]Department of Pathology, University of Cambridge, Cambridge, CB2 1QP, UK.

[14]Department of Dermatology, Royal Victoria Infirmary, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

[†]These authors contributed equally.

[*]Corresponding authors.

Abstract: Messenger RNA encodes cellular function and phenotype. In the context of human cancer it defines the identity of malignant cells and diversity of tumor tissue. We studied 72,501 single cell transcriptomes of human renal tumors and normal tissue from fetal, pediatric and adult kidneys. We matched childhood Wilms' tumor with specific fetal cell types, thus providing evidence for the hypothesis that Wilms' tumor cells are aberrant fetal cells. In adult renal cell carcinoma we identified a canonical cancer transcriptome that matched a little known subtype of proximal convoluted tubular cell. Analyses of the tumor composition defined cancer-associated normal cells and delineated a complex VEGF signaling circuit. Our findings reveal the precise cellular identity and composition of human kidney tumors.

One Sentence Summary: Single cell mRNAs of 72,501 normal and cancerous kidney cells reveals the cellular identity of childhood and adult tumors.

Main Text: Cancer cell identity is defined by morphological appearance, tissue context, and marker gene expression. Single cell transcriptomics refines this cellular identity on the basis of a comprehensive and quantitative read out of messenger RNA. Precise cellular transcriptomes may reveal a tumor's cell of origin and the transcriptional trajectories underpinning malignant transformation (*1*).

We sought to define the identity of normal and cancerous human kidney cells from a catalogue of 72,501 single kidney cell transcriptomes, integrated with tumor whole genome DNA sequencing (*2*). We studied Wilms' tumor (n=3), clear cell (ccRCC; n=3) and papillary renal cell carcinoma (pRCC; n=1) in relation to healthy fetal (n=2), pediatric (n=3), adolescent (n=2), and adult kidneys (n=5), as well as ureters (n=4; **Table S1**)

Normal tissue biopsies were taken from macroscopically normal regions of kidneys resected due to cancer (n=10) or for transplantation (n=2). We performed technical replicates of each biopsy and biological replicates, where clinically permissible (**Table S1**). We processed kidneys

immediately following resection, generating single cell solutions enriched for viable cells. We derived counts of mRNA molecules in each cell for further analyses, subject to quality control (*2*).

We split 72,501 fetal, normal, and tumor cells into immune and non-immune cell compartments (**Fig. S1**). Using a community detection algorithm (*2*), transcriptomes were further segregated into distinct clusters of cells (**Table S2**). We next generated a reference map of normal mature and fetal cells, assigning an identity to each cluster, by cross-referencing cluster-defining transcripts with canonical markers curated from the literature (**Table S3**). Ambiguous clusters were not included in the reference map and are presented in **Fig. S2-S8**. Highly specific cluster-defining transcripts (potential cell markers) are appended (**Table S4**).

Amongst 42,809 non-malignant cells, 37,951 mature kidney cells represented epithelial cells from distinct micro-anatomical regions of the nephron, with a large proportion of proximal tubular cells (**Fig. 1A-C, Fig. S4**). Furthermore, there were fibroblasts, myofibroblasts, and vascular endothelial cells, i.e. glomerular endothelium, ascending and descending vasa recta (**Fig. 1D, Fig. S2**). 4,858 fetal cells grouped into developing nephron cells (ureteric bud, cap mesenchyme, primitive vesicle) and fibroblasts, myofibroblasts, vascular endothelial and ganglion cells (**Fig. 2A-C, Fig. S5**).

To determine transcriptional programs underlying nephrogenesis, we identified differentially expressed transcription factors in ureteric bud cells against cap mesenchyme and primitive vesicle cells (**Fig. 2D**). Furthermore, we applied pseudo-timing methods to identify transcription factors that define the transition from cap mesenchyme to primitive vesicle (**Fig. 2D**). Together, these analyses identified established, and unknown, transcription factors associated with nephron development, included as a reference for subsequent analyses of malignancy (**Table S5**).

Having established the single cell landscape of healthy kidneys, we characterized the cellular identity of 6,333 non-immune (**Fig. S7**) and 17,821 immune (**Fig. S8**) tumor cells from Wilms' tumor (n=3), ccRCC (n=3) and pRCC (n=1; **Table S1**). Children had received neoadjuvant cytotoxic treatment prior to nephrectomy, as per British practice. Although this pre-treatment reduced yield (**Table S6**), recovered cells represent therapeutically relevant surviving cancer cells

that determine the degree of adjuvant cytotoxic chemotherapy required (*3*). We used logistic regression to quantify the similarity between tumor and normal cell clusters, validated through intrinsic control populations (*2*). That is, the model found myofibroblasts from tumors matched myofibroblasts from mature and fetal kidneys (**Fig. 3A**) and no match for mast cells, a negative control population inserted into the training data.

This similarity metric may be obfuscated by the phenotypic plasticity of tumor cells. We therefore developed a method to genotype individual cancer cells from mRNA reads using somatic copy number changes (**Table S7; Fig. S9**) defined by whole genome sequencing (**Fig. S10**). We validated genotyping calls by phasing single nucleotide polymorphisms across segments with altered copy number, testing for the presence of somatic single nucleotide variants, and comparison to control populations (**Fig. S11-S14**).

Integrating genotyping and similarity analyses, we found that Wilms' cells resembled fetal normal cells, evidencing that Wilms' tumor represents aberrant fetal cells. We found different populations of Wilms' tumor that matched ureteric bud and primitive vesicle cells (specific developing nephron populations) (**Fig. 3A**). One cluster (WF), composed of Wilms' cancer cells and non-cancerous ccRCC fibroblasts, exhibited a fibroblast-myofibroblast transcriptome. In one case, we obtained an anatomically separate perilobar nephrogenic rest, thought to represent a precursor lesion of Wilms'. Like Wilms' cancer cells, we observed that nephrogenic rest resembled ureteric bud or primitive vesicle. This suggests that the potential to generate the different cell states of the fetal nephron is acquired early, or was not lost, by the developing Wilms' cancer, although this conclusion is based on only one sample.

To validate the cellular identity of Wilms' cells, we interrogated bulk transcriptomes of an independent series of 124 Wilms' tumors for cellular signatures of ureteric bud and primitive vesicle (*4, 5*). We extracted specific markers expressed within ureteric bud or primitive vesicle cells and unexpressed within non-tumor cells (**Table S8**, (*2*)) and probed bulk transcriptomes for these cluster defining transcripts. As comparators to Wilms' we included fetal, pediatric, and adult normal tissue bulk transcriptomes (n=135) and other childhood kidney tumors: 17 congenital mesoblastic nephroma and 65 malignant rhabdoid tumors. Corroborating the presence of these

cells in Wilms', signatures of primitive vesicle and ureteric bud cells were seen in, and confined to, Wilms' and normal fetal tissue (**Fig. 3B**).

Placing Wilms' tumor cells in pseudo-time revealed two transcriptional programs emanating from the ureteric bud: one branch predominantly describing the development of nephrogenic rest cells and the other of Wilms' cancer cells (**Fig. 3C**). There was a significant overlap in the transcription factors underpinning these two programs (**Fig 3D; Table S9**) and normal nephrogenesis ($p<10^{-4}$; hypergeometric test). This indicates that developmental relationships exist between Wilms' tumor cells that have been adopted from normal nephrogenesis. Our analyses reveal the plasticity and fetal identity of Wilms' cells and transcriptionally defines developmental cell states and trajectories that may harbor targetable vulnerabilities.

Next we studied ccRCC and pRCC (type 1), including one case of von Hippel Lindau disease-related ccRCC (**Table S1**). Matching ccRCC and pRCC with normal mature cells, we found that they retained transcriptional features of cluster PT1, a specific subtype of convoluted proximal tubular cell (**Fig. 4A**). Most (6/7) ccRCC clusters and all pRCC cells matched this particular PT1 cell, indicating that it represents an RCC cell state that transcends the diversity of RCC cells within and across tumors. Little is known about the nearest normal cell correlate of RCC, the PT1 cell, which has been identified to become more abundant in inflamed renal tissue (*6*).

To validate the identity of the PT1 signature in RCC, we exploited the fact that they were defined by SLC17A3 and VCAM1 with absence of SLC7A13 within our data (**Fig. 4B; Fig. S2**). We measured these transcripts in an independent series of 1,019 publicly available bulk kidney tumor and normal tissue transcriptomes. High expression of SLC17A3 mRNA distinguished ccRCC and pRCC (type 1 and 2) from other types of RCC ($p<10^{-4}$; Mann-Whitney test), whereas SLC7A13 mRNA was significantly depleted in ccRCC/pRCC bulk transcriptomes versus normal ($p<10^{-4}$; Mann-Whitney test), as were mRNAs representing other regions of the nephron (**Fig. 4B**). VCAM1, specific to PT1 within proximal tubules, was also significantly elevated across RCC bulk transcriptomes ($p<10^{-4}$; Mann-Whitney test; **Fig. 4B**), with each individual RCC tumor exhibiting PT1 features (**Fig. S15**). Confocal microscopy demonstrated co-localization of VCAM1 and SLC17A3 in CA9+ cells, CA9 being a specific marker of ccRCC cells (**Fig. 4C**). Furthermore, we

studied the earliest precursor lesions of ccRCC: CA9+ proximal tubular cells residing in morphologically normal kidney, predisposed to ccRCC through pathogenic germline mutation of *VHL*. Examining tissue from three individuals, we identified CA9+/VCAM1+ clusters of proximal tubular cells (**Fig. 4D**). Similarly, tumors arising in these kidneys harbored CA9+/VCAM1+ cells (**Fig. S16**). As expected, VCAM1 was otherwise sparsely expressed on proximal tubular cells. Together these observations substantiate our proposition that PT1 cells are the nearest normal cell correlate of ccRCC cells. The presence of the PT1 signature in both ccRCC and pRCC may indicate a common origin of these tumors with divergent fates.

Apart from the PT1 signature in pRCC and ccRCC, we found that one ccRCC cell cluster (cR7) matched PT3 cells and that pRCC cells exhibited an additional, weaker match with collecting duct cells (**Fig. 4A**). Neither signal was enriched in bulk transcriptomes (**Fig. 4B**). As our study was confined to type 1 pRCC, it is possible that we missed other pRCC cell types.

Finally, we dissected the tumor microenvironment occupied by cancer-associated normal cells, comprised of immune cells, fibroblasts, myofibroblasts and vascular endothelial cells (predominately ascending vasa recta) (**Fig. S7, S8, S17**). Within these we studied VEGF signaling, an established target in RCC treatment (*7, 8*). The VEGF signaling circuit in renal tumors involves VEGFA secretion from RCC cells resulting in a response from endothelial cells (*7, 8*). Measuring expression of the key components of VEGF signaling, we identified tumor infiltrating macrophages as a further source of VEGFA (**Fig. S18A**), confirmed by confocal microscopy of ccRCC cells and flow cytometry of an independent ccRCC tumor (**Fig. S18B,C,D**). VEGF-signaling receptors (KDR, FLT1, FLT4) were mainly expressed by one population of ascending vasa recta cells (**Fig. S18A,** cluster tE1). The other ascending vasa recta cluster, tE2, (**Fig. S18A)** exhibited lymphangiogenic VEGFC and FLT1. Furthermore, tE2 endothelial cells expressed high levels of ACKR1, a marker of venular endothelium promoting tissue migration of immune cells (*9*). Overall these findings delineate a complex VEGF signaling circuit within RCC tissue.

By identifying specific normal cell correlates of renal cancer cells, our study moves our understanding of these malignancies beyond a notion of “fetalness” or an approximate micro-anatomical region to a precise cellular, molecularly quantitative resolution. Our findings portray

the peak incidence of Wilms' tumor in early childhood as a corruption of fetal nephrogenesis, in contrast to the life-long development of RCC in mature kidneys. Our study provides a scalable experimental strategy for determining the identity of human cancer cells.

5

References and Notes:

1. C. Ziegenhain *et al.*, Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell* **65**, 631-643.e634 (2017).
2. Supplementary Methods.
3. K. Pritchard-Jones *et al.*, Omission of doxorubicin from the treatment of stage II-III, intermediate-risk Wilms' tumour (SIOP WT 2001): an open-label, non-inferiority, randomised controlled trial. *Lancet (London, England)* **386**, 1156-1164 (2015).
4. E. J. Perlman *et al.*, MLLT1 YEATS domain mutations in clinically distinctive Favourable Histology Wilms tumours. *Nature communications* **6**, 10013 (2015).
5. S. Gadd *et al.*, A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nature genetics* **49**, 1487-1494 (2017).
6. D. Seron, J. S. Cameron, D. O. Haskard, Expression of VCAM-1 in the normal and diseased kidney. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* **6**, 917-922 (1991).
7. B. Ljungberg *et al.*, EAU guidelines on renal cell carcinoma: 2014 update. *European urology* **67**, 913-924 (2015).
8. S. Fernandez-Pello *et al.*, A Systematic Review and Meta-analysis Comparing the Effectiveness and Adverse Effects of Different Systemic Treatments for Non-clear Cell Renal Cell Carcinoma. *European urology* **71**, 426-436 (2017).
9. A. Thiriot *et al.*, Differential DARC/ACKR1 expression distinguishes venular from non-venular endothelial cells in murine tissues. *BMC biology* **15**, 45 (2017).
10. D. Gerrelli, S. Lisgo, A. J. Copp, S. Lindsay, Enabling research with human embryonic and fetal tissue resources. *Development (Cambridge, England)* **142**, 3073-3076 (2015).
11. I. Kozarewa *et al.*, Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295 (2009).
12. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
13. D. Jones *et al.*, cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15 10 11-15 10 18 (2016).
14. A. Menzies *et al.*, VAGrENT: Variation Annotation Generator. *Curr Protoc Bioinformatics* **52**, 15 18 11-11 (2015).
15. P. Van Loo *et al.*, Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* **107**, 16910-16915 (2010).
16. A. Lun, S. Riesenfeld, T. Andrews, T. P. Dao, T. Gomes, "On the correct detection of empty droplets in droplet-based single-cell RNA sequencing protocols," (https://github.com/TimothyTickle/hca-jamboree-cell-identification/blob/master/docs/EmptyDrops_group4_report.pdf, 2017).
17. Y. J. Chen *et al.*, Single-cell RNA sequencing identifies distinct mouse medial ganglionic eminence cell types. *Sci Rep* **7**, 45656 (2017).
18. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
19. L. J. P. v. d. Maaten, G. E. Hinton, Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

20. L. J. P. v. d. Maaten, Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221-3245 (2014).
21. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
22. A. Rajaraman, J. D. Ullman, *Mining of Massive Datasets*. (Cambridge University Press, Cambridge, 2011).
23. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309-315 (2017).
24. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
25. H. M. Zhang *et al.*, AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40**, D144-149 (2012).
26. H. M. Zhang *et al.*, AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76-81 (2015).
27. J. H. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *2010* **33**, 22 (2010).
28. N. Cancer Genome Atlas Research, Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49 (2013).
29. N. Cancer Genome Atlas Research *et al.*, Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* **374**, 135-145 (2016).
30. E. J. Perlman *et al.*, MLLT1 YEATS domain mutations in clinically distinctive Favourable Histology Wilms tumours. *Nature communications* **6**, 10013 (2015).
31. A. L. Walz *et al.*, Recurrent DGCR8, DROSHA, and SIX homeodomain mutations in favorable histology Wilms tumors. *Cancer cell* **27**, 286-297 (2015).
32. A. Colaprico *et al.*, TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).
33. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
34. W. M. Hern, Correlation of fetal age and measurements between 10 and 26 weeks of gestation. *Obstetrics & Gynecology* **63**, 26-32 (1984).
35. E. W. Brunskill *et al.*, Atlas of gene expression in the developing kidney at microanatomic resolution. *Developmental cell* **15**, 781-791 (2008).
36. S. Metsuyanim *et al.*, Expression of stem cell markers in the human fetal kidney. *PloS one* **4**, e6709 (2009).
37. J. W. Lee, C. L. Chou, M. A. Knepper, Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment-Specific Transcriptomes. *Journal of the American Society of Nephrology : JASN* **26**, 2669-2677 (2015).
38. M. Habuka *et al.*, The kidney transcriptome and proteome defined by transcriptomics and antibody-based profiling. *PloS one* **9**, e116125 (2014).
39. D. Chabardes-Garonne *et al.*, A panoramic view of gene expression in the human kidney. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13710-13715 (2003).
40. X. Han, S. Amar, Secreted Frizzled-related Protein 1 (SFRP1) Protects Fibroblasts from Ceramide-induced Apoptosis. *Journal of Biological Chemistry* **279**, 2832-2840 (2004).

41. M. Matsuyama, A. Nomori, K. Nakakuni, A. Shimono, M. Fukushima, Secreted Frizzled-related Protein 1 (Sfrp1) Regulates the Progression of Renal Fibrosis in a Mouse Model of Obstructive Nephropathy. *Journal of Biological Chemistry* **289**, 31526-31533 (2014).
42. R. Lennon *et al.*, Global analysis reveals the complexity of the human glomerular extracellular matrix. *Journal of the American Society of Nephrology : JASN* **25**, 939-951 (2014).
43. V. S. LeBleu *et al.*, Origin and Function of Myofibroblasts in Kidney Fibrosis. *Nature medicine* **19**, 1047-1053 (2013).
44. L. Wang *et al.*, NDUFA4L2 is associated with clear cell renal cell carcinoma malignancy and is regulated by ELK1. *PeerJ* **5**, e4065 (2017).
45. M. Habuka *et al.*, The Urinary Bladder Transcriptome and Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PloS one* **10**, e0145301 (2015).
46. W. C. Aird, Phenotypic Heterogeneity of the Endothelium. *Circulation Research* **100**, 174 (2007).
47. R. Nawroth *et al.*, VE-PTP and VE-cadherin ectodomains interact to facilitate regulation of phosphorylation and cell contacts. *The EMBO journal* **21**, 4885-4895 (2002).

Acknowledgments: We thank Sir Michael Stratton, Peter Campbell, David Rowitch, Manfred Gessler and Manasa Ramakrishna for review of the manuscript; Moritz Gerstung and Valentine Svensson for advice regarding logistic regression. We are indebted to our patients and their families for participating in this research.

Funding: This experiment was principally funded by the St Baldrick's Foundation (Robert J Arceci International Award to S.B.). Additional funding was received from: Wellcome (S.B., M.H., G.C., C.G.); Cambridge Biomedical Research Campus (biobanking infrastructure; M.R.C.); CRUK Cambridge Centre (biobanking infrastructure); NIHR Blood and Transplant Research Unit (M.R.C.); MRC (M.R.C.); Arthritis Research UK (M.R.C.); The Lister Institute for Preventative Medicine (M.H.); NIHR and Newcastle-Biomedical Research Centre (M.H.); ISAC SRL-EL program; A.F.); joint Wellcome Trust/MRC (S.Lis., S.Lin.); Kidney Cancer UK (M.G.B.T.); Facing up 2 Kidney Cancer (M.G.B.T.); EMBO (R.V.T); Human Frontier Science Program (R.V.T); Children with Cancer UK (S.J.F.).

Author contributions: S.B. conceived the experiment. M.D.Y., T.J.M. and S.B. analyzed the data, with contributions from F.V.B., B.S., M.D.C.V.H., G.C. and M.C. Samples were curated and / or experiments performed by: F.V.B., J.R.F., M.G.B.T., P.H.M, R.A.B., D.M.O., R.V-T., E.S., K.L., S.Far., C.G., N.R., L.M., T.A., J.N.A., A.C.P.R., I.M., S.F., C.J., D.R., J.N., A.F.,

J.B., S.Lis., S.Lin. and G.D.S. Pathological expertise was provided by A.Y.W., N.S., and N.C. A.C. created illustrations. T.J.M., M.D.Y. and S.B. wrote the manuscript. M.H., S.A.T., M.C. and S.B. co-directed the study.

Competing interests: None.

Data and materials availability: Raw sequencing data have been deposited in the European Genome-phenome Archive (EGA) under study IDs EGAS00001002171, EGAS00001002486, EGAS00001002325 and EGAS00001002553. Sample specific identifiers can be found in **Table S6,S10,** a table of mapped UMI counts for each cell and gene combination in **Data S1** and metadata about each cell in **Table S11**. The code necessary to perform the analysis and generate figures can be obtained from https://github.com/constantAmateur/scKidneyTumors.

Supplementary Materials:

Materials and Methods

Figures S1-S19

Tables S1-S12

Data S1

References (*10-47*)

Figure 1. Canonical cell types of normal human kidneys.

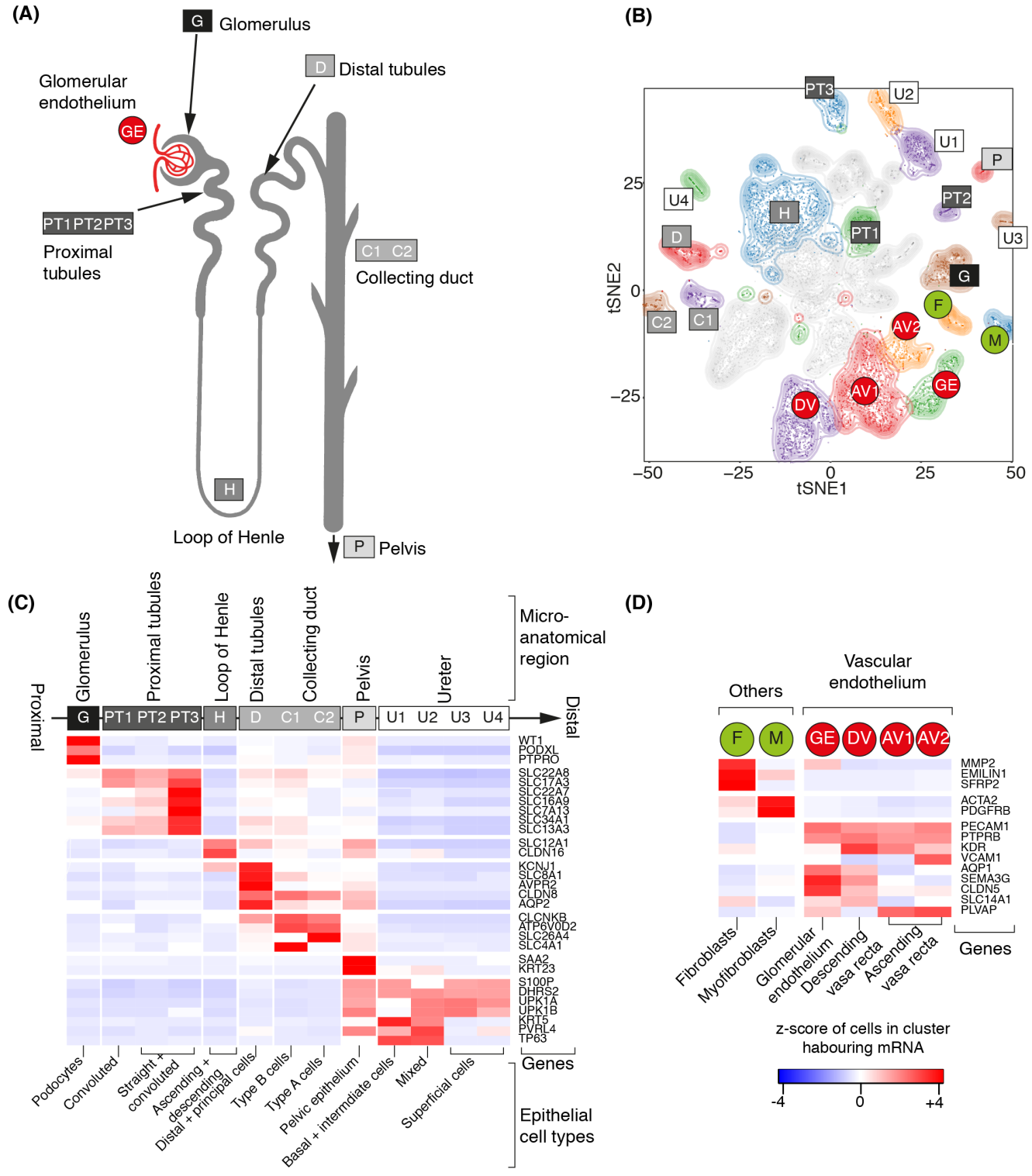


Figure 1. Canonical cell types in normal human kidneys.

(A) Illustration of nephron anatomy with cell clusters marked.

(B) tSNE representation of 8,707 normal epithelial and vascular cells. Clusters are colored, uniquely labelled and emphasized with density contours. Ambiguous clusters are de-emphasized and fully shown in **Fig. S2**.

(C) Expression of canonical nephron specific genes (**Table S3**) in clusters from **(A)**. Colors give the fraction of cells expressing each gene in a cluster, scaled to have mean 0 and standard deviation 1 across all clusters.

(D) Expression of clusters in **(A)** not shown in **(C)** and their canonical genes.

Figure 2. Fetal cell types and nephrogenesis.

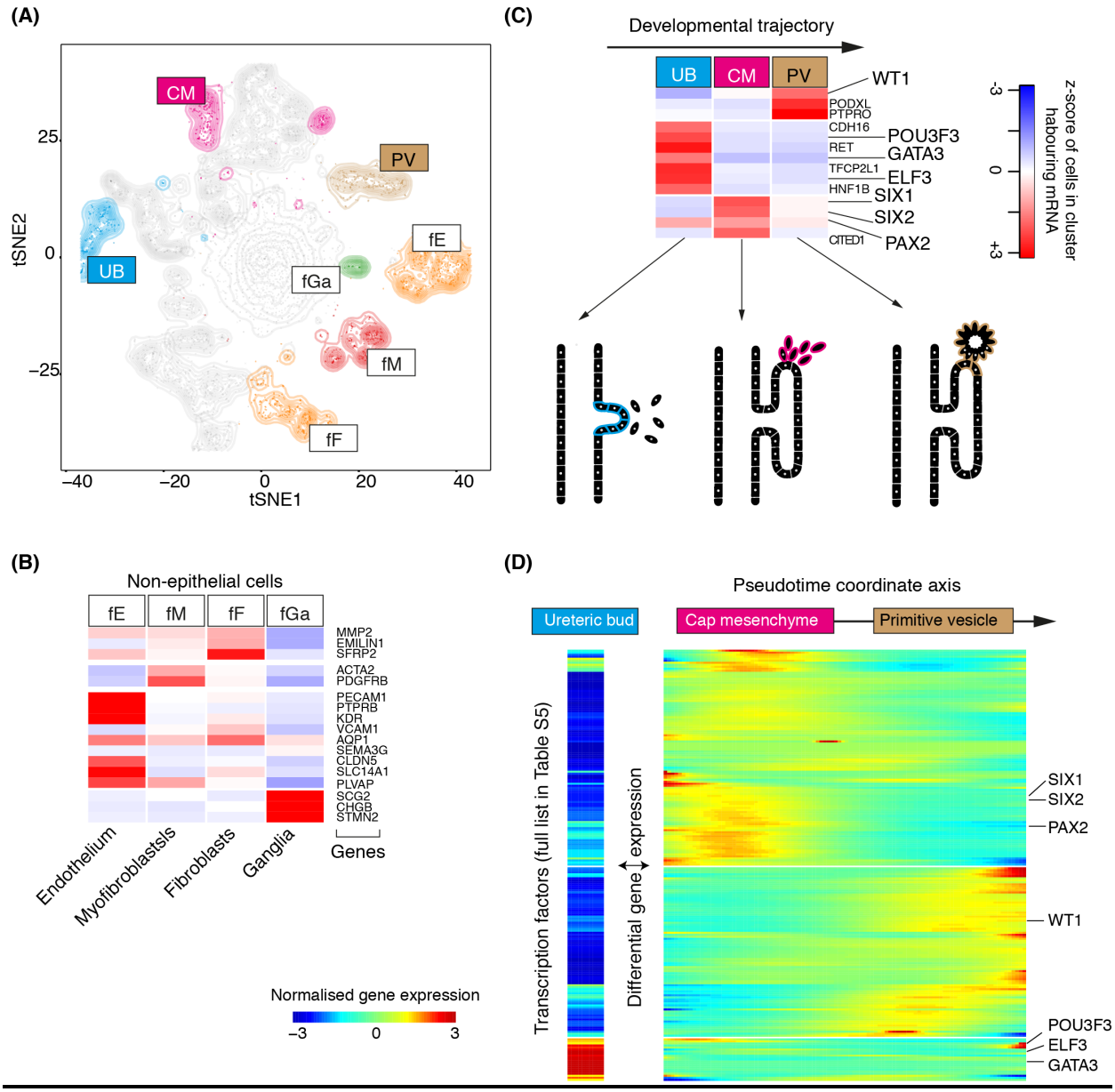


Figure 2. Fetal cell types and nephrogenesis.

(A) tSNE representation of 4,858 fetal epithelial and vascular cells, colored and labelled as in **Fig. 1A**.

(B) Expression of markers of clusters in **(A)**, colored as in **Fig. 1C**.

(C) Expression of nephrogenesis markers from clusters in **(A)** with illustration of nephron development. Formation of nephrons emanates from the ureteric bud, which induces condensation of the overlying mesenchyme into the cap mesenchyme. The cap mesenchyme then

forms the primitive vesicle, the precursor of the glomerulus. The tubular system grows out from both ends of the fetal nephron: ureteric bud and primitive vesicle.

(D) The expression of transcription factor which vary significantly ($p<0.01$; likelihood ratio test) along the pseudo-time trajectory defined using the CM and PV cells from **(C)**, or differentially expressed between UB versus CM and PV. UB expression is shown in a separated block on the left. Within the right block, pseudo-time increases from left to right and rows are clustered and grouped by hierarchical clustering with canonical transcription factors of nephrogenesis highlighted (see **Table S6**).

Figure 3. Matching childhood tumours with normal cells.

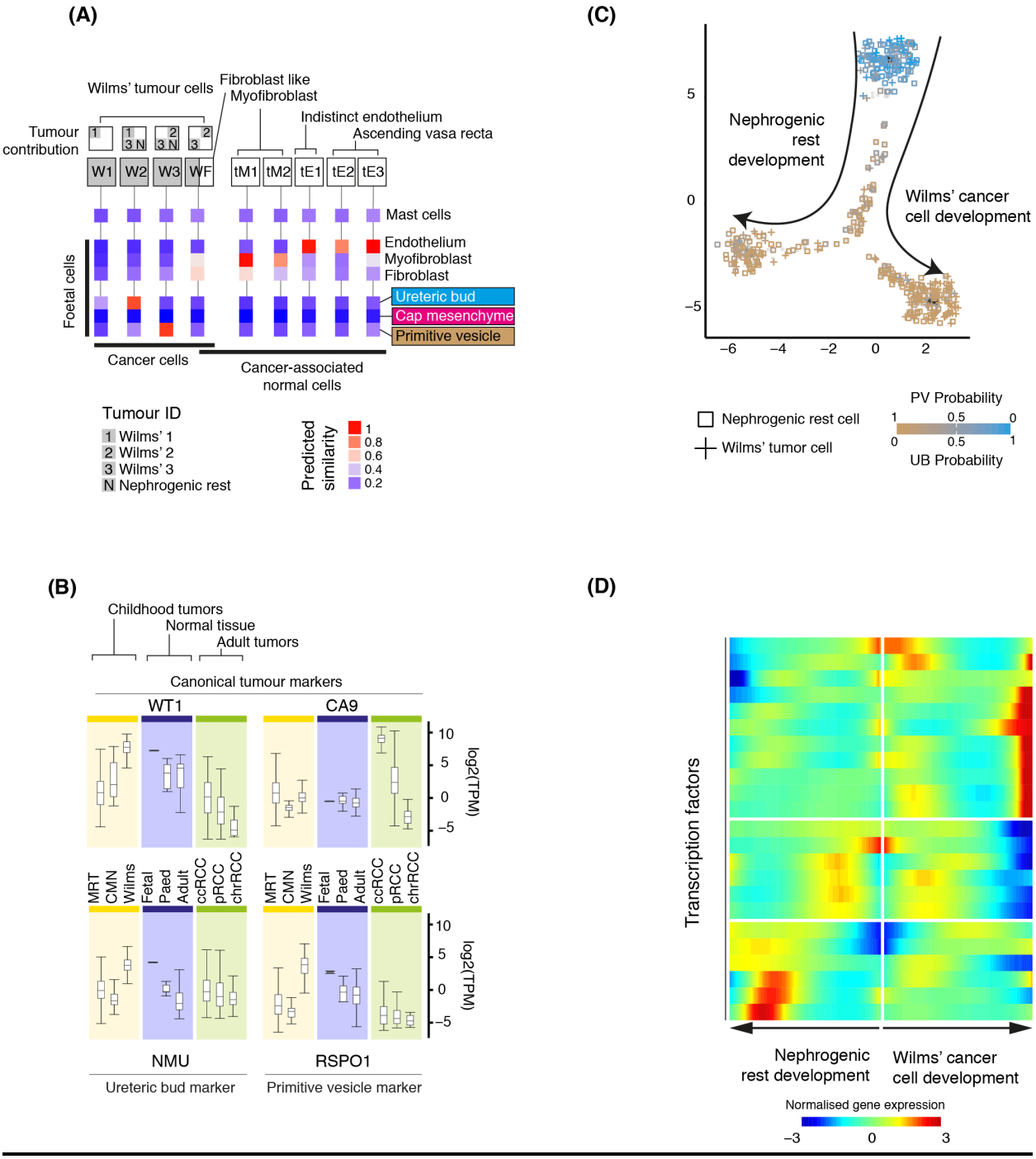


Figure 3. Matching childhood tumors with normal fetal cells.

(A) Similarity of Wilms' tumor and cancer-associated normal cells to the reference fetal kidney map (**Fig. 2A**), with mast cells added as a negative control. Square boxes indicate sample

contribution. Colors represent the probability that the cluster identified in the column header is "similar" to the fetal cluster identified by the row label (*2*).

(B) Expression of canonical tumor markers and representative UB and PV specific genes (**Table S8**) in RNA-seq from childhood cancers (yellow), normal tissue (blue) or adult cancers (green). MRT: malignant rhabdoid tumor; CMN: congenital mesoblastic nephroma. As positive controls, canonical tumor markers are shown: WT1 (Wilms'); CA9 (ccRCC).

(C) Pseudo-time trajectory of all Wilms tumor and nephrogenic rest cell. Color indicates similarity of each cell to the PV or UB fetal population. Jitter has been added to each point's position with the original position plotted underneath in black (*2*).

(D) Transcription factors identified as varying significantly along the pseudo-time trajectory in **(C).** The center of the heatmap corresponds to the cells at the top of **(C)** and then proceeding left/right along the arrows shown in **(C)**.

Figure 4. Matching adult tumours with normal cells.

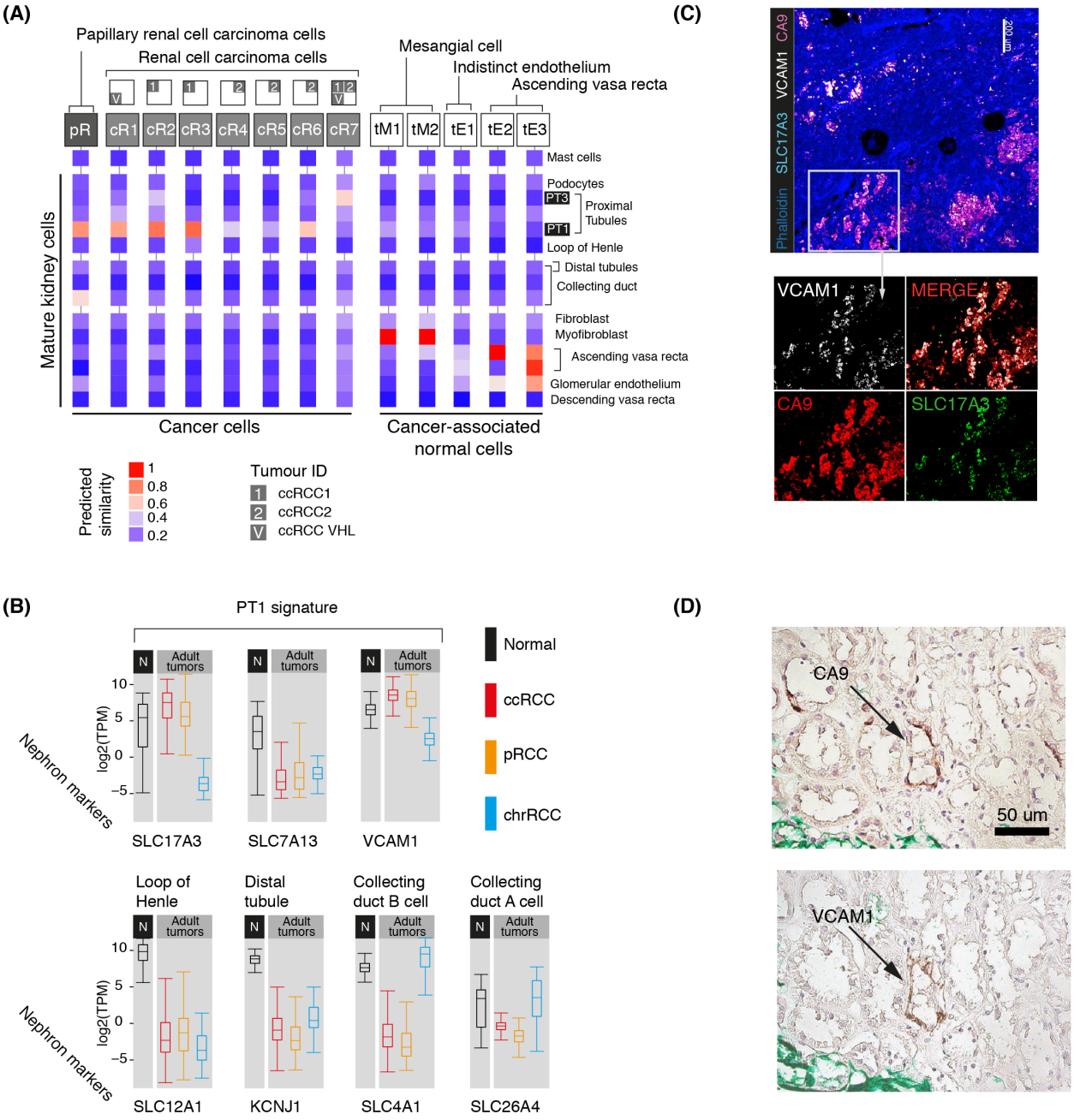


Figure 4. Matching adult tumors with normal mature kidney cells.

(A) Similarity of adult cancer and cancer-associated normal cells to the mature kidney reference map (**Fig. 1B**), with mast cells added as a negative control. Square boxes indicate sample

contribution. Colors represent the probability that the cluster identified in the column header is "similar" to the normal cluster identified by the row label ***(2)***.

(B) Expression of nephron specific genes in bulk RNA-seq as in **Fig. 3B**. pRCC samples are both type 1 and 2.

(C) Confocal microscopy showing co-localization of PT1 markers (VCAM1, SLC17A3) in ccRCC cells (CA9).

(D) Staining of a proximal tubular ccRCC precursor lesion (CA9) for the PT1 marker, VCAM1.

Science
AAAS

Supplementary Materials for

Single cell transcriptomes from human kidneys reveal the cellular identity of renal tumors

Matthew D Young, Thomas J Mitchell, Felipe A Vieira Braga, Maxine GB Tran, Benjamin J Stewart, John R Ferdinand, Grace Collord, Rachel A Botting, Dorin-Mirel Popescu, Kevin W Loudon, Roser Vento-Tormo, Emily Stephenson, Alex Cagan, Sarah Farndon, Martin Del Castillo Velasco-Herrera, Charlotte Guzzo, Nathan Richoz, Lira Mamanova, Tevita Aho, James N Armitage, Antony CP Riddick, Imran Mushtaq, Stephen Farrell, Dyanne Rampling, James Nicholson, Andrew Filby, Johanna Burge, Steven Lisgo, Patrick H Maxwell, Susan Lindsay, Anne Y Warren, Grant D Stewart, Neil Sebire, Nicholas Coleman, Muzlifah Haniffa, Sarah A Teichmann, Menna Clatworthy, Sam Behjati

Correspondence to:

sb31@sanger.ac.uk | mrc38@medschl.cam.ac.uk | st9@sanger.ac.uk | m.a.haniffa@newcastle.ac.uk

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Data S1 – Table of counts for all droplets in experiment.

Materials and Methods

Ethics statement

Adult kidneys samples were collected from patients enrolled in the DIAMOND study; Evaluation of biomarkers in urological disease (NHS National Research Ethics Service reference 03/018). Pediatric samples were acquired from patients enrolled in the 'Investigating how childhood tumors and congenital disease develop' (NHS National Research Ethics Service reference 16/EE/0394). Human fetal material was provided by the Joint MRC / Wellcome Trust-funded (grant # 099175/Z/12/Z) Human Developmental BiologyResource (HBDR, http://www.hdbr.org; (*10*)), with appropriate maternal written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee. HDBR is regulated by the UK Human Tissue Authority (HTA; www.hta.gov.uk) and operates in accordance with the relevant HTA Codes of Practice. VHL kidneys were studied under UK ethics approval, REC 16/WS/0039 and 2002/6486.

Tissue processing

Kidney biopsies were taken by pathologists from normal and tumor regions. Where clinically permissible, separate cortical, medullary, pelvic, and ureteric biopsies were obtained. Otherwise a biopsy from the interface of medulla and cortex was taken. Tissues were sliced into approximately 30mm^3 pieces of tissue and digested for 30 min at 37°C with agitation in a digestion solution containing 25µg/ml Liberase TM (Roche) and 50µg/ml DNase (Sigma) in RPMI (Gibco). Following incubation samples were transferred to a C tube (Miltenyi Biotec) and processed on a gentle MACS (Miltenyi Biotec) on programme spleen 4 and subsequently lung 2. The resulting suspension was passed through a 70µm cells strainer (Falcon), washed with PBS and live cells enriched using a Dead Cell Removal kit (Miltenyi Biotec) as per manufactures instructions. Enriched live cells were washed and counted using a hemocytometer with trypan blue. For fetal samples, whole kidneys were dissociated into single cell suspension following rapid 30 min collagenase treatment. Live, single cells were enriched for FACS-sorted DAPI- cell with further enrichment of immune cells by CD45 expression. Overall it took 5-6 hours from obtaining biopsies to generating single cell suspensions run on the Chromium 10X device.

Comparison of human kidney tissue disaggregation methods

It is likely that biases are introduced by different tissue dissociation protocols. We used a combination of mechanical and enzymatic dissociation (liberase) because mechanical dissociation alone introduced measurable biases as to which cell populations were captured (see **Fig. S18**). However, it should be noted that liberase itself can introduce transcriptional artefacts.

In assessing dissociation methods, we used kidneys donated for transplant that were subsequently deemed unsuitable. Renal cortex samples were dissected, weighed (approximately 0.64g per experiment) and sliced into 5mm^3 pieces. Samples were digested in 50 µg/ml DNase (Sigma) in RPMI (Gibco) with or without 25 µg/ml Liberase TM (Roche) for 30 min at 37°C, with agitation. Following incubation, samples were transferred to a gentleMACS C Tube (Miltenyi Biotec) and processed using a gentleMACS Dissociator (Miltenyi Biotec) on program

spleen 4 and subsequently lung 2. The resulting suspension was passed through a 70 µm cells strainer (Falcon) and washed with PBS before leukocyte enrichment using a Percoll (Sigma-Aldrich) density gradient. Cell counts per gram of tissue were calculated with the addition of 123count eBeads (Invitrogen). Fc-Receptors were blocked using saturating concentrations of FcR blocking reagent (Miltenyi Biotec) and then stained with live/dead fixable Aqua (Invitrogen). Surface staining was performed with an anti-CD45- APC-eFluor780 (clone HI30, eBioscience), All samples were acquired on a BD LSR 4 laser Fortessa and data analysed using FlowJo v10.

10X library preparation and sequencing

The concentration of single cell suspensions were manually counted using a hemocytometer and adjusted to 1000 cells/ul or counted by flow cytometry. Cells were loaded according to standard protocol of the Chromium single cell 3' kit in order to capture between 5000 cells/chip position (V2 chemistry). All the following steps were performed according to the standard manufacturer protocol. We used one lane of an Illumina Hiseq 4000 per 10x chip position.

Flow sorting and VEGFA staining

Tissue was obtained from a clear cell renal cell carcinoma immediately post nephrectomy, following histological confirmation. Samples were sliced into approximately 30mm^3 pieces and digested for 15 min at 37°C, with agitation, in a digestion solution containing 25µg/ml Liberase TM (Roche) and 50µg/ml DNase (Sigma) in RPMI (Gibco). The resulting suspension was passed through a 70µm cells strainer (Falcon) and washed with PBS before blocking Fc-Receptors using saturating concentrations of FcR blocking reagent (Miltenyi). Samples were stained with live/dead fixable Aqua (Invitrogen). Antibodies for surface staining were anti-CD3-PerCP-Cyanine5.5 (clone SK7, eBioscience), anti-CD19-BV785 (clone HIB19, BioLegend), anti-CD14-PE-Cy7 (clone 61D3, eBioscience), anti-CD45-APC-eFluor780 (clone HI30, eBioscience), anti-HLA-DR V450 (clone L243, BD Bioscience), anti-CD68-FITC (clone KP1, Dako); and for intracellular antigens anti-VEGFA-PE (clone EP1176Y, Abcam). Cells were then fixed and permeabilised using FoxP3 intracellular kit (eBioscience) for intracellular analysis as per the manufacturers instructions. All samples were acquired on a BD LSR 4 laser Fortessa and data analyzed using FlowJo v10.

Confocal microscopy

Kidney samples were fixed in 1% paraformaldehyde (Electron Microscopy Services) / L-lysine/ sodium periodate (both Sigma-Aldrich) buffer for 24h, followed by 8h in 30% sucrose in P-buffer. 30 um sections were cut on a cryostat, permeabilized and blocked in 0.1M TRIS, containing 0.1% Triton (Sigma), 1% normal mouse serum, 1% BSA (R and D systems). Samples were stained with the appropriate antibodies for 16h at 4°C in a wet chamber then washed 3 times in PBS. Where required, samples were stained with secondary antibodies for 4h at room temperature in a wet chamber and washed 3 times in PBS. They were then stained with streptavidin in PBS for 10 minutes at room temperature, washed 3 times in PBS and mounted in Fluoromount-G® (Southern Biotech). Images were acquired using a TCS SP8 (Leica, Milton Keynes, UK) confocal microscope. Raw imaging data were processed using Imaris (Bitplane).

Antibodies/dilutions for immunofluorescence microscopy

Antibodies and dilutions used are listed in **Table S12**.

Staining of precursor RCC lessions

IHC was performed on FFPE kidney tumor samples from patients with VHL (n=3) [REC 16/WS/0039; and 2002/6486]. 3-μm serial sections mounted on Snowcoat X-tra slides (Surgipath, Richmond, IL) were dewaxed in xylene and rehydrated using graded ethanol washes. For antigen retrieval (CD10 and VCAM-1), sections were immersed in preheated DAKO target retrieval solution (DAKO) and treated for 90 seconds in a pressure cooker. Sections analysed contained both tumor and adjacent normal renal parenchyma acting as an internal control; in addition, substitution of the primary antibody with antibody diluent was used as a negative control. Antigen/antibody complexes were detected using the Envision system (DAKO) according to the manufacturer's instructions. Sections were counterstained with Gill's hematoxylin for 30 seconds, dehydrated in graded ethanol washes, and mounted in DPX (Lamb, London, United Kingdom). Antibodies used were: CAIX (gift from S. Pastorekova, Institute of Virology, Bratislava, Slovak Republic), CD10 (clone: NCL-L-CD10-270, Novacastra) and VCAM-1 (clone: EPR5047, Abcam)

Bulk DNA processing

DNA was extracted from fresh frozen tissue. Peripheral blood DNA was used as a matched normal.

Short insert (500bp) genomic libraries were constructed, flowcells prepared and 150 base pair paired-end sequencing clusters generated on the Illumina HiSeq X platform according to Illumina no-PCR library protocols (*11*). The average sequence coverage was 40X and 98X for renal and matched peripheral blood samples, respectively.

Bulk DNA sequence alignment

DNA sequencing reads were aligned to the GRCh 37d5 reference genome using the Burrows-Wheeler transform (BWA-MEM) (*12*). Sequencing depth at each base was assessed using Bedtools coverage v2.24.0.

Substitution calling from bulk DNA

Single base somatic substitutions were called using an in-house version of CaVEMan v1.11.2 (Cancer Variants through Expectation Maximization) (*13*). CaVEMan compares sequencing reads from tumor and matched normal samples and uses a naïve Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan). Small insertions and deletions (indels) were called using an in-house version of Pindel (v2.2.2; github.com/cancerit/cgpPindel). Point mutation variants were annotated using VAGrENT2 (*14*) according to ENSEMBL version 58. Post-processing filters required that the following criteria were met to call a somatic substitution:

1. At least a third of the reads calling the variant had a base quality of 25 or higher.
2. If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first 2/3 of the read.
3. Less than 5% of the mutant alleles with base quality ≥ 15 were found in the matched normal.
4. Bidirectional reads reporting the mutant allele.
5. Not all mutant alleles reported in the second half of the read.
6. Mean mapping quality of the mutant allele reads was ≥ 21.
7. Mutation does not fall in a simple repeat or centromeric region.
8. Position does not fall within a germline insertion or deletion.
9. Variant is not reported by ≥ 3 reads in more than one percent of samples in a panel of approximately 400 unmatched normal samples.
10. A minimum 2 reads in each direction reporting the mutant allele.
11. At least 10-fold coverage at the mutant allele locus.
12. Minimum variant allele fraction 5%.
13. No insertion or deletion called within a read length (150bp) of the putative substitution.
14. No soft-clipped reads reporting the mutant allele.
15. Median BWA alignment score of the reads reporting the mutant allele ≥ 140.

The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event:

16. Any mutant allele reported within 150bp of another variant.
17. Mutant allele reported in >1% of the matched normal reads.

Copy number detection in bulk DNA

The ascatNGS algorithm (v4.0.1) (*15*) was used to estimate tumor purity and ploidy and to construct copy number profiles prior to running the Battenberg algorithm (v2.2.5) (github.com/cancerit/cgpBattenberg) to allow for tumor subclonality.

Mapping and quantification of single cell RNA-seq

Single cell RNA-seq data were quantified using the 10X software package cellranger (version 1.3.1) to map sequencing data to version 1.2.0 of the build of the GRCh38 reference genome supplied by 10X.

This software produces a table of counts of unique molecular identifiers (UMIs) for each gene in roughly 1 million droplets per cell. Next, we used the EmptyDrops package (https://github.com/TimothyTickle/hca-jamboree-cell-identification/tree/master/src/poisson_model), which identifies droplets that have an expression profile that differs significantly from the expression profile of the unambiguously empty droplet with fewer than 100 UMIs (*16*). We retained for further QC those droplets that either:

- Were identified by EmptyDrops as having expression significantly different from the background with 5% FDR.
- Had total number of UMIs in excess of the minimum of:
    - 100,000.
    - The UMI cut-off identified by cellranger.
    - The UMI cut-off identified by the findKnee function in EmptyDrops.

This method was used in preference to relying on the cellranger UMI cut-off alone (or the findKnee cut-off, which produces almost the same value for our data) as we found that there were a number of biologically plausible clusters of cells, which passed all our QC filters and were only identified by the EmptyDrops method.

Quality control of single cell data

To filter lower quality cells, we removed any cells that had greater than 20% expression originating from mitochondrial genes or expressed fewer than 200 distinct transcripts. We used more liberal cut-offs than studies involving tissue that could be collected and prepared in ideal conditions (e.g., model organism studies such as (*17*)) in order to capture as broad a range of cell types as possible.

Because our base line QC filters were fairly permissive, we further filtered our data by removing any clusters that had hemoglobin genes as one of its cluster defining genes. This was done on the basis that such cells were either reticulocytes (and not of biological interest) or were heavily contaminated. Finally, we removed clusters with biologically implausible combinations of markers that were suggestive of having captured doublets.

In constructing a reference map of the mature and fetal kidneys and comparing this map to the tumor cells, we ignored any cluster that could not be unambiguously assigned using well established marker genes. The full characterization and identification of marker genes for the all clusters (ambiguous and unambiguous) can be found in **Fig. S2 to S8** and **Table S2**.

Normalisation of single cell data

Data was first normalized for sequencing depth by dividing by the total number of UMIs in each cell and then transformed to a log scale for each cell using the Seurat (http://satijalab.org/seurat/) NormalizeData function. That is, the transformed data, y, is given by:

$$y_{gc} = \log\left(1 + F\frac{x_{gc}}{\sum_g x_{gc}}\right)$$

where x is the UMI count matrix with g indexing gene and c indexing the cell. F is the Seurat "scale.factor" parameter (which we left at the default value of 10,000).

Batch correction was performed using COMBAT (*18*) to regress out the variability introduced by individual 10X channels. To prevent the batch correction from imputing expression from genes that have no expression, we forced all entries which were 0 before batch correction to remain 0 post batch correction. Following this correction, we re-normalized the data such that it was consistent with being derived from an expression vector which sums to 1. That is, after re-normalization the data have the property that:

$$\sum_g (e^{y_{gc}} - 1) = F$$

where y is the final data matrix and cell.

Feature selection and dimension reduction

Following normalization, we identified genes with high variability using the Seurat FindVariableGenes function. This function calculates the mean expression and dispersion for each gene, then groups genes into bins (of size 20) by their mean expression and identifies any gene for which the z-score calculated from the dispersion exceeds some cut-off. We used a permissive cut-off of z=0.5 and mean expression in the range 0.0125 to 3 to ensure a large number of biologically variable genes were captured.

The normalized data was scaled to have mean 0 and standard deviation 1 and principle component analysis was performed using the variable genes identified together with any gene that we identified as being potentially biologically interesting (regardless of its variability in the data).

We then used the first 30 principle components (PCs) to construct a two-dimensional representation of the data using tSNE (*19, 20*) with perplexity 30 and using the Barnes-Hut approximation with opening angle theta=0.5 to speed up the calculation. This representation was then used only to visualize the data.

Cluster identification

Clusters were identified using the community identification algorithm as implemented in the Seurat "FindClusters" algorithm (*21*). We used the first 30 PCs as input to this method and set the resolution parameter to 1. We chose this value of the resolution parameter as it produced a number of clusters that was large enough to capture most of the important biological variability but not so large as to make detailed manual scrutiny of each cluster impractical. All other parameters were set to the function defaults.

Allocation of cells to cellular compartments

Cells were first split into tumor, normal and fetal based on the tissue of origin. Within each of these three groups, cells were initially assigned to either immune or epithelial and vascular based on the average PTPRC (CD45) expression within each cluster. Clusters of cells were then

annotated within each of these compartments and clusters of cells that were deemed to be misplaced based on marker gene expression were moved to the correct compartment.

In the case of the normal mature kidney, we found that the majority of the cells (80%) were proximal tubular cells of indistinct type. In order to obtain sufficient detail to identify all cell types, we created a further sub-compartment that excluded this mass of proximal tubular cells (see **Fig. S2**). This sub-compartment (**Fig. 1, Fig. S2**) allowed us to capture the full diversity of normal cell types present in the normal epithelium and vasculature.

Annotation of clusters

Annotation of clusters to cell types was done by manual inspection of the genes defining each cluster and comparison to the literature. To identify which genes were most important in defining each cluster we ranked genes by an adaption of the "tf-idf" metric widely used in natural language processing (*22*). From this ranked list, we inspected the minimum of 100 or all those genes for which the genes presence was enriched within a cluster relative to all other clusters in its cellular compartment (e.g., cluster 7 in the tumor immune map versus all cells in the tumor immune map not in cluster 7) by a hypergeometric test with a 0.01 p-value cut-off after multiple hypothesis correction.

The purpose of the tf-idf ranking was to provide a computationally efficient way of identifying all those genes with expression specific to a cluster. The tf-idf value was calculated for each gene as,

$$-f_{gc}\log(F_g)$$

where f is the fraction of cells in cluster c with non-zero expression of gene g and F is the fraction of cells in the entire map expressing gene g. Note that the p-values derived from the hypergeometric test are mathematically equivalent to calculating p-values by permuting the labels and recalculating the tf-idf score an infinite number of times.

Pseudo-time analysis

To place cells onto pseudo-time trajectories, cells were re-processed using the monocle R package (*23*) by excluding all genes expressed in fewer than 3 cells and normalizing for library size using the "estimateSizeFactors" function.

Following this, negative binomial over-dispersion was estimated for each gene using the "estimateDispersions" function. Dimension reduction was then performed using the DDRTree method on genes selected to have a mean expression value > 0.1 and variance greater than the empirical dispersion (the best fit mean-dispersion trend-line).

Finally, cells were placed onto a pseudo-time trajectory using the "orderCells" function (*24*).

Identification of genes differentially expressed along pseudo-time trajectory

The pseudo-time trajectory inferred from the cells comprising the mesenchymal population of the developing nephron had minimal branching structure and so we considered only the distance from the root node (representing the cap mesenchyme) for the purpose of identifying changes in expression pattern.

To identify genes which changed steadily along the identified trajectory we performed a likelihood ratio test for a negative binomial model with and without a term given by spline smoothing of the pseudo-time using the "differentialGeneTest" function. We restricted our test to the 1691 transcription factors identified by the AnimalTFDB project (*25, 26*) and identified all genes that were significant at a strict 0.01 p-value cut-off after multiple hypothesis correction.

For trajectories that exhibited significant branching structure, where testing for differences along a linearized pseudo-time axis would not be valid, we instead tested for differences along each branch relative to the other branch and the root. These statistical tests were performed using the BEAM algorithm implemented in monocle and we retained any transcription factor (from the same list as above) with a p-value less than 0.01 after multiple hypothesis correction.

Differential expression between ureteric bud and cap mesenchyme/primitive vesicle

To identify genes differentially expressed between ureteric bud and cap mesenchyme/primitive vesicle we labelled all cells belonging to these populations as either ureteric bud (UB) or cap mesenchyme/primitive vesicle (CM/PV). The same cut-offs and test (negative binomial model based likelihood ratio test) as was used to identify transcription factors differentially expressed along pseudotime were used to compare a model with and without this label as a covariate and to identify differentially expressed transcription factors distinguishing these two populations.

Cell type similarity inference

To measure the similarity of tumor cells to either normal mature kidney cells or fetal cells, we trained a logistic regression model using elastic net regularization on the cellular identities defined by the clusters in normal mature and fetal epithelial and vascular compartments. In training this model we set alpha=0.99 to produce strong regularization but to prevent the exclusion of strongly co-linear genes.

To obtain regression coefficients specific to each cluster in our training data we fit a series of N binomial logistic regression models, where N is the number of clusters in the training data (i.e., one-versus-rest binomial logistic regression). To prevent the observed frequencies of cells (which we do not expect to accurately reflect the true abundances in situ) from biasing the regression coefficients we use an offset for each model given by,

$$\log\left(\frac{f}{1-f}\right)$$

where f is the fraction of cells in the cluster being trained.

In each case, we performed 10-fold cross validation and selected the regularization coefficient, lambda, to be as large as possible (i.e., as few non-zero coefficients as possible) such that the cross validated accuracy was within 1 standard deviation of the minimum.

These models were then used to calculate a predicted similarity to each of the fetal and normal clusters for each cell in the tumor map. In calculating the predicted values, an offset of 0 was used. Softmax normalization was not used to allow for the possibility that tumor cells do not resemble any of the fetal or normal cells in the training set. Predicted logits were then averaged within each tumor cluster and converted to probabilities for visualization.

This approach was implemented using the "glmnet" R package (*27*) .

Identification of tissue specific genes

To identify genes that were specific to one cell type and no other in the kidney (e.g., that NMU is only expressed in ureteric bud in the fetal kidney), we utilized a series of cut-offs on fraction of cells expressing a given gene of interest. Specifically, we identified all genes for which the percentage of cells expressing a gene in the target cluster (the cluster the gene is specific for) was 20% higher than all other clusters and that no other cluster had more than 10% of cells expressing this gene. To avoid intermediate populations in the developing nephron excluding useful genes, we required for this population only that a candidate gene be expressed most frequently in the target population and that no other non-nephrogenesis cluster had more than 10% of cells expressing the candidate gene.

Differential expression of marker genes in bulk data

To measure the expression of cell specific markers in bulk TCGA and TARGET data, a table of counts was retrieved from TCGA (*28, 29*) and TARGET (*30, 31*) using the R TCGAbiolinks package (*32*). We filtered these samples to retain only primary tumors from Wilms, malignant Rhabdoid tumors (MRT), ccRCC, pRCC, chromophobe RCC (chrRCC) or normal tissue biopsies. These counts were then normalized to produce log (TPM) for each gene/sample pair for visualization.

To determine differentially expressed transcripts between different groupings of tumor and normal, we used a Mann-Whitney test to compare the distribution of log2 transformed TPM values between the two groups and kept genes with a p-value less than 0.01. For a gene to be considered differentially expressed, we further required that the median log2(TPM) values of the two distributions differ by at least 1.

Genotyping of tumor cells

Genomic variation present in individual cells is captured by 10X single cell RNA-seq. However, the regions of the genome which are covered is limited by the fact that only highly expressed genes are captured, and within the genes that are captured the sequence is enriched for 3' mRNA. To overcome these coverage limitations and detect genomic variation in single cells we aggregated information across as large a genomic region as possible within each cell.

Rather than attempt to detect genomic changes *de novo*, we utilized bulk sequencing of tumor DNA and peripheral blood to identify the dominant clonal genomic changes present in each cancer. In the case of single base substitutions, we considered only those changes with a VAF of at least 20% in the tumor to enrich for clonal substitutions. We then calculated a global "Mutant Allele Frequency" (MAF) for each cell by aggregating the number of UMIs supporting the mutant and wild type alleles across all substitutions in the genome. For each molecule we determined the allele by taking the consensus base across all the reads covering that base with the same UMI. If there was no consensus call, the UMI was not included in the calculation of the MAF. We excluded any variant that did not fall within a gene body.

To detect the presence of clonal copy number (CN) changes, we considered only clonal losses of heterozygosity (LOH), except for the pRCC tumor where no LOH events could be found. In the case of the pRCC, we instead used the 2-to-1 gains on chromosomes 3,12, and 17. In each patient we identified all SNPs heterozygous in the peripheral blood in the regions where the clonal CN changes occur using a B allele frequency cut-off of 0.4 to 0.6. We then phased these SNPs by designating the allele with the highest B allele frequency in the tumor (i.e., the allele on the gained or retained allele) as the "Major" allele. We considered only those alleles for which the B allele frequency in the tumor was significantly different from 50/50 as measured by a binomial test followed by multiple hypothesis correction.

For all samples, we used a corrected p-value cut-off of 0.1, except for RCC1 where we used a cut-off of 0.5. We elected to use a weaker cut-off for RCC1 as it had a much higher level of genomic contamination and the stronger cut-off left us with too few informative SNPs to call CN changes. The effect of using a weaker cut-off is to mis-phase a larger number of SNPs, which pulls the signal in the single cell data away from the expected allelic ratio, but does not remove it entirely.

Next, we calculated the frequency of the major allele across all informative SNPs from the single cell data on a cell-by-cell basis. In each case we aggregated the information across all informative SNPs in the genome to calculate one allele frequency for each cell. To prevent allele specific expression cofounding our signal, we excluded any gene that showed consistent expression of one allele across all cell types (i.e., including those from normal tissue we expect to be genotypically normal) as measured by an allele frequency greater than 0.7 (or less than 0.3) and a corrected p-value from a binomial test of less than 0.05.

These two approaches provided two independent methods of measuring the presence of clonal genomic variations in individual cells. To provide further evidence of the degree to which clusters of cells identified by their single cell transcriptomes carried the clonal tumor genotype, we aggregated counts across groups of cells to calculate the mutation and CN allele frequency for each cluster of cells (**Fig S11-S14)**

Comparison of fetal and Wilm's tumor TF program

To test if the transcription factors (TFs) that define the different Wilm's tumor cell types overlap significantly with the transcription factors that define normal nephron development we

extracted all TFs that were significantly altered along the pseudo-time trajectory in fetal nephron development or along the branched trajectory defined by Wilm's tumor cells. A hypergeometric test was then performed to compare the null hypothesis of no enrichment of the fetal group of TFs for the Wilm's TFs against a background of 1691 TFs (*25, 26*).

Doublet mitigation strategy

Droplet-based technologies are known to produce "doublets", where two cells are captured within a single droplet. At the concentrations of cells used in our experiment, we expect the stochastic doublet rate to be very low (~1%) (*33*). However, we cannot exclude the possibility of "biological doublets", where cells have a propensity to be captured together due to mutual affinity or failure to dissociate. At present, we know of no informatic tools to reliably remove such doublets. We controlled for this problem by removing clusters of cells with a transcriptomic signature of multiple cell types, but cannot exclude the possibility that we discarded interesting populations of cells in the process.

Limitations of single cell assay

The aim of this experiment was to provide an unbiased survey of the cell types comprising human kidney (in health and malignancy) at a molecular level. For this reason, we chose to use an assay that maximised the number of cells we were able to profile (droplet-based sequencing). The trade-off of this approach is that in each cell, a fraction of the transcriptome is not captured. We believe this approach is justified for a broad survey such as ours as most cell types can be unambiguously distinguished by the presence of only a handful of genes (we were able to distinguish all parts of the nephron using ~30 genes). To further compensate for this limitation, we have limited ourselves to making statements about groups of cells. By averaging across the cell-to-cell transcriptomic stochasticity, a more complete picture of each cell type's transcriptome is obtained. However, we cannot exclude the possibility that an assay with greater transcriptomic coverage would allow for greater segregation of the cell types identified here.

Although at present our approach probably represents the "gold standard" for surveying tissues, it is likely that systematic biases against capturing certain cell types exist. For example, we found no convincing fibroblast population in mature kidneys, whilst in ureter and fetal kidneys these were consistently seen. This may be due to the rarity of fibroblasts in mature kidney or due to technical biases preventing isolation of fibroblasts from mature tissue. Similarly, WT1 positive podocytes were more prevalent in kidneys of the two youngest children in our data, leading us to hypothesise that WT1 podocyte abundance decreases with age. However, immunohistochemical staining showed this population to be present across a range of ages.

Data availability

Raw sequencing data have been deposited in the European Genome-phenome Archive (EGA) under study IDs EGAS00001002171, EGAS00001002486, EGAS00001002325, EGAS00001002553 and EGAS00001002534. Sample specific identifiers can be found in **Table**

S6. The raw table of counts for all droplets identified as potentially containing cells (125,139) is included as a supplementary data file **Data S1**.

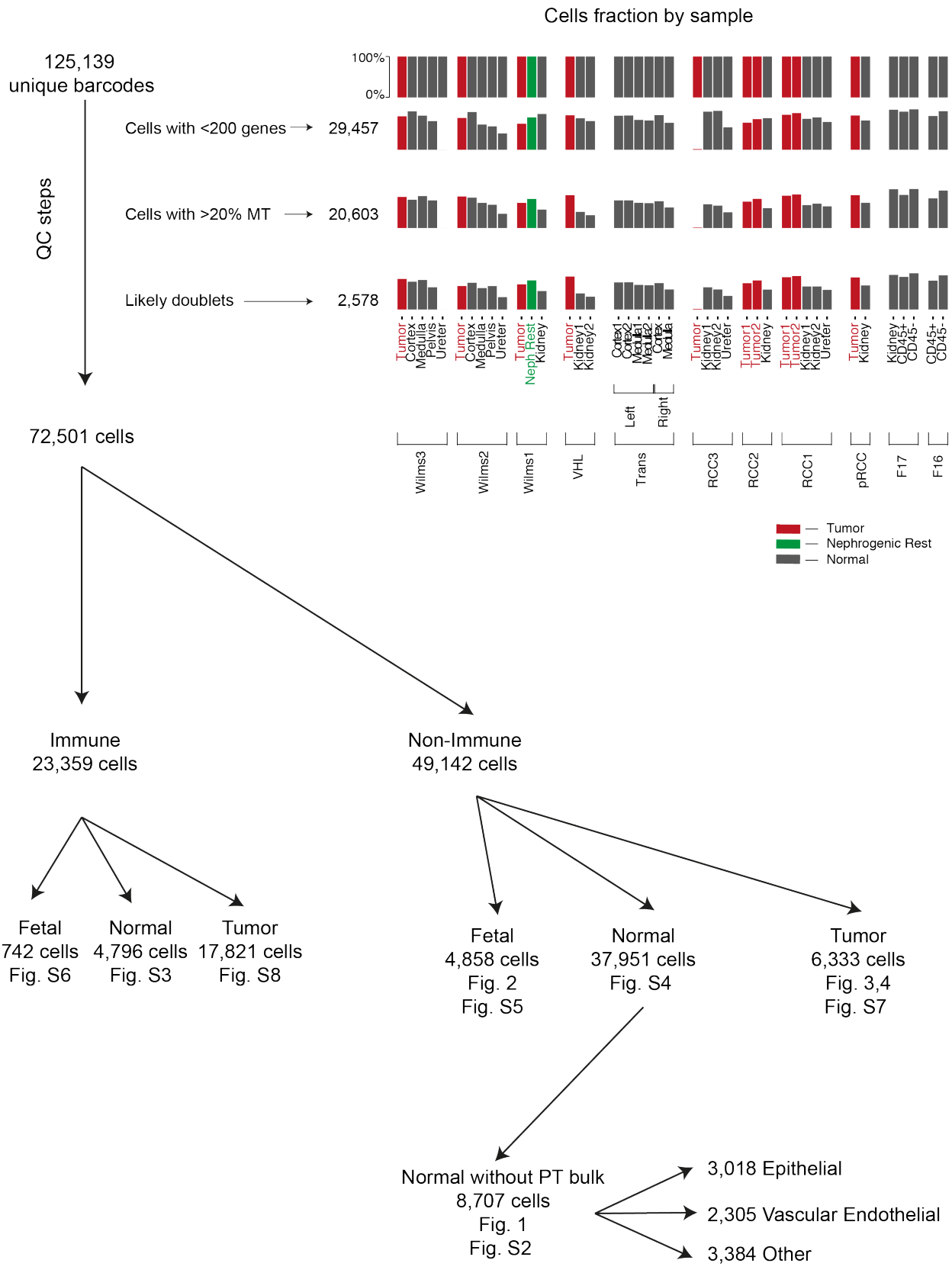

Cells fraction by sample
125,139
unique barcodes
QC steps
Cells with <200 genes
29,457
Cells with >20% MT
20,603
Likely doublets
2,578
72,501 cells
Tumor
Nephrogenic Rest
Normal
Immune
23,359 cells
Non-Immune
49,142 cells
Fetal
742 cells
Fig. S6
Normal
4,796 cells
Fig. S3
Tumor
17,821 cells
Fig. S8
Fetal
4,858 cells
Fig. 2
Fig. S5
Normal
37,951 cells
Fig. S4
Tumor
6,333 cells
Fig. 3,4
Fig. S7
Normal without PT bulk
8,707 cells
Fig. 1
Fig. S2
3,018 Epithelial
2,305 Vascular Endothelial
3,384 Other

Figure S1. Overview of droplet quality control and cell allocation to compartments

Flow chart indicating how many candidate droplets are removed by quality control and how the resulting 72,501 cells are divided into different compartments. The bar plot on the right shows the fraction of droplets remaining after each QC step for each biopsy in our experiment.

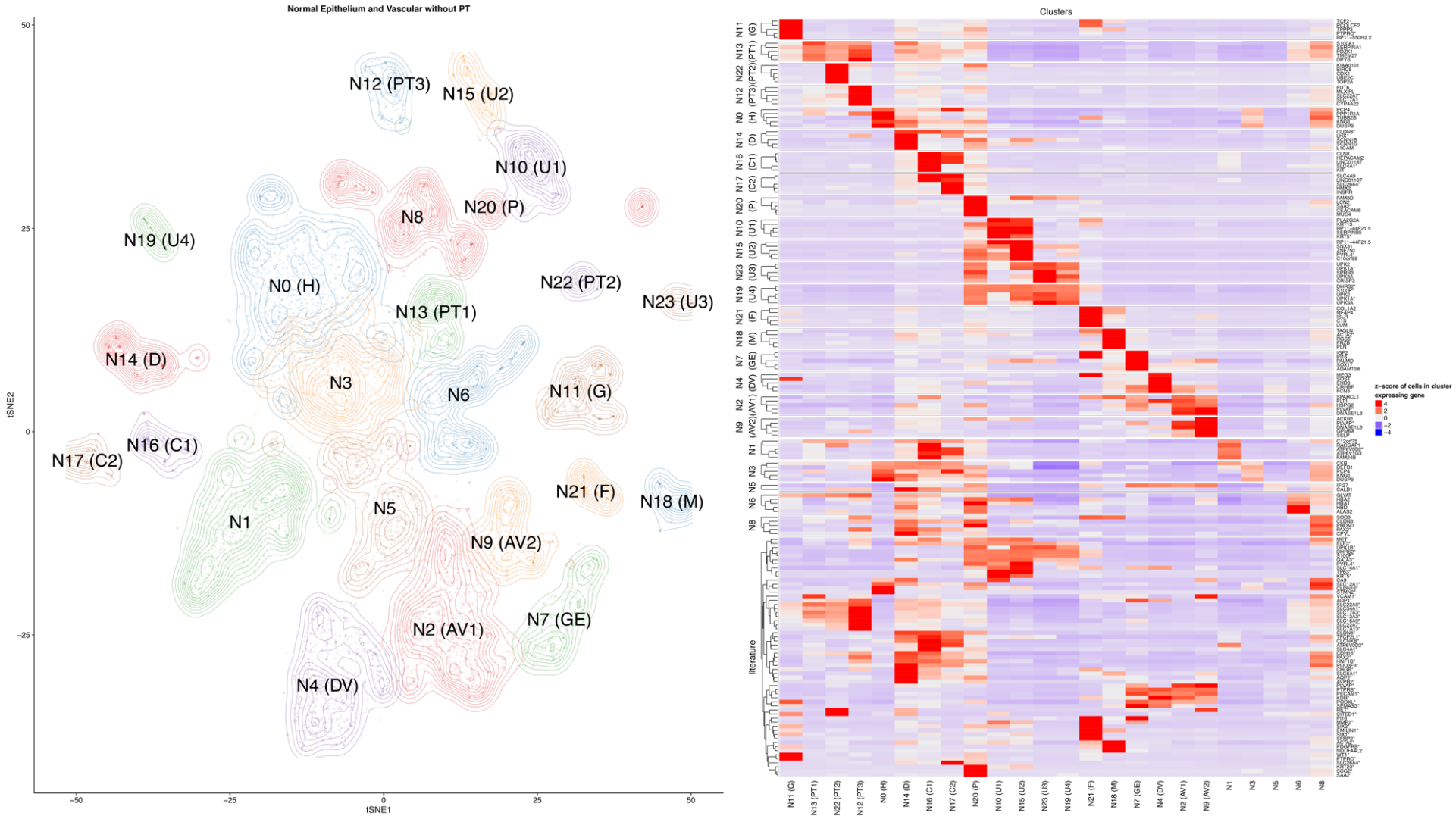


Figure S2. Full characterization of non-immune mature kidney cells

The left panel shows a tSNE representation of the 8,707 cells in the normal mature kidney map (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following cell types: “G” for Glomerulus, “PT1” for convoluted proximal tubular, “PT2” and “PT3” for straight and convoluted proximal tubular, “H” for loop of Henle, “D” for distal tubules, “C1” for Type B collecting duct, “C2” for Type A collecting duct, “P” for renal pelvic epithelium, “U1” for basal and intermediate ureter, “U2” for mixed ureter, “U3” and “U4” for superficial cells, “F” for fibroblasts, “M” for myofibroblasts, “GE” for glomerular endothelium, “DV” for ascending vasa recta and “AV1” and “AV2” for ascending vasa recta.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Cluster are organized from proximal to distal as in **Fig. 1C,D** with ambiguous clusters on the far right. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a “*” in the row labels on the right.

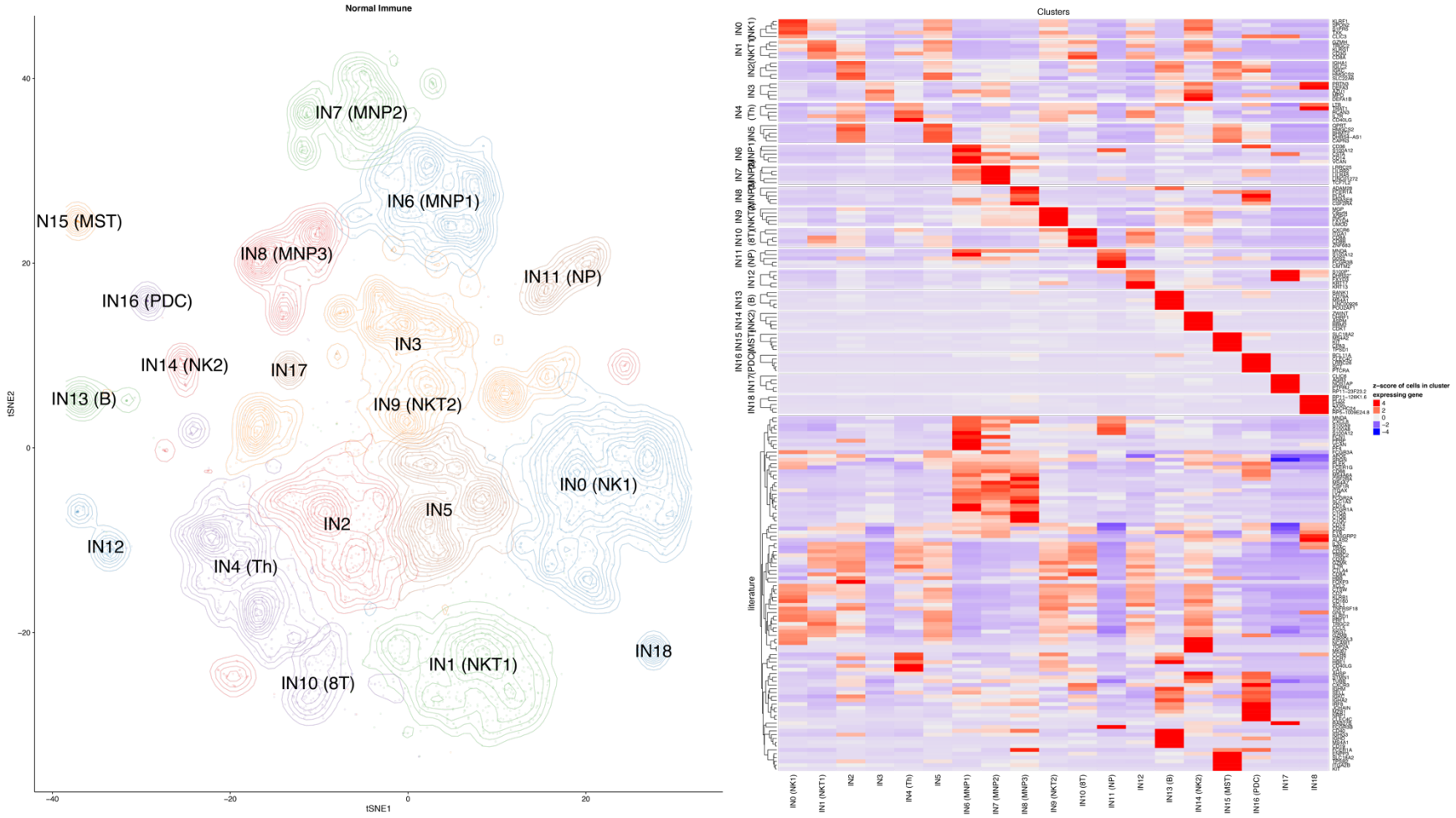


Figure S3. Full characterization of immune cells in the mature kidney

The left panel shows a tSNE representation of the 4,796 immune cells in the normal mature kidney map (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following cell types: “NK1” and “NK2” for natural killer cell, “NKT1” and “NKT” for natural killer T-cell, “Th” for T helper cell, “MNP1-3”

for Mononuclear phagocyte, “8T” for CD8+ T-cell, “NP” for neutrophil, “B” for B-cell, “MST” for Mast cell and “PDC” for plastmacytoid dendritic cell.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a “*” in the row labels on the right.

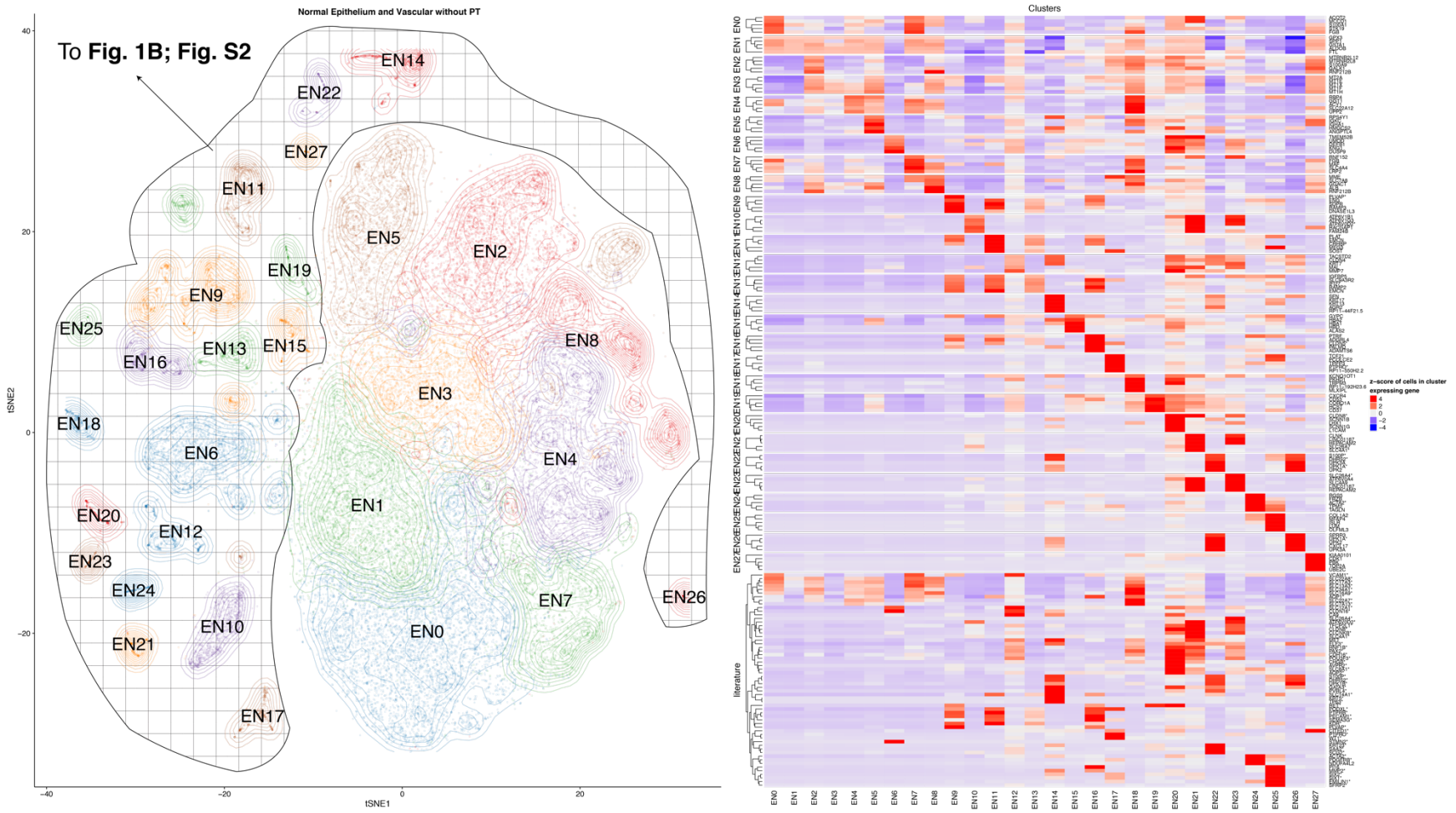


Figure S4. Full characterization of all non-immune mature kidney cells

The left panel shows a tSNE representation of the 37,951 non-immune cells in the normal mature kidney map (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The tSNE map is annotated to indicate which clusters of cells are retained for the reference map of the non-immune mature kidney.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a "*" in the row labels on the right.

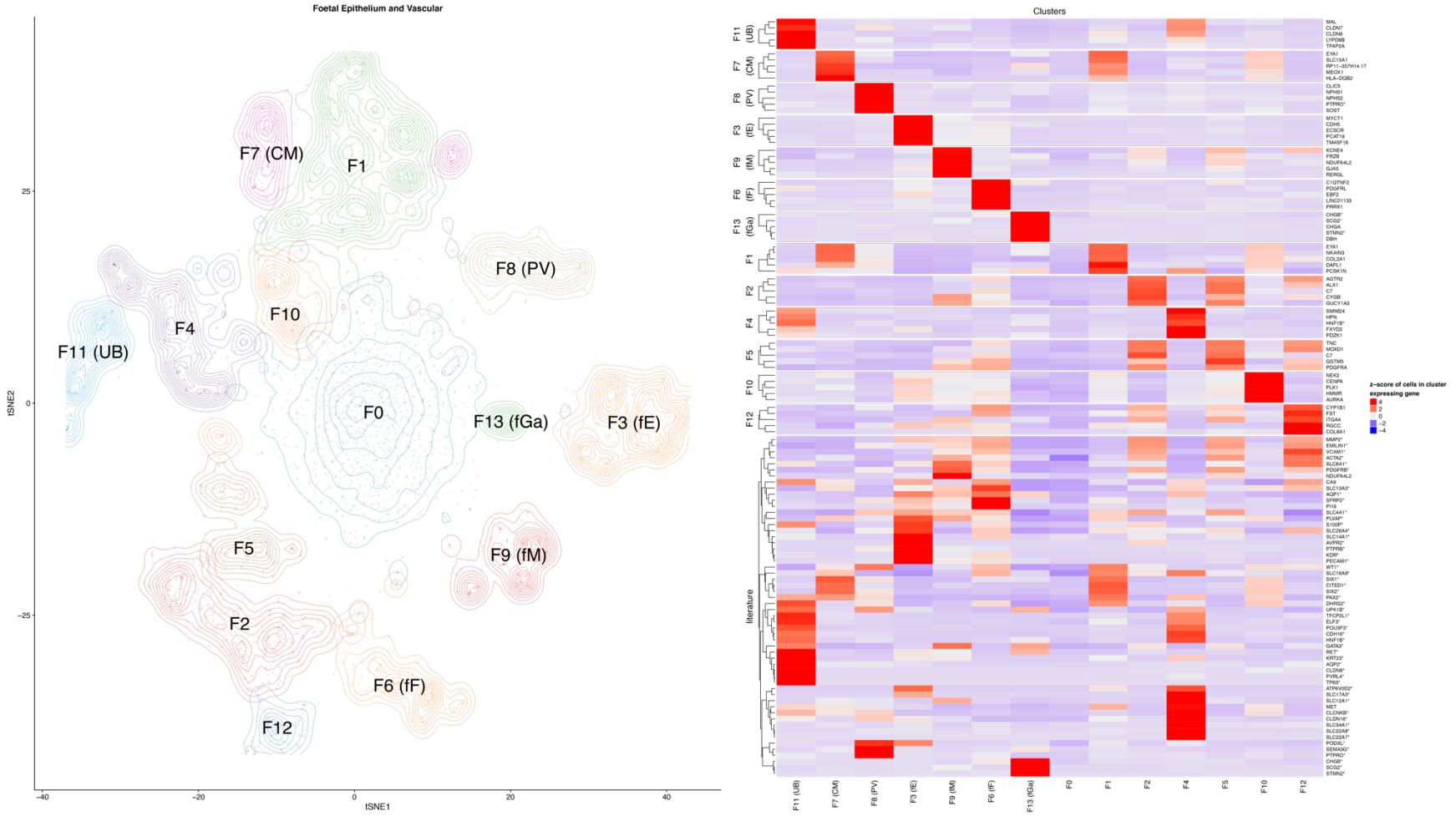


Figure S5. Full characterization of non-immune fetal kidney cells

The left panel shows a tSNE representation of the 4,858 non-immune cells of the normal fetal kidney map (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following

cell types: “fE” for endothelium, “fF” for fibroblasts, “fM” for myofibroblasts, “fGA” for ganglia, “UB” for ureteric bud, “CM” for cap mesenchyme and “PV” for primitive vesicle.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a “*” in the row labels on the right.

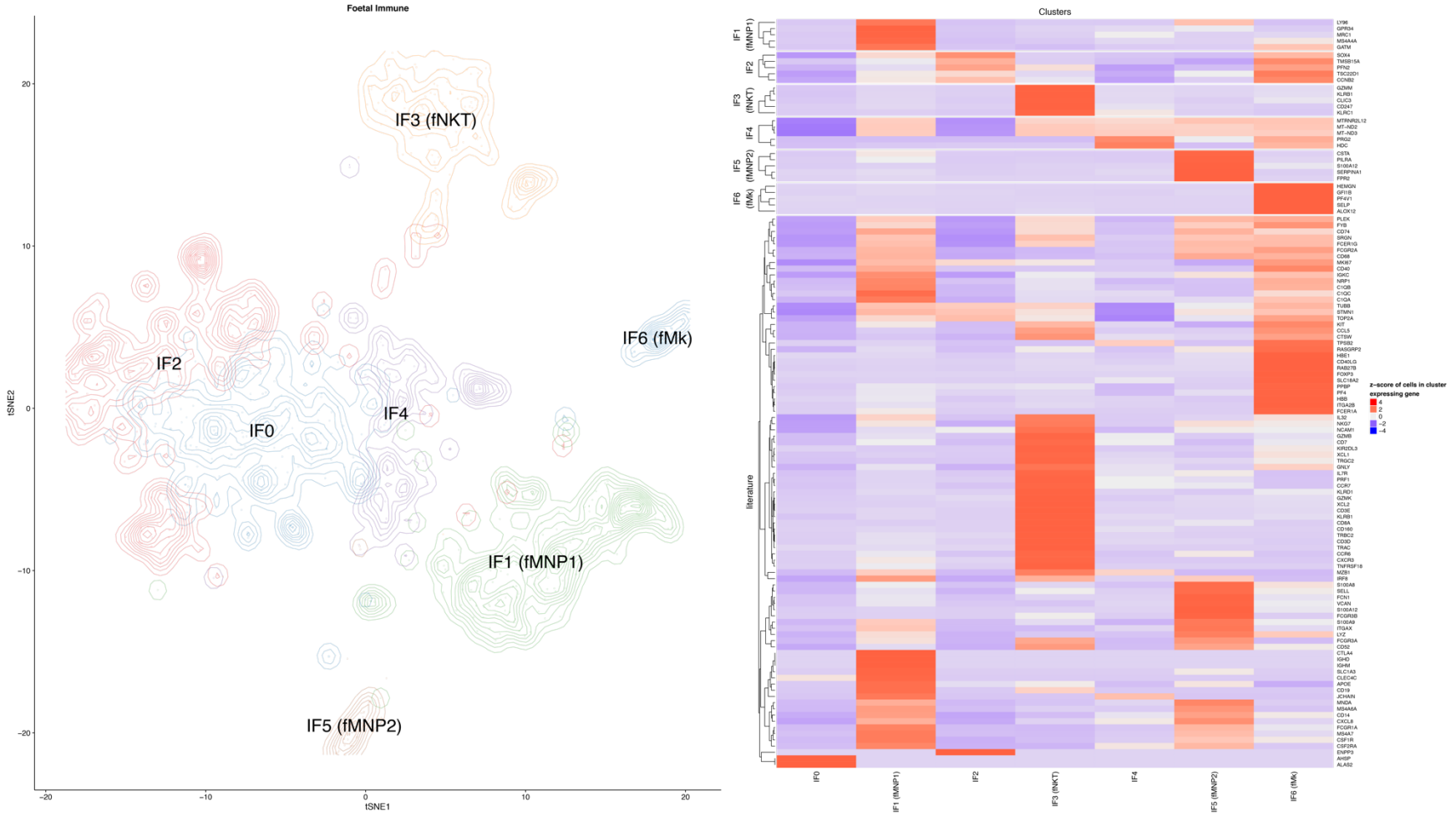
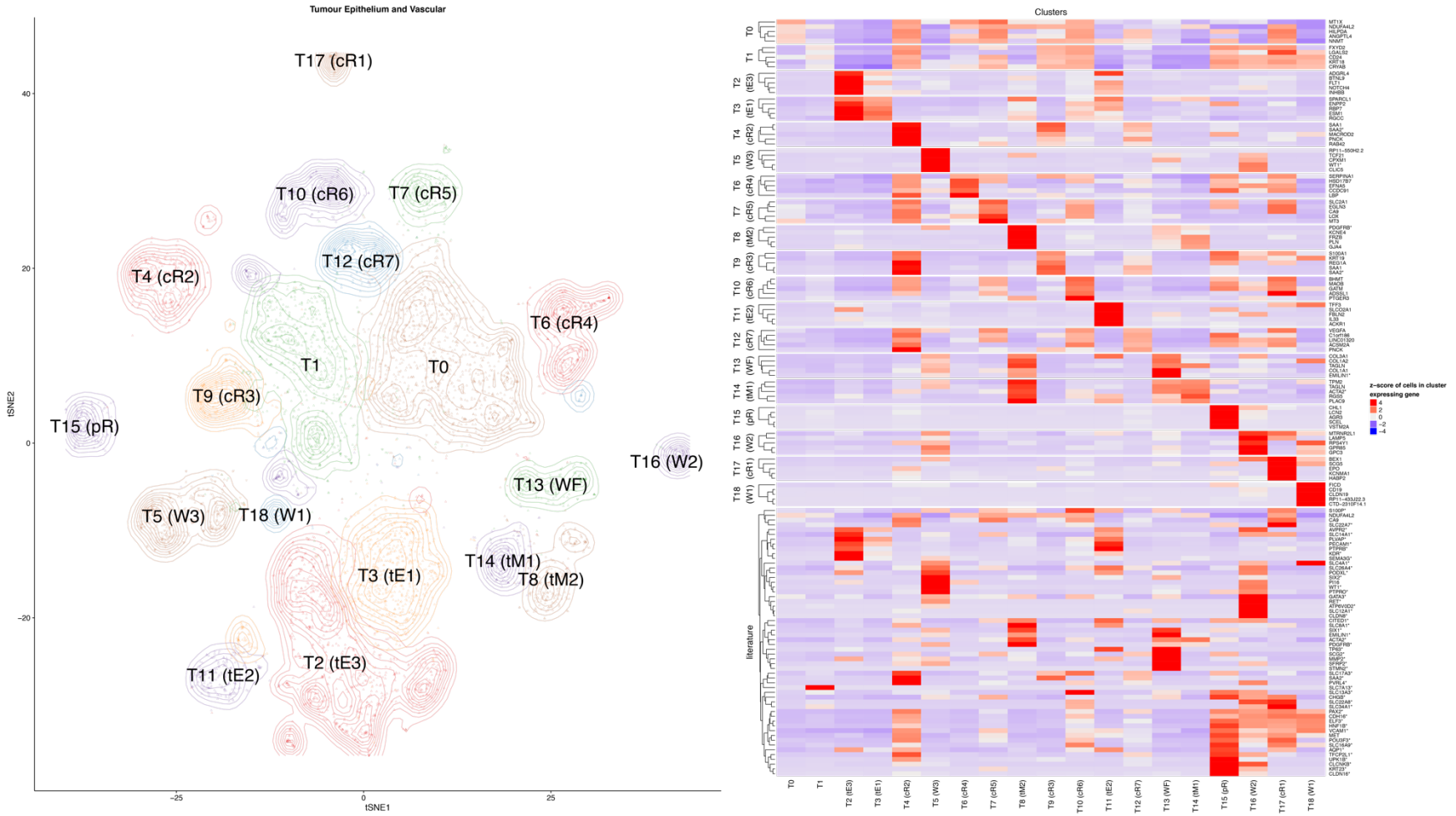


Figure S6. – Full characterization of immune cells in the fetal kidney

The left panel shows a tSNE representation of the 742 immune cells in the normal mature kidney map (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following cell types: “fMNP1” and “fMNP2” for mononuclear phagocytes, “fNKT” for natural killer T cells and “fMk” for megakaryocytes.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a "*" in the row labels on the right.

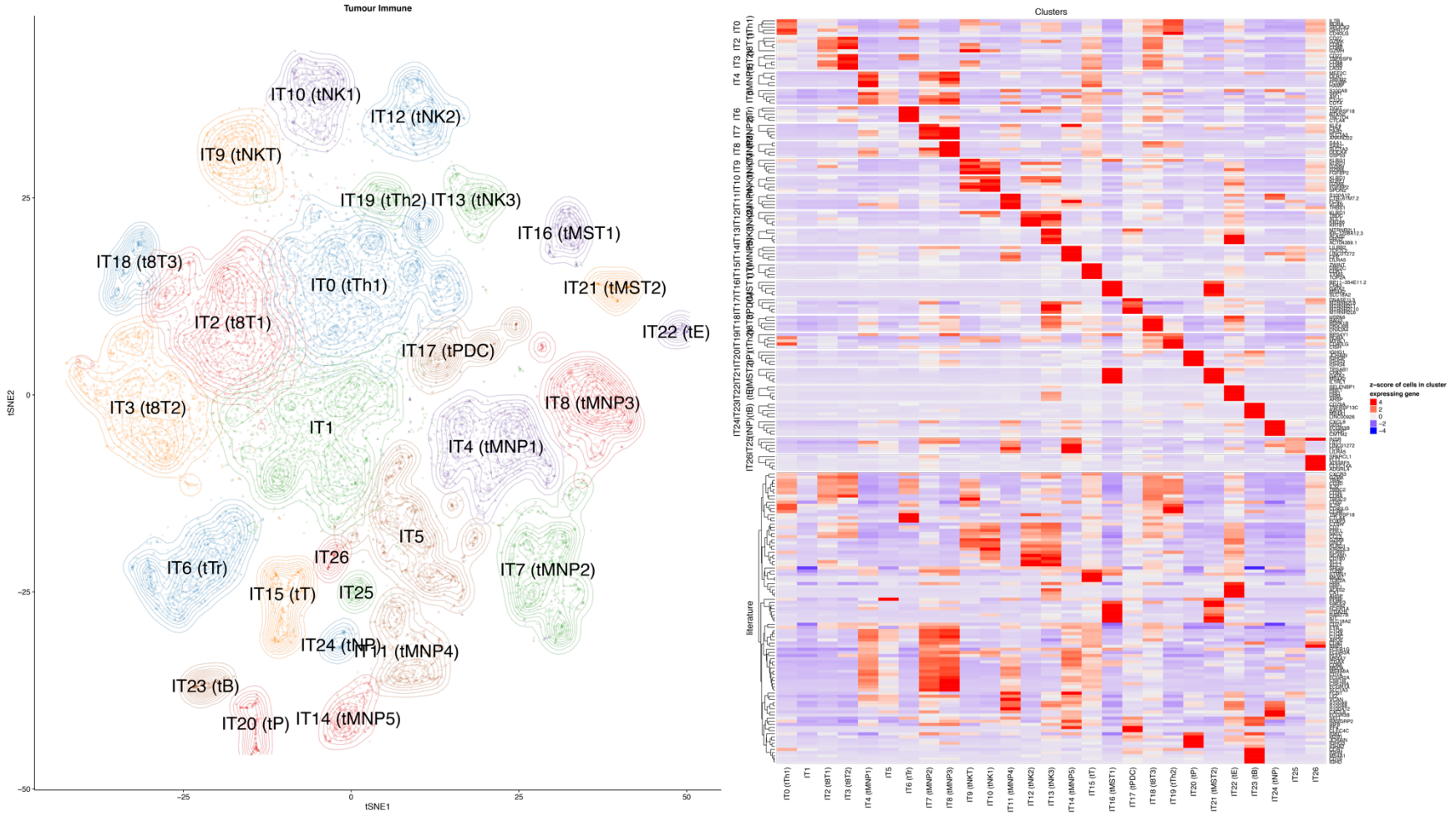


Supplementary Figure S7. Full characterization of non-immune kidney tumor cells

The left panel shows a tSNE representation of the 6,333 non-immune cellsfrom all kidney tumor biopsies (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following

cell types: “tE1-3” for endothelium, “cR1-7” for ccRCC, “W1-3” for Wilms’ tumor, “tM1” and “tM2” for myofibroblasts, “WF” for Wilms’ tumor and fibroblasts and “pR” for ppRCC.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified as being markers of different types of kidney cell in the literature are marked with a “*” in the row labels on the right.

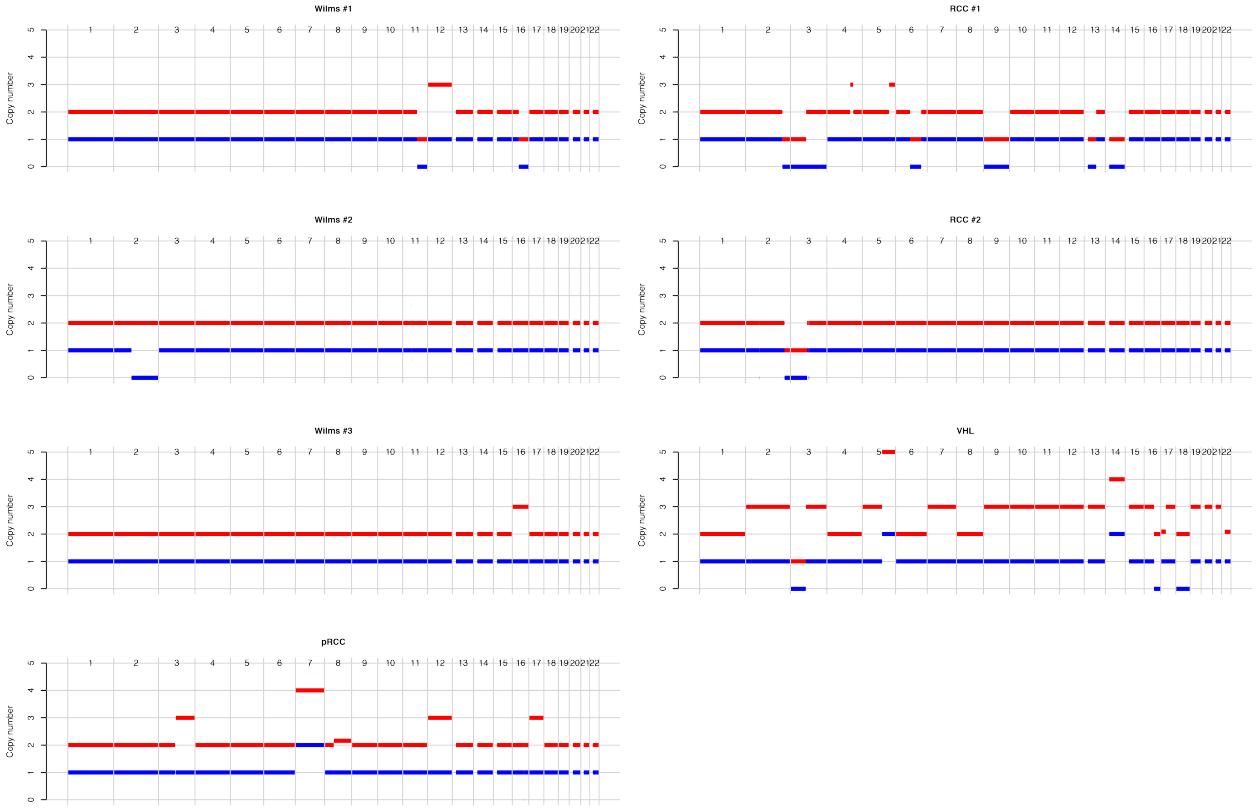


Supplementary Figure S8. Full characterization of kidney tumor immune cells

The left panel shows a tSNE representation of the 17,821 immune cells from all kidney tumor biopsies (see **Fig. S1**), with clusters labeled (with aliases in brackets where applicable), colored and emphasized with density contours. The aliases represent the following cell types: “tTh1-2” for T helper cells, “t8T1-3” for CD8+ T-cells, “tMNP1-5” for mononuclear phagocytes, “tTr” for T regulatory

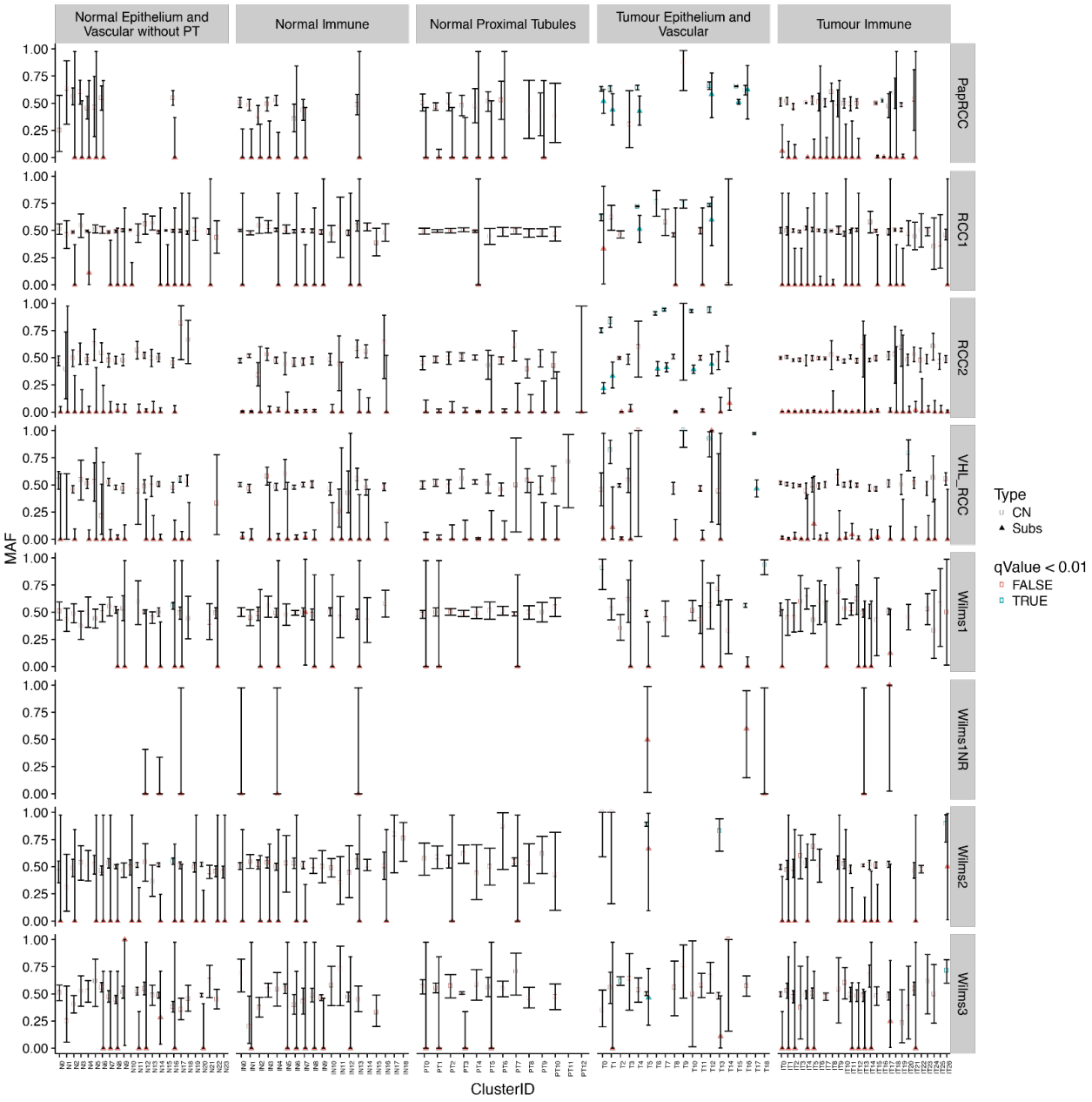
cell, “tNKT” for natural killer T-cells, “tNK1-3” for natural killer cells, “tT” for proliferating T-cells, “tMST1-2” for mast cells, “tPDC” for plasmacytoid dendritic cells, “tP” for plasma cell, “tE” for erythroblast, “tB” for B-cell and “tNP” for neutrophils.

The right panel shows the expression of the top 5 most specific marker genes for each cluster (see **Methods**) and the expression of canonical markers curated from the literature (see **Table S3**). Within each group of genes, rows are hierarchically clustered and expression (fraction of cells expressing the gene in each cluster) is row normalized to have mean 0 and standard deviation 1. Genes that are identified in the literature as being markers of different types of kidney cell in the literature are marked with a “*” in the row labels on the right.



Supplementary Figure S9. Copy number profiles of tumors from bulk DNA

Copy number profiles for each tumor as determined by Battenberg on match tumor/normal whole genome sequencing (see **Methods**). The red line indicates the total copy number and the blue line the copy number of the minor allele.



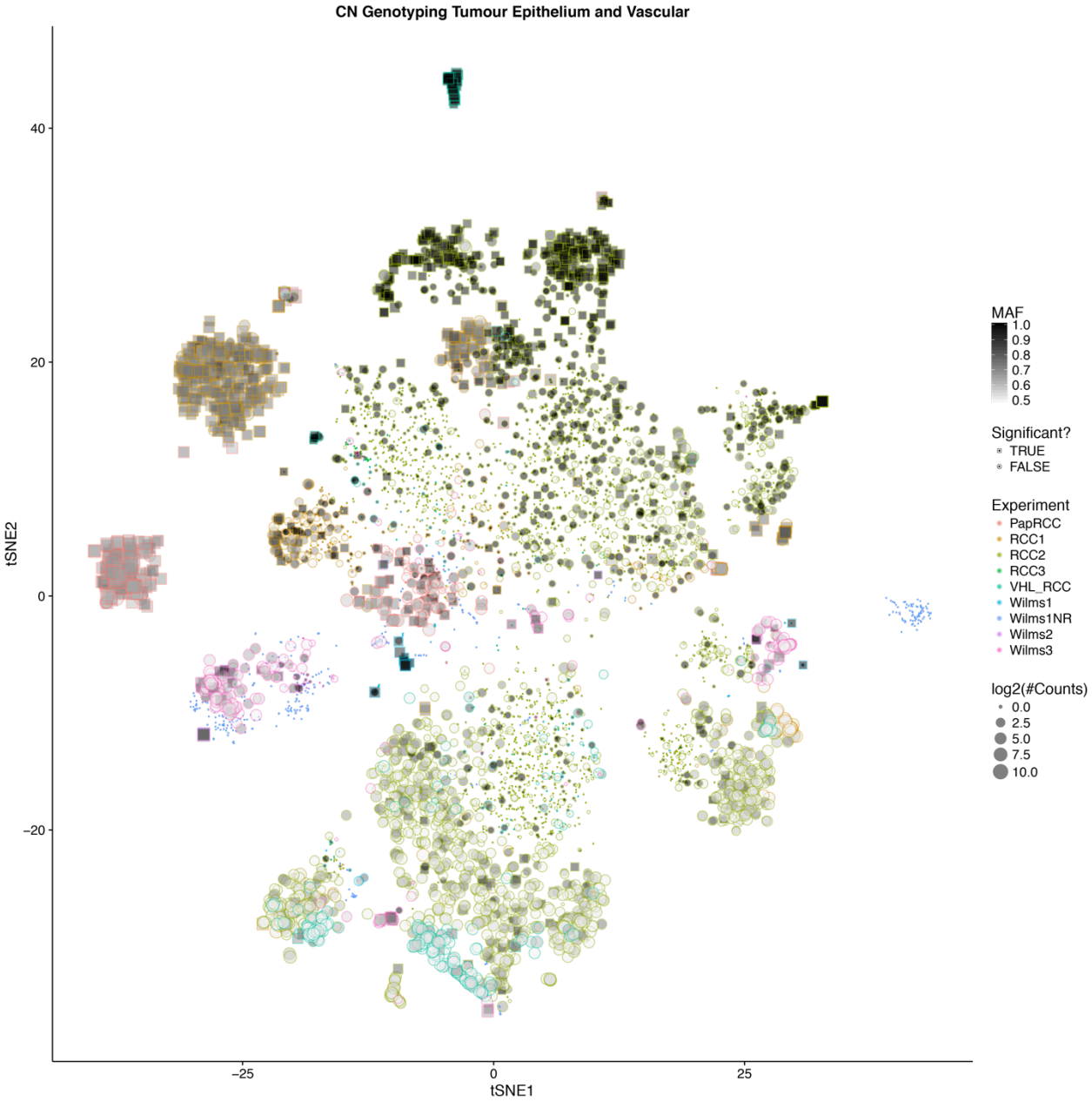
Supplementary Figure S10. Overview of the genotyping of individual cells using single cell RNA-seq data

A plot summarizing the evidence supporting the presence in clusters of single cells of the somatic genomic changes (CN changes and single nucleotide variants) identified as clonally present in each tumor using bulk DNA (see **Methods**). Each column shows a different cellular compartment (i.e., a different grouping of cells as shown in **Fig. S1**). Within each column, each entry on the x-axis represents an individual cluster. Within each cluster, the cells derived from each of the different tumor samples are compared to the somatic genotype for their corresponding tumor, which is what each row represents. For example, looking in the second column (Normal Immune), at the third row (RCC2) at the first entry (N0) summarizes the

evidence for those cells in cluster N0 in the normal immune map (**Fig. S3**) that are taken from the normal biopsy from patient RCC2 having the clonal CN change and single nucleotide variants found in the DNA of the tumor biopsy taken from RCC2.

For each cluster/sample combination, two points are shown: a triangle representing substitutions and a circle representing CN. For the substitutions, the mutant allele frequency (MAF) is calculated as the total fraction of reads supporting the mutant allele derived from the DNA, summed across all substitutions in the genome and all cells in this grouping. If the cells in this cluster are not tumor cells, we expect a MAF of close to 0. For the CN, the MAF is calculated as the fraction of reads supporting the major allele, which is defined to be the allele of heterozygous SNPs that falls on the chromosome with the highest copy number (see **Methods**). If the cells in this cluster are not tumor cells, we expect a MAF of close to 0.5. Error bars showing the 95% confidence interval on the estimate of the MAF for each point are shown and points that are significantly different from the expected MAF (0 for substitutions, 0.5 for CN; $q<0.01$) are colored green. Clusters where one or both of the symbols are missing indicate that there were no cells with informative reads for that particular combination of cluster, tumor genotype and variant type (SNV or CN).

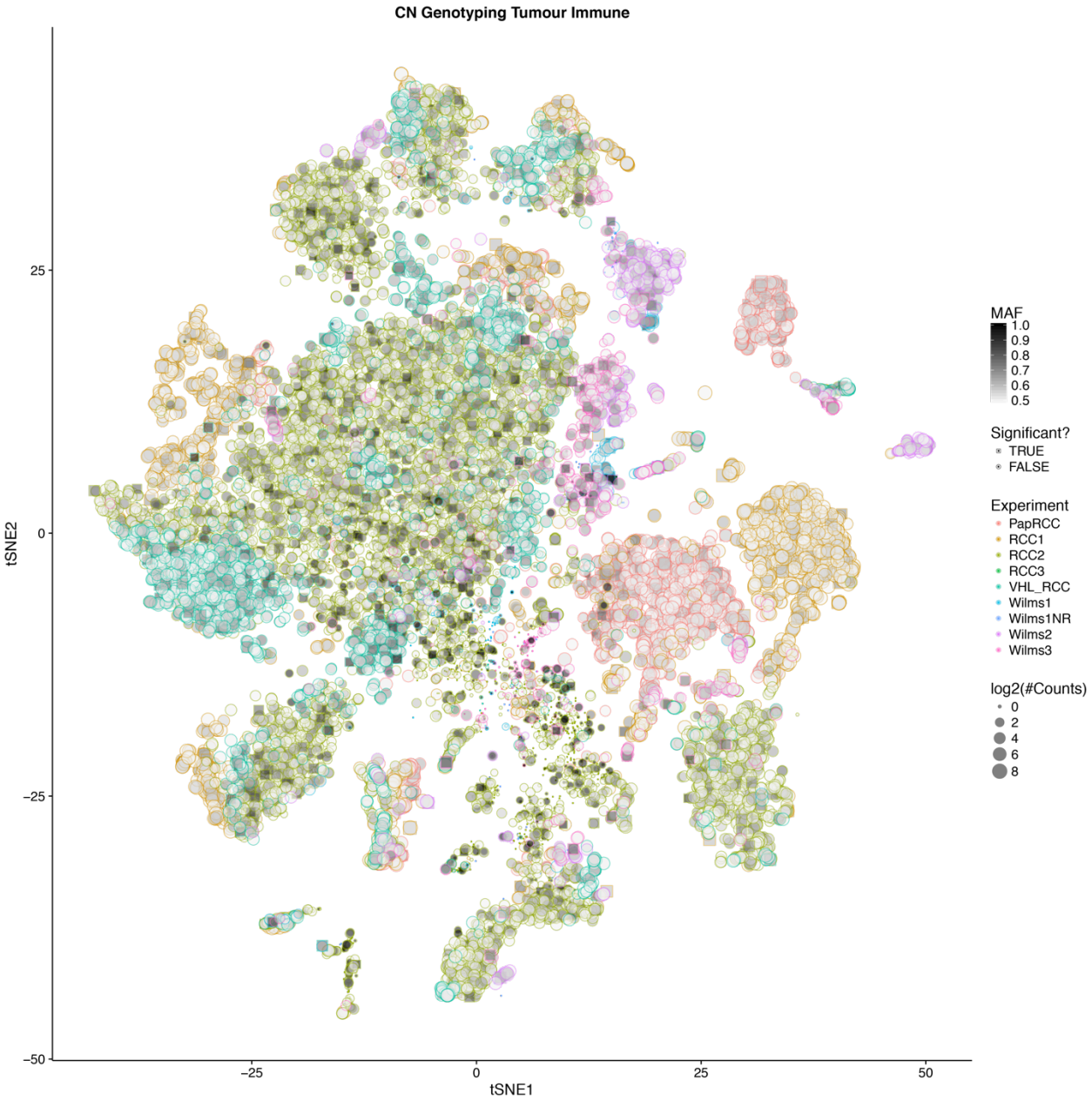
Clusters of cells that contain the tumor genotype can be clearly identified in the clusters of single cells in the tumor epithelial and vascular compartment. There is no cluster outside this compartment where the CN and SNV data both support the tumor genotype being present.



Supplementary Figure S11. – CN genotyping of every non-immune tumor cell

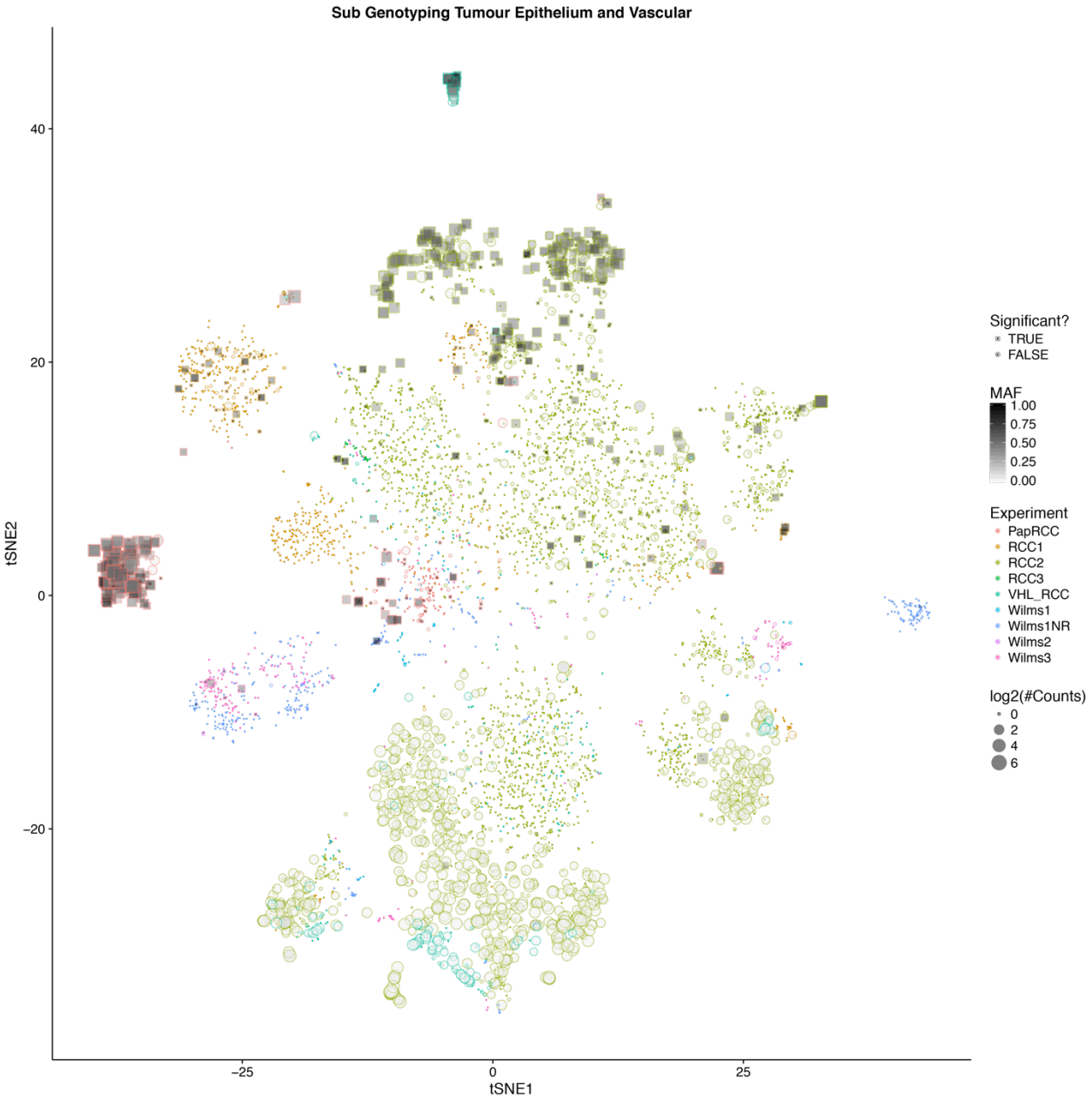
Each point in this map represents a tumor cell (see **Fig. S7**) and is colored by the sample it was generated from. For each cell, an allele frequency is calculated by averaging across the allele with the largest CN in the bulk DNA data for the relevant sample. The size of the point indicates the number of informative reads that are available for this cell and the shape indicates if the deviation of the allele frequency from 0.5 (the expected value if the cell is CN neutral) is significant (FDR < 0.01; binomial test). The shading of each cell indicates the observed allele frequency, with any allele frequency less than 0.5 shown in white. Cells for which no informative reads were available are plotted as a single dot. As an example, a square with a red outline indicates a cell from the pRCC sample for which the allele frequency significantly

exceeds 0.5. In this case the allele frequency is the sum across SNPs belonging to the allele with the largest number of copies in the clonal CN change identified in the bulk DNA from the pRCC sample. The shading indicates the exact value of this allele frequency and the size of the dot indicates how many reads it is based on, with larger dots indicating more reads on a log2 scale.



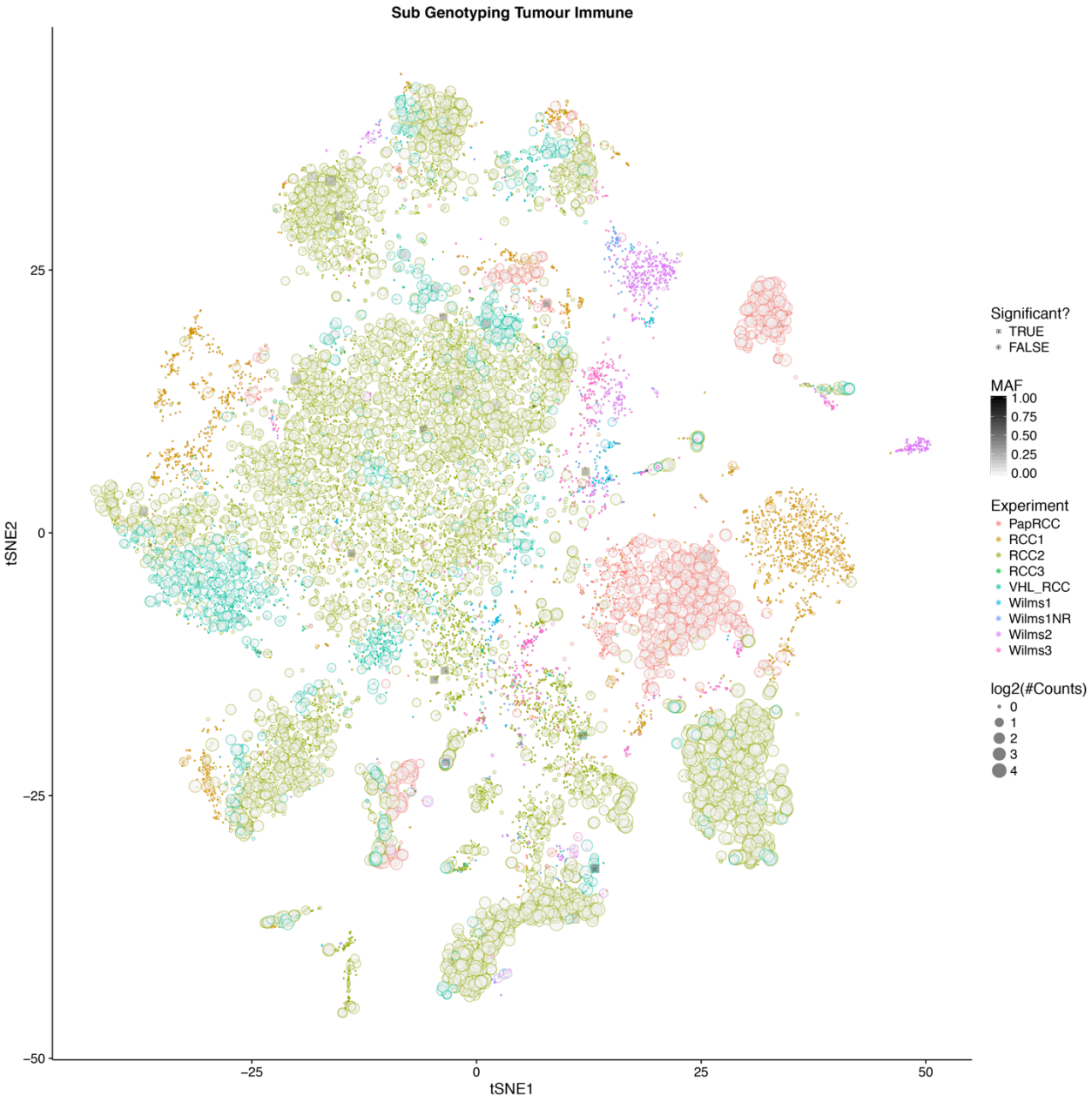
Supplementary Figure S12. CN genotyping of every immune tumor cell

As in **Fig. S11** except the genotyping of CN of tumor immune cells (**Fig. S8**) is shown as a negative control.



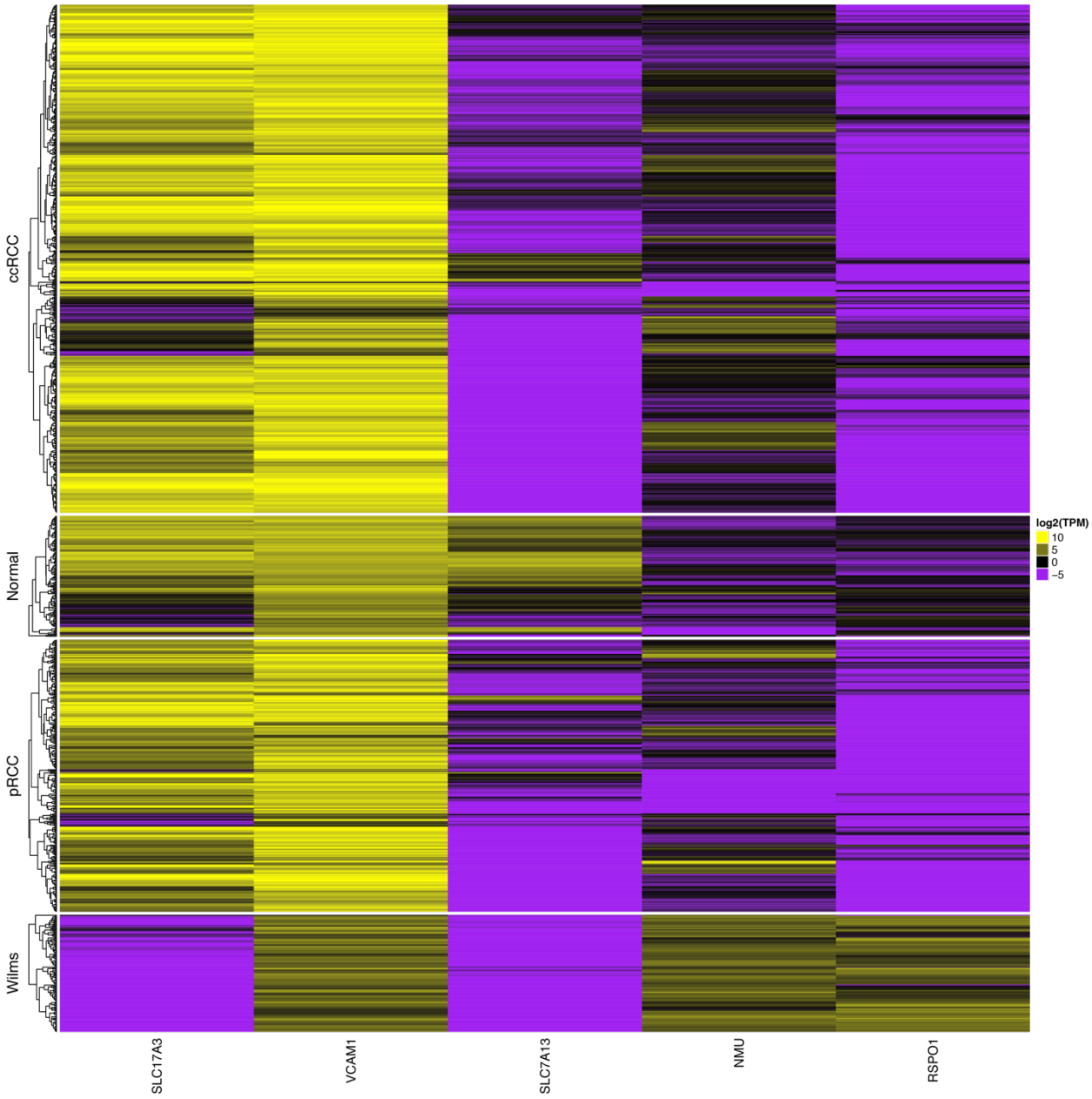
Supplementary Figure S13. Substitution genotyping of every non-immune tumor cell

As in **Fig. S11** except here the allele frequency represents the fraction of reads supporting the mutant allele at all single nucleotide variants identified in the bulk DNA. As the expected allele frequency for a cell with the wild type genotype is 0, squares show those cells for which the allele frequency significantly exceeds 0 and the shading covers all values of the MAF rather than being truncated at 0.5 as was done in **Fig. S11**.



Supplementary Figure S14. – Substitution genotyping of every immune tumor cell

As in **Fig. S11** except the genotyping of substitutions of tumor immune cells (**Fig. S8**) is shown as a negative control.



Supplementary Figure S15. – Expression of key genes in all bulk RNA-seq data for studied tumor types

A heatmap showing the expression of the key markers of the PT1 population of proximal tubular cells (**Fig. 1C; Fig. S2**) and of the ureteric bud and primitive vesicle (**Fig. 2C; Fig. S5**). Specifically, PT1 cells are SLC17A3+, SLC7A13- and VCAM1+, UB cells are NMU+ and PV cells are RSPO1+. The color scheme shows the log2(TPM) expression of each sample (row) where rows have been split into the three cancer types and normal and then heirachically clustered.

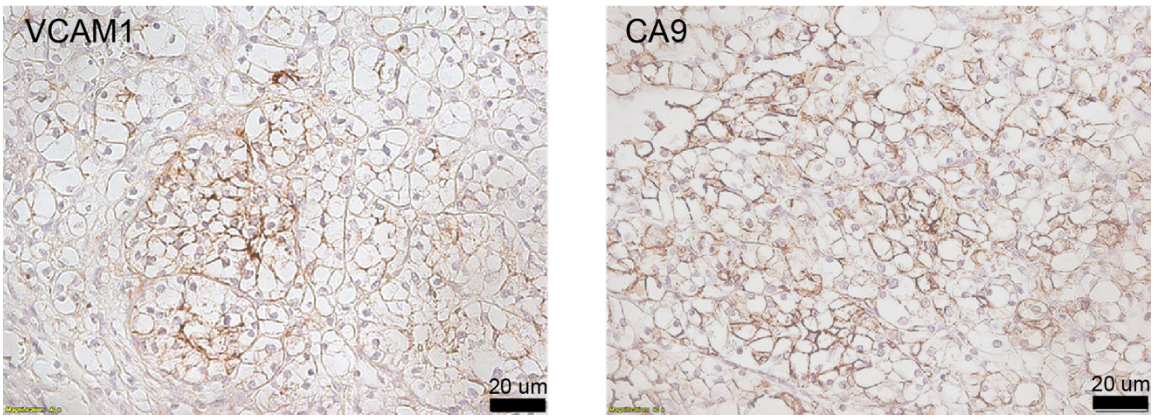
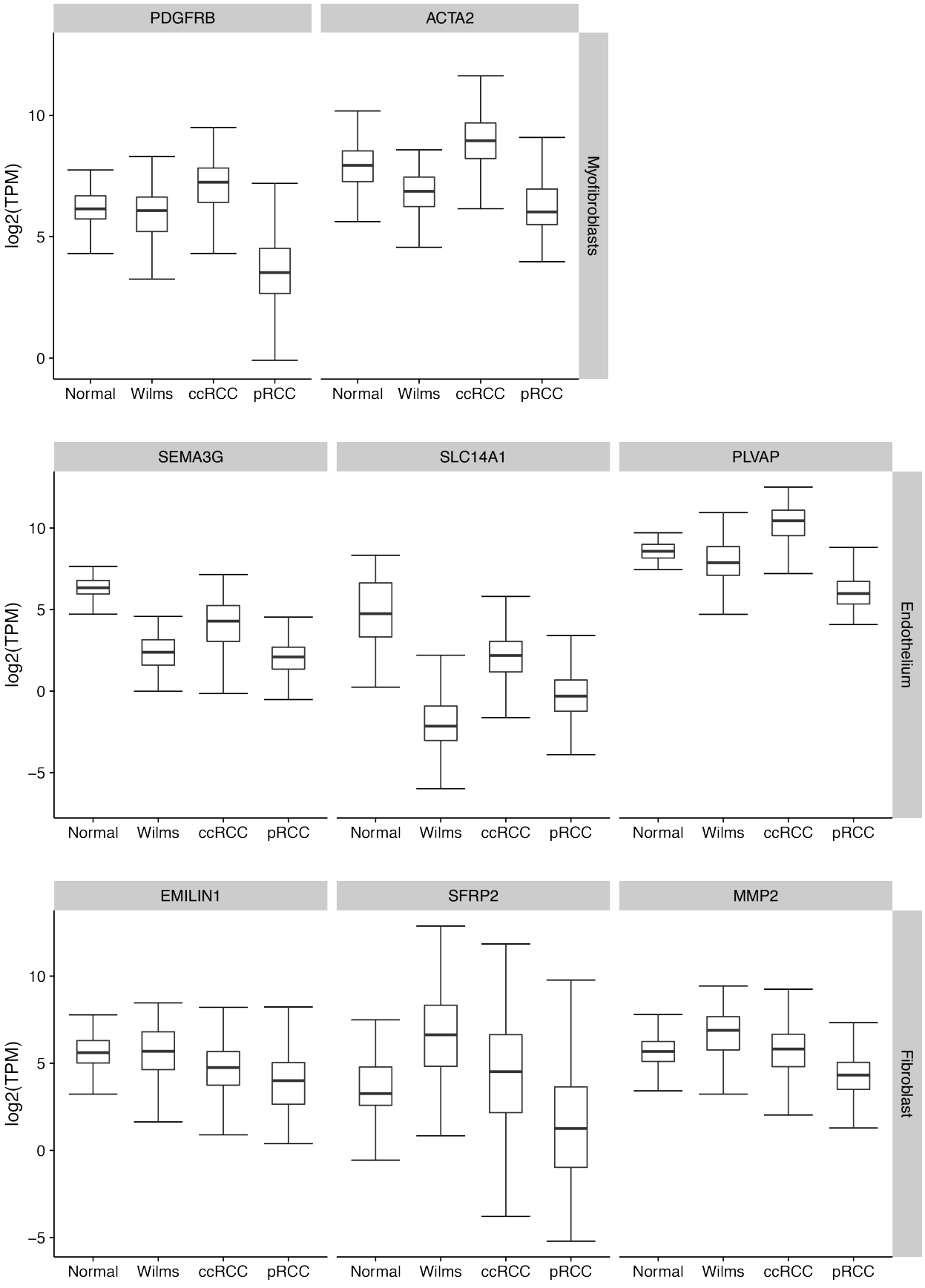


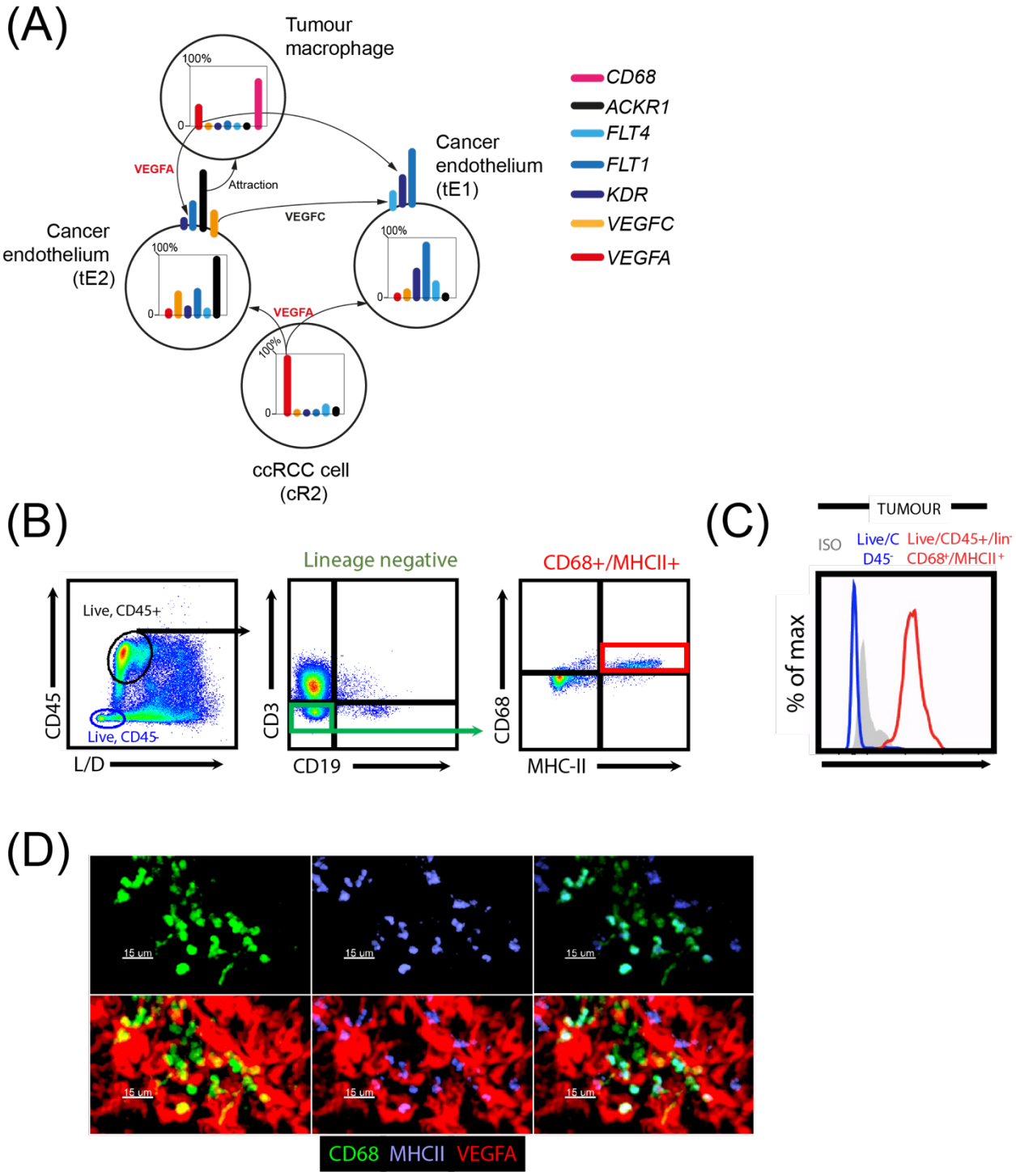
Figure S16. VCAM1 staining in tumors of patients predisposed to clear cell renal cell carcinoma.

Tumor sections of a ccRCC from a patient with a pathogenic germline mutation in the VHL gene are shown to demonstrate that VCAM1, a PT1 marker, is expressed by cancer cells.


PDGFRB
ACTA2
Myofibroblasts
SEMA3G
SLC14A1
PLVAP
Endothelium
EMILIN1
SFRP2
MMP2
Fibroblast
log2(TPM)
10
5
0
−5
Normal
Wilms
ccRCC
pRCC

Supplementary Figure S17. Expression of markers of tumor associated genes in bulk RNA-seq

Boxplots showing the expression of key markers of myofibroblasts (first row), vascular endothelium (second row) and fibroblasts (third row) for bulk RNA-seq data from normal kidney, Wilms tumor, ccRCC and pRCC respectively (see **Methods**).



Supplementary Figure S18. – Full characterization of kidney tumor immune cells

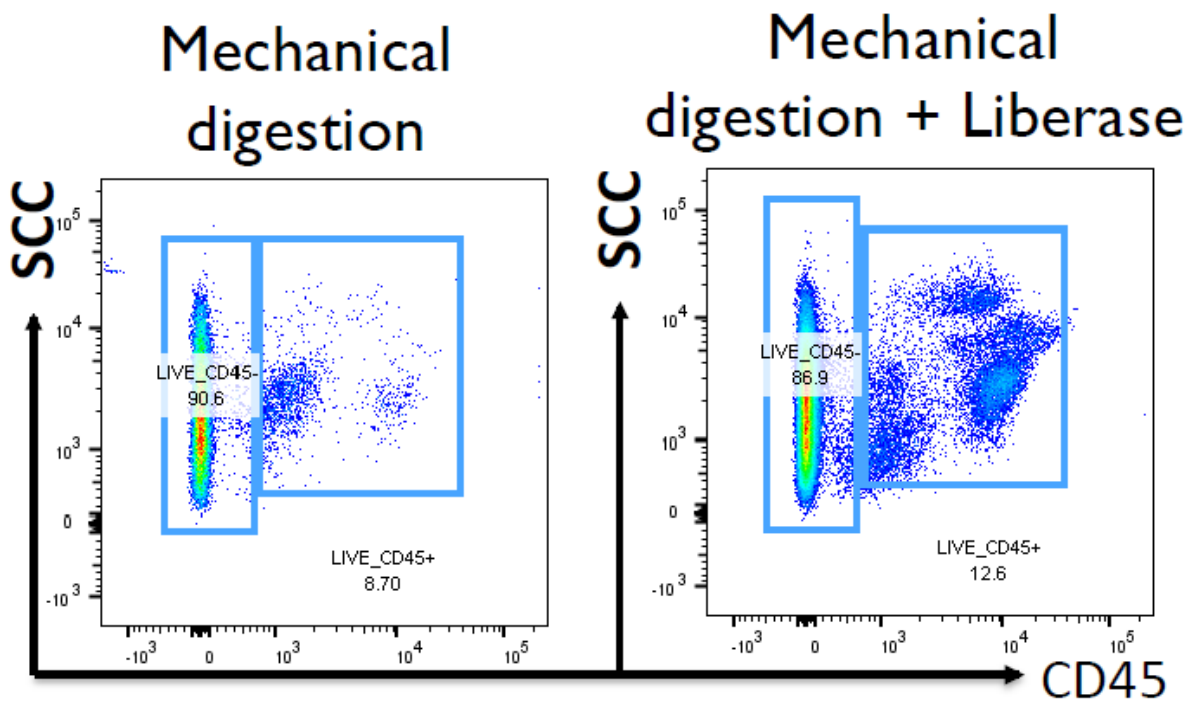
(A) VEGF signaling derived from single cell data. Expression of VEGFA/VEGFC, its receptors (FLT1, FLT4, and KDR), and ACKR1 in clusters of ccRCC, cancer endothelium and

macrophages. Circles represent clusters, bar plots fractional expression and arrows receptor/ligand interactions or the facilitation of migration (ACKR1).

(B) Validation of VEGFA bearing macrophages by flow cytometry. We gated on live, CD45+ cells (following exclusion of doublets), CD19 and CD3 negative cell within this gate were then analyzed for their expression of MHCII and CD68 to identify macrophages.

(C) CD68+ VEGFA+ cells. Representative histogram showing VEGFA staining in tumor macrophages (Live/lin-negative, CD68/MHCII+ cells, red line), in CD45 negative cells (blue line). Isotype control staining shown in grey filled histogram.

(D) Confocal microscopy showing co-localization of CD68, MHCII, and VEGFA.



Supplementary Figure S19. Comparison of dissociation protocols

Flow cytometric assessment of mechanical dissociation alone or in combination with liberase shows that some cell populations, here CD45+ cells, are missed when mechanical dissociation is used alone. The experimental protocol is described in **Methods**.

Donor (study ID)	Experiment	Age	Normal kidney	Renal pelvis	Ureter	Tumor	Tumor type
Foetus1	Foetus16	8 PCW	1	-	-	-	-
Foetus2	Foetus17	9 PCW	1	-	-	-	-
Child1	Wilms1	4 years 2 months	2 (1 Nephrogenic Rest)	-	-	1	Wilms'
Child2	Wilms2	8 months	2	1	1	1	Wilms'
Child3	Wilms3	2 years 6 months	2	1	1	1	Wilms'
Adolescent	Declined_transplant	12 years	6 (2 kidneys)	-	-	-	-
Adult1	PapRCC	70	1	-	-	1	Papillary cell carcinoma
Adult2	RCC1	67	2	-	1	2	Clear cell carcinoma
Adult3	RCC2	63	1	-	-	2	Clear cell carcinoma
Adult4	VHL_RCC	49	2	-	-	1	Clear cell carcinoma
Adult5	RCC3	72	2	-	1	1	-

Table S1. Patient manifest

Clinical features of the data. Age for fetal samples is given in post conception weeks (*34*). Numbers give the number of biopsies taken from each sample.

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS2 – Cluster info" or in TableS2.xlsx

Table S2. Summary of clusters

Annotation and other information about clusters of cells. Each cluster is given a unique Cluster ID and optionally a more informative alias. The annotation of each cluster is giving in increasing level of detail in the columns "Cell_type1", "Cell_type2" and "Cell_type3". The column "Positive_marker_mRNA" gives the names of those genes that were used to assign a cell type to this cluster. The genotype column indicates if the majority of cells in this cluster carry the genotype of tumor, normal or nephrogenic rest cells identified from bulk DNA sequencing. Those clusters which could not be unambiguously identified as one cell type were labelled as "Junk".

Gene	Marker_of	Reference
MET	pRCC	(*29*)
SIX2	CapMesenchyme	(*35, 36*)
CITED1	CapMesenchyme	(*35, 36*)
PAX2	CapMesenchyme	(*35, 36*)
SIX1	CapMesenchyme	(*35, 36*)
CA9	ccRCC	(*28*)
ATP6V0D2	Collecting_duct	(*37-39*)
CLCNKB	Collecting_duct	(*37-39*)
SLC26A4	Collecting_duct_A	(*37-39*)
SLC4A1	Collecting_duct_B	(*37-39*)
AVPR2	Distal	(*37-39*)
SLC8A1	Distal	(*37-39*)
KCNJ1	Distal	(*37-39*)
CLDN8	Distal_collecting	(*37-39*)
SFRP2	Fibroblast	(*40, 41*)
EMILIN1	Fibroblast	(*42*)
MMP2	Fibroblast	(*42*)
CLDN5	Glom_vascular	(*37-39*)
SEMA3G	Glom_vascular	(*37-39*)
AQP1	Glom_vascular	(*37-39*)
PTPRO	Glomerulus	(*37-39*)
PODXL	Glomerulus	(*37-39*)
WT1	Glomerulus	(*37-39*)
PTPRO	Glomerulus_PrimitiveVesicle	(*37-39*)
PODXL	Glomerulus_PrimitiveVesicle	(*37-39*)
WT1	Glomerulus_PrimitiveVesicle	(*37-39*)
CLDN16	Henle_ascending	(*37-39*)
SLC12A1	Henle_descending	(*37-39*)
PDGFRB	Myofibroblasts	(*43*)
ACTA2	Myofibroblasts	(*43*)
NDUFA4L2	ccRCC	(*44*)
KRT23	Pelvic	(*37-39*)
SAA2	Pelvic	(*37-39*)
TP63	Pelvic_ureter	(*45*)
PVRL4	Pelvic_ureter	(*45*)
KRT5	Pelvic_ureter	(*45*)
UPK1B	Pelvic_ureter	(*45*)
UPK1A	Pelvic_ureter	(*45*)
DHRS2	Pelvic_ureter	(*45*)
S100P	Pelvic_ureter	(*45*)
AQP2	Principal_cells	(*38, 39*)
SLC13A3	Proximal	(*38, 39*)
SLC34A1	Proximal	(*38, 39*)
SLC17A3	Proximal_convoluted	(*38, 39*)
SLC22A8	Proximal_convoluted	(*38, 39*)
SLC7A13	Proximal_straight	(*38, 39*)
SLC16A9	Proximal_straight	(*38, 39*)
SLC22A7	Proximal_straight	(*38, 39*)
HNF1B	UretericBud	(*35, 36*)
RET	UretericBud	(*35, 36*)
GATA3	UretericBud	(*35, 36*)
ELF3	UretericBud	(*35, 36*)
POU3F3	UretericBud	(*35, 36*)
TFCP2L1	UretericBud	(*35, 36*)
CDH16	UretericBud	(*35, 36*)
PLVAP	Vascular	(*46*)
SLC14A1	Vascular	(*46*)
VCAM1	Vascular	(*46*)
KDR	Vascular	(*46*)
PTPRB	Vascular	(*47*)
PECAM1	Vascular	(*46*)

Table S3. Markers curated from the literature

Genes used as markers of different cell types and the reference that justifies their use.

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS4 –Algorithmic markers" or in TableS4.xlsx

Table S4. Algorithmically identified marker genes of each cluster

Genes specific to different clusters of cells and the evidence for their specificity. Genes listed here have a FDR <0.05 from a hypergeometric test (see **Methods**) or are one of the 7 key genes that are listed for all clusters. The column, "isKeyGene" indicates if a gene is a key gene that is always listed or listed because it is statistically significant. The fraction of cells expressing a gene (prefix "geneFrequency") and average normalized expression (prefix "geneExpression") within each cluster, outside of each cluster and across all cells is given. Within each cluster, genes are sorted by their tf-idf value (see **Methods**). The unique cluster ID is given in the "Cluster" column (see **Table S2**. for aliases).

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS5 – Nephrogenesis TFs" or in TableS5.xlsx

Table S5. Transcription factors important to nephrogenesis

Transcription factors that are differentially expressed between UB and CM/PV or change significantly along the pseudotime axis joining CM and PV cells (**Fig. 2D**). The p-values for these two tests is given along with the average expression of each gene in each of the three populations (UB, CM, and PV).

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS6 – Sample manifest" or in TableS6.xlsx

Table S6. Sample manifest

Manifest describing each of the 10X channels processed. There is a one-to-one correspondence between 10X channels and SangerID. The column "Label" gives a compact summary of each channel's meta-data in the format:
<Experiment>_<Organ>_<Location>_<SortUsed>_<BioRepNo>_<TechRepNo>
All other columns are self explanatory except for "BulkID" which gives the ID of the bulk DNA samples that correspond to this channel of data.

Chr	Start	End	TotalCN	MinorCN	Experiment	PD_ID
2	199481900	242985493	1	0	RCC1	PD37104a
3	60197	84808933	1	0	RCC1	PD37104a
3	84809350	197846280	2	0	RCC1	PD37104a
9	209134	141068960	1	0	RCC1	PD37104a
13	19455253	65486222	1	0	RCC1	PD37104a
14	20433516	107283886	1	0	RCC1	PD37104a
2	211334242	242985493	1	0	RCC2	PD37228c
3	65982	90362964	1	0	RCC2	PD37228c
3	93505756	197811124	3	1	PapRCC	PD35918h
12	188285	133812333	3	1	PapRCC	PD35918h
17	9034	79998834	3	1	PapRCC	PD35918h
2	55984	242985493	3	1	VHL_RCC	PD36793a
3	60596	83348972	1	0	VHL_RCC	PD36793a
3	83352258	197606877	3	1	VHL_RCC	PD36793a
5	850203	107560911	3	1	VHL_RCC	PD36793a
5	107562191	180687907	5	2	VHL_RCC	PD36793a
7	115401	159122682	3	1	VHL_RCC	PD36793a
9	203937	141068960	3	1	VHL_RCC	PD36793a
10	266373	135235890	3	1	VHL_RCC	PD36793a
11	196944	134944142	3	1	VHL_RCC	PD36793a
12	188285	133839356	3	1	VHL_RCC	PD36793a
13	19455957	114999838	3	1	VHL_RCC	PD36793a
15	20021973	102431166	3	1	VHL_RCC	PD36793a
16	84170	53627229	3	1	VHL_RCC	PD36793a
16	53628412	89997381	2	0	VHL_RCC	PD36793a
17	27074565	79998834	3	1	VHL_RCC	PD36793a
18	125371	78017073	2	0	VHL_RCC	PD36793a
19	226776	59097308	3	1	VHL_RCC	PD36793a
20	20000786	62954871	3	1	VHL_RCC	PD36793a
21	15345102	48101335	3	1	VHL_RCC	PD36793a
11	78383801	134937738	1	0	Wilms1	PD36165d
12	203339	133839356	3	1	Wilms1	PD36165d
16	35272667	89973832	1	0	Wilms1	PD36165d
2	96057011	242852391	2	0	Wilms2	PD37272a
16	86084	89998157	4	1	Wilms3	PD37276a
16	32156537	33767707	3	1	Wilms1NR	PD36165e

Table S7. Clonal copy number changes in tumor DNA

Table of copy number changes identified in bulk DNA for each of the tumor samples in our experiment

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS8 – Markers of UB and PV" or in TableS8.xlsx

Table S8. Genes globally specific to ureteric bud or primitive vesicle

Statistics summarizing how specific each gene is to the ureteric bud or primitive vesicle cluster in the fetal single cell data ("MarkerClusterFrequency" column), relative to other clusters in the fetal nephron ("MaxFetalNephFrequencyExcluingMarker" column) and all other non-tumor single cell clusters ("MaxOutOfClusterFrequency" column). Genes that meet the criteria for being globally specific markers of the UB or PV defined in **Methods** are indicated by the "MeetsCriteria" column.

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS9 –Wilms development TFs" or in TableS9.xlsx

Table S9. Transcription factors important to Wilms' development

The transcription factor changes that change significantly along either of the two branches from right to left in (**Fig. 3C**). The last three columns give the average expression from cells in the different nodes of the pseudotime trajectory in **Fig. 3B**.

SangerID	PD_ID	EGA_ID	EGA_StudyID	StudyTitle	Experiment
4496STDY6908196	PD36165b	EGAN00001547798	EGAS00001002171	Orphan Tumour Study	Wilms1
4766STDY6993201	PD35918g	EGAN00001561561	EGAS00001002486	Kidney tumour_DNA	PapRCC
4766STDY7061100	PD35918g	EGAN00001586322	EGAS00001002486	Kidney tumour_DNA	PapRCC
4766STDY6993199	PD35918h	EGAN00001561559	EGAS00001002486	Kidney tumour_DNA	PapRCC
4766STDY7061101	PD35918h	EGAN00001586323	EGAS00001002486	Kidney tumour_DNA	PapRCC
4766STDY6993197	PD36165d	EGAN00001561556	EGAS00001002486	Kidney tumour_DNA	Wilms1
4766STDY6993198	PD36165e	EGAN00001561558	EGAS00001002486	Kidney tumour_DNA	Wilms1
4766STDY6993204	PD36793a	EGAN00001561564	EGAS00001002486	Kidney tumour_DNA	VHL_RCC
4766STDY6993203	PD36793c	EGAN00001561563	EGAS00001002486	Kidney tumour_DNA	VHL_RCC
4766STDY6993195	PD37104a	EGAN00001561555	EGAS00001002486	Kidney tumour_DNA	RCC1
4766STDY6993196	PD37104b	EGAN00001561557	EGAS00001002486	Kidney tumour_DNA	RCC1
4766STDY7061105	PD37228c	EGAN00001586327	EGAS00001002486	Kidney tumour_DNA	RCC2
4766STDY7061102	PD37228f	EGAN00001586324	EGAS00001002486	Kidney tumour_DNA	RCC2
4766STDY7061111	PD37272a	EGAN00001586333	EGAS00001002486	Kidney tumour_DNA	Wilms2
4766STDY7061117	PD37272g	EGAN00001586339	EGAS00001002486	Kidney tumour_DNA	Wilms2
4766STDY7061116	PD37276a	EGAN00001586338	EGAS00001002486	Kidney tumour_DNA	Wilms3
4766STDY7061120	PD37276g	EGAN00001586342	EGAS00001002486	Kidney tumour_DNA	Wilms3

Table S10. DNA manifest

Table relating sample IDs used in experiment to other identifiers, including EGA accession number and study name. This table pertains only to the bulk DNA samples.

Please note that this table has been uploaded as a separate file named SupplementaryTables.xlsx and can be found on the sheet labeled "TableS11 – Cell manifest" or in TableS11.xlsx

Table S11. Cell manifest

Summary of each cell identified in this experiment, which sample it originated from and which compartment and cluster it has been assigned to. The column "Barcode" gives the 10X barcode for this cell, while the column "DropletID" prepends the sample ID to produce an ID which uniquely identifies each cell across all experiments. nUMI gives the total number of UMIs detected in each cell, nGenes the total number of genes with non-zero expression and MTfrac gives the fraction of expression in each cell that comes from genes on the mitochondria. Finally, "QCpass" indicates if this cell has passed all QC filters (see **Methods**).

PE Rabbit anti-human VEGFA, clone EP1176Y from abcam® (Cat# ab209439)	1/1000 dilution
FITC Mouse anti-human CD68, clone KP1 from Dako antibodies (Cat# F7135)	1/200 dilution
V450 Mouse anti-human HLA-DR, clone L243 from BD Biosciences (Cat# 655874)	1/100 dilution
Alexa Fluor™ 647 Phalloidin (Cat# A22287)	1/100 dilution
Primary Rabbit anti-human SLC17A3, polyclonal from abcam® (Cat# ab23332)	1/100 dilution
Biotinylated Mouse anti-human CD106 (VCAM-1), clone STA from BioLegend® (Cat# 305804)	1/100 dilution
PE Mouse anti-human CA9, clone 303123 from R&D systems (Cat# FAB2188P)	1/100 dilution
Alexa Fluor® 488 Donkey anti-rabbit IgG (H+L), polyclonal from ThermoFisher Scientific (Cat# A21206)	1/200 dilution
Streptavidin APC from ThermoFisher Scientific (Cat# 17-4317-82)	1/200 dilution

Table S12. Antibodies and dilutions

A table of the antibodies and dilutions used in performing the immunofluorescence microscopy.

Please note that these data have been uploaded

as a separate file named DataS1.tar.gz

Data S1 – Table of counts

A table of counts for all candidate droplets (i.e., the 125,139 unique barcodes at the top of **Fig. S1**). Stored in matrix mart format along with two extra tables describing the column labels (droplets) and row labels (genes). Counts represent number of unique UMIs for each gene/droplet pair (see **Methods**).