Integration and disentanglement of single-cell and spatial transcriptomics in health and disease

Novella Rausell, C.

2

A COMPREHENSIVE MOUSE KIDNEY ATLAS ENABLES RARE CELL POPULATION CHARACTERIZATION AND ROBUST MARKER DISCOVERY

The kidney's cellular diversity is on par with its physiological intricacy, yet identifying cell populations and their markers remains challenging. Here, we created a comprehensive atlas of the healthy adult mouse kidney (MKA: Mouse Kidney Atlas) by integrating 140.000 cells and nuclei from 59 publicly-available single-cell and single-nuclei RNA-sequencing datasets from eight independent studies. To harmonize annotations across datasets, we built a hierarchical model of the cell populations. Our model allows the incorporation of novel cell populations and the refinement of known profiles as more datasets become available. Using MKA and the learned model of cellular hierarchies, we predicted previously missing cell annotations from several studies. The MKA allowed us to identify reproducible markers across studies for poorly understood cell types and transitional states, which we verified using existing data from micro-dissected samples and spatial transcriptomics.

This chapter is published as ***Claudio Novella-Rausell***, *Magda Grudniewska, Dorien J.M. Peters, and Ahmed Mahfouz. A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. iScience, 26(6):106877, June 2023.*

2.1 INTRODUCTION

KIDNEYS are organs with a high degree of cellular complexity reflected in an array of different renal functions: from filtering the blood, regulating water homeostasis, production of hormones, to excretion of waste products. These diverse functions are driven by distinct anatomical structures called nephrons. Each nephron comprises more than tens of highly specialized cell types, including abundant epithelial cells supported by vascular, stromal and immune cells[76]. Notably, the function and nomenclature of cells that assemble the nephron depend on their location relative to the main tubular structures: the proximal tubule, loop of Henle, distal convoluted tubules, and the collecting duct[77].

More than 150 litres of filtrate are reabsorbed by the nephrons in a day. Most of this reabsorption occurs in the proximal tubules, which are primarily located in the cortex, the outermost portion of the kidney. Sodium gradient, generated by the activity of numerous Na^+/K^+-ATPase channels, drives the transport of salts, water, glucose and amino acids, back to the bloodstream. This process requires large amounts of energy, supplied by the abundant mitochondria. Proximal tubules are thus the most metabolically active structures in the nephron[78–80]. The filtrate then enters the loop of Henle that connects the proximal and distal tubule and is most notably involved in extracellular fluid volume and blood pressure regulation as well as Ca2+, Mg2+, and acid-base homeostasis. Through the activation of several processes, required to generate a gradient of increasing osmolality from cortex to medulla, this segment also contributes to urine concentration[81]. Finally, the filtrate travels through the distal convoluted tubule and collecting duct system, where water is reabsorbed and urine is concentrated[82] (**Fig. 1A**).

Although outstanding efforts to characterize the transcriptomic profiles of the different cell types present in the kidney have been supported by the recent advances in single-cell technologies[83–90], the identification of markers for distinct and, in particular, rare cell types remains elusive. A substantial portion of the transcriptomic data comes from the proximal tubule and loop of Henle cells, which are one of the largest structures of the nephron[91] . Rare cell populations usually remain undetected and the costs to profile these cells often precludes the studies of less abundant cell populations[92,93]. The aforementioned challenges could be addressed by creating a reference atlas of the kidney, that leverages the vast collection of available cells and nuclei profiled, which can then be used to annotate specific kidney cell types in a supervised manner.

Integrating different datasets into a common space can overcome batch effects which occur due to the differences in library preparation protocols and data processing steps. However, in most studies to date, the annotation of cell populations is performed in an unsupervised manner, a process that is time-consuming and involves several refinement iterations[94]. The subjective nature of this approach limits the ability to compare populations across studies due to annotation inconsistencies. Ideally, the reference atlas would account for the different resolutions at which cell populations have been annotated and include a harmonized annotation that allows for a better characterization of rare cell populations.

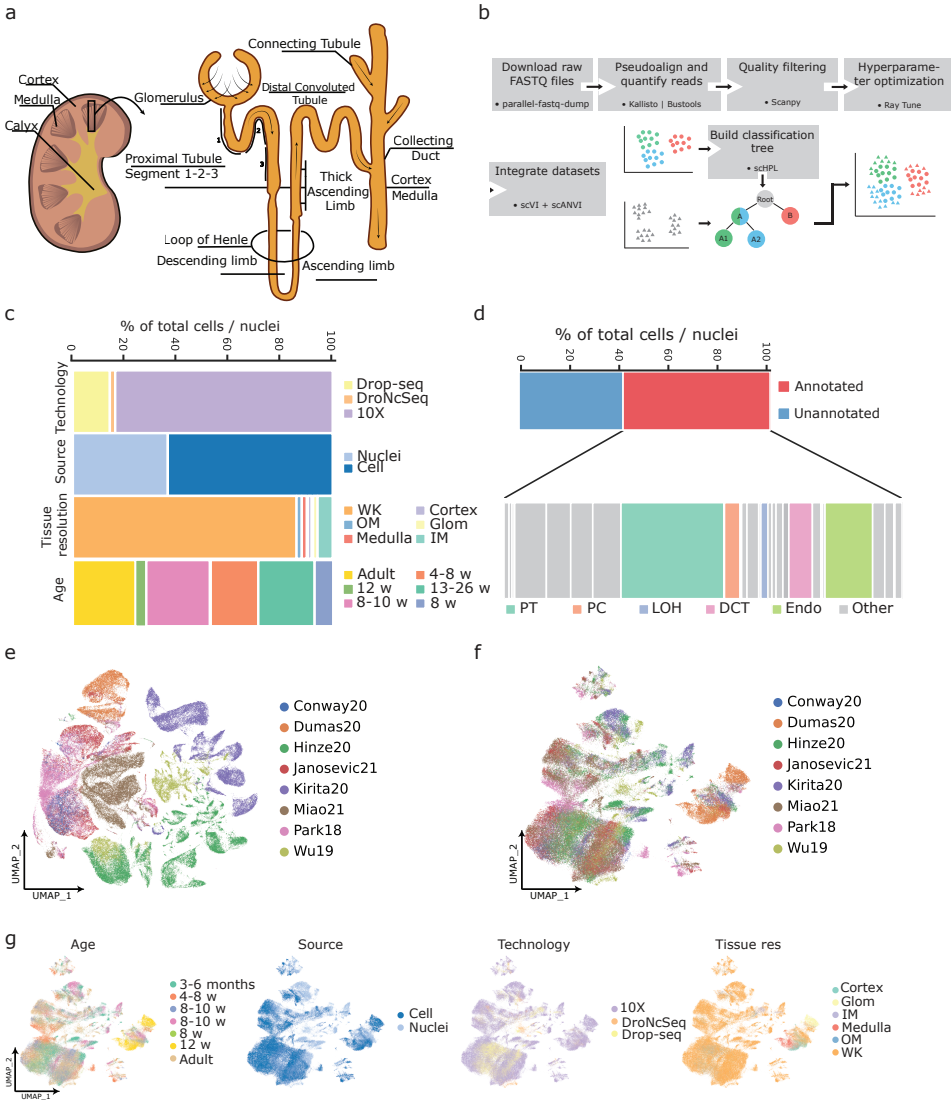

Figure 2.1: *Generation of the mouse kidney atlas from eight independent studies.* **a** Schematic of a kidney and a nephron. Arrows indicate the flux of glomerular filtrate through the tubular segments. **b** Workflow guiding the generation of the mouse kidney atlas. Colours represent different hypothetical cell type annotations from two independent studies (A, B and A1, A2), whereas shapes depict originally annotated (circle) or unannotated (triangles) cells and/or nuclei. **c** Metadata information across all datasets. Age of the animals is represented in weeks or months (*w, m*) when available; otherwise, an overall age estimator is provided (*Adult*). Tissue resolution varied from Whole Kidney (*WK*), Cortex, Medulla to more selected regions such as Outer Medulla (*OM*), Inner Medulla (*IM*) or Glomerulus (*Glom*). The suspension type was either single-cell or single-nuclei sequenced using Drop-seq, DroNc-Seq, or 10x Genomics. **d** Proportion of all annotated cell types across all datasets. Relevant cell types in the nephron are highlighted, namely Proximal Tubule cells (*PT*), Principal Cells (*PC*), Loop of Henle cells (*LOH*), Distal Convoluted Tubule cells (*DCT*) and Endothelial cells (*Endo*). **e** Uniform Manifold Approximation and Projection (UMAP) embedding of all used datasets prior to integration. Colours correspond to the different datasets. **f** UMAP visualization of merged datasets following integration and batch correction (see Methods). *sCell*: single-cell, *sNuc*: single-nuclei. **g** UMAP representations of the 140K cells and nuclei after integration. Relevant metadata was extracted for each of the datasets. Age of the animals is represented in weeks or months when available; otherwise, an overall age estimator is provided (*Adult*).

Here we create an atlas of the adult healthy mouse kidney (MKA: Mouse Kidney Atlas) by integrating and harmonizing annotations from publicly available single-cell and single-nuclei transcriptomic studies. We integrated 140.000 cells and nuclei from 59 healthy samples sequenced in eight different studies[83–90] to generate an atlas that reflects the biological component of the different samples, while accounting for technical differences. We built a hierarchical model of the cell populations present in the healthy mouse kidney, that accurately predicts cell annotations in unlabelled datasets. In addition, MKA allows further integration of new datasets as they become available by relying on a progressive learning approach[34] (**Fig. 1B**). We show and verify novel and robust markers for both known cell types and previously unexplored rare cell populations.

2.2 RESULTS

2.2.1 INTEGRATED ATLAS ACCOUNTS FOR TECHNICAL DIFFERENCES AMONG SEVEN INDEPENDENT STUDIES

To create a comprehensive atlas of the healthy mouse kidney we downloaded the raw sequencing data (FASTQ) from eight different studies including a total of 59 samples[83–90] (**Table 1**). To reduce variability in alignment rates between different genetic make-ups, we only included healthy samples with a C57BL/6 background. The raw reads of all samples were processed using the same pipeline and we recovered 140,000 cells and nuclei after filtering low quality cells and nuclei (see Methods section for details). The samples included in this study differ in single-cell technology, source of material, tissue resolutions and age of sacrifice (**Fig. 1C**). Approximately 40% of the cells and nuclei included in this study were missing computer-readable annotations (**Fig. 1D**). These differences can be visualized in the Uniform Manifold Approximation and Projection (UMAP) of the data (**Fig. 1E**), where source-specific populations were identified. To resolve these batch effects, we evaluated the performance of five batch correction methods (Seurat, Harmony, Scanorama, scVI and scVI-scANVI[51,59,95–97]) using their respective default parameters (**Fig. S1A** and **Table S1**). The best performing method was scVI (overall score of 0.72). Notably, whilst algorithms such as Seurat efficiently correct the batch effects (batch effect removal score of 0.84) compared to scVI (batch effect removal score of 0.71), the latter better maintains the biology of each individual dataset after integration (biological conservation scores of 0.72 and 0.38 for scVI and Seurat, respectively). We also observed that methods such as Seurat overcorrected for batch differences by aligning all datasets to a common latent space. This is especially evident in the case of Dumas20, a dataset that only contains endothelial cells. Its cells are overcorrected by Seurat and hence aligned with all other datasets and cell types (Figs. S1B, S1C).

Based on these evaluation results, we built an integration pipeline in which we first use a tuned (see Methods) version of scVI to integrate all eight datasets. Second, we apply scANVI[51,59] to the integrated latent space results from scVI together with cell type labels to refine the integration. We computed the same integration metrics as before for our tuned

version of scVI-scANVI (**Fig. S1A**). Notably, tuning scVI's hyperparameters to maximize both batch separation and cell type similarity (see Methods) improves the performance of scVI-scANVI considerably in both batch correction and biological conservation metrics.

After integration, the aligned compendium demonstrates that the different data sources have been properly aligned and no metadata is driving the differences observed in the UMAP space (**Figs. 1F and G**).

2.2.2 INTEGRATION HIGHLIGHTS ANNOTATION INCONSISTENCIES ACROSS STUDIES

After integration, we investigated cell population annotations across the four datasets for which annotations were available or were manually annotated (Park18, Wu19, Kirita20, Miao21, Dumas20, and Janosevic21). The six datasets varied significantly in the resolution and the ontology used to annotate distinct cell populations. Only two cell populations were common between the five studies (Dumas20 only surveyed endothelial cells) based on the set of author's annotated terms, with most of the annotations being dataset-specific (**Fig.2A**). For example, collecting duct intercalated cells (IC) can be further classified into Type A (ICA) or Type B (ICB), depending on the expression and localization of *Slc4a1* in the membrane and the presence of a transport protein called pendrin, encoded by the *Slc26a4* gene. Whereas ICA cells lack pendrin and acidify the urine by excreting H, ICB cells have pendrin and secrete OH^- equivalents[98]. Another example are Proximal Tubule Cells (PT). While certain studies identify PT cells, some others further classify these cells in three different segments (PTS1, PTS2, or PTS3) depending on their location along the nephron (**Fig.1A and 2B**). These differences in ontology, together with the distinct annotation resolutions, highlight the subjective nature of unsupervised cell type annotation and the need for an integrated and comprehensive view of cell heterogeneity in the kidney.

To overcome these challenges, we used single-cell Hierarchical Progressive Learning (scHPL)[34], a method that automatically infers cell hierarchies from annotated datasets and builds a classification tree that can be used to classify unlabeled cells. We used scHPL to build a cell hierarchy and capture the relationships between kidney cell populations from the six annotated datasets. Perfect matches were found between cell populations across five (Principal Cells, PC), four (Endothelial and Podocytes), three (Distal Convoluted Tubule, DCT and T lymphocytes), and two datasets (PTS1, PTS3, Ascending Thick Limb of Henle, ATL, Macrophages, ICA, ICB, and CNT). Not a single cell population matched across all six datasets (**Fig. 2C**). On the other hand, some cell populations are misplaced in the tree. For example, the fibroblasts from Park18 (hereafter cell type$_{\text{dataset}}$) are placed under Macrophages$_{\text{Kirita20, Miao21}}$, another example are PTS1$_{\text{Janosevic21}}$ cells which are placed under Endothelial$_{\text{Kirita20, Wu19, Park18, Miao21}}$ cells. Other populations are lacking available resolution. For example, IC$_{\text{Park18, Miao21}}$ cells, which appear as a parent node of ICA$_{\text{Kirita20, Wu19}}$ and ICB$_{\text{Kirita20, Wu19}}$ cells. Peri Stromal$_{\text{Janosevic21}}$ cells. Moreover, some cell populations are missing in the final tree because they have been rejected (e.g., PT$_{\text{Park18}}$, PCT$_{\text{Miao21}}$,

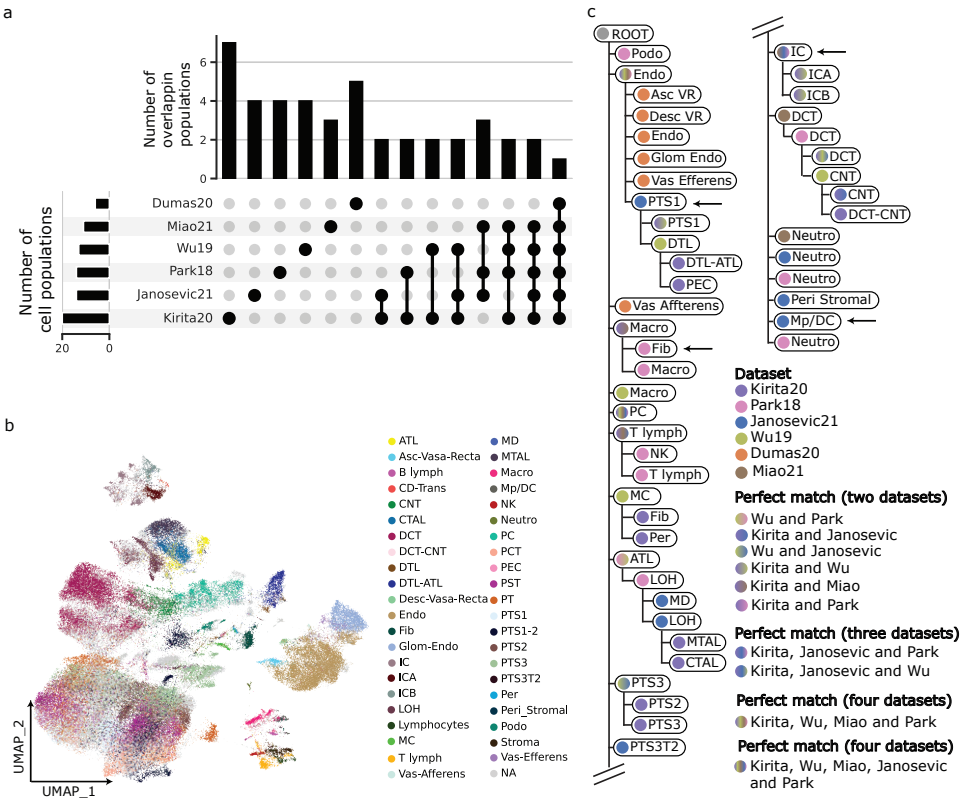

Figure 2.2: *Learned classification tree from independently annotated datasets.* **a** UpSet plot visualizing the intersection and the number of common cell type annotations between the different datasets. Disconnected dots correspond to the number of unique cell types in each dataset, while connected dots represent the intersection between the datasets. Additionally, the number of cell types identified in each study is plotted alongside each dataset. **b** UMAP representation of the originally annotated cell types across all datasets. **c** Learned classification tree applying a k-Nearest Neighbor (kNN) classifier on the six annotated datasets. The colour(s) of the tree nodes correspond to the supporting dataset(s). Missing populations: $PTS1_{Janosevic21}$, $PTS2_{Janosevic21}$. Arrows mark both inaccurate placements in the tree (Fib, Stroma) and cell types that can be further annotated to increase resolution.

PST_{Miao21}, $PTS1_{Janosevic21}$, and $PTS2_{Janosevic21}$ cells could not be classified).

2.2.3 MANUAL CURATION OF ANNOTATIONS SIGNIFICANTLY IMPROVES HIERARCHY LEARNING

To refine the cell tree constructed by scHPL and reduce the number of rejected populations, we performed a manual curation of the original cell population annotations (**Figs. S2 to S4**). The initial tree constructed by scHPL indicates that $Stroma_{Miao21}$ cells have similar transcriptomic profiles to T $lymphocytes_{Park18}$ (**Fig. 2C**), which is supported by their overlap in the UMAP and the high similarity of their average expression profile (**Figs. S2A to S2C**). This observation was supported by the expression of T lymphocyte markers[99,100] (*Cd4*, *Cd8a*, *Cd28*) in cells annotated as Stroma (**Fig. S2D**). In addition, we compared the expression of *Cd4*, *Cd8a*, and *Cd28* in $Stroma_{Miao21}$ cells, T $lymphocytes_{Park18}$, and Kirita20 non-immune cell types (**Fig. S2E**). As expected, $Stroma_{Miao21}$ cells share the expression of these markers with T $lymphocytes_{Park18}$ but not with non-immune populations. A similar scenario applies to $Fibroblasts_{Park18}$, which are placed under the Macrophages node (**Figs. 2C and S2E**). We checked whether these cells might have been mislabelled by visualizing the expression of M1-M2 Macrophage markers[99,101] (*Cd68*, *H2-Ab1*, and *Il4r*) (**Figs. S2F and S2G**). We also plotted the expression of markers for all cell types present in the MKA in both $Stroma_{Miao21}$ and $Fibroblasts_{Park18}$ (**Fig. S2H**). This confirmed that $Stroma_{Miao21}$ mainly express T lymphocyte markers (*Cd247*, *Cd4*, and *Cd8a*) whereas $Fibroblasts_{Park18}$ express Macrophage markers (*Cd68*, *H2-Ab1*, and *Cd74*). Based on these observations, we re-annotated $Stroma_{Miao21}$ and $Fibroblasts_{Park18}$ to T lymphocytes and Macrophages, respectively.

We then evaluated the location of PT cells in the tree, which can be further classified into different segments (Segments 1, 2, 3, and 3 type 2; PTS1, PTS2, PTS3, PTS3T2). The proximal tubule is the first nephron segment after the glomerulus where numerous transporters regulate reabsorption and excretion[80]. Janosevic21 specified the different PT cell types (i.e., PTS1, PTS2, PTS3, PTS3T2), while Park18 included the lower resolution term PT (Fig. S3A) and Miao21 annotations included the terms Proximal Straight Tubule (PST) and Proximal Convoluted Tubule (PCT) (**Fig. S3B**). Wu19 grouped PTS1 and PTS2 cells together (**Fig. S3C**), and Kirita20 did not include PTS3T2 (**Fig. S3D**). To re-annotate these cells as PTS1, PTS2, PTS3, or PTS3T2, we used unsupervised clustering and visualized known markers to rename the resulting cell populations. The visualized markers were *Slc5a12*, *Cyb5a*, *Slc27a2*, and *Cyp7b1* for PTS1, PTS2, PTS3, and PTS3T2, respectively[90,102]. In the case of $PTS1\text{-}2_{Wu19}$, the population was matched to $PTS1_{Kirita20}$ during training of scHPL. We re-annotated $PTS1\text{-}2_{Wu19}$ as $PTS1_{Wu19}$.

We refined the annotation of IC_{Park18}, IC_{Miao21}, and $Endothelial_{Park18}$ cells following the same strategy as described above (**Figs.S4A and S4B**). IC_{Park18} and IC_{Miao21} cells were re-annotated as either ICA or ICB based on the expression of *Slc4a* (ICA marker) and *Insrr* (ICB marker) in the unsupervised clusters (**Fig. S4C**). $Endothelial_{Park18}$ cells were

originally re-annotated as Descending Thin Limb of Henle (DTL) by the original authors in the manuscript, but this correction was missing from the annotations provided with the dataset. Therefore, we similarly refined the annotation based on the expression of *Slc14a2* (DTL marker) and *Adgrl4* (Endothelial marker) (**Fig.S4D**).

Following these annotation refinements, we applied scHPL to rebuild the cell tree and re-trained the classifier on the new tree. The new tree correctly captures the expected cell hierarchy, placing similar populations within the same node (**Fig. 3A**). This is exemplified by the identification of the Medullary and Cortically Thick Ascending Limb of Henle (MTAL and CTAL) as child nodes of Loop of Henle (LOH) cells (**Fig. 1A**). This shows the ability of the hierarchical model to group functionally and morphologically related cell types in the nephron. In this case, perfect matches between datasets were more easily identified, likely due to the lower number of rejected cells while training.

Next, we used the refined classification tree from scHPL (**Fig. 3A**) to predict the cell type annotations for cells and nuclei from the remaining two unlabelled studies (Conway20 and Hinze20) (**Figs. 3B to 3D**). After merging the predicted and original labels, we obtained the final fully annotated adult healthy kidney atlas (MKA) (**Fig. 3E**). The complete overview of the cell population shows that the integration process preserved the shared biological component between the different studies.

Due to the lack of labels for these two datasets, we could not perform a quantitative analysis of the obtained labels. To confirm our annotations, we visualized known markers for the major cell types in the nephron, namely PT, Podocytes, DTL, ATL, MTAL, CTAL, DCT, CNT, ICA, ICB, and Endothelial cells (**Figs. S5A to S5D**). Moreover, we compared the cellular composition of each dataset to that reported in the original studies. We found that the proportion of all predicted Proximal Tubule Cells, i.e., PTS1, PTS2, PTS3, and PTS3T2, matches the proportion described in the original publications, 77% in Conway20 and approximately 60% in Hinze20 (**Fig. S5E**). The same applies to MTAL, CTAL, and DCT in the MKA. These cell populations were described as LOH/DCT in Conway20 and TAL in Hinze20 with a proportion of approximately 7% and 10%, respectively (**Fig. S5E**).

The MKA allowed us to annotate these datasets at a higher resolution than originally reported. For example, in Conway20, they annotated 15 cell types. We now identify 28 distinct populations, providing further resolution for annotations such as LOH/DCT (MTAL, CTAL, and DCT in the MKA) or CD (CD-Trans, ICA, ICB, and CD in the MKA). We also identify previously overlooked important cell populations such as PC (**Fig. S5B**). Another example is Hinze20, in which MKA identified 25 subpopulations among the original set of 10 cell types, including PTS1, PTS2, PTS3, and PTS3T2 instead of PT; ICA and ICB instead of CD-IC; and DTL and ATL instead of TL (Thin Limbs) (**Fig. S5A**). In summary, these annotations provide a comprehensive collection of cell types in the healthy kidney and are supported by at least one published dataset. Moving forward, we keep this resolution and cell type set. However, we could identify other cell types that were missing in the set of input annotations, such as Vascular Smooth Muscle Cells (**Fig. S6A**). Unfortunately, we couldn't find higher-resolution populations for other cell types, such as Macrophages. Markers for

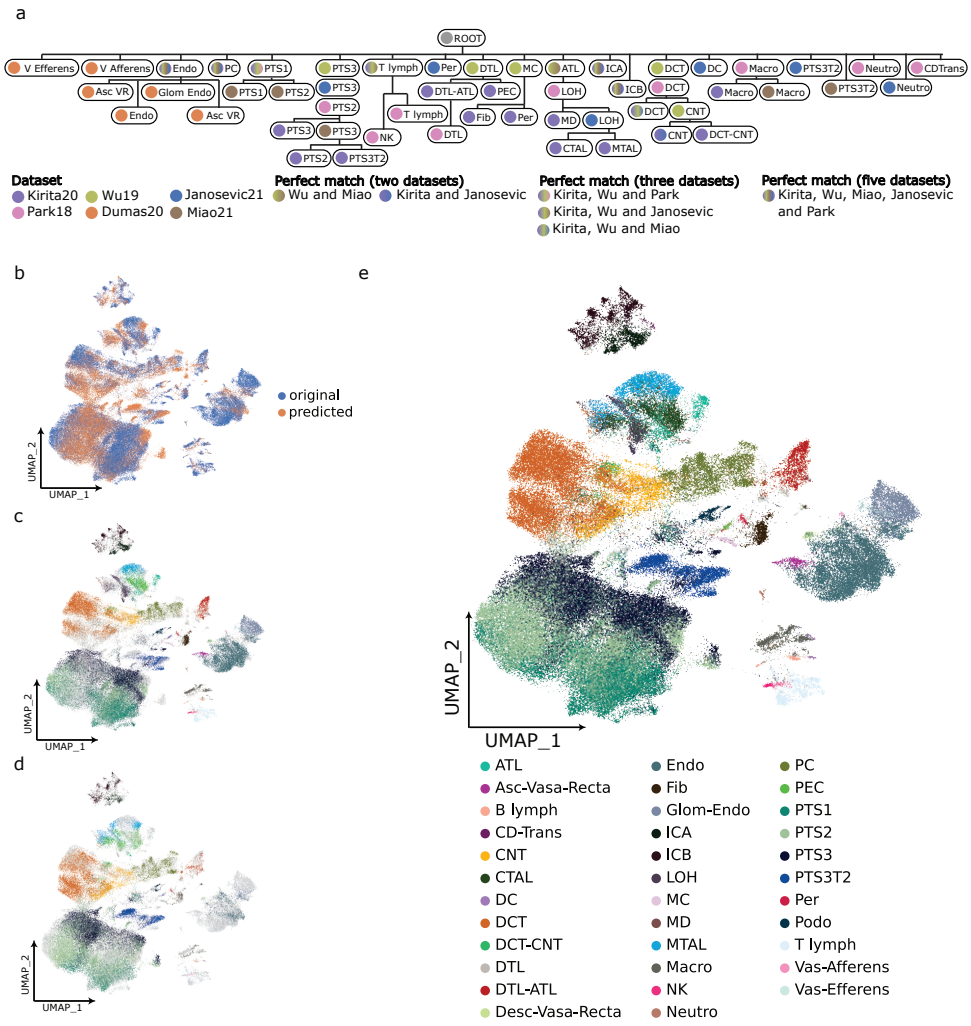

Figure 2.3: *Hierarchically defined kidney mouse atlas.* **a** Learned classification tree on the four annotated datasets after manually harmonizing annotations. Tree nodes are coloured by the supporting dataset or datasets in case of two or more cell populations matching. **b** UMAP plot visualizing cells used to train the classification tree (blue) and cells for which cell type was not known (orange). **c** UMAP representation of annotated cell typesPark18, Kirita21, Dumas20, Miao21, Janosevic21 and Wu19. **d** UMAP embedding of predicted cell types for Hinze20, Conway20 and cells labelled as 'unknown' or 'missing'. **e** UMAP plot combining predicted and available annotations resulting in the integrated mouse kidney atlas.

infiltrating monocytes did not show any expression pattern in the latent space that could indicate their presence or any other subpopulations (**Fig. S6B**).

Given the two different suspension types present in the MKA (i.e., single-cell and single-nuclei), we investigated whether there are sampling differences between them at the cell type level. Most cell types have an equal contribution of single-cell and single-nuclei datasets (**Fig. S7A**). CNT, DCT-CNT, DTL-ATL, Fibroblasts, ICA, PC, Podocytes, and PTS3T2 have a significantly higher contribution from single-nuclei datasets when compared to single-cell ones. We observed the biggest effect size in the PTS3T2 population. To understand if these differences are due to variability in sampling between single-cell and single-nuclei datasets, we compared the total number of detected PTS3T2 cells or nuclei in each of the datasets (**Fig. S7B**). We observed that single-cell datasets had a very similar size in total to their single-nuclei counterparts, indicating that the lack of PTS3T2 cells in single-cell datasets is not due to overall under-sampling.

We sought to further explore the impact of suspension type in the cell types present in the MKA at the gene expression level. To this end, we correlated the batch-corrected expression for a given cell type between single-cells and single-nucleus (**Fig. S7C**). All cell types had a significant correlation between cells and nuclei, with their correlation coefficient being higher than 0.65 in all of them.

2.2.4 MKA ACCURATELY CLASSIFIES UNSEEN CELLS

To evaluate the accuracy of cell type classification using MKA as a reference, we performed leave-one-dataset-out cross-validation experiments with two different classifiers. In the first experiment, we chose one of the annotated datasets as a test set and trained the scHPL classifier on the remaining datasets (hereafter MKA*) at each iteration, with the same parameters as defined earlier in this manuscript. We then compared the performance of scHPL with Azimuth [74], a widely used pipeline to automatically annotate cells based on Seurat, in a second experiment. Following a similar approach as in the first experiment, we chose one of the datasets as a query dataset and set the rest of the annotated datasets as our partial reference (MKA*) and performed Azimuth's workflow with default parameters. Finally, we evaluated the performance of Azimuth to predict mouse cell types using the available human reference [103]. To do this, we performed a third experiment in a similar fashion to the previous two. At each iteration, we submitted each test set's raw counts (namely Park18, Miao21, Kirita20, Wu18, Dumas20, and Janosevic21) as a query to the Azimuth web application using the human kidney reference. The median F1 score of all folds across the three experiments (**Fig. 4A**) highlighted the importance of using the MKA when transferring labels to mouse datasets, regardless of the classifier. The human reference is also more likely to have outliers in terms of label transfer performance in a human-to-mouse scenario. For example, when using the human reference available, Park18 seems to be predicted at an accuracy closer to that obtained by using MKA (F1 score of 0.73). On the contrary, Janosevic21 is predicted with very poor accuracy (F1 score of 0.03).

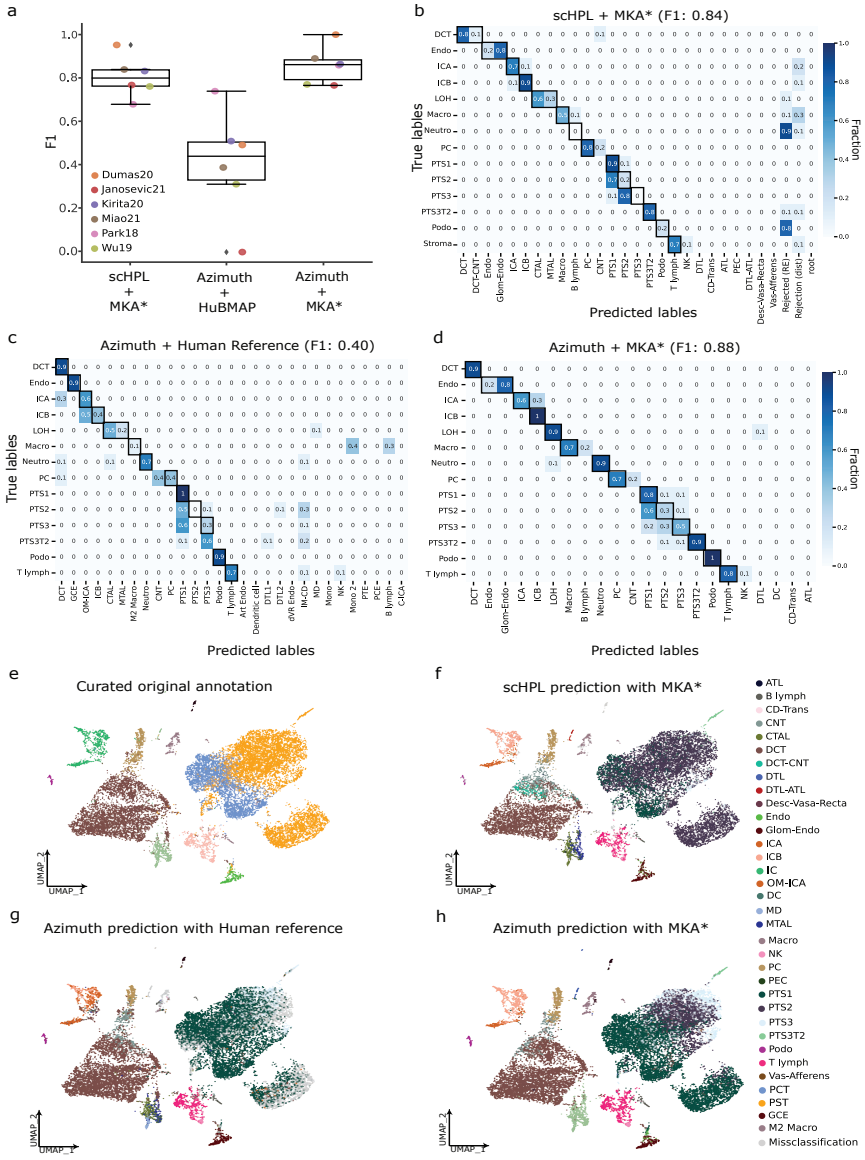

Figure 2.4: *Evaluation of the scHPL classifier.* **a** Boxplot of median F1 scores (y-axis) computed over six folds in three different scenarios (x-axis). From left to right, scHPL trained with MKA*, Azimuth's label transfer using the HubMAP reference available [104], and Azimuth's label transfer using MKA* as reference. Each dot corresponds to the median F1 score computed across cell populations for a given training and validation set (i.e., MKA* and each of the annotated datasets in MKA, respectively). **b-d** Confusion matrices normalized by class support size, computed using the predicted annotations by scHPL and our Atlas reference (**b**), the transferred labels from Azimuth's human kidney reference [103] (**c**), or the transferred labels from Azimuth using our reference (**d**). Higher values indicate higher agreement between predicted and true cell labels. **e-h** UMAP plot of the Miao21 dataset coloured by the original cell types (after manual re-annotation) (**e**), by the predicted cell types from the learned classification tree (**f**), by the transferred cell types from the Azimuth human reference (**g**), and by the transferred labels using Azimuth with our Atlas reference (**h**).

To further highlight the value of the MKA, we tested the cell-type label transfer accuracy when using single-dataset references. Overall, the accuracy greatly depends on the query dataset for which we are trying to predict the labels (**Fig. S8**), something that is mitigated by using the MKA as a reference (**Fig. 4A**). Despite matching the predicted and original sets of labels to account for inconsistencies in annotation resolution (i.e., original labels included PTS1, PTS2, and PTS3 but the predicted labels from using a single-dataset reference include only PT), the accuracy of the predictions for a given query greatly depends on the reference used. An example of such a case is Park18. Miao21 is the best reference in this case, with other datasets quickly dropping in accuracy (F1 scores of 0.68, 0.46, and 0 for Wu19, Kirita21, Janosevic21, and Dumas20, respectively). Another important pitfall of using single-dataset references is exemplified by Kirita21. In this dataset, the authors define novel cell states that lie in-between known cell types (i.e., DCT-CNT or ATL-DTL) and annotate other cell types at a great resolution. These cell types are not captured by other datasets, which affects their performance as references when predicting labels from Kirita21.

In order to understand the contribution of different cell types to the overall F1 scores, we chose Miao21 as a test set and trained the scHPL classifier on the MKA*. The resulting tree (**Fig. S9**) was then used to predict the labels of cells from Miao21. Most of the original annotations from the dataset were accurately predicted by the scHPL classifier with a median F1 score of 0.84 (**Figs. 4B and 4F**). Moreover, scHPL further classified cells at a higher resolution compared to the low-resolution labels present in the original dataset (**Fig. 4E**). For example, in the original study, Miao21 identified LOH cells, which scHPL can classify into MTAL and CTAL, the two major cell types present in the thick ascending limb of the Loop of Henle. Notably, some cells and nuclei were assigned to the root node (i.e., unclassified). For example, neutrophils were mostly rejected (**Fig. 4B**). This is not surprising, since there were only 26 neutrophil cells in the training data (i.e., MKA*), which inevitably led to poor performance in predicting neutrophils in Miao21. On the other hand, one of the most abundant cell types in the training data (DCT with 4559 cells and nuclei) is correctly predicted 90% of the time (DCT and DCT-CNT). scHPL can reject cells due to the lack of cells from a specific population during training, e.g., neutrophils. But rejection can also mean that the query dataset includes novel cell populations not seen during training. In the latter case, rejected cells can be further characterized and annotated to update the cellular knowledge stored in MKA.

As in our cross-validation experiments, we used Azimuth to predict the labels of our query dataset (Miao21) using both the human reference and our MKA*. In the case of the human reference, despite having a wider array of cell populations (46 populations), Azimuth misclassified many cells with a median F1 score of 0.40 (**Figs. 4C and 4G**). For instance, PC cells were classified as CNT or DCT 80% of the time. Previous studies have identified a transitional CNT-PC subpopulation of cells in healthy human kidney samples[105]. This finding suggests that the mislabelled cells may not be a distinct cell type, but rather in a transitional stage, given their transcriptomic overlap. This is a high rate of misclassification considering that these are two very distinct cell types specialized in different functions in the nephron. These misclassifications can be due to the lack of a rejection option in Azimuth, differences in the cell type-specific transcriptomic profiles between human and mouse kidney,

or a combination of both factors. When using our partial reference atlas (MKA*), we were able to accurately classify cells in the query data with a median F1 score of 0.88 (**Figs. 4D and 4H**). This result indicates that the low performance of Azimuth compared to scHPL is mainly due to the use of a human reference to classify mouse cells.

In order to understand how different populations contributed to the F1 score, we computed a median F1 score per cell type, model, and for each fold in the cross-validation experiments (**Fig. S10**). Two of the fourteen populations included were accurately classified (F1 > 0.8) across the different validation experiments. In the case of the MKA*+scHPL experiment, the number of accurately classified populations increases to seven out of fourteen. Despite proximal tubule cells (PTS1, PTS2, PTS3, and PTS3T2) being the most abundant cell type in the nephrons[99], we saw a lot of variation in the classification accuracy of these populations. In the training data (MKA*), PT cells account for 45% of the total number of cells and nuclei. PTS3T2 cells (3%) are accurately predicted when using MKA* and either scHPL or Azimuth. This can be explained by the lack of this population in the human kidney reference available. PTS1 and PTS2 cells (10% and 17%) display a high degree of F1 score variability across the different experiments (**Fig. S10**). This is expected, as segments 1 and 2 can be identified morphologically but have almost identical functionality in the nephron[106]. As a consequence, their transcriptomic profiles are highly overlapping, which has led to several authors considering them a single cell type[86,89]. None of the reference and classifier combinations we tested accurately classify both segments. PTS3 is the least abundant cell type in the nephron and has the highest accuracy score when using the MKA* with Azimuth. Even in this case, 50% of the time PTS3 cells are misclassified as either PTS1 or PTS2 (**Fig. 4D**).

2.2.5 MOUSE KIDNEY ATLAS FACILITATES THE IDENTIFICATION OF ROBUST CELL POPULATION MARKERS

Technological limitations in single-cell transcriptomics result in a high proportion of unmeasured genes leading to low replicability of cell type markers across different studies. We capitalized on the large collection of cells and nuclei from diverse samples in MKA to identify replicable cell population markers.

Based on MKA, we identified meta-markers, which are genes that have a high detection rate and are reliable markers for a given cell population across different datasets (see Methods for details). The resulting set of meta-markers per cell type included previously known markers (e.g., *Slc12a1* for both MTAL and CTAL, and *Slc12a3* for DCT) as well as novel candidates (e.g., *Bst1* for DTL or *Rhcg* for CNT) (**Fig. 5A** and Table S2).

To verify our newly identified meta-markers, we verified their expression in the respective cell types using: (i) bulk gene expression data from microdissected samples and (ii) 10x Visium spatial transcriptomic data. First, we used the gene expression profiles of 64 bulk RNA-seq samples obtained from microdissected kidney segments generated by Chen et

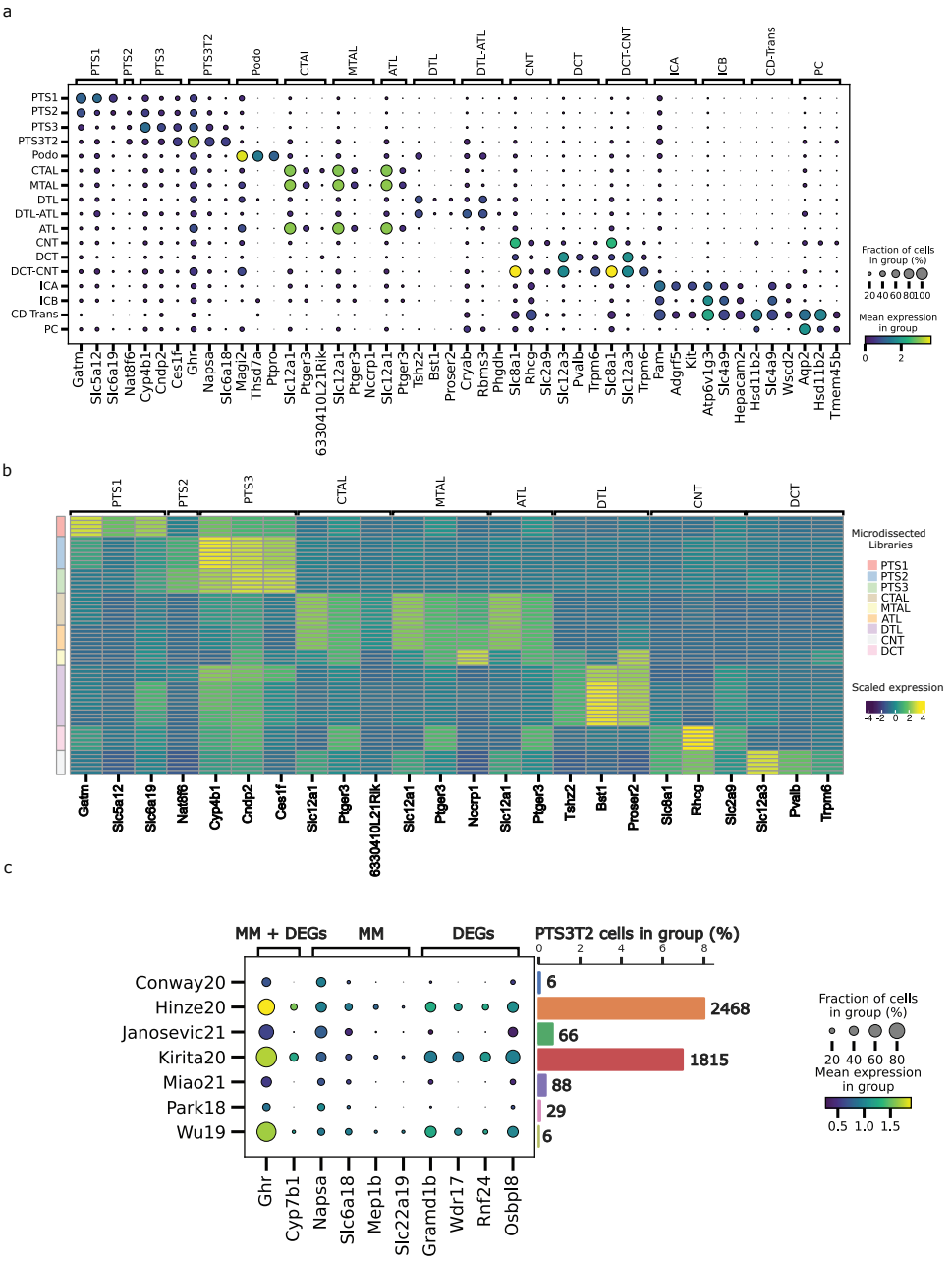

Figure 2.5: *Joint downstream analyses highlight known cell type markers and help define meta-markers across studies.* **a** Dotplot of the top 3 meta-markers (when their recurrence is equal or greater than 2) per cell type across datasets. Values sorted by fold change and auroc. **b** Heatmap showing the scaled and normalized transcript per million (TPM) expression of the top meta-markers in the microdissected kidney bulk-RNAseq libraries[39]. Only matching cell types between the two experiments were kept. Columns represent the RNA-seq libraries, rows correspond to genes. Both rows and columns are annotated by cell type. In the case of rows, the annotation corresponds to the cell type from which these markers were identified in the atlas. For the columns, the annotations are the different regions from which the RNA-seq libraries are derived. **c** Dotplot of the top 5 meta-markers (when their recurrence is equal or greater than 2; sorted by fold change of detection rate and auroc) and the top 5 DEGs for PTS3T2. Barplots show the number of PTS3T2 cells in each dataset both in relative (% of total cells in the dataset) and absolute terms (total number of cells on top of each bar).

al.[102]. These segments (excluding CD and Glomerulus) are identified morphologically and ideally contain a single cell population each. We confirmed the expression of the top three meta-markers that appeared in at least two per cell population in the bulk RNA-seq samples. These correspond to PTS1, PTS2, PTS3, CTAL, MTAL, ATL, CNT, DCT, and DTL (**Fig. 5B**). Furthermore, we found a significant overlap between the MKA-based meta-markers and the microdissection-defined markers (Table S3). Second, we used 10x Visium spatial transcriptomic data of the healthy mouse kidney from GSE171406[107]. We plotted the gene expression of two meta-markers for each of the following populations: PTS3T2, DCT-CNT, DTL-ATL, and CD-Trans (**Fig. S11A to S11D**). *Napsa* and *Slc6a18* expression in the corticomedullary junction of the kidney co-localizes with what others have previously described as a PTS3T2 cluster[90]. DCT-CNT (*Slc12a3-Trpm6*) and DTL-ATL (*Cryab-Phgdh*) meta-markers follow a characteristic cortical and medullary expression pattern, respectively. This is expected given that the individual cell types are mainly localized in the cortex (DCT and CNT) or medulla (DTL and ATL). CD-Trans meta-markers (i.e., *Slc4a9* and *Wscd2*) display a heterogeneous pattern across the tissue slide. This suggests a location similar to CD and IC cells in the kidney. The low number of spots with detectable gene expression of both meta-markers is in agreement with the low fraction of cells labeled as CD-Trans in MKA (less than 0.5%) and what others have reported[83].

To highlight the value of MKA and the meta-markers we identified, we investigated rare, understudied cell populations. First, we characterized a recently described cell type, PTS3T2 cells[90,105]. Together with PTS3, PTS3T2 cells are thought to play an important role in the kidney injury process[104]. However, the few available marker genes for PTS3T2 are based on unsupervised clustering of single-cell RNA-seq studies[90] and are yet to be validated. Within our MKA-based meta-markers for PTS3T2, we identified previously known markers, such as *Slc22a13*, as well as novel markers: *Ghr* or *Mep1b* (**Fig. S10A**). *Ghr* has been previously associated with chronic kidney disease[108], whereas *Mep1b* plays a role in acute kidney injury, with $Mep1b^{-/-}$ mice showing improved renal function compared to WT mice[109]. We compared the expression of the top five PTS3T2 meta-markers with the top five PTS3T2 differentially expressed genes in the MKA (**Fig. 5C and Table S4**). Meta-markers such as *Slc6a18* and *Napsa* displayed a robust expression pattern across the non-endothelial datasets (excluding Dumas20). However, DEGs such as *Gramd1b*, *Wdr17*, *Rnf24*, and *Osbpl8* were expressed mostly in datasets with the highest number of PTS3T2 cells, lacking replicability across studies. This was the case for the meta-markers *Mep1b* and *Slc22a19* too. *Ghr*, which encodes the growth hormone receptor, was identified as both a meta-marker and a DEG with detectable expression in all datasets. However, *Ghr* is a significant DEG in 30 of the 36 cell populations included in the MKA (Table S2), indicating that *Ghr* expression is not specific. *Cyp7b1* is also identified as both a meta-marker and DEG but its expression pattern is biased towards Hinze20, Kirita20, and Wu19.

Although single-cell studies usually aim to describe discrete cell types, kidney nephrons are tubular structures formed by a continuum of epithelial cells. Due to this, cells with mixed transcriptomic profiles are likely to be sequenced[84]. We set out to define meta-markers that are known for the cell types that are part of the mixed population but also to identify novel markers of transitional cell types. In the case of DCT-CNT (**Fig. S11B**), meta-markers

included known markers for both DCT[102] (*Slc12a3* and *Slc8a1*) and CNT[30] (*Trpm6*) cells. Novel markers for this mixed population included *Acss3* and *Ltc4s*.

Cryab, a known marker for ATL and LOH cells, is identified as a meta-marker for ATL-DTL cells (**Fig. S11C**). Other previously unknown meta-markers include *Rbms3*, *Phgdh*, and *Slc4a11*. Some of these genes have already been implicated in kidney biology. For instance, *Phgdh* has been identified as a treatment target in kidney cell carcinoma in patients resistant to HIF2α antagonists[110]. *Slc4a11* is known to be expressed in DTL cells, although expression has been described only in the medullary part of the kidney[111].

Next, we investigated the novel collecting duct transitional cell population (CD-Trans), which was described by Park and colleagues[83] and by Chen et al.[112], who labeled these as 'hybrid cells'. CD-Trans cells have been described as an intermediate state between PC and IC cells, expressing markers for both cell types[83,112]. While ICA and ICB cells play a role in the regulation of acid-base homeostasis[106], PCs' main function is salt and water transport. In the latter case, sodium (epithelial sodium channel, *Scnn1a/b*) and water (Aquaporin 2, *Aqp2*) channels control the levels of Na^+ and K^+ in plasma, blood pressure, and extracellular fluid osmolality. Further understanding of CD-Trans cells has been hampered by their low abundance in the kidney, often being masked by other cell types, such as proximal tubule cells. In MKA, CD-Trans cells were identified in four datasets (Park18, Miao21, Janosevic21, and Conway20) after annotation of the full atlas with 60 cells in total. Our meta-marker list for CD-Trans cells includes *Hsd11b2*, *Slc4a9*, *Wscd2*, and *B3gnt7*, which were found to be highly accurate and able to confidently classify cells as CD-Trans (**Fig. S11D**). Kidney-specific *Hsd11b2*$^{-/-}$ mice show systemic salt-dependent hypertension[113]. Moreover, CD-Trans cells in the MKA express both *Aqp2* and *Slc4a9* (meta-marker for ICB), further confirming the transitional state between PC and IC of these cells.

2.3 DISCUSSION

The maturity of single-cell and single-nuclei transcriptomics becomes apparent by the ever-increasing number of publications applying these technologies[114,115]. Although this has given rise to a vast collection of publicly available cellular transcriptomes, researchers continue to analyze their work in an isolated environment, often without considering the data from other reports. As it has been recently noted in the literature[105], the relationships between the populations defined in kidney single-cell studies are not clear, and integrative studies are needed. Here, we integrate cells and nuclei from eight independent studies (Table 1) to create the first mouse kidney atlas (MKA). We demonstrate that, despite between-sample biological and technical differences, our atlas establishes a robust and comprehensive view of the cell heterogeneity present in the mouse kidney.

A major challenge in single-cell analyses is cell type annotation. Usually, cell types are annotated based on the expression of marker genes in unsupervised clusters. Clustering algorithms require the tuning of hyperparameters, leading to a subjective choice on the

number of clusters. This is aggravated by the possible presence of new (sub)cell types in the dataset, which usually causes over-clustering [116]. This introduces subjectivity to the analysis, ultimately leading to incomplete and ambiguous annotations between studies. We highlight these inconsistencies in the case of the mouse kidney using scHPL, a supervised hierarchical machine learning model. By refinement of these annotations and further cell type learning, we improve the atlas reference transcriptome, accurately capturing consensus cell identities across studies. An important feature of such a model is its ability to capture the different resolutions at which cell types have been annotated. For example, some studies limit their labeling to LOH cells, while others further classify these cells as MTAL or CTAL [84,89]. In our work, we convey a hierarchically defined atlas, further characterizing the variety of cell types present in the healthy mouse kidney (**Fig. 3E**). Consequently, we identify 35 distinct cell types, including both high- and low-resolution annotations. We have shown that most of these cell types are equally detected in both single-cell and single-nuclei studies (**Figs. S7A-B**). Despite single-cell studies having a similar number of cells, PTS3T2 cells are detected in higher proportions in single-nuclei studies. We hypothesize that PTS3T2 cells are harder to detect in single-cell studies, possibly due to differences in their survival in cell and nuclei isolation protocols. Differences in cell type composition between single-cell and single-nuclei studies have been reported before [117]. As noted by Wu et al., single-nuclei RNA-seq was able to detect 20-fold more Podocytes than the proportions reported by single-cell studies. In addition, Mesangial cells were completely missing from their single-cell dataset, further elucidating the differences in detection between dissociation protocols.

Unfortunately, our atlas cannot predict, with full accuracy, all cell types in the kidney. This limitation is not exclusive to this organ, as supervised cell classification remains a challenge for all tissues. It is often due to the lack of a precise definition of cell types, lack of robust markers, technical limitations, and sampling variability [117]. In addition, renal plasticity and the ability of renal cells to switch cell types might generate some less defined cells [118,119]. In our work, we highlight the common misclassification of PTS1, PTS2, and PTS3 cells by different methods (**Figs. 4B to 4D**). Although functional differences between the segments are known, the different segments have traditionally been identified based on cell ultrastructure [120]. This results in their transcriptomes being too similar, rendering these cells hard to classify computationally. As indicated by Shanley and colleagues [121], the third segment of the proximal tubule is particularly vulnerable to ischemic damage. It is not yet clear whether what we and others [90] have identified as PTS3T2 constitutes a genuine cell type or rather a damaged state of PTS3 cells. We would like to note, however, that by extensive integration of datasets we can largely overcome these shortcomings, as we have demonstrated in the present work. As the field develops, and clearer definitions are proposed, the inclusion of more datasets into our atlas will further enhance cell type identities and classification. For example, if including larger unannotated healthy samples in the MKA results in more cells being classified as PTS3T2, there would be *in silico* evidence that this cell type is indeed present in healthy proximal tubules. However, to confirm its identity as a cell type, a complete understanding of its origin from a developmental perspective is probably needed.

The importance of our work is further highlighted by the pressing need to develop novel

therapies for kidney failure. Kidneys are the most frequently transplanted organ. Due to the increasing prevalence of chronic kidney diseases in the population, demand exceeds the number of available donors, and strategies based on renal (stem) cells are being investigated. On grounds of these and other shortcomings, as noted in the literature [122], an understanding of the cell heterogeneity present in the kidney is needed in order to develop much-needed therapies. The efficacy of these will depend on the cell type-specific expression and activity of pathways [123]. Despite this, the knowledge of cell types, their markers, and the molecular mechanisms and pathology underlying these diseases at the single-cell level is still incomplete. For example, a recent study shows that CNT cells can display a partial DCT phenotype [124]. However, this transitional cell type (DCT-CNT) is usually not identified or masked by more abundant cell types in single-cell studies. Consequently, most reports identify individual CNT and DCT clusters [125]. To this end, the kidney atlas can aid the discovery of robust novel markers for DCT-CNT cells. These markers are detected across the different datasets and can accurately classify DCT-CNT cells. As demonstrated by the above example, we identify meta-markers for the cell types present in our atlas, including previously known and novel genetic markers. When compared to markers obtained without accounting for each individual dataset in an integrated space, meta-markers with a high detection rate can provide replicability that generalizes the cell type identities defined in our atlas. A clear example of such a scenario is DTL-ATL cells. As has been described previously, one of the meta-markers identified for this population is *Slc4a11*, which expresses a membrane transporter involved in water, ammonia, and H^+ transport. Its expression has been located in DTL cells within the outer medulla and the outer stripes of the inner medulla in mice [111].

These findings will benefit the broader kidney research community, for example, by aiding the robust *in vivo* identification of cell types. Since human and mouse kidneys show important physiological differences at the cellular level [126], we believe our work is especially relevant in mouse models. The discovery potential of our atlas, however, is much broader and largely not explored. We acknowledge that, although statistically robust and *in silico* verified with microdissected nephron segments [102] and spatial transcriptomics tissue slides [107], these compendium-wide markers need further *in vivo* validation.

Cellular knowledge of the kidney is likely to change in the coming years. As technologies improve and innovative studies are published, novel cell types will be described. Likewise, cell identities will be re-defined in newer contexts. We aim to incorporate these changes within the atlas in a continuous fashion. We provide a learned transcriptome-based cell hierarchy that can be easily updated and improved with newer studies, updating the cellular knowledge captured in the compendium. In addition, our atlas is missing cell types that were not present in the original set of annotations provided by the authors. We've shown that the MKA can potentially detect previously unannotated cell types, such as vascular smooth muscle cells. Because we used scVI and scANVI as our integration model, we can leverage scArches [127] to update the latent space of our atlas without retraining. For example, the recently published dataset by Song et al. [128] could be used to enhance the MKA with rich immune annotations. In other instances, updating the latent space and scHPL's classification tree will allow us to annotate matching cell types and identify potential novel populations that arise from treatment or disease state. To account for the technical variation

of new datasets in the context of the MKA, one can make use of the pre-trained and optimized model we present to obtain an updated latent space that we then use to update the classifier. To make this easily accessible to the community, we share our atlas via a user-friendly web interface, hosted at Cellxgene (`https://cellxgene.cziscience.com/e/42bb7f78-cef8-4b0d-9bba-50037d64d8c1.cxg/`).

In summary, we leverage the large collection of publicly available single-cell and single-nuclei studies and establish a dynamic atlas of the mouse kidney. We demonstrate the extraordinary power of such an approach by providing robust markers for elusive cell types. However, the full potential of the created compendium is yet to be explored.

2.4 LIMITATIONS OF THE STUDY

This is the first complete single-cell and single-nuclei atlas of the mouse healthy kidney that harmonizes annotations across several publicly available datasets. However, there are several limitations in our study. Firstly, the lack of in vitro or in vivo validation of the computed metamarkers. Secondly, lack of more recent studies and different single-cell technologies. All the cells and nuclei in our atlas come from droplet-based libraries and short read sequencing. Adding plate-based technologies in the atlas might prove beneficial in the future, as more low-abundant transcripts are detected, a more accurate cell type classification will be achieved.

2.5 ACKNOWLEDGEMENTS

We thank the Chan Zuckerberg Initiative (CZI) for their help with integrating our MKA atlas in cellxgene. We thank investigators for making their data publicly accessible, as it has been essential to this work. We also thank Lieke Michielsen for her help to set up scHPL and useful feedback. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860977.

2.6 AUTHOR CONTRIBUTIONS

C.N.R, M.G, D.P and A.M designed the study. C.N.R performed the analyses and experiments. A.M provided feedback on the computational biology results. D.P provided feedback on the kidney physiology side of the project. All authors discussed the results and commented on the manuscript.

2.7 METHODS

2.7.1 KEY RESOURCES TABLE

RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python 3.7.12	Python	python.org
R 4.0.5	R	r-project.org
scvi-tools 0.19.0	GitHub	github.com/scverse/scvi-tools
scHPL 1.0.0	GitHub	github.com/lcmmichielsen/scHPL
Custom code	This paper	github.com/nrclaudio/MKA
Datasets		
Wu19	10.1681/ASN.2018090912	GSE119531
Miao21	10.1038/s41467-021-22266-1	GSE157079
Park18	10.1126/science.aar2131	GSE107585
Kirita20	10.1073/pnas.2005477117	GSE139107
Dumas20	10.1681/ASN.2019080832	E-MTAB-8145
Conway20	10.1681/ASN.2020060806	GSE140023
Hinze21	10.1681/ASN.2020070930	GSE145690
Janosevic21	10.7554/eLife.62270	GSE151658

2.7.2 RESOURCE AVAILABILITY

LEAD CONTACT

Further information and requests for analyses or method details will be fulfilled by the lead contact, Ahmed Mahfouz (a.mahfouz@lumc.nl).

MATERIALS AVAILABILITY

This study did not generate new unique reagents.

DATA AND CODE AVAILABILITY

- All single-cell and/or single-nuclei RNA-seq datasets used in this study are publicly available. Their accession numbers are listed in the key resources table.
- Jupyter notebooks and scripts used in the analyses as well as supplemental data are available on GitHub (github.com/nrclaudio/MKA). Interactive visualization and downloading of the kidney mouse atlas are available at cellxgene (cellxgene.cziscience.com/e/42bb7f78-cef8-4b0d-9bba-50037d64d8c1.cxg/).

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

2.7.3 METHOD DETAILS

COLLECTING RAW DATA AND QUANTIFICATION OF READS

All raw fastq files were downloaded from the Sequence Read Archive (SRA) using parallel-fastq-dump (v0.6.7; `github.com/rvalieris/parallel-fastq-dump`). Accession numbers and other relevant metadata are provided in **Table 1**. Single-cell and single-nuclei droplet-based sequencing data were aligned and quantified using `kallisto/bustools`[129,130] (`kb-python` v0.26.4) `ref` and `count` wrappers, specifying `-workflow nucleus` in the case of single-nuclei sequencing experiments. Reads were pseudo-aligned to the mouse reference genome GRCm38 downloaded from Ensembl[131].

PRE-PROCESSING OF SEQUENCING DATA AND NORMALIZATION

Filtered count matrices from `Kallisto/bustools` were used when the cell count was within a 10k margin from the matrices deposited by the authors. Otherwise, the unfiltered count matrices were loaded, and barcodes were matched between the author's and the unfiltered set of cells. Count matrices were pre-processed using `Scanpy`[132] (v1.8.1). We applied quality filters to all samples, specifically, we filtered out cells with more than 50% of counts derived from mitochondrial genes. Furthermore, we applied dataset-specific quality filters based on the number of detected genes. These filters are available in **Table S5**. Samples were merged and normalized for plotting purposes with `Scanpy`'s `normalize_total`.

INTEGRATION BENCHMARK

We compared several integration methods for our use case, including `scVI`, `scANVI`[51,59], `Harmony`, `Scanorama`[95,97], and `Seurat`'s[96]. Based on the evaluation results (see below in the **Results** section), we chose to use a hybrid approach in which we start with fully unsupervised integration using `scVI` followed by a refinement step using `scANVI` (**Fig. S12**, steps 1 – 2). `scANVI` uses cell type labels to inform the manifold-learning process such that cells with the same label are explained by similar low-dimensional features. This improves the representation learnt by `scVI` by incorporating biological information (such as cell types) in the model. This workflow (i.e. improving the latent representation of `scVI` with cell type labels using `scANVI`) is denoted as `scVI-scANVI` from now on. As different hyperparameter combinations and model configurations can affect the performance of deep learning models, we used `Ray tune`[133] to optimize `scVI`'s model. Raw counts and batch information were used to test 1000 different hyperparameter combinations. Our search space consisted of model configurations such as continuous and categorical covariates; model hyperparameters such as dropout rate, number of layers and number of latent dimensions;

learning hyperparameters such as learning rate and preprocessing steps such as highly variable genes (HVG) filtering and number of HVGs. The objective function to optimize was the silhouette score of both batch and cell type information as implemented in `scib`[134]. Detailed information and the scripts used to perform these analyses are available at `https://github.com/nrclaudio/MKA`.

INTEGRATION METRICS

Batch and biological conservation metrics were computed using `scib` (v1.0; `https://github.com/theislab/scib`). Note that some of the metrics are scaled to range from 0 to 1, for details refer to the original publication[134]. Batch conservation metrics include graph Local Inverse Simpson's Index[95] (LISI), kBET, Average Silhouette Width (ASW) and Principal Component Regression (PCR)[135]. In short, these metrics quantify the alignment between the different batch labels in the data. Specifically, kBET examines to what extent the different batches are mixed when neighbourhoods are randomly sampled, LISI captures the diversity of batches within a local neighbourhood of cells and PCR explains the total variance attributed to the batch variable when regressed on the Principal Components of the data. Biological conservation metrics include some of the previous metrics applied to cell type labels (cell-type ASW and LISI), Adjusted Rand-Index (ARI)[136], Normalized Mutual Information (NMI)[137], trajectory, cell cycle and variable gene conservation. These metrics quantitatively assess how much of the original biological variation is kept in the integrated space. The final score for each evaluation was computed as a weighted average of biological conservation and batch removal scores, with weights 0.6 and 0.4 respectively.

DATASET INTEGRATION

We used the method with the highest overall score in the benchmark to integrate the different studies (i.e. our tuned version of `scVI-scANVI`). The hyperparameter configuration with the highest silhouette score obtained in our tuning experiment (see Integration benchmark) was then used to train `scVI`. In our case, we reduced the feature space of our atlas to the top 3000 HVGs. Variable genes were obtained using `Scanpy`'s `highly_variable_genes` with the flavour set to `seurat_v3` and the `batch_key` set to the different datasets of origin. We included the percentage of mitochondrial reads as a continuous covariate in the model and the source of the material (cells or nuclei) as a categorical covariate. The model was initialized with 2 hidden layers, 26 latent dimensions, a dropout rate of 0.096 and the gene likelihood set to a Negative Binomial distribution. The model was then trained for a total of 111 epochs with a learning rate of 0.0013. The obtained model was then used as input for `scANVI` in order to further improve the latent space representation. We included available cell type annotations and set the `unlabeled_category` to the set of cells with missing annotations. The `scANVI` model was trained to a maximum of 20 epochs and with 100 cell subsamples per label class per training epoch.

DIMENSIONALITY REDUCTION

After integration and batch-correction, 26 latent dimensions were obtained from the model. These were used as input for the Nearest Neighbor graph calculation using `Scanpy`'s `neighbors` function. We further reduced the dimensionality to visualize the data in a 2D UMAP using 26 latent dimensions.

SIMILARITY METRICS

To assess cell population similarity across studies, pairwise similarity measures were computed using `sklearn`[137] `pairwise_distances` with the correlation metric. The similarity between two cell populations is reported as $1-$ correlation distance between their average normalized transcriptomic profile. Correlations between single-cell and single-nuclei profiles were computed using `scipy`'s `pearson_r` (`scipy.stats.pearsonr`). The input vectors per cell type and suspension type were obtained using `scANVI`'s `get_normalized_expression` with the `transform_batch` option set to the list of datasets in the atlas. The counts were then scaled by a factor of 1000.

CELL TYPE LEARNING AND CLASSIFICATION

All 26 latent dimensions from the annotated datasets (Park18, Wu19, Miao21, Kirita20, Dumas20 and Janosevic21) along with their original (**Fig. S12** step 3) or curated (**Fig. S12** step 4) cell type labels were used as input for single-cell Hierarchical Progressive Learning[34] (`scHPL`, v1.0.0). For both the original and the curated labels, the classification tree was learnt using a `kNN` and default values. To classify the cells that were missing annotations, the learnt tree and the latent dimensions from Hinze20 and Conway20 were used as input for `scHPL`'s `predict_labels` function.

EVALUATION OF THE CLASSIFIER

We used leave-one-dataset-out cross-validation experiments to evaluate the classifiers performances. At each iteration, we select one of the six datasets as a test set and treat the rest of the dataset as our training set.

To evaluate the performance of `scHPL`, the classification tree was learnt as described in the previous section. At each iteration, the test set labels were predicted and compared to the original curated labels. To measure the accuracy of the prediction, the F1 score (harmonic mean of the precision and recall) was computed, for every cell population, using `scikit-learn` v1.0.1 `f1_score` function with the `average` set to `micro`. The overall F1 score per dataset was computed as the median of F1 scores across cell populations.

To compare the performance of scHPL trained in our reference with other methods and references, we submitted each test set's raw count data as a query to Azimuth with the Human Kidney Reference atlas [74,103]. We kept the quality control filters we applied in our own preprocessing. The l2.annotation labels were transferred to the query using Azimuth. We also tested the performance of Azimuth's workflow (In Seurat [96]: Seurat::SCTransform, Seurat::FindTransferAnchors and Seurat::MapQuery) to transfer the labels from our reference to the test query dataset. As described previously, the annotations predicted by Azimuth with the Human Kidney Reference and our own reference were compared to the original curated labels of the query. The accuracy of this prediction was computed as an overall F1 score.

For each evaluation experiment (i.e. scHPL trained with our reference, Azimuth trained with our reference and Azimuth trained with the available Human reference) the median F1 score across all folds was computed. In the case of Miao21, for each pair of predicted and original labels, confusion matrices were computed using scHPL's confusion_matrix function. The row vectors of these matrices were normalized to sum up to 1.

To compare the label-transfer accuracy of using single-dataset references against using our atlas as a reference, we performed evaluation experiments on each of the studies in our atlas that have available annotations (i.e. Miao21, Park18, Kirita21, Janosevic21, Wu19 and Dumas20). For each dataset, we predicted the original labels using Azimuth's label transfer workflow treating each of the remaining studies as a reference. For example, in the case of Miao21, we predicted its labels from five different references, corresponding to each of the remaining datasets (i.e. Park18, Kirita21, Janosevic21, Wu19 and Dumas20). We then computed a median F1 score across references for a given query.

DIFFERENTIAL GENE EXPRESSION ANALYSES, META-MARKER DISCOVERY AND VERIFICATION

Cell type markers were computed using Scanpy's rank_gene_groups function with the Wilcoxon rank-sum test. Meta-markers were computed using the MetaMarkers R package [138] (v0.0.1; https://github.com/gillislab/metamarkers). Raw counts were converted to CPM values (as in original work). Markers were computed for each dataset with compute_markers to then obtain meta-markers using make_meta_markers. These two functions are using a Mann-Whitney test per dataset and an aggregation based on meta-analysis of the obtained p-values, respectively. Pareto boundary markers (i.e. markers with high precision and detection rate) were visualized using plot_pareto_markers.

IN SILICO VERIFICATION OF META-MARKER EXPRESSION IN KIDNEY TISSUE

To verify that the expression of the computed meta-markers agrees with the spatial location of their cell types, we plotted their log-normalized gene expression values in a healthy mouse kidney spatial transcriptomics tissue slide [107] from Gene Expression Omnibus (GEO,

GSE171406). Spots with less than 2000 unique genes expressed or higher than 50% of mitochondrial reads were removed.

COMPARISON WITH MICRODISSECTED KIDNEY BULK RNA-SEQ

TPM values were downloaded from GSE150338. From a total of 96 cell type bulk RNA-seq libraries, we kept 64 corresponding to the matching cell types in our atlas. TPM values were normalized as $\log_2(\text{TPM } 1)$. We then visualized the normalized expression of the previously computed cell type markers in the bulk RNA-seq context using `pheatmap` (`https://github.com/raivokolde/pheatmap`). To test the significance of the overlap between the lists of differentially expressed genes (i.e. cell type markers defined by either our kidney atlas or the microdissection study), we used `scipy`'s v1.5.4 Fisher exact test (`fisher_exact`) in every cell type present in both the atlas and the microdissection study. We used the list of genes present in the atlas as background in the test. In both cases, we considered as significant those genes with an adjusted (using Benjamini-Hochber's FDR correction) p-value 0.01. 99% Confidence intervals were computed for the odds ratios obtained in the test. This test evaluates whether a list of significant markers is independent of the list of markers that it is being compared to.

2.8 SUPPLEMENTARY MATERIALS

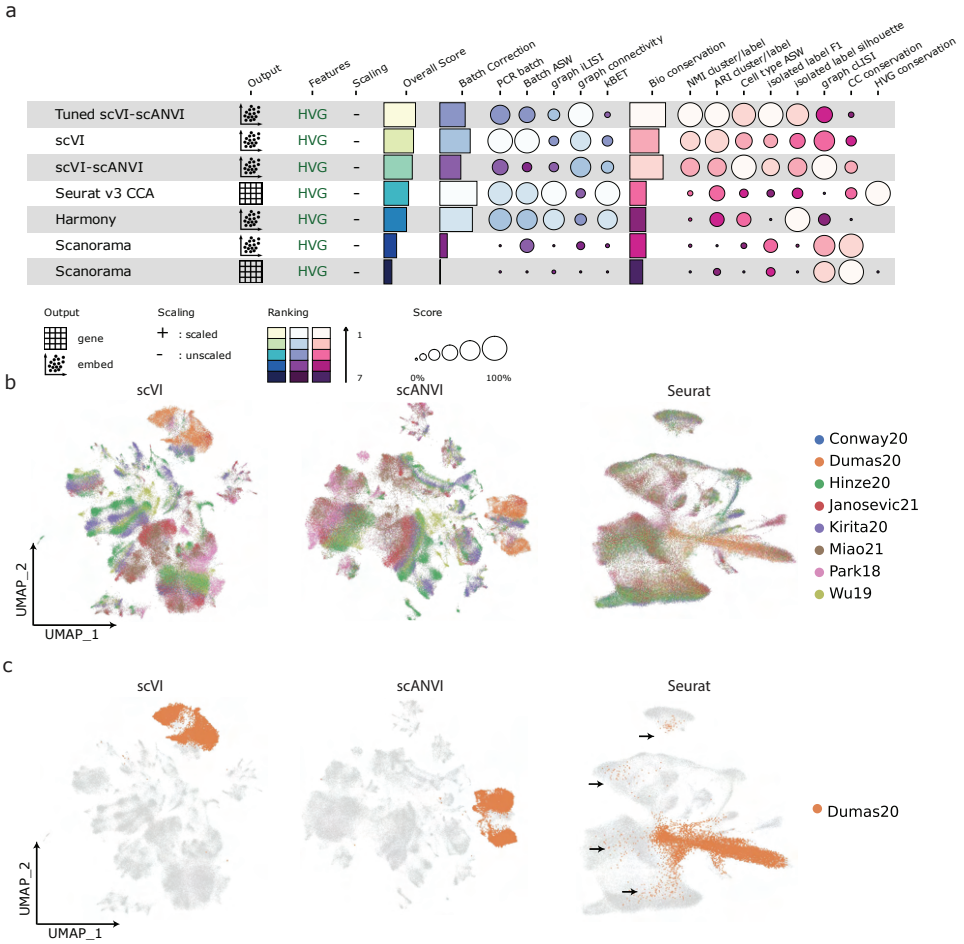

Figure 2.S1: Comparison of integration methods. a Integration benchmark summary plot generated using default parameters for all methods compared. In all cases, 3000 highly variable genes were selected before integration. Methods are ranked based on overall score, computed with a weight of 0.6 and 0.4 for biological conservation and batch correction scores respectively (see Methods for details). **b** UMAP plots generated using batch-corrected latent features (scVI and scANVI) or batch-corrected expression matrices (Seurat). Coloured by dataset of origin. **c** UMAP plots of the different integration methods highlighting in orange cells from the Dumas20[13] dataset, containing exclusively endothelial cells. Arrows roughly indicate cells whose signal is overcorrected for batch differences, diluting the biological signal coming from the Dumas20 dataset.

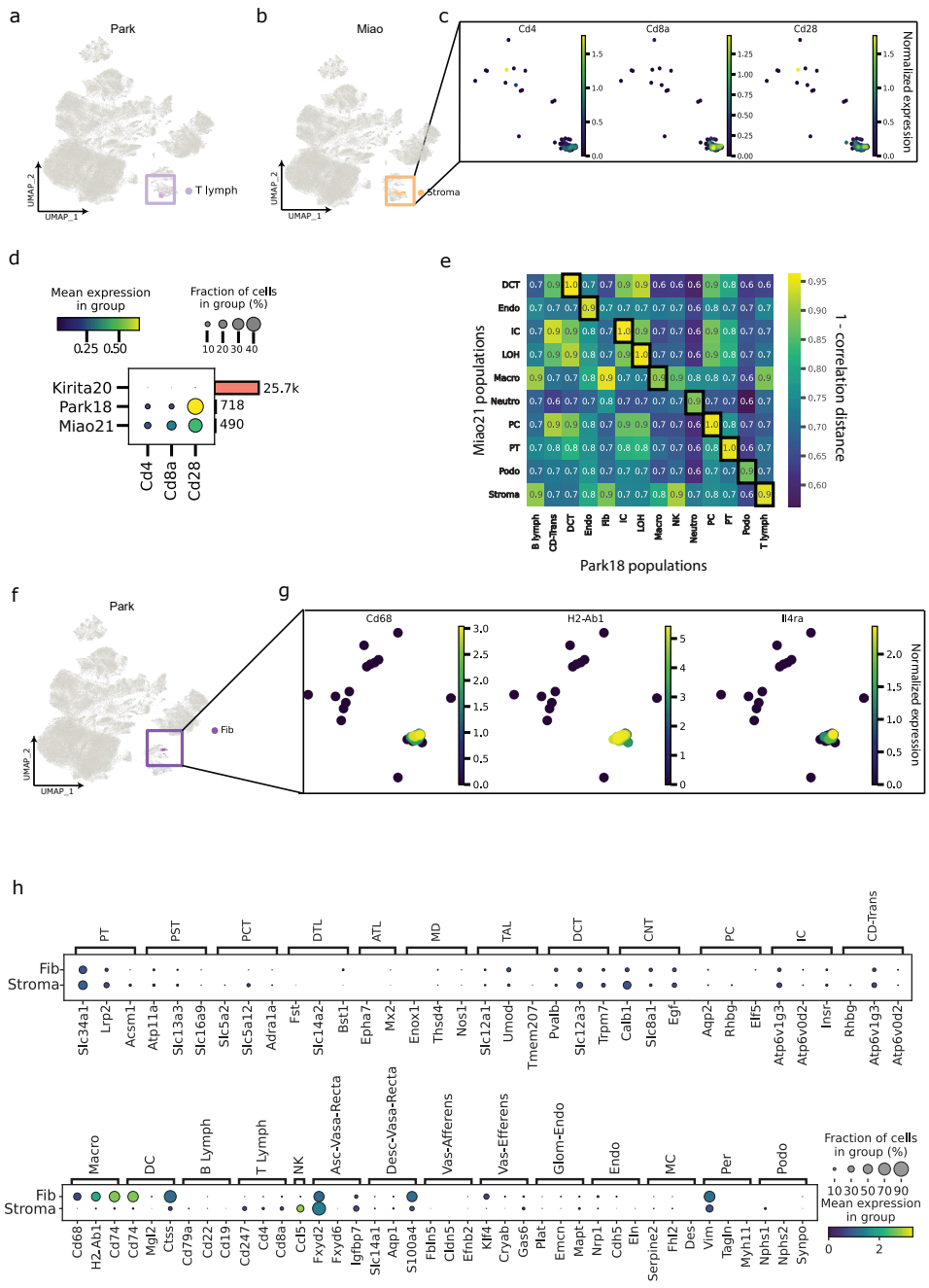

Figure 2.S2: *Manual annotation of dissenting cell populations based on scHPL.* (**a-b**) UMAP plots, coloured by both T lymphocytes$_{\text{Park18}}$ (**a**) and Stroma$_{\text{Miao21}}$ cells (**b**). (**c**) UMAP plots of Stroma$_{\text{Miao21}}$ cells coloured by scaled T lymphocyte marker expression (*Cd4, Cd8a, Cd28*). (**d**) Dot plot showing the frequency and average level of expression in selected T lymphocyte markers (*Cd4, Cd8a, Cd28*) in Stroma$_{\text{Miao21}}$ cells, T lymphocytes$_{\text{Park18}}$ as positive control, and Kirita20 non-immune cells as a negative control. The bar plot (right) shows the total cell/nuclei count in each row. (**e**) Heatmap of the similarity (expressed as $1-$ correlation distance) of cell populations between Miao21 and Park18. (**f**) UMAP plot, coloured by Fibroblasts$_{\text{Park18}}$. (**g**) UMAP plots of Fibroblast cells coloured by scaled Macrophage marker expression (*Cd68, H2-Ab1, Il4ra*). (**h**) Dotplot of canonical markers$_{44}$ per cell type in both T lymphocytes$_{\text{Park18}}$ and Stroma$_{\text{Miao21}}$ cells.

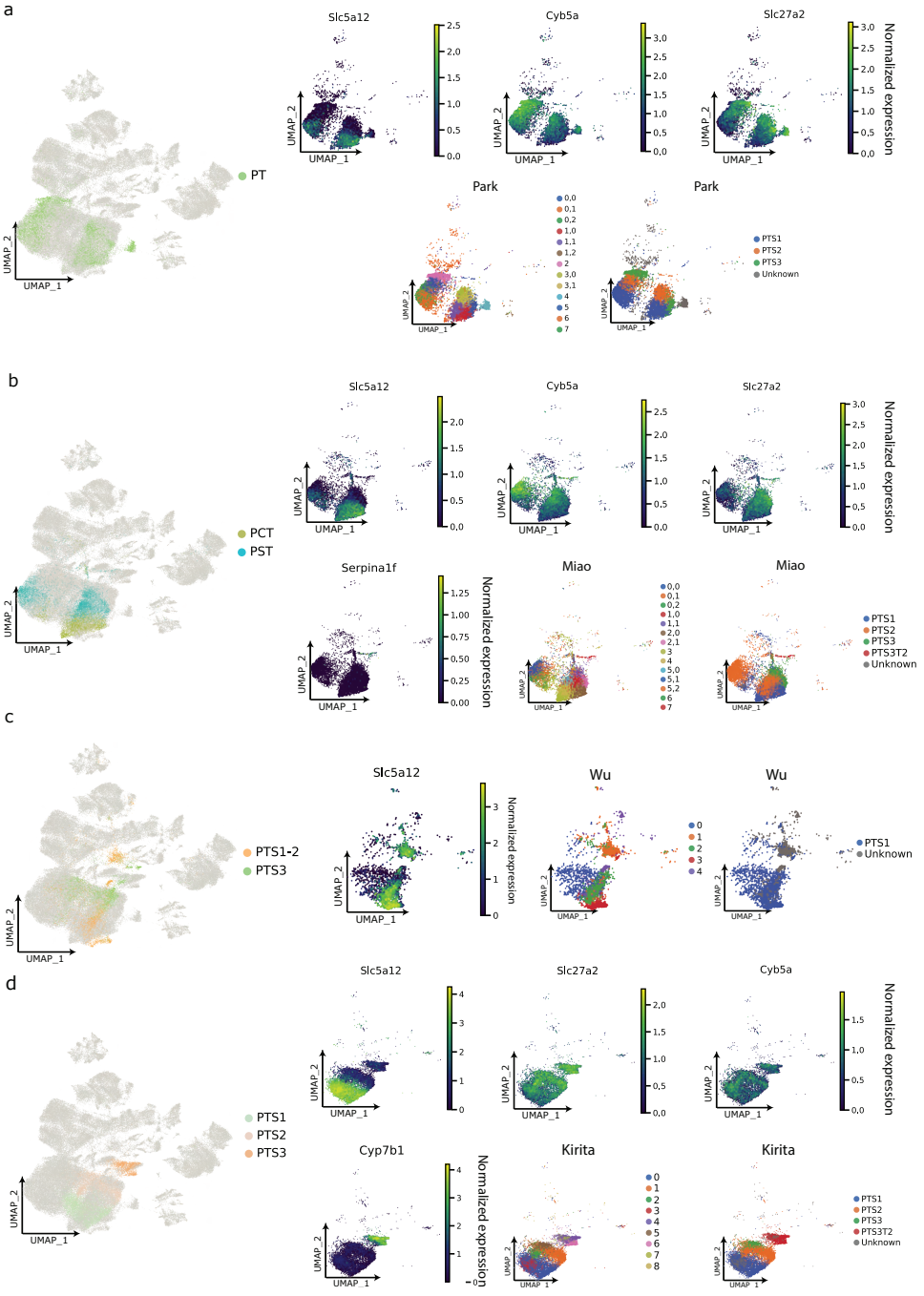

Figure 2.S3: *Integrated atlas allows further division of proximal tubule cells.* (**a-d**) UMAP plots for Park18 (**a**), Wu19 (**b**), Miao21 (**c**) and Kirita20 (**d**) highlighting *PT, PTS1, PTS2* and/or *PTS3* (top left), their marker expression (top right; *PTS1*: *Slc5a12*; *PTS2*: *Cyb5a*; *PTS3*: *Slc27a2*; *PTS3T2*: *Cyp7b1*) and the unsupervised clusters (bottom left). Clusters are renamed to *PTS1, PTS2, PTS3* or *PTS3T2* according to the marker expression overlay (bottom right). If the given marker has no expression on the dataset, the corresponding UMAP plot is omitted. **PT**: Proximal Tubule, **PTS1**: Proximal Tubule Segment 1, **PTS2**: Proximal Tubule Segment 2, **PTS3**: Proximal Tubule Segment 3, **PTS3T2**: Proximal Tubule Segment 3 Type 2.

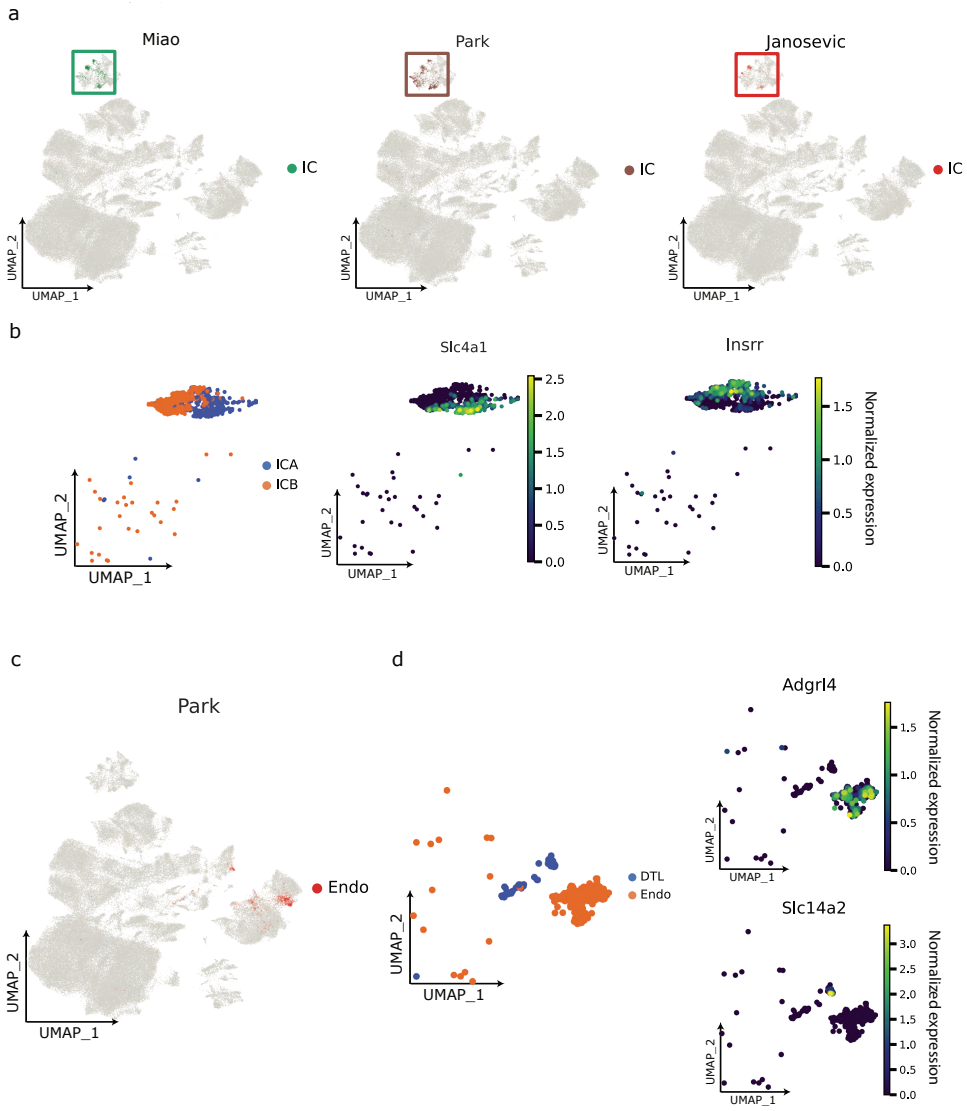

Figure 2.S4: *Further division of collecting duct intercalated cells and endothelial cells.* (**a-b**) UMAP plots coloured by IC_{Miao21}, IC_{Park18} and $IC_{\text{Janosevic21}}$ (**a**) and $Endothelial_{\text{Park18}}$ cells (**b**). (**c**) UMAP plot coloured by the Endothelial cluster. (**d**) UMAP plots coloured by renamed cluster (left) and marker expression (right) of *IC* (top) or *Endothelial* cells (bottom). Clusters are renamed to *ICA*, *ICB*, *Endothelial* or *DTL* according to the marker expression overlay (*ICA*: *Slc4a1*; *ICB*: *Insrr*; *Endo*: *Adgrl4*; *DTL*: *Slc14a2*).

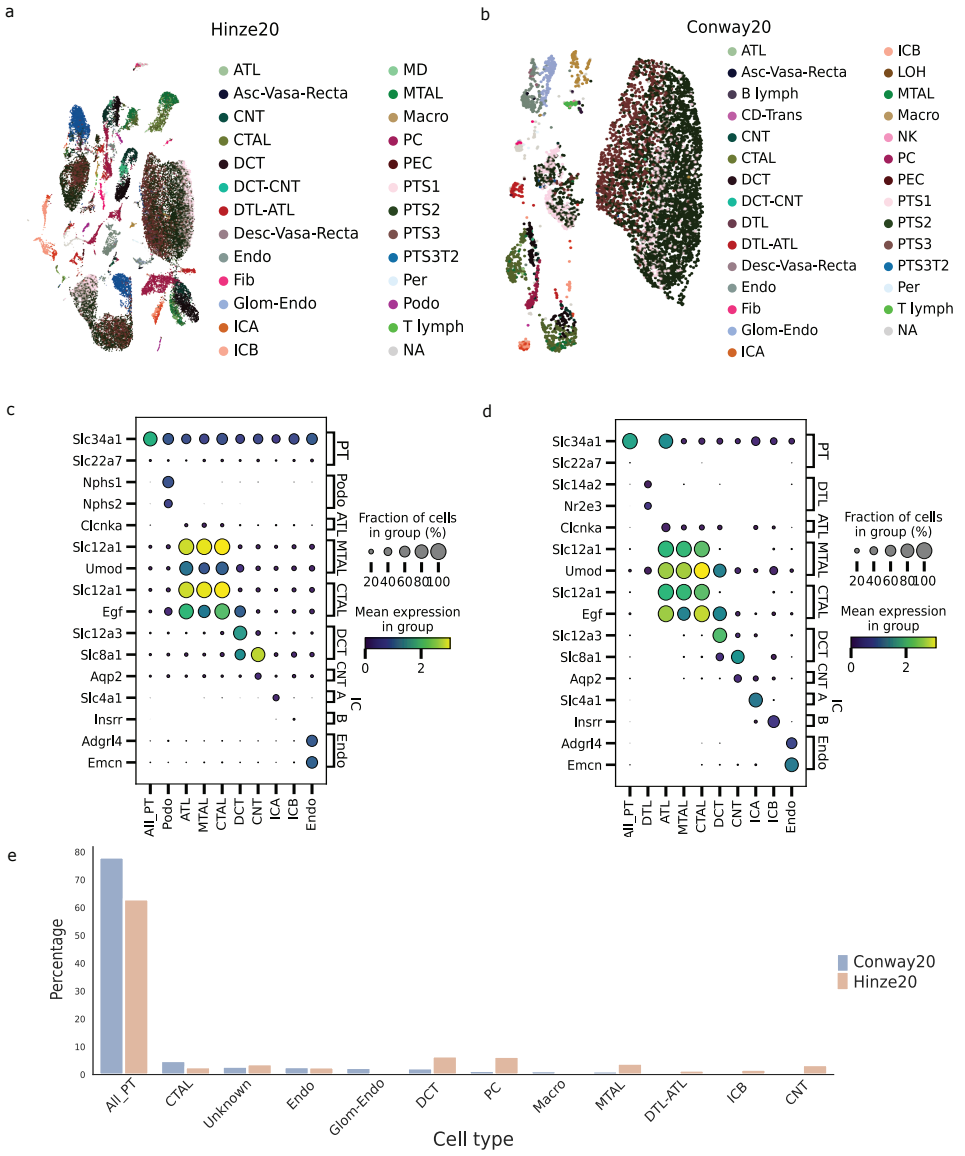

Figure 2.S5: *Cell type prediction per dataset.* (**a-b**) The final learned hierarchy (classification) tree was used to individually predict the cell types of the unannotated cells in the used datasets in the atlas. (**c-d**) Dotplot of predefined set of standard markers in every predicted population per dataset. *IC*: Intercalated Cell, *ICA*: Intercalated Cell Type A, *ICB*: Intercalated Cell Type B, *Endo*: Endothelial Cell, *Fib*: Fibroblast, *Macro*: Macrophage, *B lymph*: B lymphocyte, *Stroma*: Stroma cell, *NK*: Natural Killer, *T lymph*: T lymphocyte, *PT*: Proximal Tubule, *PTS1*: Proximal Tubule Segment 1, *PTS2*: Proximal Tubule Segment 2, *PTS3*: Proximal Tubule Segment 3, *PC*: Principal Cell, *PEC*: Parietal Epithelial Cell, *Per*: Pericyte, *DCT*: Distal Convoluted Tubule, *ATL*: Ascending Thin Limb of Henle, *MD*: Macula Densa, *LOH*: Loop of Henle, *CTAL*: Thick Ascending Limb of Henle in Cortex, *MTAL*: Thick Ascending Limb of Henle in Medulla, *CNT*: Connecting Tubule, *Podo*: Podocyte, *DTL*: Descending Thin Limb of Henle, *MC*: Mesangial Cell, *Neutro*: Neutrophil, *Asc-Vas-Recta (Asc VR)*: Ascending Vasa Recta, *Desc-Vas-Recta (Desc VR)*: Descending Vasa Recta, *Glom Endo*: Glomeruli Endothelial.

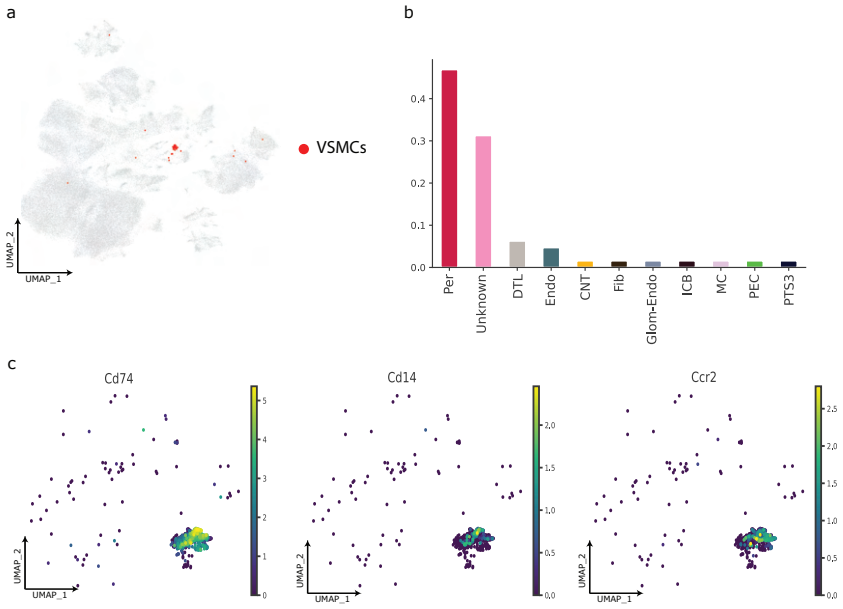

Figure 2.S6: *Identification of additional subpopulations.* (**a**) UMAP plot coloured by VSMCs, defined as cells expressing both *Acta2* and *Myh11*. (**b**) Bar plot with the % of cell type labels present in the 689 detected VSMCs. (**c**) UMAP plots of Macrophages coloured by the log-normalized expression of Macrophage marker (*Cd74*) and Infiltrating monocyte markers (*Cd14* and *Ccr2*).

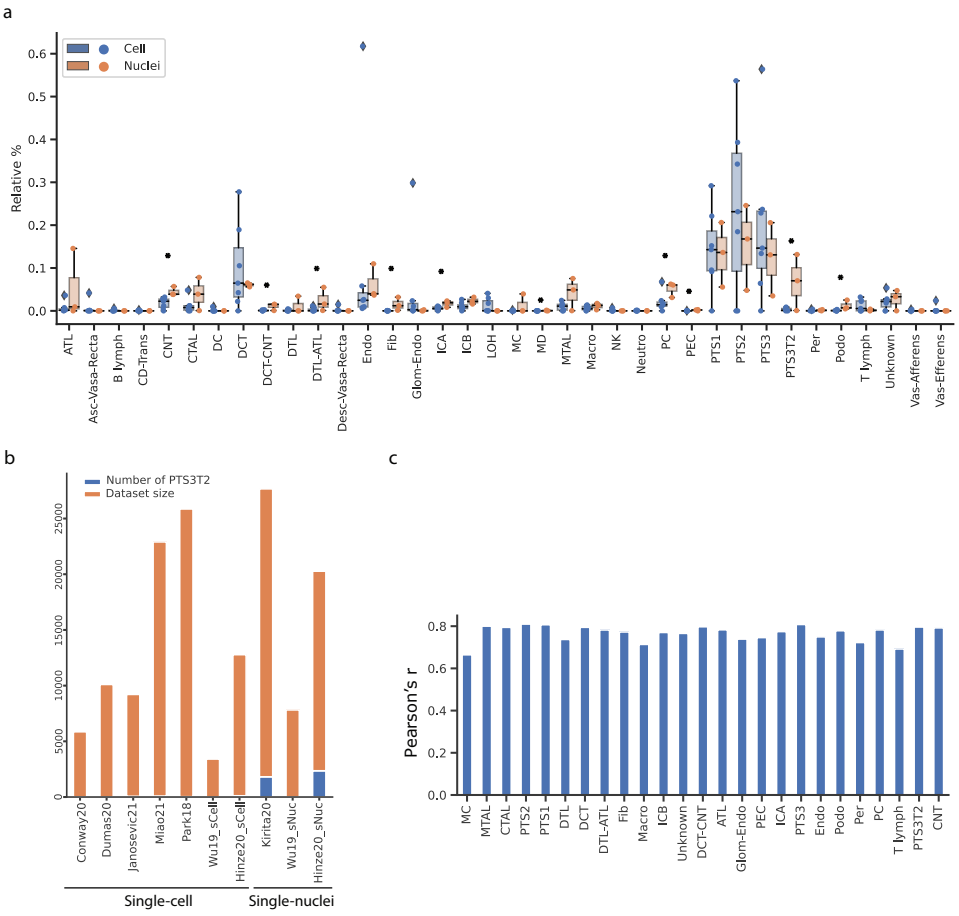

Figure 2.S7: *Single-cell and single-nuclei contribution to the MKA.* (**a**) Boxplot showing the relative contribution (compared to the total amount of cells or nuclei present in each dataset) of single-cells and single-nucleus to each of the cell types present in the MKA. Each dot represents a dataset in the MKA. Blue dots correspond to single-cell datasets whereas orange dots correspond to single-nuclei datasets. $*$: p-value 0.05, two-sided T-test. (**b**) Stacked barplot showing the total amount of PTS3T2 cells or nuclei detected (blue) per dataset, with their respective total size (orange). (**c**) Barplot showing Pearson's coefficient of correlation between the batch-corrected expression profile of single-cells and single-nucleus for a given cell type.

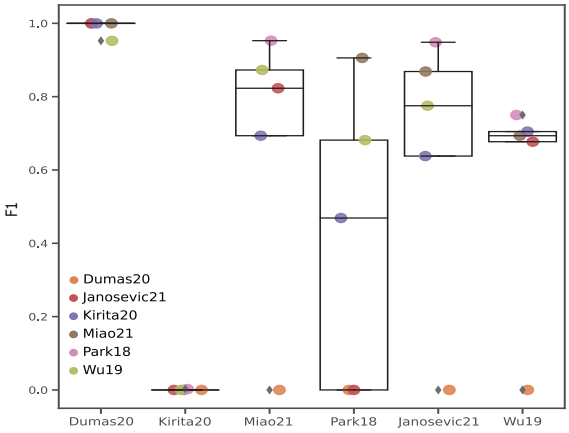

Figure 2.S8: *Accuracy of predictions using single-dataset references.* Boxplots of median F1 scores (**x-axis**) computed over five single-dataset references for all annotated datasets in MKA (**y-axis**). Colours indicate the single dataset used as reference in Azimuth's label transfer workflow.

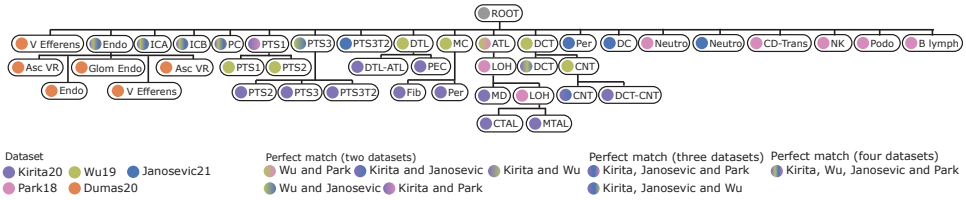

Figure 2.S9: *MKA*'s learned classification tree.* Tree built by training a k-Nearest Neighbor (kNN) classifier on five of the six annotated datasets (Miao21 dataset was excluded, MKA*). The colour(s) of the tree nodes represent the agreement with the supporting dataset(s).

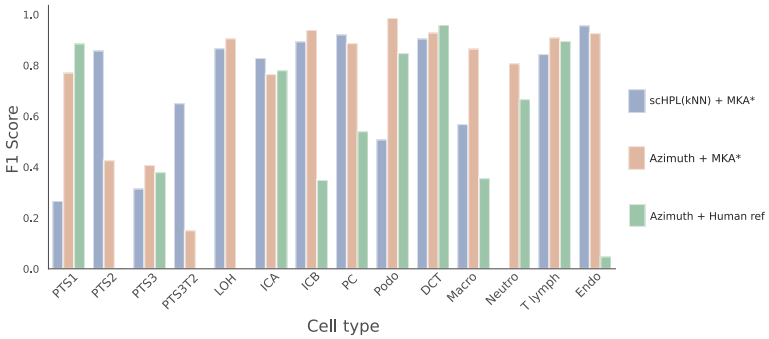

Figure 2.S10: *F1 scores per cell type and evaluation experiment.* For each cell type, the F1 score is plotted for the different Miao21 classification tasks. Namely, scHPL trained with our partial reference mouse atlas, Azimuth trained with a human reference, and Azimuth trained with our partial reference mouse atlas.

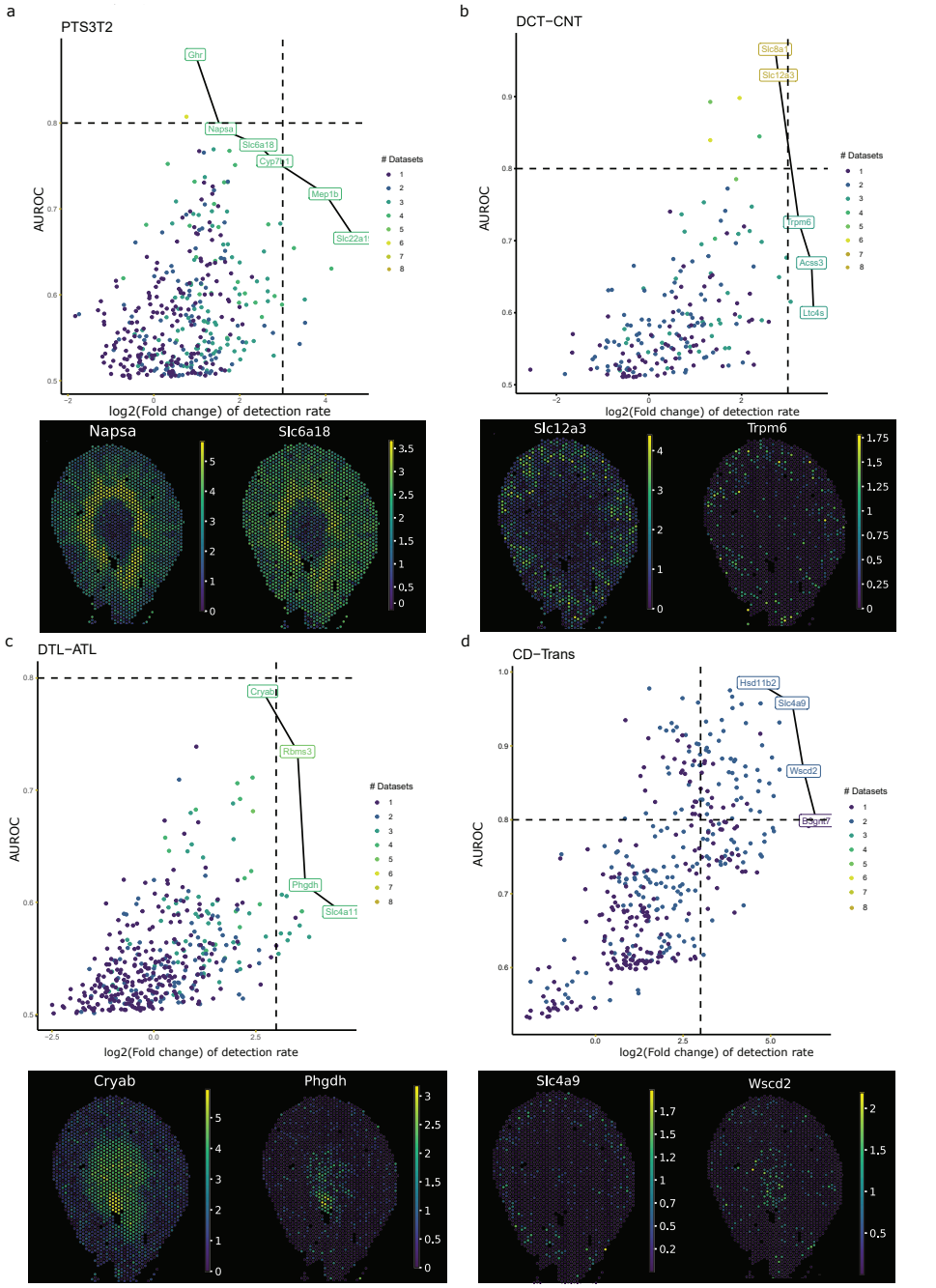

Figure 2.S11: *Meta-markers of transitional cell types and rare populations.* (**a-d**) Each dot represents a significant meta-marker (FDR < 0.05). Top markers with respect to detection rate (expressed as log2 Fold change) and precision (area under the receiver-operator curve; AUROC) for poorly described cell types are highlighted with a connected boundary line. Bottom panels show the log-normalized expression of two of these meta-markers in a healthy spatial transcriptomics kidney slide for every cell type. PTS3T2: Proximal Tubule Segment 3 Type 2, DCT: Distal Convoluted Tubule, ATL: Ascending Thin Limb of Henle, CNT: Connecting Tubule, DTL: Descending Thin Limb of Henle; CD-Trans: Collecting duct transitional cell population.

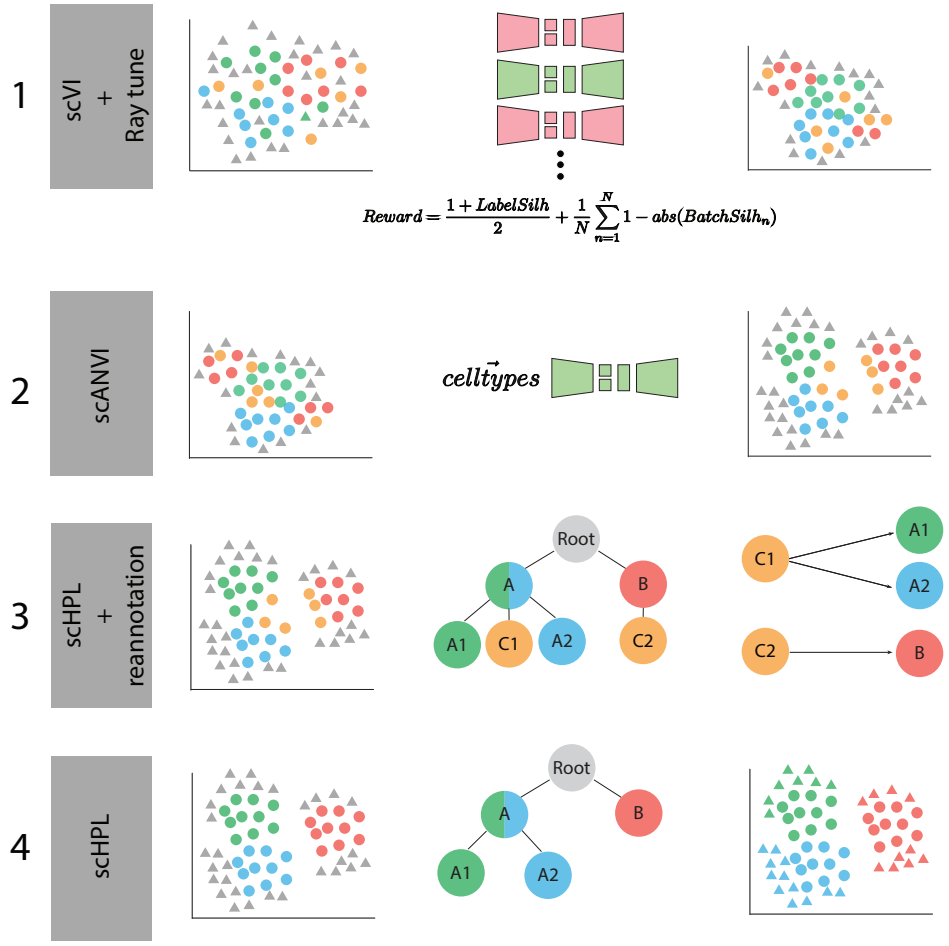

Figure 2.S12: *MKA pipeline.* Colours represent different hypothetical cell type annotations from two independent studies (A, B, A1, A2) and two mislabelled cell types (C1 and C2). Shapes depict originally annotated (circle) or unannotated (triangles) cells and/or nuclei. Funnel illustrations represent a Variational Autoencoder architecture. Red funnels have suboptimal hyperparameter combinations according to the reward function. The green funnel indicates the VAE architecture and hyperparameter combination that yielded the best batch mixing and cell type separation in the latent space. The vector cell types is a 1-dimensional array with cell type labels.