scientific **data**

OPEN

ANALYSIS

Meta-analysis of single-cell and single-nucleus transcriptomics reveals kidney cell type consensus signatures

Marceau Quatredeniers[1 ✉], **Alice S. Serafin**[1], **Alexandre Benmerah**[1], **Antonio Rausell**[2], **Sophie Saunier**[1] **& Amandine Viau**[1]
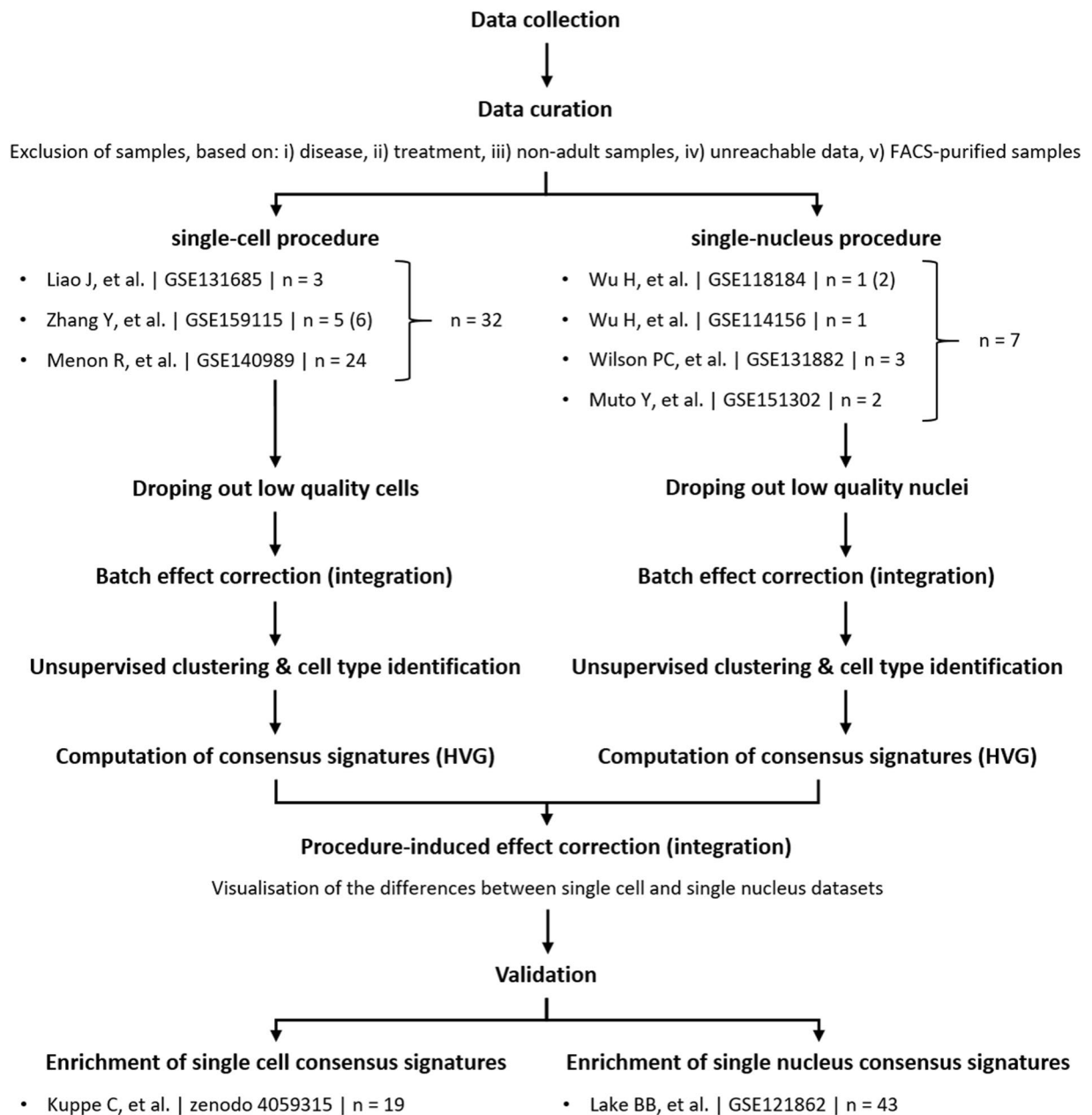
While the amount of studies involving single-cell or single-nucleus RNA-sequencing technologies grows exponentially within the biomedical research area, the kidney field requires reference transcriptomic signatures to allocate each cluster its matching cell type. The present meta-analysis of 39 previously published datasets, from 7 independent studies, involving healthy human adult kidney samples, offers a set of 24 distinct consensus kidney cell type signatures. The use of these signatures may help to assure the reliability of cell type identification in future studies involving single-cell and single-nucleus transcriptomics while improving the reproducibility in cell type allocation.

Introduction

The kidney is a vital and highly complex organ responsible for blood filtration, elimination of metabolites and waste products, and body homeostasis (oxygen sensing, pH balance, electrolyte levels, systemic blood pressure regulation). These critical functions are enabled through its functional subunit, the nephron, an intricate inter-play between the nephron sub-structures and some mesenchymal cells, involving the coordinated action of more than 20 cell types (immune cells, stromal cells, nephron epithelial cells, and cells from rare populations such as the juxtaglomerular apparatus) from the entry of the blood in glomerular capillaries to the urine compartment[1]. Nephron sub-structures are organized following the urine flow: blood arrives to the glomerulus and filtered through fenestrated capillaries and the slit-diaphragm (*n.b.* this active fenestration process is controlled by podocytes), then primary urine passes first through the Bowman capsule lined by parietal epithelial cells (PEC) and then through the tubule per se, including the proximal tubule (PT; reabsorption of water, sodium, calcium, potassium, chloride, phosphate, proteins and glucose), the loop of Henle (LoH; reabsorption of water, sodium and chloride; urine concentration), the distal tubule consisting in the distal convoluted tubule (DCT) and the connecting tubule (CNT; fine tuning of sodium and chloride levels, regulation of H + and HCO3- assuring pH balance), and finally flows into the collecting ducts in the cortex (CCD) and medulla (MCD; water reabsorption; urine concentration) until its storage in the bladder[2,3].

Although bulk transcriptomics have critically promoted the understanding of kidney development, physiology and diseases[4–6], such approach is not suitable for investigating renal cell type-specific features at a single cell scale. More recently, advances in high-throughput single-cell (scRNA-seq) and single-nucleus (snRNA-seq) transcriptomics allowed to evaluate cell populations and biological processes of different tissues at the single-cell/nucleus level[7–9]. Except for liquid samples (*e.g.* blood, urines…), scRNA-seq involves tissue dissociation, single-cell emulsion and encapsulation, passage in the microfluidics one cell at a time, creation of a library through high-throughput sequencing, and finally data analysis. As it may be performed on frozen tissue, snRNA-seq may overcome some issues observed with scRNA-seq, such as dissociation-induced stress response leading to the expression of specific set of genes, poor viability and loss of rare and fragile cell types[10,11]. Although the heterogeneity between individuals and cell states have been demonstrated[12,13], studies are often performed on a reduced number of samples due to the limited availability of human tissue and the cost of

[1]Université de Paris Cité, Imagine Institute, Laboratory of Hereditary Kidney Diseases, Paris, INSERM UMR 1163, F-75015, France. [2]Université de Paris Cité, Imagine Institute, Laboratory of Clinical Bioinformatics, Paris, INSERM UMR 1163, F-75015, France. ✉e-mail: quatredeniers.marceau@gmail.com

Data collection

↓

Data curation

Exclusion of samples, based on: i) disease, ii) treatment, iii) non-adult samples, iv) unreachable data, v) FACS-purified samples

single-cell procedure

- Liao J, et al. | GSE131685 | n = 3
- Zhang Y, et al. | GSE159115 | n = 5 (6)   } n = 32
- Menon R, et al. | GSE140989 | n = 24

↓

single-nucleus procedure

- Wu H, et al. | GSE118184 | n = 1 (2)
- Wu H, et al. | GSE114156 | n = 1
- Wilson PC, et al. | GSE131882 | n = 3   } n = 7
- Muto Y, et al. | GSE151302 | n = 2

↓

Droping out low quality cells → **Droping out low quality nuclei**

↓

Batch effect correction (integration) (both columns)

↓

Unsupervised clustering & cell type identification (both columns)

↓

Computation of consensus signatures (HVG) (both columns)

↓

Procedure-induced effect correction (integration)

Visualisation of the differences between single cell and single nucleus datasets

↓

Validation

Enrichment of single cell consensus signatures

- Kuppe C, et al. | zenodo 4059315 | n = 19

Enrichment of single nucleus consensus signatures

- Lake BB, et al. | GSE121862 | n = 43

Fig. 1 Meta-analysis workflow. All available scRNA-seq and snRNA-seq data were retrieved and downloaded from the Gene Expression Omnibus (GEO) or zenedo repository. Single cell RNA-seq and snRNA-seq samples were analysed separately, quality control metrics were measured and poor quality cells were filtered out in accordance. Then scRNA-seq samples and snRNA-seq samples were integrated independently. High resolution unsupervised clustering followed by visualisation of the expression of specific transcriptomic markers allowed to attribute each cluster a clear cell type (certain cell types were attributed to several clusters), or a cell type followed by « na » (*i.e.* not attributed) for the cells that did not show strong enough differenciation markers expression. Consensus signatures were computed using the FindAllMarkers() function in Seurat. Single cell RNA-seq and snRNA-seq were then integrated together to evaluate the matching between both datasets annotations. Finally, consensus signatures were used for cell type enrichments on previously published and annotated datasets.

scRNA-seq and snRNA-seq experiments. Other critical confounding factors in both experimental and analytical settings may affect scRNA-seq and snRNA-seq data, including low sequencing depth, context-dependent cell states, clustering settings, or markers checked for cell type identification.

Still, scRNA-seq and snRNA-seq are very powerful techniques increasingly used within the biomedical field in general, and in the kidney field in particular[14]. To date, scRNA-seq or snRNA-seq studies of human kidneys have involved different technologies and different data preprocessing and analysis workflows. Cell type labelling in particular suffers from the lack of a universal definition of known nephron segments and cell types, as well as

	Ref.	GEO accession number		Donor characteristics								Healthy tissue from…
		Serie (= batch)	Sample	Gender	Age	Smoker	Alcohol use	GFR	IFTA (%)	Serum creatinine (mg/dL)	BUN (mg/dL)	
snRNA-seq	Wilson PC, et al.[19]	GSE131882[26]	GSM3823939	M	54	NA	NA	58.4	1–10	1.28	NA	Non-tumor tissue in a patient undergoing nephrectomy for renal mass
			GSM3823940	M	62	NA	NA	60.7	1–10	1.21	NA	Non-tumor tissue in a patient undergoing nephrectomy for renal mass
			GSM3823941	F	61	NA	NA	68.5	1–10	0.89	NA	Non-tumor tissue in a patient undergoing nephrectomy for renal mass
	Wu H, et al.[20]	GSE118184[27]	GSM3320197-8	M	62	NA	NA	NA	NA	1.03	12	Renal cortex from a nephrectomy
	Wu H, et al.[21]	GSE114156[28]	GSM3135714	M	70	NA	NA	NA	NA	1.10	NA	Discarded renal donation
	Muto Y, et al.[22]	GSE151302[29]	GSM4572195	M	50	NA	NA	78	1–10	1.10	NA	Non-tumor tissue in a patient undergoing nephrectomy for renal mass
			GSM4572196	F	52	NA	NA	98	1–10	0.89	NA	Non-tumor tissue in a patient undergoing nephrectomy for renal mass
	Liao J, et al.[24]	GSE131685[31]	GSM4145204	M	57	Y	Y	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing nephrectomy for clear RCC
			GSM4145205	F	59	N	N	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing nephrectomy for clear RCC
			GSM4145206	M	65	Y	N	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing nephrectomy for clear RCC
	Zhang Y, et al.[23]	GSE159115[30]	GSM4819726	M	70	NA	NA	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing a partial nephrectomy for ccRCC
			GSM4819728	M	69	NA	NA	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing a partial nephrectomy for ccRCC
			GSM4819730-1	F	65	NA	NA	NA	NA	NA	NA	/
			GSM4819733	M	76	NA	NA	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing a partial nephrectomy for ccRCC
			GSM4819735	M	74	NA	NA	NA	NA	NA	NA	Non-tumor tissue in a patient undergoing a partial nephrectomy for ccRCC
scRNA-seq			GSM4191941	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191942	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191943	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191944	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191945	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191946	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191947	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191948	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191949	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191950	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191951	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
	Menon R, et al.[25]	GSE140989[32]	GSM4191952	NA	NA	NA	NA	NA	NA	NA	NA	Living donor
			GSM4191953	NA	NA	NA	NA	NA	NA	NA	NA	Living donor
			GSM4191954	NA	NA	NA	NA	NA	NA	NA	NA	Living donor
			GSM4191955	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191956	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191957	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191958	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191959	NA	NA	NA	NA	NA	NA	NA	NA	Non-tumor tissue from a tumor nephrectomy
			GSM4191960	NA	NA	NA	NA	NA	NA	NA	NA	Surveillance biopsy after kidney transplantation
			GSM4191961	NA	NA	NA	NA	NA	NA	NA	NA	Surveillance biopsy after kidney transplantation
			GSM4191962	NA	NA	NA	NA	NA	NA	NA	NA	Surveillance biopsy after kidney transplantation
			GSM4191963	NA	NA	NA	NA	NA	NA	NA	NA	Surveillance biopsy after kidney transplantation
			GSM4191964	NA	NA	NA	NA	NA	NA	NA	NA	Surveillance biopsy after kidney transplantation

Table 1. Clinical characteristics of the samples included in the meta-analysis. **Keys: GEO:** gene expression omnibus; **M:** male; **F:** female; **Y:** yes; **N:** no; **GFR:** glomerular filtration rate; **IFTA:** interstitial fibrosis and tubular atrophy; **BUN:** blood urea nitrogen; **NA:** not available.

		Batch ID	Sample ID	Number of cells	Mean number of features per cell	Mean number of counts per cell	% mitochondrial genes
snRNA-seq	Before filtering	GSE118184	GSM3320197-8	4524	1801.88	3933.58	0.27
		GSE131882	GSM3823939	6905	2328.92	6722.17	0.61
		GSE131882	GSM3823940	4236	1124.52	2090.37	0.62
		GSE131882	GSM3823941	6599	1671.81	3684.27	0.09
		GSE114156	GSM3135714	4297	1163.80	2028.99	0.92
		GSE151302	GSM4572195	4495	1559.68	3417.31	0.11
		GSE151302	GSM4572196	4708	1165.47	2194.50	0.11
	After filtering	GSE118184	GSM3320197-8	4226	1644.03	3395.02	0.28
		GSE131882	GSM3823939	5520	1835.33	4454.69	0.71
		GSE131882	GSM3823940	4179	1088.10	1951.29	0.56
		GSE131882	GSM3823941	6274	1539.50	3176.72	0.10
		GSE114156	GSM3135714	4234	1118.01	1895.09	0.92
		GSE151302	GSM4572195	4307	1447.19	2972.75	0.11
		GSE151302	GSM4572196	4672	1142.99	2109.04	0.11
scRNA-seq	Before filtering	GSE131685	GSM4145204	8098	959.47	2582.13	14.84
		GSE131685	GSM4145205	6449	1017.39	2690.40	14.12
		GSE131685	GSM4145206	10732	751.89	1843.62	4.08
		GSE159115	GSM4819726	839	2586.06	13765.11	35.30
		GSE159115	GSM4819728	777	2309.92	11523.01	38.88
		GSE159115	GSM4819730-1	1591	1433.38	4667.69	14.86
		GSE159115	GSM4819733	1538	884.67	2260.69	9.50
		GSE159115	GSM4819735	1854	2203.91	9366.49	20.67
		GSE140989	GSM4191941	1229	1005.80	3477.85	15.13
		GSE140989	GSM4191942	2456	860.88	2294.43	19.52
		GSE140989	GSM4191943	6525	693.38	1553.54	11.42
		GSE140989	GSM4191944	412	803.57	2467.24	7.43
		GSE140989	GSM4191945	2444	830.86	2315.52	8.91
		GSE140989	GSM4191946	6101	571.20	1749.37	32.41
		GSE140989	GSM4191947	1193	773.52	2367.96	13.12
		GSE140989	GSM4191948	4848	452.26	1065.57	21.27
		GSE140989	GSM4191949	607	576.69	1516.98	7.55
		GSE140989	GSM4191950	4666	820.94	2518.52	16.76
		GSE140989	GSM4191951	430	707.04	1983.04	8.48
		GSE140989	GSM4191952	5683	1023.92	4294.04	58.21
		GSE140989	GSM4191953	7671	946.94	3914.46	59.53
		GSE140989	GSM4191954	4344	861.71	3766.08	57.63
		GSE140989	GSM4191955	3519	618.15	1755.23	19.64
		GSE140989	GSM4191956	3055	593.53	1661.66	15.96
		GSE140989	GSM4191957	3107	597.92	1632.59	15.46
		GSE140989	GSM4191958	1221	489.10	1131.03	13.27
		GSE140989	GSM4191959	596	561.41	1344.56	9.12
		GSE140989	GSM4191960	762	1085.96	3913.73	26.40
		GSE140989	GSM4191961	1027	989.01	3418.46	31.97
		GSE140989	GSM4191962	1071	1163.96	3972.85	20.18
		GSE140989	GSM4191963	771	1483.22	4985.44	26.86
		GSE140989	GSM4191964	5815	835.83	2360.71	27.72
	After filtering	GSE131685	GSM4145204	7285	998.90	2662.66	11.86
		GSE131685	GSM4145205	5612	1067.32	2762.65	10.27
		GSE131685	GSM4145206	10605	754.70	1847.35	3.65
		GSE159115	GSM4819726	289	2260.57	8076.82	9.53
		GSE159115	GSM4819728	254	2280.94	7394.10	12.56
		GSE159115	GSM4819730-1	1275	1373.87	4012.67	8.59
		GSE159115	GSM4819733	1389	896.88	2203.29	5.42
		GSE159115	GSM4819735	1170	2017.44	6879.66	5.19
		GSE140989	GSM4191941	937	791.90	2247.86	7.79
		GSE140989	GSM4191942	1850	856.84	2149.69	12.73
Continued							

		Batch ID	Sample ID	Number of cells	Mean number of features per cell	Mean number of counts per cell	% mitochondrial genes
		GSE140989	GSM4191943	5915	645.66	1327.69	9.19
		GSE140989	GSM4191944	368	817.72	2356.84	5.10
		GSE140989	GSM4191945	2207	872.55	2332.26	4.51
		GSE140989	GSM4191946	2953	603.88	1504.32	18.17
		GSE140989	GSM4191947	992	711.93	1934.63	6.87
		GSE140989	GSM4191948	3420	463.18	1027.92	11.65
		GSE140989	GSM4191949	541	556.00	1351.68	5.16
		GSE140989	GSM4191950	3675	763.50	1960.04	8.58
		GSE140989	GSM4191951	391	666.83	1718.86	5.25
		GSE140989	GSM4191952	531	2049.43	6304.25	23.51
		GSE140989	GSM4191953	649	1994.28	6019.57	23.39
		GSE140989	GSM4191954	504	1662.21	5190.07	21.80
		GSE140989	GSM4191955	2756	635.38	1684.71	12.40
		GSE140989	GSM4191956	2521	589.11	1370.54	9.10
		GSE140989	GSM4191957	2582	579.20	1316.96	9.09
		GSE140989	GSM4191958	1087	494.10	985.52	9.36
		GSE140989	GSM4191959	534	572.36	1220.78	6.18
		GSE140989	GSM4191960	475	956.46	2781.59	14.44
		GSE140989	GSM4191961	523	1026.92	2979.92	18.13
		GSE140989	GSM4191962	782	903.26	2358.95	13.12
		GSE140989	GSM4191963	457	1318.73	3882.06	18.11
		GSE140989	GSM4191964	3499	827.35	2026.18	17.35

Table 2. QC metrics, before and after filtering of low quality cells/nuclei.

	Ref.	GEO accession number	Technical characteristics		
			Single cell/nucleus platform	Sequencer	Data pre-processing
sn	Wilson PC, *et al.*	GSE131882	10X Chromium	Illumina NovaSeq 6000	zUMIs v2.0
	Wu H, *et al.*	GSE118184	10X Chromium	Illumina HiSeq 2500	zUMIs v1
	Wu H, *et al.*	GSE114156	inDrop	Illumina HiSeq 2500, NextSeq	dropTag, dropEst
	Muto Y, *et al.*	GSE151302	10X Chromium	Illumina NovaSeq 6000	CellRanger v3.1.0
sc	Liao J, *et al.*	GSE131685	10X Chromium	Illumina Hiseq Xten	CellRanger v3.0
	Zhang Y, *et al.*	GSE159115	10X Chromium	Illumina HiSeq 2500	CellRanger v2.1.1
	Menon R, *et al.*	GSE140989	10X Chromium	Illumina HiSeq 4000	CellRanger

Table 3. Technical characteristics of the datasets included in the meta-analysis.

standard lists of RNA markers depicting each kidney cell types, which may therefore lower the comparison reliability between studies. Chen *et al.* already reported this issue and proposed a nomenclature for kidney epithelial cells to better compare studies[15]. However, no consensus list of human kidney cell type transcriptomic markers has been published so far. Thus, the establishment of consensus transcriptomic kidney cell type signatures might be of utmost importance considering significant batch effects within scRNA-seq and snRNA-seq datasets[16–18]. Here we present a meta-analysis of publicly available scRNA-seq and snRNA-seq datasets from 39 healthy adult kidneys, consisting in 68,028 single cells and 33,412 single nuclei. As data were taken from different sources, data were normalized following the SCTransform analysis pipeline in Seurat v4 and batch effects were mitigated by integration of scRNA-seq and snRNA-seq samples, respectively. Cell types were attributed to clusters using broad cell type markers, consensus cell type signatures were computed, and labelled scRNA-seq and snRNA-seq samples were integrated together to map cell types depending on the method used (sc/snRNA-seq). Finally, single-cell and single-nucleus consensus signatures were benchmarked by enrichment in previously published and annotated datasets.
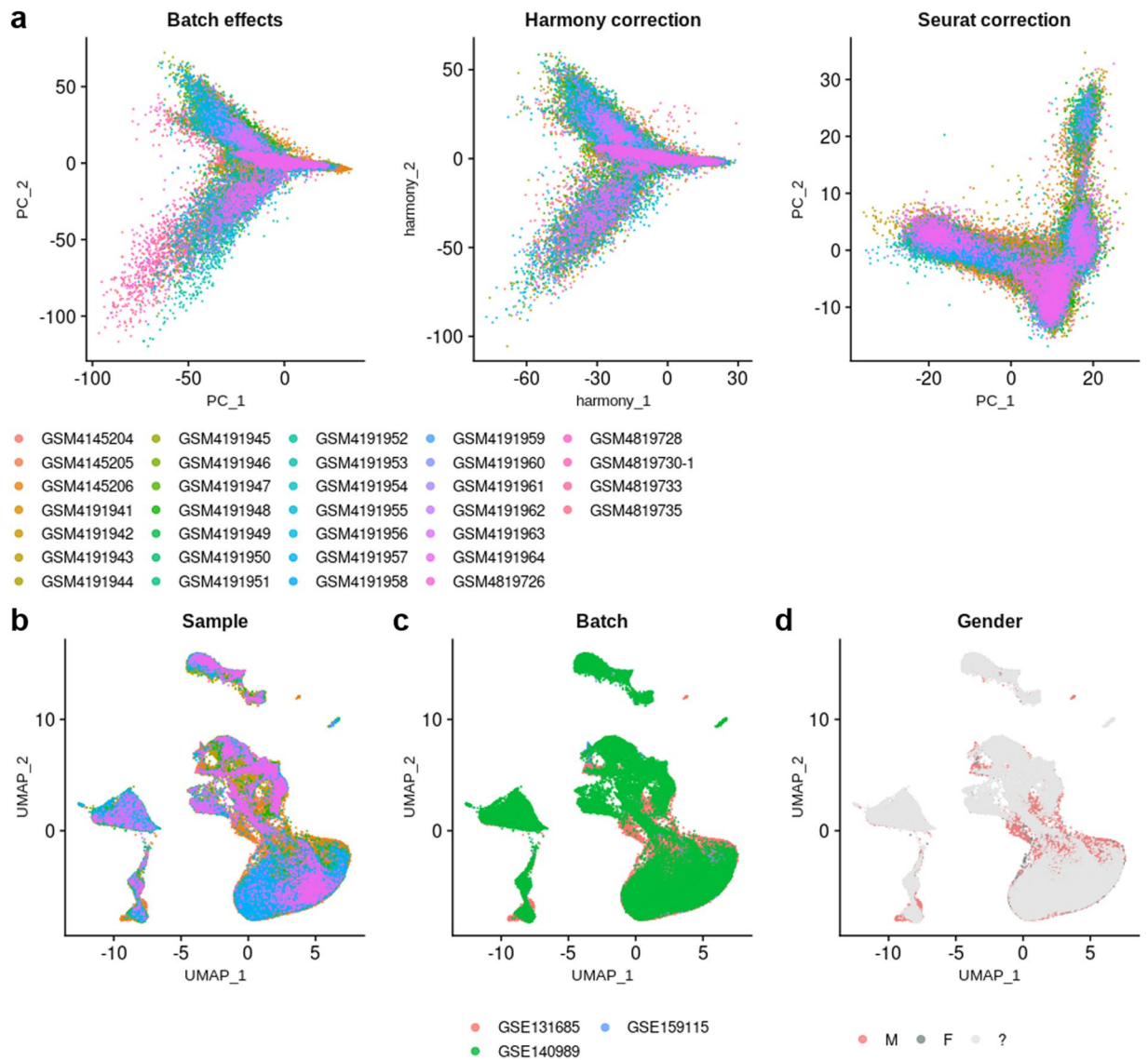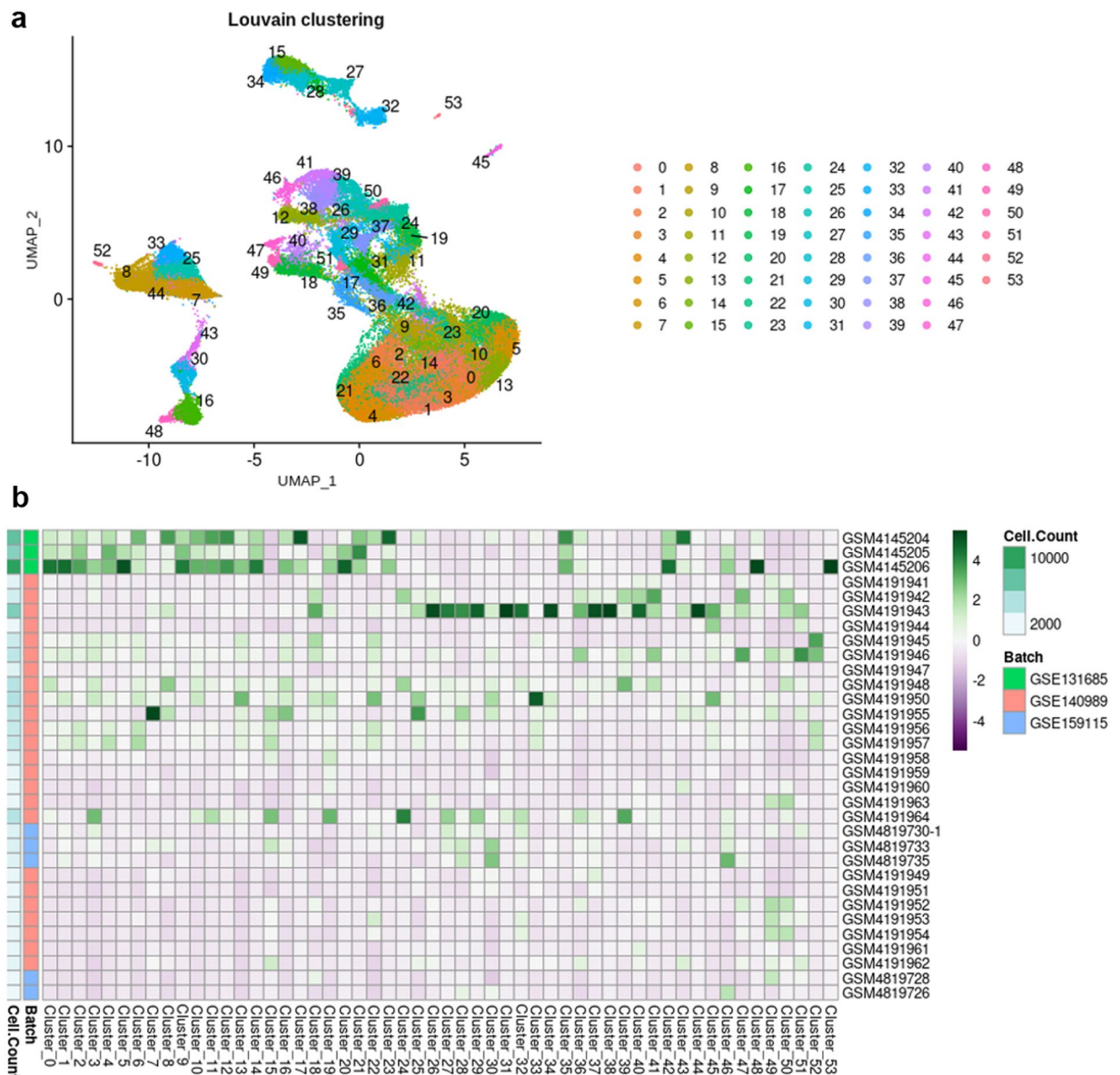
Fig. 2 Integration of scRNA-seq datasets. (**a**) PCA plots of scRNA-seq samples before any integration (left), after Harmony integration (middle), and after Seurat v4 integration (right). (**b**) UMAP plot showing the dispersion of cells after Seurat v4 integration, according to their sample of origin. (**c**) UMAP plot showing the dispersion of cells after Seurat v4 integration, according to their batch of origin (*i.e.* the publication). (**d**) UMAP plot showing the dispersion of cells after Seurat v4 integration, according to the gender; grey shade indicates that the gender is not known.

Results

Analysis workflow.   To determine consensus gene signature associated to each kidney cell type, we first aimed to collect kidney scRNA-seq and snRNA-seq data available on public databases according to the workflow presented in Fig. 1. Data collection ended up with 3 scRNA-seq and 4 snRNA-seq datasets publicly available[19–32], encompassing a total of 101,431 cells and 35,764 nuclei, from 32 and 7 healthy adult kidneys, respectively (Table 1). Samples from the different datasets were pre-processed with Seurat v4, and cells expressing between 200 and 3500 genes were kept for analysis (discarding cell debris and cell doublets). As some kidney cell populations highly express mitochondrial genes, the percentage of mitochondrial gene expression threshold to use in kidney tissue is debating (varying between 20% and 50% across studies)[33,34]. We chose to keep cells with less than 30% mitochondrial genes expressed[24] (Table 2). Despite nuclei should not express mitochondrial genes, nuclei with less than 5% mitochondrial genes expressed were kept to limit the waste due to possible little contamination. Since confounding variables may affect the different samples from the different studies and further analysis (Tables 1, 3), scRNA-seq and snRNA-seq samples were integrated separately using Seurat IntegrateData() function to mitigate the batch effects, following the newly implemented SCTransform framework for normalization and count data variance stabilization[35]. Finally, the integrated datasets consisted in 68,028 single cells and 33,412 single nuclei.

Fig. 3 Unsupervised clustering of scRNA-seq dataset. (**a**) UMAP plot of 32 integrated scRNA-seq samples showing the scattering of the cells and the distribution of the 54 clusters. (**b**) Heatmap displaying the number of cells per sample, and the number of cells from each sample in each cluster (scaled by cluster).

Generation of a healthy human kidney consensus scRNA-seq dataset. To generate a healthy human kidney consensus scRNA-seq dataset, we first assessed the quality of the integration by comparing the distribution of cells on Principal Component Analysis (PCA) plot, before and after integration using both Harmony and Seurat v4 correction (Fig. 2a). The correction of PC1 and PC2 by Harmony did not look as good as the one obtained with Seurat for which PC1 and PC2 did not depend anymore from the origins of the samples after integration. Thus the Seurat v4 correction was used for further computations. Uniform Manifold Approximation and Projection (UMAP) of the integrated dataset showed a very good scattering of cells from the different samples and from the different batches (Fig. 2b,c). In addition, it has been suggested that kidney cells express subsets of genes that are regulated in a sex-dependent manner in mice[36]. However, as the sex was not known for 24 samples among 32, we could not evaluate whether a sex bias may occur in cell type attribution in humans (Fig. 2d).

Unsupervised clustering (Louvain, resolution = 3.4) resulted in 54 distinct clusters (Fig. 3a). Despite a satisfying correction of the batch effects, certain clusters were driven by a few samples, which may reflect individual differences rather than cell types or cell states (Fig. 3a,b and Supp. Table 1). In particular, cluster 17 mostly belongs to sample GSM4145204 (50.13% of the cells), clusters 20, 48 and 53 to sample GSM4145206 (54.4%, 61.37 and 100%, respectively), and clusters 31, 38 and 44 to sample GSM4191943 (77.93%, 71.28% and 57.57%, respectively). Of note, these are the top 3 most abundant samples of the dataset (Table 2 and Supp. Table 1). These clusters were automatically labelled « not-attributed >> (na).
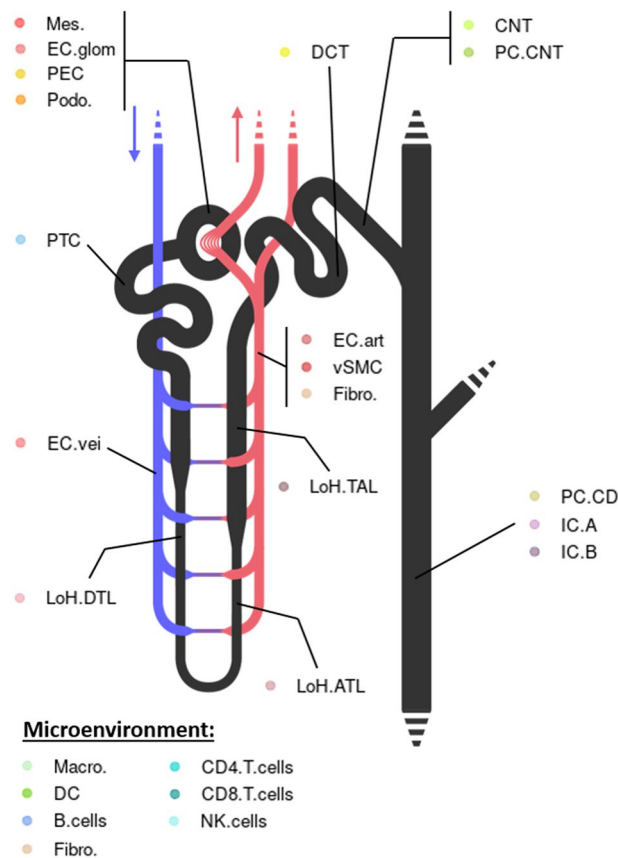
Fig. 4 Schematic representation of a nephron and its associated cell types. Scheme of a nephron, locating anatomical structures the cell types described in the study belong to. See Table 4 for more details about the nomenclature. Keys: Macro.: Macrophages; DC: Dendritic cells; B.cells: B cells; CD4.T.cells: CD4+ T cells; CD8.T.cells: CD8+ T cells; NK.cells: Natural killer cells; EC.vei: Veinous endothelial cells; EC.glom: Glomerular endothelial cells; EC.art: Arterial endothelial cells; vSMC: Vascular smooth muscle cells; Mes.: Mesangial cells; Fibro.: Fibroblasts; PEC: Parietal epithelial cells; Podo.: Podocytes; PTC: Proximal tubule cells; LoH.DTL: Descending thin limb of the loop of Henle cells; LoH.ATL: Ascending thin limb of the loop of Henle cells; LoH. TAL: Thick ascending limb of the loop of Henle cells; DCT: Distal convoluted tubule cells; CNT: Connecting tubule cells; PC.CNT: Principal cells, connecting tubule; PC.CD: Principal cells, collecting duct; IC.A: Intercalated cells, A-type; IC.B: Intercalated cells, B-type.

Broad cell type markers expression was studied to allocate cell types to clusters[6,9,22,37–54] (Table 4, Figs. 4, 5a,b). A total of 29 cell types were retrieved, including nephron epithelial cells, kidney mesenchymal cells, and immune cells from both myeloid and lymphoid lineages. The nomenclature from Chen *et al.*[15] was used for nephron epithelial cell labelling, with minor modifications to match our findings (*e.g.* we were not able to allocate several sub-populations of the descending thin limb nor of the thick ascending limb of the loop of Henle, and we did not find macula densa cells; please refer to Table 4 and Fig. 4 for a description of the adapted nomenclature). Considering the potential differences between sample origins (peritumoral, healthy donor, surveillance biopsy; Table 1), sampling methods, data pre-processing (Table 3) and distribution among clusters, some clusters were labeled « not-attributed » for ambiguous populations of proximal tubule cells (PTC.na), loop of Henle cells (LoH.na), principal cells (PC.na) and endothelial cells (EC.na) (Fig. 5b,c; populations in lightgrey). Thus, the computation of the signatures of PTC, LoH, PC, and EC did not depend on these ambiguous cells. Highly variable genes (HVG) were computed for every cell type: the corresponding gene lists define the consensus transcriptomic cell type signatures of kidney cells from healthy adult individuals (Fig. 5c and Supp. Table 2).

Generation of a healthy human kidney consensus snRNA-seq dataset. The evaluation of the mitigation of the batch effects for snRNA-seq integrated dataset was not as good as the one obtained for scRNA-seq dataset, but PC1 and PC2 distribution was more satisfying after integration using Seurat v4 compared to Harmony (Fig. 6a). Hence Seurat correction was adopted to pursue the analysis. When nuclei are displayed according to the origin of the sample or the origin of the batch the sample comes from (*i.e.* the publication), it is clear that sample GSM3135714 from batch GSE114156 is not well integrated to the dataset (Fig. 6b,c). As there are only 7 samples, and some of the nuclei from this sample do not mix with the rest of the nuclei from the other samples, we chose to keep the nuclei from this sample in the analysis and exclude only the non-mixed ones after clustering. By contrast to scRNA-seq dataset, the gender was known for the 7 snRNA-seq samples and allowed to appreciate differences
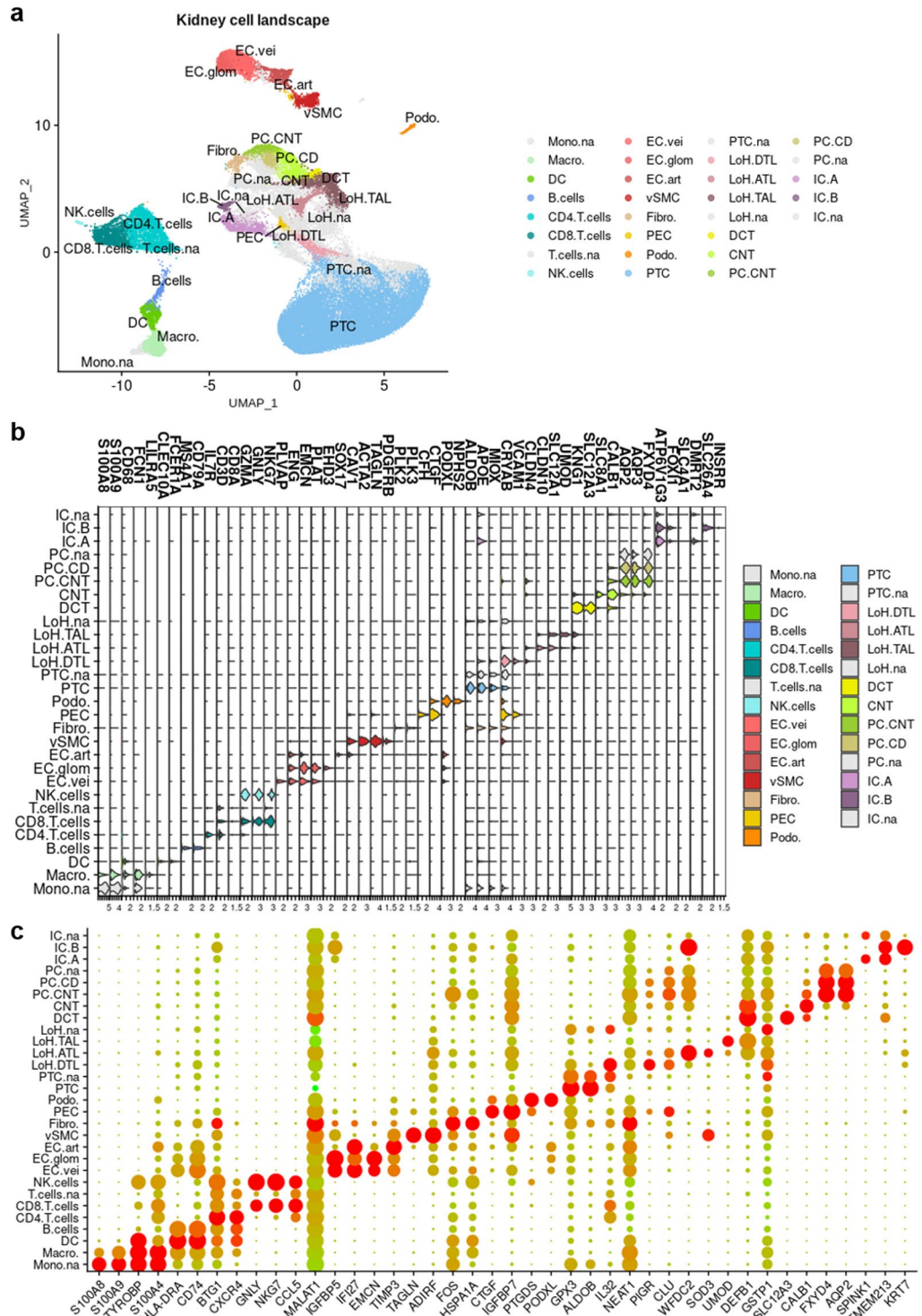
Fig. 5 Healthy human kidney landscape at the single cell level. (**a**) Cell type attribution to clusters based on the expression of specific transcriptomic markers. (**b**) ViolinPlot showing the expression of some of the kidney cell type-specific transcriptomic markers used to allocate cell types to clusters. (**c**) Dot plot presenting the expression of the 3 first genes of each computed signature, for all the signatures; this plot illustrates the list of consensus signatures for scRNA-seq samples.

in sex representation within each identified population, in particular for the principal cells of the collecting duct (PC.CD) and the cells from both the ascending thin limb (LoH.ATL) and the thick ascending limb of the loop of Henle (LoH.TAL; Fig. 6d). Unfortunately, with only 5 men and 2 women, we could not assess whether these differences were due to a real gender bias rather than inter-individual differences or some remaining batch effects.
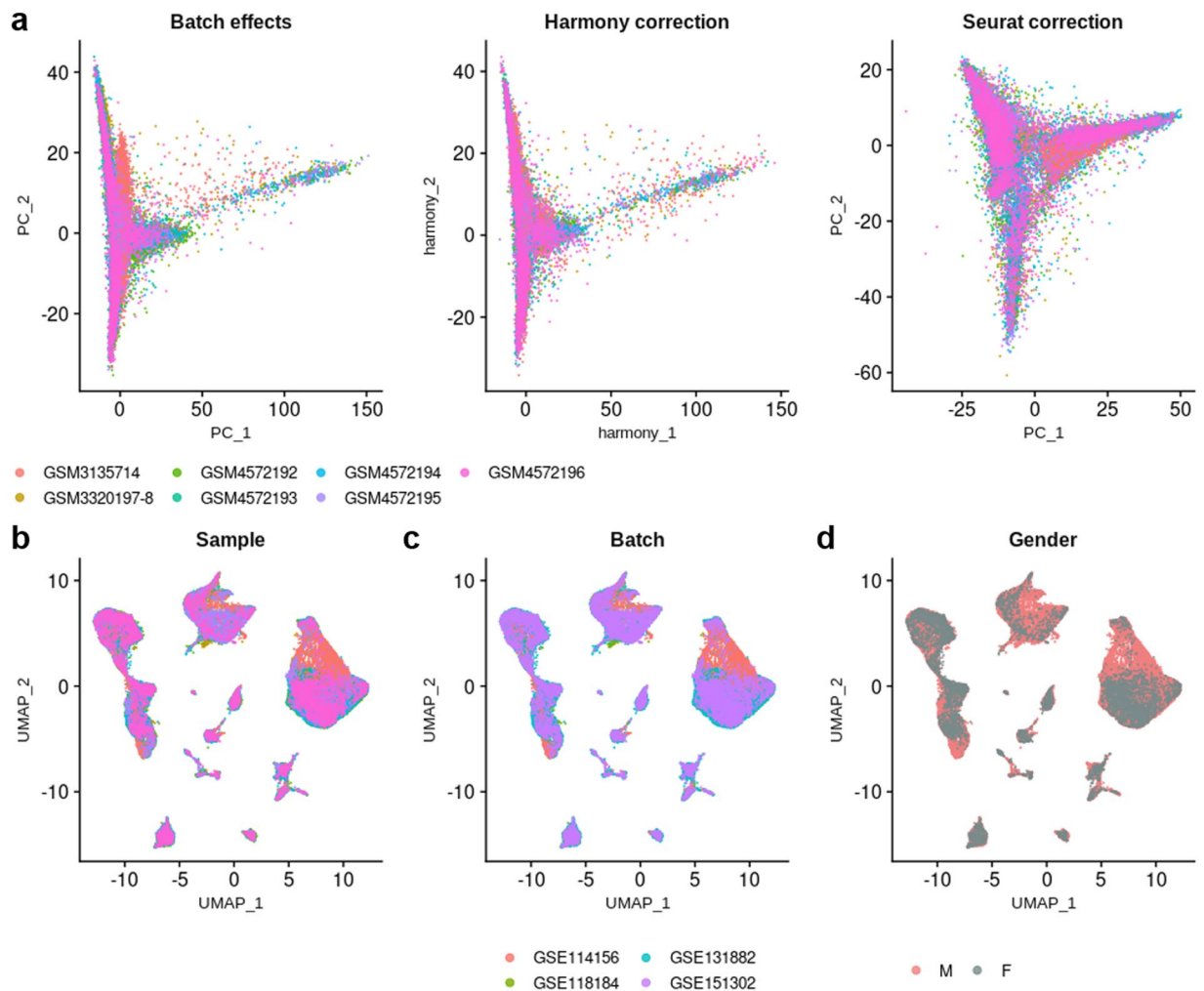
Fig. 6 Integration of snRNA-seq datasets. (**a**) PCA plots of snRNA-seq samples before any integration (left), after Harmony integration (middle), and after Seurat v4 integration (right). (**b**) UMAP plot showing the dispersion of nuclei after Seurat v4 integration, according to their sample of origin. (**c**) UMAP plot showing the dispersion of nuclei after Seurat v4 integration, according to their batch of origin (*i.e.* the publication). (**d**) UMAP plot showing the dispersion of nuclei after Seurat v4 integration, according to the gender; grey shade indicates that the gender is not known.

Besides this potential sex bias was different from what was shown in mice, where the authors observed discrepancies in the PTC populations while comparing 2 males to 2 females[36].

Unsupervised clustering (Louvain, resolution = 3.0) resulted in 53 distinct clusters (Fig. 7a). As expected, several clusters (*i.e.* clusters 7, 23, 30, 33 and 34) consisted mainly in nuclei from sample GSM3135714 (accounting for 73.4%, 86.5%, 71.6%, 64% and 86.9%, respectively) (Fig. 6b,c, Fig. 7a,b and Supp. Table 1). We also observed that clusters 3 and 17 mainly belonged to sample GSM4572195 (58.6% and 63%, respectively), cluster 39 to sample GSM3320197-8 (52.5%) and cluster 48 to sample GSM4572192 (51.8%) (Fig. 7b and Supp. Table 1). Again, classical markers were studied to allocate cell types to clusters[6,9,22,37–54] (Fig. 8a,b, Fig. 4 and Table 4). Sticking as much as possible to the same nomenclature used for scRNA-seq dataset, a total of 22 cell types were retrieved among nuclei, including nephron epithelial cells, kidney mesenchymal cells, and 4 populations of PTC, LoH, PC and T cells labeled « not attributed » (Fig. 8a,b).

The computation of HVG for every cell type has been performed, and these gene lists correspond to the consensus transcriptomic cell type signatures of kidney nuclei from healthy adult individuals (Fig. 8c and Supp. Table 2).

Joint analysis of scRNA-seq and snRNA-seq labeled datasets. To study the similarities and discrepancies between the results obtained with the two procedures, scRNA-seq and snRNA-seq samples were integrated together. The mitigation of the batch effects for the integration of 39 samples was overall acceptable, as attested by the correction of PC1 and PC2 (Fig. 9). Again, Harmony correction was not as satisfying as Seurat v4 one. Samples looked well merged, but nuclei and cells did not colocalize everywhere (Fig. 10a,b). The allocated cell types were highly consistent between cells and nuclei (Fig. 10c). Of note, we cannot rule out whether selecting
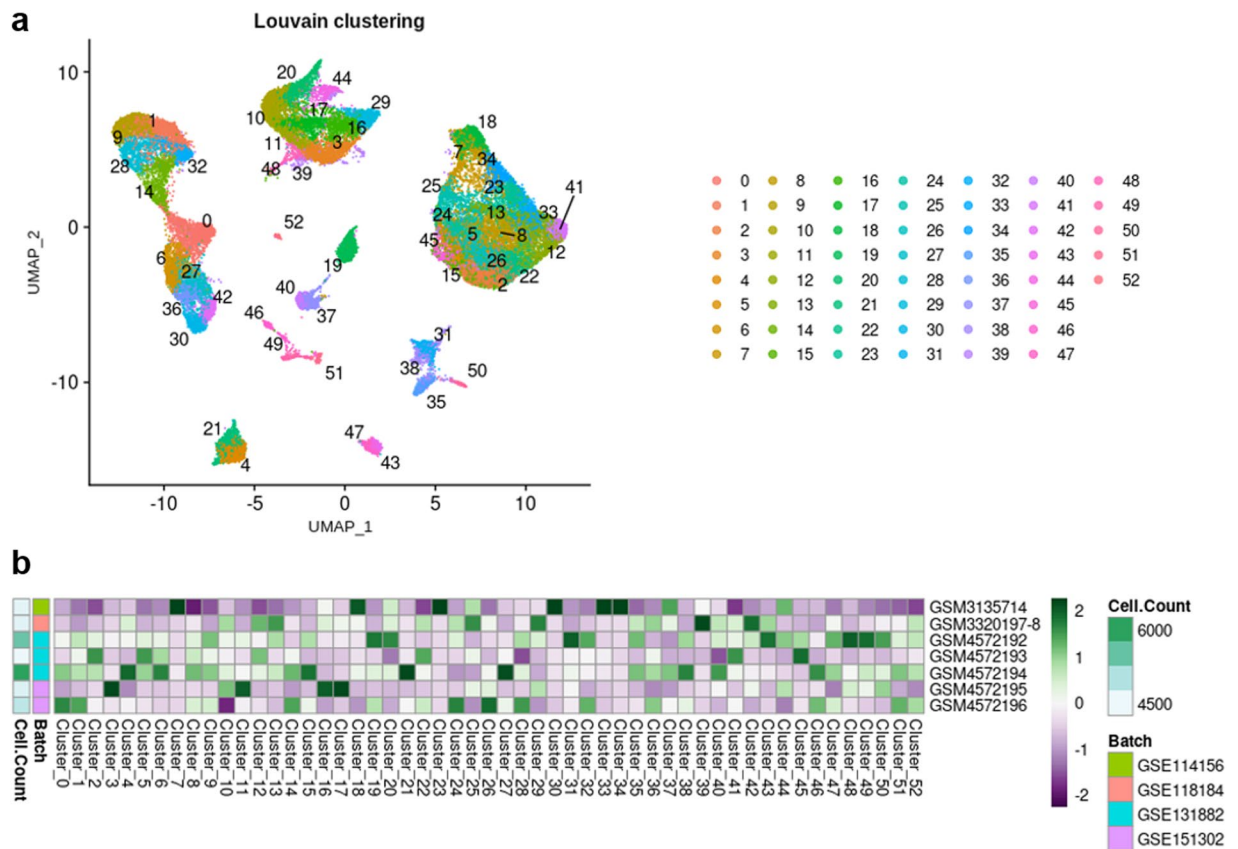
Fig. 7 Unsupervised clustering of snRNA-seq dataset. (**a**) UMAP plot of 7 integrated snRNA-seq samples showing the scattering of the nuclei and the distribution of the 53 clusters. (**b**) Heatmap displaying the number of nuclei per sample, and the number of nuclei from each sample in each cluster (scaled by cluster).

viable cells on the basis of mitochondrial genes expression may influence this observation, since we cannot filter nuclei on the same basis. However the fact that overall, cell types were allocated at the same coordinates in cells and nuclei may give further confidence in the identified cell populations in both scRNA-seq and snRNA-seq datasets. In light of these results, we would not recommand to integrate scRNA-seq and snRNA-seq datasets before cell types have been allocated to cells and nuclei. Overall, these results demonstrated that snRNA-seq and scRNA-seq consensus signatures should be used to enrich for cell types within snRNA-seq and scRNA-seq datasets, respectively.

Validation of the enrichment of consensus signatures for automatic cell type annotation. To test wether enrichment of consensus signatures are suitable for the identification of cell types within scRNA-seq and snRNA-seq datasets, we downloaded publicly available annotated datasets from Kuppe *et al.*[55,56] (scRNA-seq) and Lake *et al.*[9,57] (snRNA-seq). Sample expression matrices were processed and integrated as previously. Then CelliD v1.0[58] was used to perform enrichment analysis for scRNA-seq or snRNA-seq consensus signatures in every single cell or nucleus respectively, and the results were compared to the original labelling of the authors. To better evaluate automatic cell type annotation on test datasets, original labels were adapted to match consensus signatures nomenclature (Table 5).

After filtering out poor quality cells and cell doublets (less than 200 or more than 3500 expressed genes with more than 30% of mitochondrial genes), scRNA-seq dataset from Kuppe *et al.*[55,56] consisted in 81,239 cells from 19 samples, representing a total of 13 chronic kidney disease patients (hypertensive nephroscle-rosis)[55] (Fig. 11 and Fig. 12a). Enrichment of consensus scRNA-seq signatures was performed following Multiple Correspondence Analysis (MCA), and UMAP was computed on the residues of the MCA using the RunMCUMAP() function implemented in CelliD[58]. However, to avoid annotation of cells with the « na » label that is not informative, signatures for « na » annotated cell types were not tested. Enrichment retrieved cell labels closely related to the original labels (Fig. 12b,c). Some differences were observed, in particular the non-attributed endothelial cells were recognized as B cells, a population of macrophages was recognized as dendritic cells, and the cells of the thick ascending limb of the Loop of Henle labeled as distal tubule cells. As only cell types belonging to the list of consensus signatures may be attributed, we did not find any schwann cell, urothelial cell, monocyte or mast cell (dendritic cells instead), myofibroblast or pericyte (vascular smooth muscle cells instead), macula densa cell (thin ascending limb of the loop of Henle instead) (Fig. 12c). Overall, automatic cell type annotation using scRNA-seq consensus signatures pretty matched the original labels from Kuppe *et al.*[55], demonstrating its suitability and reliability to help in cell type allocation (Fig. 12b,c).
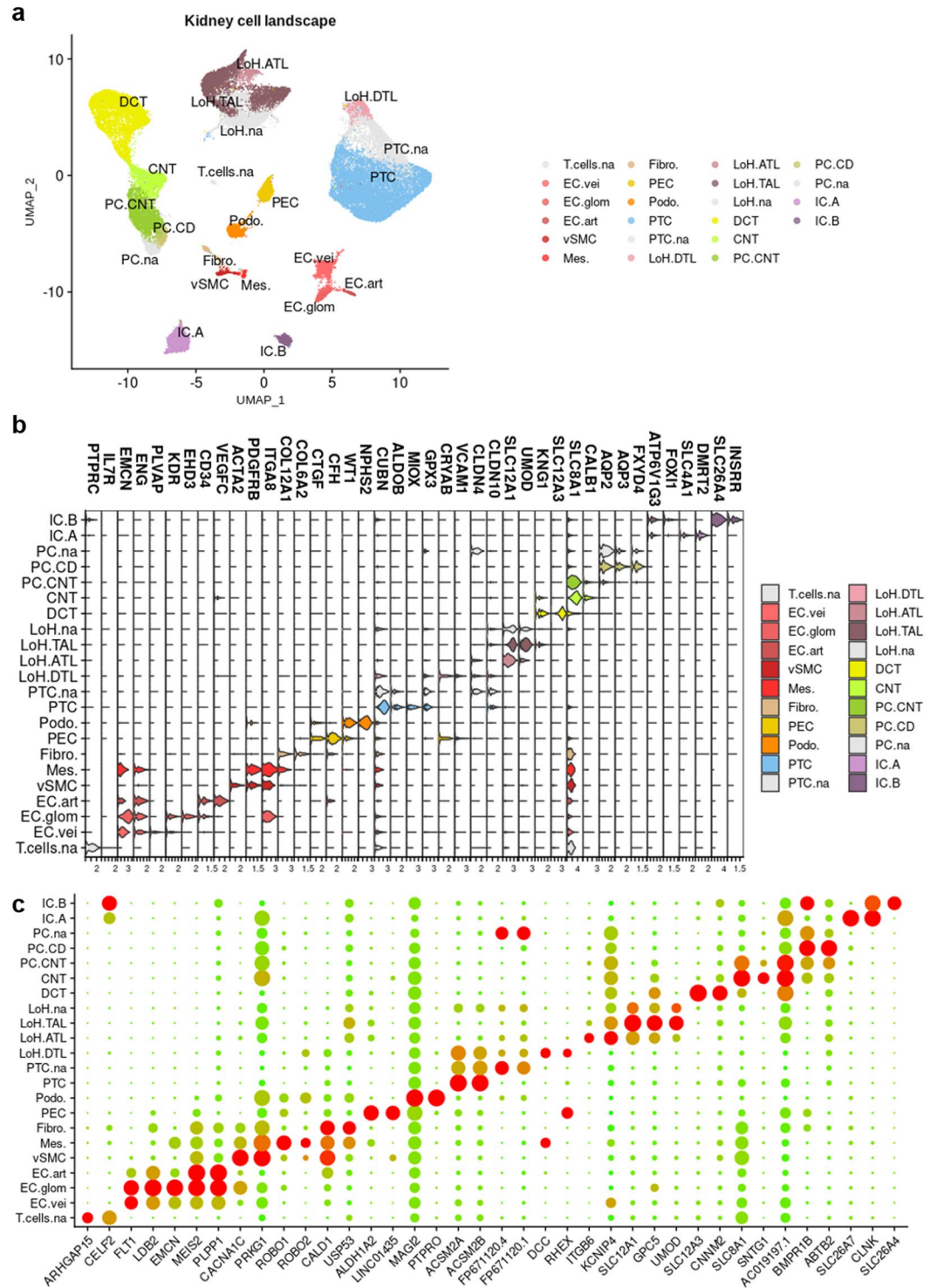
Fig. 8 Healthy human kidney landscape at the single nucleus level. (**a**) Cell type attribution to clusters based on the expression of specific transcriptomic markers. (**b**) ViolinPlot showing the expression of some of the kidney cell type-specific transcriptomic markers used to allocate cell types to clusters. (**c**) Dot plot presenting the expression of the 3 first genes of each computed signature, for all the signatures; this plot illustrates the list of consensus signatures for snRNA-seq samples.

Concerning the validation of our identified snRNA-seq signature, Lake *et al.*[9,57] snRNA-seq dataset was processed as previously described to remove poor quality nuclei, and finally consisted in 17,375 nuclei from 43 samples that belonged to 16 individuals, including 14 tumor-free regions of nephrectomies and 2 deceased donor kidneys[9] (Fig. 13). The nomenclature of the original labels was modified as previously described for scRNA-seq,

A. scRNA-seq markers								
Cell type			**Markers**					
Immune	Macro.	Macrophage	CD68[37,45,46]	S100A8[46]	S100A9[46]	FCN1[52]	LILRA5[46]	
	DC	Dendritic cells	CD68[46]	FCER1A[45]	CLEC10A[46]			
	B.cells	B cells	CD79A[37,45,46]	MS4A1[45,46]				
	CD4.T.cells	CD4 T cells	CD3D[45,46]	IL7R[45,46]				
	CD8.T.cells	CD8 T cells	CD3D[45,46]	NKG7[46]	GZMA[37,47]	GNLY[48]		
	NK.cells	Natural killer cells	NKG7[46]	GZMA[47]	GNLY[48]			
Vascular	EC.vei	Endothelial cells, veinous	EMCN[9,22,37]	ENG[22,37,38]	PLAT[22]	PLVAP[22,37]		
	EC.glom	Endothelial cells, glomerular	EMCN[9,22,37]	ENG[22,37,38]	PLAT[22]	EHD3[38]		
	EC.art	Endothelial cells, arterial	EMCN[9,22,37]	ENG[22,37]	CAV1[38]			
	vSMC	Vascular smooth muscle cells	ACTA2[9,41]	TAGLN[37]	CAV1[42]	PDGFRB[39]		
	Fibro.	Fibroblasts	PLK2[50]	PLK3[51]				
Nephron epithelail cells	Podo.	Podocytes	NPHS2[6,9,22,37]	PODXL[6,22,37]	CTGF[22]	CTGF[22]		
	PEC	Parietal epithelial cells	CRYAB[22]	CFH[22]	CTGF[22]	VCAM1[22]		
	PTC	Proximal tubule cells	CRYAB[22]	MIOX[22]	ALDOB[22]	APOE[22]		
	LoH.DTL	Descending thin limb of the loop of Henle cells	CRYAB[6,9]	VCAM1[9]	CLDN4[49]			
	LoH.ATL	Ascending thin limb of the loop of Henle cells	CLDN10[9,49]	SLC12A1[22]	CLDN4[49]			
	LoH.TAL	Thick ascending limb of the loop of Henle cells	SLC12A1[6,9,22,37]	UMOD[6,9,22,37]	KNG1[6,9,22]	CLDN10[54]		
	DCT	Distal contourned tubule cells	SLC12A3[6,9,37]	CALB1[9,22]	KNG1[22]			
	CNT	Connecting tubule cells	CALB1[6,9,22,37]	SLC8A1[6,9,22]	KNG1[22]			
	PC.CNT	Principal cells, connecting tubule	CALB1[22]	AQP2[22]	AQP3[6,22]	FXYD4[6]		
	PC.CD	Principal cells, collecting duct	AQP2[6,9,22,37]	AQP3[6,9,22,37]	FXYD4[6,9,22,37]			
	IC.A	Intercalated cells, A-type	SLC4A1[6,9,22,37]	FOXI1[6,9,22]	DMRT2[9,22]	ATP6V1G3[22]	APOE[53]	
	IC.B	Intercalated cells, B-type	SLC26A4[9,22,37]	INSRR[22,37]	ATP6V1G3[22]	FOXI1[22]		
B. snRNA-seq markers								
Vascular	EC.vei	Endothelial cells, veinous	EMCN[9,22,37]	ENG[22,37,38]	PLVAP[22,37]			
	EC.glom	Endothelial cells, glomerular	EMCN[9,22,37]	ENG[22,37,38]	KDR[22,37,38]	EHD3[38]	CD34[37]	ITGA8[22]
	EC.art	Endothelial cells, arterial	EMCN[9,22,37]	ENG[22,37,38]	CD34[22,38]	VEGFC[44]		
	vSMC	Vascular smooth muscle cells	ACTA2[9,41]	PDGFRB[39]	ITGA8[43]			
	Mes.	Mesangial cells	PDGFRB[9,39]	ITGA8[9,22,40]	EMCN[22]	ENG[22]	COL12A1[22]	
	Fibro.	Fibroblasts	COL12A1[22]	COL6A2[22]				
Nephron epithelail cells	Podo.	Podocytes	NPHS2[6,9,22,37]	WT1[6,22,37]				
	PEC	Parietal epithelial cells	CTGF[22]	CFH[22]	WT1[22]	CRYAB[22]		
	PTC	Proximal tubule cells	MIOX[6,22,99]	GPX3[6,9]	CUBN[6,9]	ALDOB[22]		
	LoH.DTL	Descending thin limb of the loop of Henle cells	CRYAB[6,9]	VCAM1[9]	CLDN4[49]	CUBN[54]		
	LoH.ATL	Ascending thin limb of the loop of Henle cells	SLC12A1[22]	CLDN4[49]	UMOD[54]			
	LoH.TAL	Thick ascending limb of the loop of Henle cells	SLC12A1[6,9,22,37]	UMOD[6,9,22,37]	KNG1[6,9,22]			
	DCT	Distal contourned tubule cells	SLC12A3[6,9,37]	KNG1[22]				
Continued								

A. scRNA-seq markers							
Cell type			Markers				
	CNT	Connecting tubule cells	CALB1[6,9,22,37]	SLC8A1[6,9,22]	KNG1		
	PC.CNT	Principal cells, connecting tubule	AQP2[22]	SLC8A1[22]	CALB1[22]		
	PC.CD	Principal cells, collecting duct	AQP2[6,9,22,37]	AQP3[6,9,22,37]	FXYD4[6,9,22,37]		
	IC.A	Intercalated cells, A-type	SLC4A1[6,9,22,37]	FOXI1[6,9,22]	DMRT2[9,22]	ATP6V1G3[22]	
	IC.B	Intercalated cells, B-type	SLC26A4[9,22,37]	INSRR[22,37]	FOXI1[22]	ATP6V1G3[22]	

Table 4. List of broad cell type marker genes used to allocate cell types to clusters.
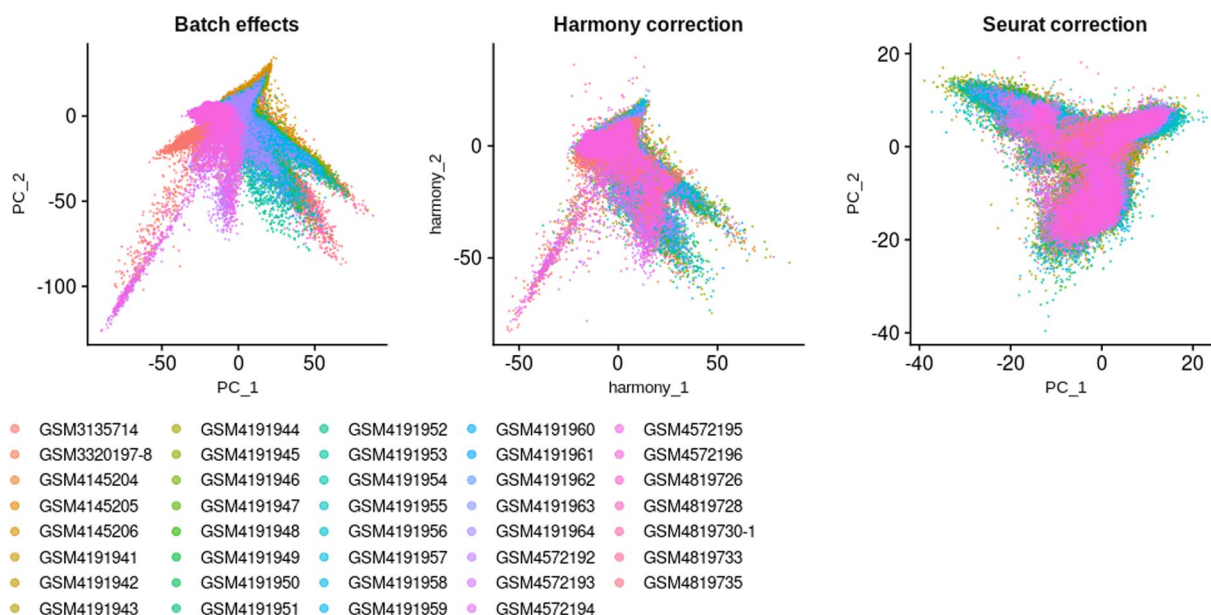
Fig. 9 Integration of scRNA-seq and snRNA-seq datasets. PCA plots of scRNA-seq and snRNA-seq samples before any integration (left), after Harmony integration (middle), and after Seurat v4 integration (right).

to match the nomenclature of the consensus cell type signatures (Fig. 14a, Table 5). Again, enrichment of snRNA-seq consensus signatures was done after computation of MCA and UMAP and signatures for « na » annotated cell types were not included for enrichment. The annotations were overall conserved between original labelling and consensus signature-based labelling (Fig. 14b,c). However, a subpopulation of proximal tubule cells was enriched for the descending thin limb of the loop of Henle (LoH.DTL) and some parietal epithelial cells (PEC) in the automatic annotation (Fig. 14b). In addition, cells originally labelled as LoH.DTL and some cells labelled as ascending thin limb of the loop of Henle (LoH.ATL) from samples NK37, NK38, NK45 and NK46, were still unassigned after consensus signatures enrichment (which means, there is no cell type enriched with a FDR < 0.01). This important unassigned population, which belonged to 4 samples among 43, may be considered « non-conventional » cells (although it may be due to remaining batch effects, as the samples were collected and conserved differently). In an original study, such nuclei would benefit from an in-depth analysis, since they could belong to non-tested cell types or non-steady cell states.

In conclusion, enrichment of consensus signatures allows the automatic and reliable annotation of kidney cell types in scRNA-seq and snRNA-seq datasets, and may be used to help in the decision of cell type allocation.

Discussion

Single cell RNA-seq and snRNA-seq are exponentially used within the kidney field. While major kidney cell types are generally retrieved, cell type identification is unconsistent between studies and therefore lacks reproducibility. It seems legit that every batch of samples, or every single sample, would not allow to identify the exact same transcriptomic signatures. It is mainly due to the small sample size of most human single-cell studies, subject to the cost of the technology and the limited availability of healthy human samples, that is in general not sufficient to generalize the conclusions to the overall population. In addition, scRNA-seq and snRNA-seq samples still suffer from a very low sequencing depth that may give rise to false positives or false negatives, within
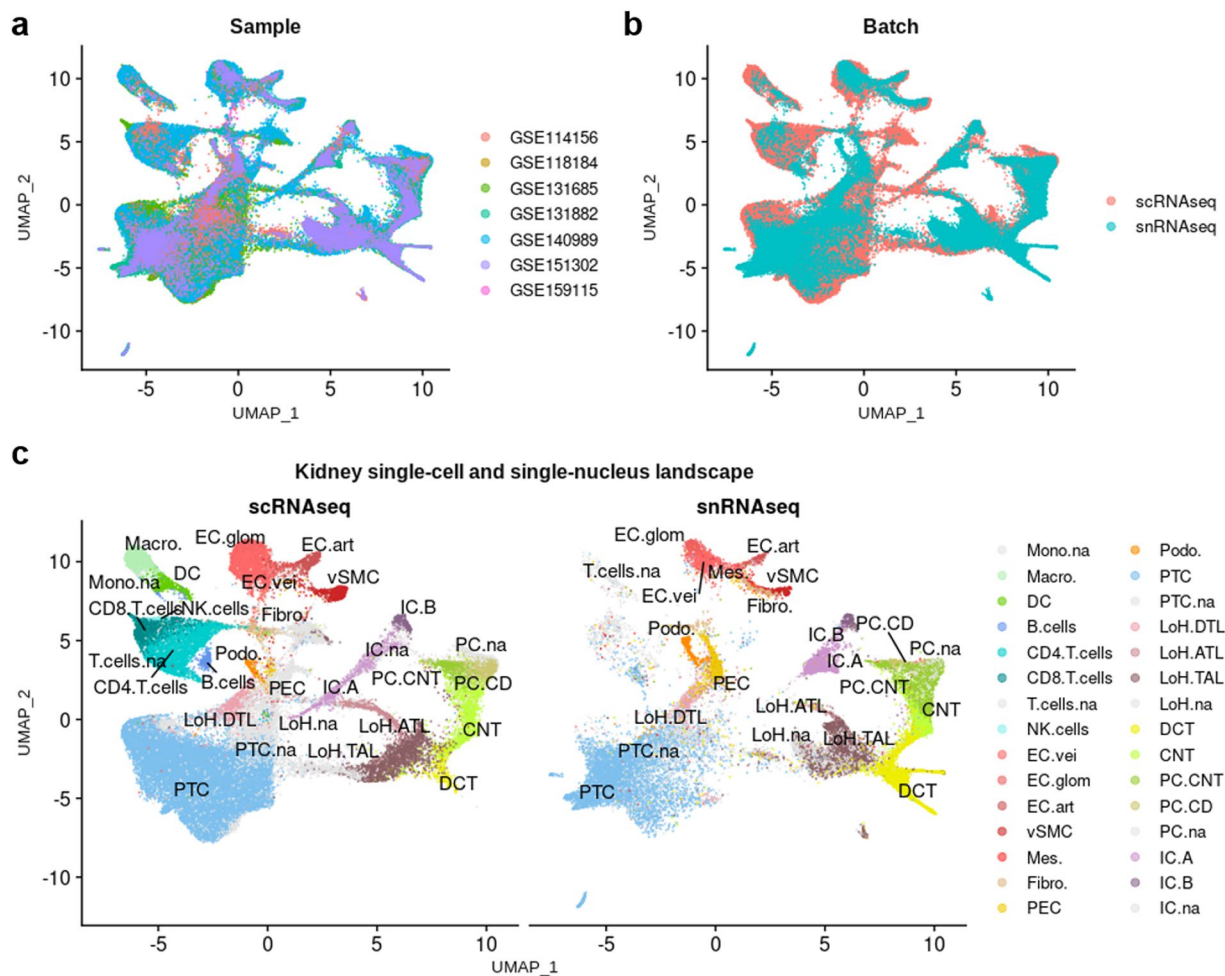
Fig. 10 Joint analysis of scRNA-seq and snRNA-seq datasets. (**a**) UMAP plot showing the dispersion of cells and nuclei after Seurat v4 integration, according to their sample of origin. (**b**) UMAP plot presenting the batch effects related to the procedure (scRNA-seq *vs* snRNA-seq). (**c**) UMAP plot showing the matching of allocated cell types between cells and nuclei.

the differentially expressed genes computed between two populations (not to mention that these differentially expressed genes are calculated on the basis of the raw data, not the matrix normalized by the integration). Mapping reads to different versions of the reference genome, as well as the pre-processing of the data are also critical factors participating in batch effects between studies. This results in limited reproducibility and reliability between the different studies involving human kidney scRNA-seq and snRNA-seq. In this meta-analysis, we performed integration of 32 scRNA-seq and 7 snRNA-seq samples, from 3 and 4 different studies respectively[26–32]. After assessing the mitigation of batch effects, we ran high resolution unsupervised clustering and allocated cell types to clusters, based on the expression of known markers, before computing consensus cell type signatures. Despite scRNA-seq and snRNA-seq samples did not equally distribute everywhere on the UMAP, cell type allocation was highly consistent across the two datasets. Finally, we showed that enrichment of consensus signatures achieved cell type allocation consistent with previously annotated datasets[56,57]. These consensus signatures may thus help increasing reproducibility and reliability between future studies involving scRNA-seq or snRNA-seq in the kidney field.

Our present study tried to standardize cell type nomenclature by the way of meta-analysis. Even though proximal nephron is functionally and anatomically divided in three segments (S1 to S3), our study could not discriminate these segments and all proximal tubule data is lumped into one proximal category. Same is true for the three descending thin limbs segments and distal convoluted tubule that is not resolved into DCT1 and DCT2. One plausible explanation is that these subpopulations are part of the unattributed population, i.e. PTC. na and LoH.na. There is also a possibility that we did not find these subpopulations because of the sparsity of the data, especially with such heterogeneity in the data sources. The datasets used in this study are the first published in the field, they were pre-processed with different tools and suffer from strong batch effects that prevent to detect many markers. However, the signatures that we proposed still allow to identify their cell types in the validation step we did.
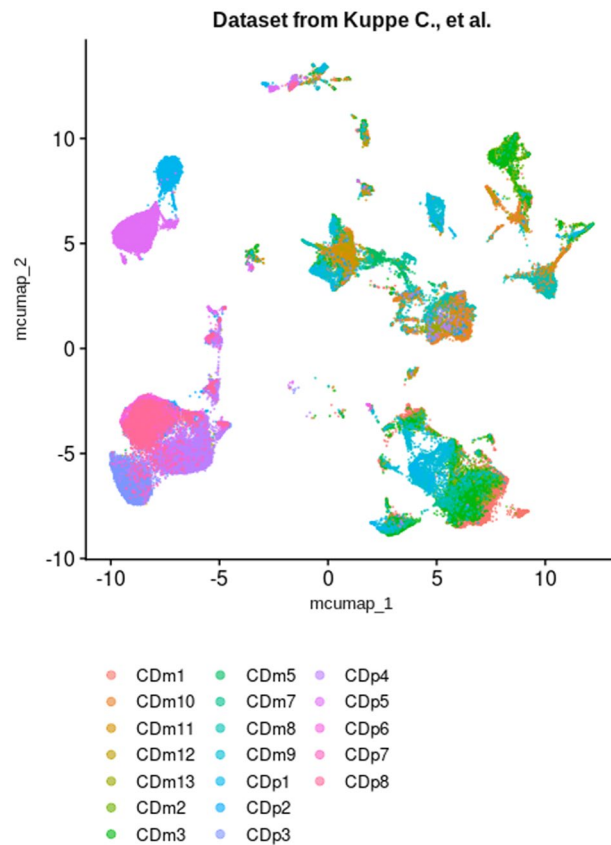
Fig. 11 Scattering of cells depending on each sample in test datasets. UMAP plot showing cell spreading according to their sample of origin within Kuppe *et al*. dataset.

Single cell and single nucleus transcriptomics allow to study the expression of every detected gene (*i.e.* read count) in every single cell, or every single nucleus, within a suspension of cells. The advantages of sn- over scRNA-seq have been studied in healthy and fibrotic mouse kidney tissue, and include the following: less dissociation bias, less stressed or dead cells, and the possibility to use frozen tissue pieces which may allow to process more and better quality biological samples[11]. Furthermore, scRNA-seq and snRNA-seq samples may present too different transcriptomes in certain cell populations to allow a good detection of every cell type in both kind of experiment. This is not striking since the set of RNA expressed in the nucleus differs from the set of RNA expressed in an entire cell. In other words, scRNA-seq samples contain nuclear, mitochondrial and cytosolic RNA, while snRNA-seq samples only contain nuclear and cytosolic RNA. Therefore, cell type assignment depending on the expression of canonical marker genes, which is the standard in single cell/nucleus transcriptomics analyses, may differ between scRNA-seq and snRNA-seq. Our analysis confirmed these assumptions and as a consequence provides two sets of cell type signatures, obtained by scRNA-seq or snRNA-seq strategies. Besides, we were not able to find immune cells within snRNA-seq datasets except few T cells, which is consistent with previous reports attesting that snRNA-seq in the kidneys failed to detect immune cells in general[9,11,59]. Interestingly, we detected a population of parietal epithelial cells in both scRNA-seq and snRNA-seq datasets that express pluripotent cell, tubular epithelium and podocyte markers (*e.g.* SOD2, KRT8, KRT18, WT1, CD24, PAX2, SOX4, VIM, RACK1, NUPR1…; Supp. Table 2) and may actually correspond to previously described parietal epithelial stem cells[60]. These cells look very different from the other clusters, express self-renewal markers (e.g. CD24, PAX2) and match with the PAX2[+] CD24[+] population previousuly observed in the glomerulus[60]. By contrast, we failed to detect any CD133[+] mesenchymal stem cell-like population.

To introduce biological heterogeneity and mitigate technical variability, we encourage authors who would use previously published healthy kidney datasets as control datasets for their purpose, to integrate several samples from different studies instead of using the samples from a single study. For those who would add new healthy human kidney samples to their single-cell or single-nucleus studies, we would advice to compare the cell type signatures from control cells with the consensus signatures we provide, and to assign cell types in their dataset using enrichment of consensus signatures (*e.g.* CelliD[58]).

However this approach is biaised in the sense that cell type enrichment depends on the tested cell types, and if a cell type is not tested it could not be attributed to cells/nuclei, even if it should. Thus, one of the main limitation of this method is that every cell or nucleus will be attributed a cell type from the tested list: the enriched cell type with the lowest p-value will be attributed, which can be misleading (if there is no enriched cell type, then cell/nucleus is labelled « unassigned »). This further means the consensus signatures we provide only define the cell types identified in the current meta-analysis. Therefore, depending on the settings, it could make sense to
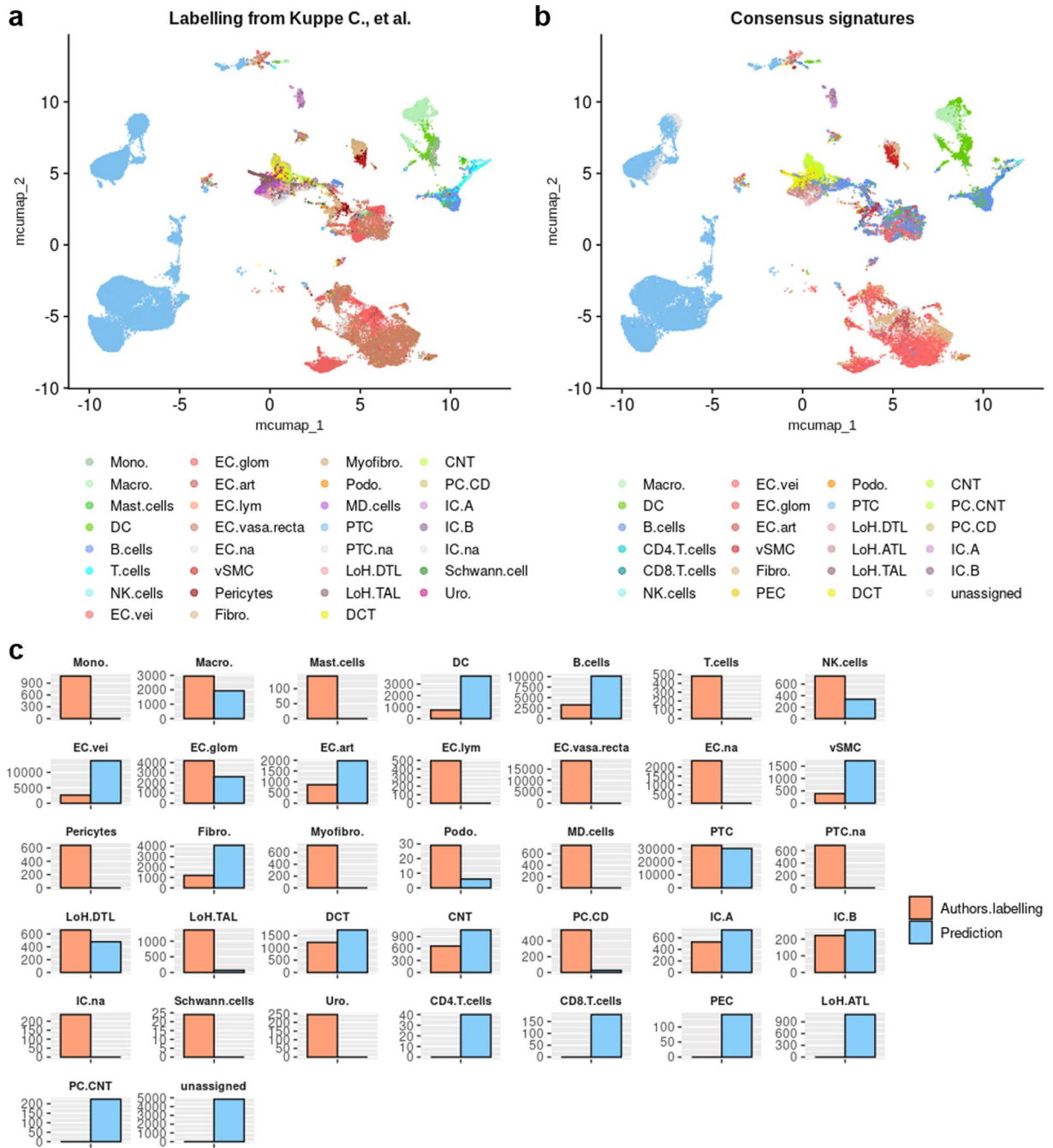
Fig. 12 Enrichment of consensus signatures automatically reveals cell type identities within scRNA-seq datasets. (**a**) UMAP plot showing the authors' original cell type annotations. These original annotations were adapted to match the nomenclatures used for consensus signatures. See also Table 5. (**b**) UMAP plot presenting the automatic cell type allocation performed by enrichment of scRNA-seq consensus signatures. (**c**) Cell count comparison for each cell type, from both original annotations by authors and consensus signature enrichment. Of note, certain labels are present in only one of the two labellings, raising counts of 0 for this label in the other labelling.

use only certain consensus signatures, for instance if the studied cells or nuclei populations have been purified by FACS prior to the transcriptomics. For the same reason, such cell type enrichment may be used as a decision helper instead of a decision maker in cell type attribution to cells/nuclei. However, a more unbiased approach is possible for original studies, based on unsupervised clustering followed by extraction of the cell-specific signatures using CelliD, and finally enrichment of functional terms or pathways of these signatures. A more general limitation of such single cell studies is the statistical power for the computation of HVG (Wilcoxon Rank Sum test), that depends on the number of cells allocated for every single cell type. Indeed, the statistical power is higher for the computation of PTC signature (computed on 29,246 PTC cells *vs* 38,782 cells within the rest of the
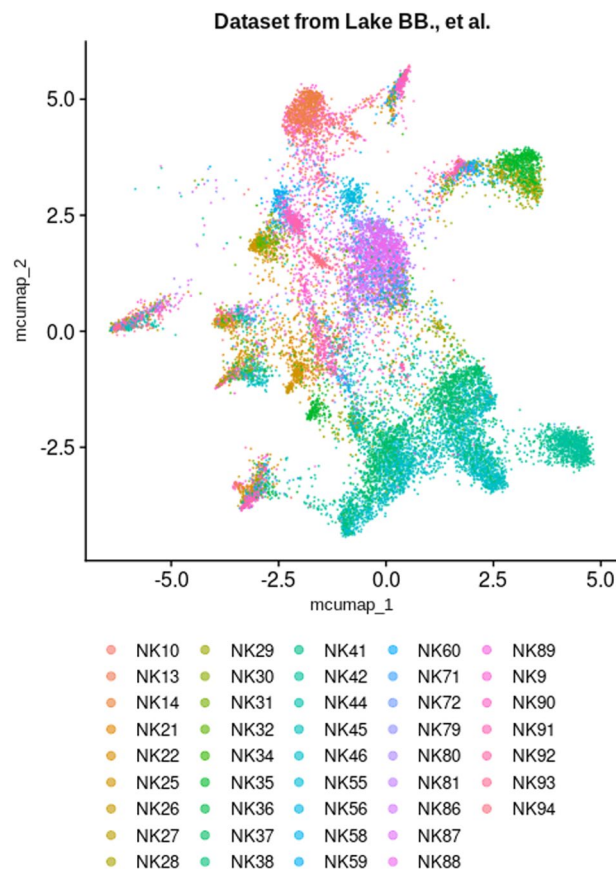
Fig. 13 Scattering of nuclei depending on each sample in test datasets. UMAP plot showing nucleus spreading according to their sample of origin within Lake *et al*. dataset.

....

dataset) than for the one of DCT cells (computed on 248 DCT cells *vs* 67,780 cells) in the scRNA-seq dataset, for instance. In the future, these signatures may benefit from being updated by integrating newly published healthy human kidney single cell datasets that may increase the biological variability and the number of cells for every population while mitigating the batch effects even better. In addition, the very low sequencing depth of these experiments implies that the results should be interpreted with caution. To solve this issue in cell type identification while specifically working with kidney tissue and validate the identified cell type-specific signatures, a bulk transcriptomic analysis of micro-dissected healthy human nephron segments would be really helpful, as it has been performed in rodents[6]. Nevertheless, cell type allocation by enrichment of consensus signatures may depend on the size of the signatures – *i.e.* the size of the gene lists, spanning between 27 (scRNA-seq signatures, LoH.TAL) and 311 (snRNA-seq, EC.art) genes in the present meta-analysis. Thus, we recommand to perform such enrichment with both the complete signatures, and truncated signatures that are close in size.

Studies involving scRNA-seq and snRNA-seq technologies in the kidney are barely comparable, because of a lack of standardized workflow (technically and analytically) and a diversity in the references used for cell type recognition. In this meta-analysis, 32 scRNA-seq samples from 3 studies, and 7 snRNA-seq samples from 4 studies, were integrated and analysed. This resulted in the computation of 30 consensus cell type signatures for kidney cell types. Future studies in the field may benefit from the use of these signatures to automatically allocate cell types to cells/nuclei.

Methods

Data acquisition.   Single-cell RNA-seq and snRNA-seq datasets generated from healthy adult kidney samples were downloaded from the Gene Expression Omnibus database (GEO; https://www.ncbi.nlm.nih.gov/geo/) as count matrices[26–32]. The collection consists of 7 snRNA-seq samples from 4 independent studies (GEO Accession ID: GSE114156, GSE118184, GSE131882, GSE151302) and 32 scRNA-seq samples from 3 independent studies (GEO Accession ID: GSE131685, GSE140989, GSE159115)[26–32]. The clinical and technical informations regarding the samples gathered from these studies are provided in Tables 1 and 3. Expression matrices of scRNA-seq samples GSM4819730 and GSM4819731 from batch GSE159115 were merged together prior to the analysis since they belong to the same individual, as well as snRNA-seq samples GSM3320197 and GSM3320198 from batch GSE118184[21,23]. Data downloaded from GEO were already pre-processed for each dataset, in different ways across the different studies involved (Table 3). This heterogeneous pre-processing of the samples may bias the analysis. However since our goal is to provide widely usable and consensus cell type signatures, this technical variation is important to retain.
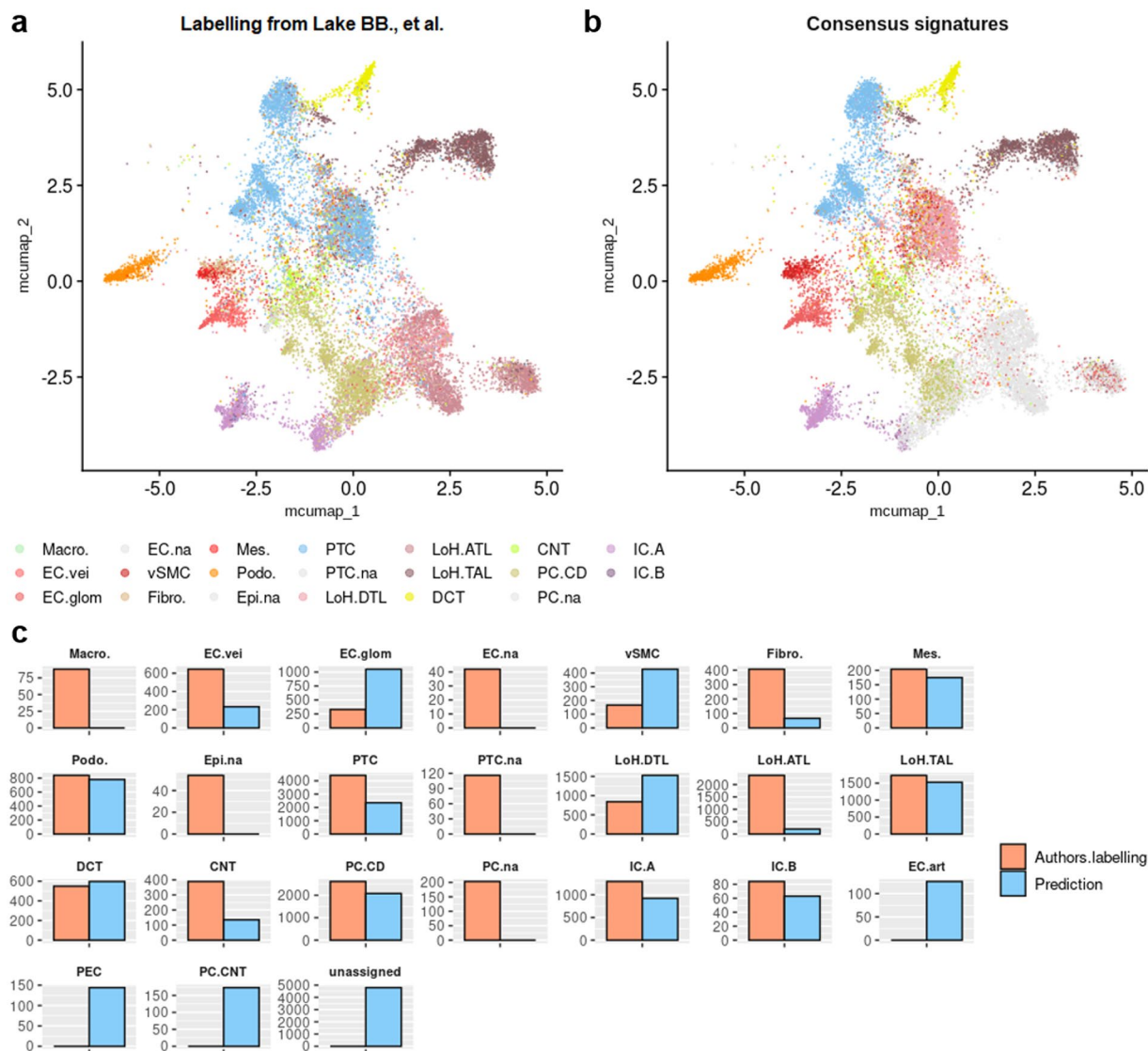
Fig. 14 Enrichment of consensus signatures automatically reveals cell type identities within snRNA-seq datasets. (**a**) UMAP plot showing the authors' original cell type annotations. These original annotations were adapted to match the nomenclatures used for consensus signatures. See also Table 5. (**b**) UMAP plot presenting the automatic cell type allocation performed by enrichment of snRNA-seq consensus signatures. (**c**) Nucleus count comparison for each cell type, from both original annotations by authors and consensus signature enrichment. Of note, certain labels are present in only one of the two labellings, raising counts of 0 for this label in the other labelling.

To test whether the computed consensus signatures may be useful to automatically allocate cell types to clusters, we also downloaded available annotated datasets. Thus, Kuppe *et al*. (#4059315)[55,56] scRNA-seq dataset was obtained from zenodo repository (https://zenodo.org/), and Lake *et al*. snRNA-seq dataset was downloaded from GEO under accession number GSE121862[9,57]. These datasets consisted in 19 chronic kidney disease samples and 43 healthy samples, respectively.

Quality control and filtering out of poor quality cells and nuclei. We used R software v4.1.0 (https://www.r-project.org/) and Seurat v4.0.5 package[61] (https://satijalab.org/seurat/) to perform the analysis. As observed in previous studies, human kidney scRNA-seq datasets generally present with high mitochondrial gene counts, which may be attributed to the processing time of human kidney samples as well as the processing itself. Moreover kidney tissue notoriously contains a lot of mitochondria, consistent with the high levels of energy needed for a proper filtration process. Therefore the standard filtering out of cells with >5% mitochondrial gene expressed was not suitable for the processing of these scRNA-seq data. Cells with <200 or >3500 (cell debris and doublets) expressed genes, and >30% mitochondrial gene expressed, were filtered out, whereas nuclei with <200 or >3500 expressed genes, and >5% mitochondrial gene expressed, were filtered out. In total, 68,028 high quality cells and 33,412 high quality nuclei were obtained after applying these thresholds. Table 2 presents quality control

scRNA-seq test dataset: Kuppe C, *et al*. Nature. 2021[37,38]			snRNA-seq test dataset: Lake BB, *et al*. Nat Commun. 2019[9,39]		
Original labels		Replacement	Original labels		Replacement
C1	Arteriolar Endothelium	EC.art	C1	Epithelial Cells (unassigned)	Epi.na
C2	B Cells	B.cells	C2	Podocytes	Podo.
C3	Collecting Duct Principal Cells	PC.CD	C3	Proximal Tubule Epithelial Cells (S1)	PTC
C4	Connecting Tubule	CNT	C4	Proximal Tubule Epithelial Cells (S2)	PTC
C5	Dendritic Cells	DC	C5	Proximal Tubule Epithelial Cells - Stress/Inflam	PTC.na
C6	Descending Thin Limb	LoH.DTL	C6	Proximal Tubule Epithelial Cells - Fibrinogen + (S3)	PTC
C7	Distal Convoluted Tubule	DCT	C7	Proximal Tubule Epithelial Cells (S3)	PTC
C8	Fibroblast 2	Fibro.	C8	Decending Limb	LoH.DTL
C9	Fibroblast 4	Fibro.	C9	Thin ascending limb	LoH.ATL
C10	Fibroblast 6	Fibro.	C10	Thin ascending limb	LoH.ATL
C11	Glomerular Capillaries	EC.glom	C11	Thin ascending limb	LoH.ATL
C12	Injured Endothelial Cells	EC.na	C12	Thick Ascending Limb	LoH.TAL
C13	Injured Proximal tubule	PTC.na	C13	Thick Ascending Limb	LoH.TAL
C14	Intercalated Cells 3	IC.na	C14	Distal Convoluted Tubule	DCT
C15	Intercalated Cells 4	IC.na	C15	Connecting Tubule	CNT
C16	Intercalated Cells 5	IC.na	C16	Collecting Duct - Principal Cells (cortex)	PC.CD
C17	Intercalated Cells 6	IC.na	C17	Collecting Duct - PCs - Stressed Dissoc Subset	PC.na
C18	Intercalated Cells 7	IC.na	C18	Collecting Duct - Principal Cells (medulla)	PC.CD
C19	Intercalated Cells 8	IC.na	C19	Collecting Duct - Intercalated Cells Type A (medulla)	IC.A
C20	Intercalated Cells A	IC.A	C20	Collecting Duct - Intercalated Cells Type A (cortex)	IC.A
C21	Intercalated Cells B	IC.B	C21	Collecting Duct - Intercalated Cells Type B	IC.B
C22	Lymph Endothelium	EC.lym	C22	Endothelial Cells - glomerular capillaries	EC.glom
C23	Macrophages 1	Macro.	C23	Endothelial Cells - AVR	EC.vei
C24	Macrophages 2	Macro.	C24	Endothelial Cells - AEA & DVR	EC.vei
C25	Macrophages 3	Macro.	C25	Endothelial Cells (unassigned)	EC.na
C26	Macula Densa Cells	MD.cells	C26	Mesangial Cells	Mes.
C27	Mast Cells	Mast.cells	C27	Vascular Smooth Muscle Cells and pericytes	vSMC
C28	Monocytes	Mono.	C28	Interstitium	Fibro.
C29	Myofibroblast 1a	Myofibro.	C29	Unknown - Novel PT CFH + Subpopulation (S2)	PTC
C30	Myofibroblast 1b	Myofibro.	C30	Immune Cells - Macrophages	Macro.
C31	Natural Killer Cells	NK.cells			
C32	Pericytes 1	Pericytes			
C33	Pericytes 2	Pericytes			
C34	Plasma Cells	B.cells			
C35	Podocytes	Podo.			
C36	Proximal Tubule	PTC			
C37	S1	PTC			
C38	S1/2 1	PTC			
C39	S1/2 2	PTC			
C40	S1/2 3	PTC			
Continued					

scRNA-seq test dataset: Kuppe C, *et al*. Nature. 2021[37,38]			snRNA-seq test dataset: Lake BB, *et al*. Nat Commun. 2019[9,39]		
Original labels		Replacement	Original labels		Replacement
C41	S3 1	PTC			
C42	S3 2	PTC			
C43	S3 3	PTC			
C44	Schwann Cells	Schwann.cells			
C45	T Cells	T.cells			
C46	Thick Ascending Limb 2	LoH.TAL			
C47	Thick Ascending Limb 3	LoH.TAL			
C48	Thick Ascending Limb 4	LoH.TAL			
C49	Uroethlial Cells	Uro.			
C50	Vasa Recta 1	EC.vasa.recta			
C51	Vasa Recta 2	EC.vasa.recta			
C52	Vasa Recta 3	EC.vasa.recta			
C53	Vasa Recta 4	EC.vasa.recta			
C54	Vasa Recta 5	EC.vasa.recta			
C55	Vasa Recta 6	EC.vasa.recta			
C56	Vascular Smooth Muscle Cells	vSMC			
C57	Venular Endothelium	EC.vei			

Table 5. Nomenclature for test datasets.

metrics of every sample (*i.e.* number of cells/nuclei, mean number of reads per cell/nucleus, mean number of features expressed per cell/nucleus, % mitochondrial genes, % ribosomal genes), prior to and after filtering. Data were normalized and scaled (regressing out % mitochondrial genes), and highly variable genes computed using the SCTransform() function[35] (Seurat v4) for every scRNA-seq and snRNA-seq sample. Identified HVG were then used to compute PCA for every sample. SCTransform is a newly implemented statistical method in Seurat v4, pooled from the sctransform R package (https://github.com/satijalab/sctransform), that aims to better resolve the technical variability and sequencing depth differencies between cells/nuclei across datasets[35]. It is particularly interesting when working with datasets obtained from different sources, which induce important variability.

Integration and dimensional reduction. Single-cell and single-nucleus samples always depend on confounding variables and may thus present differences that are called batch effects. To allow any comparison between samples, batch effects need to be mitigated as much as possible, which is done by the integration process (*i.e.* normalization step). Because further computations depend on this process, the quality of the integration deserves to be evaluated. Thus, two integration approaches were considered: the Seurat v4 method that outputs a corrected expression matrix for a list of genes to consider, and the Harmony v0.1.0 method that directly corrects the residues of the PCA for each sample.

Integration of 32 scRNA-seq samples on one hand, and 7 snRNA-seq samples on the other, was achieved by running consecutively PrepSCTIntegration(), FindIntegrationAnchors() and IntegrateData() functions from Seurat, with 2,500 integration features. Then PCA was computed and the first 30 PCs were inputed for uniform manifold approximation and projection (UMAP) of integrated scRNA-seq and snRNA-seq datasets. Harmony ran as well and UMAP was computed on the 30 first corrected PCs of both dataset. The distribution of the cells or nuclei from the different samples was compared between the two methods. Of note, Seurat and Harmony are among the best batch effect correction methods to date[17,18].

Clustering and cell type annotation. High resolution clustering is important in such meta-analysis: since there are still notable batch effects, small batch-dependent clusters may be identified. In addition, more clusters may identify more cell types when closely related, thus more consensus cell type signatures if so. Unsupervised clustering was performed using FindClusters() function with Louvain algorithm in both dataset (resolution = 3.4 and 3.0 in scRNA-seq and snRNA-seq datasets, respectively). Distribution of samples across clusters was studied thanks to the pheatmap v1.0.12 R package. The cells were then labelled according to the expression of specific markers (Table 4). To match the nomenclature adopted for consensus cell type signatures, original labels from Kuppe C, *et al*. and Lake BB, *et al*. were changed (Table 5).

Data availability

The single-cell and single-nucleus datasets generated in the study have been deposited on Figshare[62,63]. These files contain 4 assay slots (raw counts matrix, sample-dependent SCT-transformed values, post-integration SCT-corrected values, and the secondary integration SCT-corrected values) and some meta-data slots, including the dataset of origin (GEO sample accession number), the batch of origin (GEO series accession number), the method used (scRNA-seq vs snRNA-seq), the clusters, and the cell type labelling. The Figshare repository also contains supplementary Tables 1 and 2[64,65].

Code availability

The R script allowing to reproduce the entire study is available on Github (https://github.com/MarceauQuatredeniers/Meta-analysis-of-healthy-human-kidney-single-cell-transcriptomics).

References

1. Schumacher, A. *et al.* Defining the variety of cell types in developing and adult human kidneys by single-cell RNA sequencing. *NPJ Regen Med* **6**, 45 (2021).
2. Knepper, M. & Burg, M. Organization of nephron function. *American Journal of Physiology-Renal Physiology* **244**, F579–F589 (1983).
3. Verschuren, E. H. J. *et al.* Sensing of tubular flow and renal electrolyte transport. *Nat Rev Nephrol* **16**, 337–351 (2020).
4. Chabardès-Garonne, D. *et al.* A panoramic view of gene expression in the human kidney. *Proc Natl Acad Sci USA* **100**, 13710–13715 (2003).
5. Habuka, M. *et al.* The Kidney Transcriptome and Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PLOS ONE* **9**, e116125 (2014).
6. Lee, J. W., Chou, C.-L. & Knepper, M. A. Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment-Specific Transcriptomes. *J Am Soc Nephrol* **26**, 2669–2677 (2015).
7. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics* **10** (2019).
8. Fischer, J. & Ayers, T. Single nucleus RNA-sequencing: how it's done, applications and limitations. *Emerg Top Life Sci* **5**, 687–690 (2021).
9. Lake, B. B. *et al.* A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nat Commun* **10**, 2832 (2019).
10. Grindberg, R. V. *et al.* RNA-sequencing from single nuclei. *Proc Natl Acad Sci USA* **110**, 19802–19807 (2013).
11. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J Am Soc Nephrol* **30**, 23–32 (2019).
12. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
13. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
14. Stewart, B. J., Ferdinand, J. R. & Clatworthy, M. R. Using single-cell technologies to map the human immune system - implications for nephrology. *Nat Rev Nephrol* **16**, 112–128 (2020).
15. Chen, L. *et al.* Renal-Tubule Epithelial Cell Nomenclature for Single-Cell RNA-Sequencing Studies. **12** (2019).
16. Dal Molin, A. & Di Camillo, B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief Bioinform* **20**, 1384–1394 (2019).
17. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* **21**, 12 (2020).
18. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* **16**, 43–49 (2019).
19. Wilson, P. C. *et al.* The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc Natl Acad Sci USA* **116**, 19619–19625 (2019).
20. Wu, H. *et al.* Single-Cell Transcriptomics of a Human Kidney Allograft Biopsy Specimen Defines a Diverse Inflammatory Response. *J Am Soc Nephrol* **29**, 2069–2080 (2018).
21. Wu, H. *et al.* Comparative Analysis and Refinement of Human PSC-Derived Kidney Organoid Differentiation with Single-Cell Transcriptomics. *Cell Stem Cell* **23**, 869–881.e8 (2018).
22. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat Commun* **12**, 2190 (2021).
23. Zhang, Y. *et al.* Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. *Proc Natl Acad Sci USA* **118**, e2103240118 (2021).
24. Liao, J. *et al.* Single-cell RNA sequencing of human kidney. *Sci Data* **7**, 4 (2020).
25. Menon, R. *et al.* Single cell transcriptomics identifies focal segmental glomerulosclerosis remission endothelial biomarker. *JCI Insight* **5**, 133267 (2020).
26. Wilson, P. C. *et al. GEO.* https://identifiers.org/geo/GSE131882 (2019).
27. Wu, H. *et al. GEO.* https://identifiers.org/geo/GSE118184 (2018).
28. Wu, H. *et al. GEO.* https://identifiers.org/geo/GSE114156 (2018).
29. Muto, Y. *et al. GEO.* https://identifiers.org/geo/GSE151302 (2021).
30. Zhang, Y. *et al. GEO.* https://identifiers.org/geo/GSE159115 (2021).
31. Liao, J. *et al. GEO.* https://identifiers.org/geo/GSE131685 (2020).
32. Menon, R. *et al. GEO.* https://identifiers.org/geo/GSE140989 (2020).
33. Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).
34. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
35. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).
36. Ransick, A. *et al.* Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Dev Cell* **51**, 399–413.e7 (2019).
37. Clark, J. Z. *et al.* Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. *Kidney Int* **95**, 787–796 (2019).
38. Brunskill, E. W. & Potter, S. S. Gene expression programs of mouse endothelial cells in kidney development and disease. *PLoS One* **5**, e12034 (2010).
39. Fetting, J. L. *et al.* FOXD1 promotes nephron progenitor differentiation by repressing decorin in the embryonic kidney. *Development* **141**, 17–27 (2014).
40. Volkert, G. *et al.* Contribution of the α8 integrin chain to the expression of extracellular matrix components. *Cell Commun Adhes* **21**, 89–98 (2014).
41. Zeisberg, M. & Kalluri, R. Physiology of the Renal Interstitium. *Clin J Am Soc Nephrol* **10**, 1831–1840 (2015).
42. Je, H.-D., Gallant, C., Leavis, P. C. & Morgan, K. G. Caveolin-1 regulates contractility in differentiated vascular smooth muscle. *Am J Physiol Heart Circ Physiol* **286**, H91–98 (2004).
43. Warthi, G. *et al.* Generation and Comparative Analysis of an Itga8-CreER T2 Mouse with Preferential Activity in Vascular Smooth Muscle Cells. *Nat Cardiovasc Res* **1**, 1084–1100 (2022).

44. Leppänen, V.-M. *et al.* Structural determinants of growth factor binding and specificity by VEGF receptor 2. *Proc Natl Acad Sci USA* **107**, 2425–2430 (2010).
45. Su, C. *et al.* Single-Cell RNA Sequencing in Multiple Pathologic Types of Renal Cell Carcinoma Revealed Novel Potential Tumor-Specific Markers. *Front Oncol* **11**, 719564 (2021).
46. Conway, B. R. *et al.* Kidney Single-Cell Atlas Reveals Myeloid Heterogeneity in Progression and Regression of Kidney Disease. *J Am Soc Nephrol* **31**, 2833–2854 (2020).
47. Grossman, W. J. *et al.* Differential expression of granzymes A and B in human cytotoxic lymphocyte subsets and T regulatory cells. *Blood* **104**, 2840–2848 (2004).
48. Tewary, P. *et al.* Granulysin activates antigen-presenting cells through TLR4 and acts as an immune alarmin. *Blood* **116**, 3465–3474 (2010).
49. Muto, S. Physiological roles of claudins in kidney tubule paracellular transport. *Am J Physiol Renal Physiol* **312**, F9–F24 (2017).
50. Ma, S., Charron, J. & Erikson, R. L. Role of Plk2 (Snk) in mouse development and cell proliferation. *Mol Cell Biol* **23**, 6936–6943 (2003).
51. Donohue, P. J., Alberts, G. F., Guo, Y. & Winkles, J. A. Identification by targeted differential display of an immediate early gene encoding a putative serine/threonine kinase. *J Biol Chem* **270**, 10351–10357 (1995).
52. Yang, Q. *et al.* Single-Cell RNA Sequencing Reveals the Heterogeneity of Tumor-Associated Macrophage in Non-Small Cell Lung Cancer and Differences Between Sexes. *Front Immunol* **12**, 756722 (2021).
53. Saxena, V. *et al.* Publisher Correction: Whole Transcriptome Analysis of Renal Intercalated Cells Predicts Lipopolysaccharide Mediated Inhibition of Retinoid X Receptor alpha Function. *Sci Rep* **10**, 5090 (2020).
54. Chen, L., Chou, C.-L. & Knepper, M. A. A Comprehensive Map of mRNAs and Their Isoforms across All 14 Renal Tubule Segments of Mouse. *J Am Soc Nephrol* **32**, 897–912 (2021).
55. Kuppe, C. *et al.* Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281–286 (2021).
56. Kuppe, C. *et al. Zenodo.* https://doi.org/10.5281/zenodo.4059315 (2021).
57. Lake, B. B. *et al. GEO.* https://identifiers.org/geo/GSE121862 (2019).
58. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat Biotechnol* **39**, 1095–1102 (2021).
59. O'Sullivan, E. D., Mylonas, K. J., Hughes, J. & Ferenbach, D. A. Complementary Roles for Single-Nucleus and Single-Cell RNA Sequencing in Kidney Disease Research. *J Am Soc Nephrol* **30**, 712–713 (2019).
60. Huang, J., Kong, Y., Xie, C. & Zhou, L. Stem/progenitor cell in kidney: characteristics, homing, coordination, and maintenance. *Stem Cell Research & Therapy* **12**, 197 (2021).
61. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
62. Quatredeniers, M. *et al. Figshare.* https://doi.org/10.6084/m9.figshare.21587670 (2022).
63. Quatredeniers, M. *et al. Figshare.* https://doi.org/10.6084/m9.figshare.21587679 (2022).
64. Quatredeniers, M. *et al. Figshare.* https://doi.org/10.6084/m9.figshare.21587697 (2022).
65. Quatredeniers, M. *et al. Figshare.* https://doi.org/10.6084/m9.figshare.21587709 (2022).

Acknowledgements

Author contributions

M.Q., A.S.S., S.S. and A.V. designed the study. M.Q. performed all the analyses and built the figures. A.R. provided bioinformatics advices and resources. M.Q., A.S.S. and A.V. drafted the manuscript. M.Q., A.S.S., A.B., A.R., S.S. and A.V. discussed and reviewed the manuscript. A.V. and S.S. jointly supervised the work. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.