Unlocking the complexities of human kidney and heart disease

Unlocking the Complexities of Human Kidney and Heart Disease:

An Integrative Single Cell and Spatial Analysis

Christoph Kuppe

Maastricht University

Unlocking the Complexities of Human Kidney and Heart Disease

An Integrative Single Cell and Spatial Analysis

ISBN: 978-94-6419-834-8

Unlocking the Complexities of Human Kidney and Heart Disease:

An Integrative Single Cell and Spatial Analysis

Dissertation

to obtain the degree of Doctor
at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Pamela Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday 4th of Octobre 2023, at 10.00 hours

by

Christoph Kuppe

Supervisors
Prof. Dr. L.J. Schurgers
Prof. Dr. R. Kramann, RWTH Aachen, Germany

Assessment Committee
Prof. Dr. Judith Sluimer (Chair)
Dr. Marian Clahsen-van Groningen (Erasmus University Rotterdam)
Prof. Dr. Paul Shiels (Glasgow University)
Prof. Dr. Monica Stoll
Prof. Dr. Kevin Vernooy

To Emma, Frieda and Jennifer

CONTENTS

1

General Introduction

1.1. Glomerulosclerosis and interstitial fibrosis in chronic kidney disease.

Chronic kidney disease (CKD) represents a complex disease characterized by persistent structural (glomerulosclerosis and fibrosis) and functional (proteinuria, deterioration of glomerular filtration rate) renal changes (Zoccali et al. 2017). Worldwide, approximately 10% of the population suffers from CKD and it is predicted that approximately 14 out of every 100,000 people worldwide will die from it per year in 2030 (GBD 2019 Diseases and Injuries Collaborators 2020). For patients with end-stage CKD, treatment options include chronic hemodialysis, which is associated with significantly increased excess mortality, or kidney transplantation (Levey and Coresh 2012). However, waiting lists for kidney transplantation are currently very long, so many CKD-patients die on dialysis before a suitable donor organ can be transplanted. Current therapeutic options are mostly limited to systemic immunosuppression and blood pressure control using ACE inhibitors or systemic glucocorticoid therapy (Romagnani et al. 2017). Specific treatment options to counteract the progression of chronic renal failure are not currently available.

The kidney is composed of approximately 1-1.5 million functional and anatomically distinct subunits called nephrons (see Figure 1A). These in turn can be differentiated into a glomerular, tubular, or tubulo-interstitial compartment, each of which has a specific physiological role in renal function. Damage from, for example, inflammatory or hypoxic influences in any one of these compartments can lead to irreversible nephron loss, and thus to loss of renal function (Ruggenenti, Cravedi, and Remuzzi 2012) (Figure 1B-C). Two stereotypic pathological responses in the different compartments are associated with progressive renal function loss: tubulointerstitial fibrosis (TIF) (Zeisberg and Neilson 2010) and glomerulosclerosis or secondary focal segmental glomerulosclerosis (GS, FSGS) (Kriz, Hartmann, and Hosser 2001). A detailed and in-depth understanding of the early molecular changes in the different renal cell types is necessary to develop new therapeutic approaches. These therapeutic approaches aim to halt the progression of chronic renal failure as early as possible.

1.2. Cellular mechanisms of interstitial scarring in CKD

Following acute renal failure, reparative mechanisms characterized by consecutive, sometimes overlapping tissue responses begin in the kidney (Coelho et al. 2018; Basile, Anderson, and Sutton 2012). As part of the reparative process, an inflammatory response usually results in extracellular matrix production (fibrogenesis), resolution, and regeneration (proliferation of intact tubule epithelium) (Figure 1C). Often, the reparative

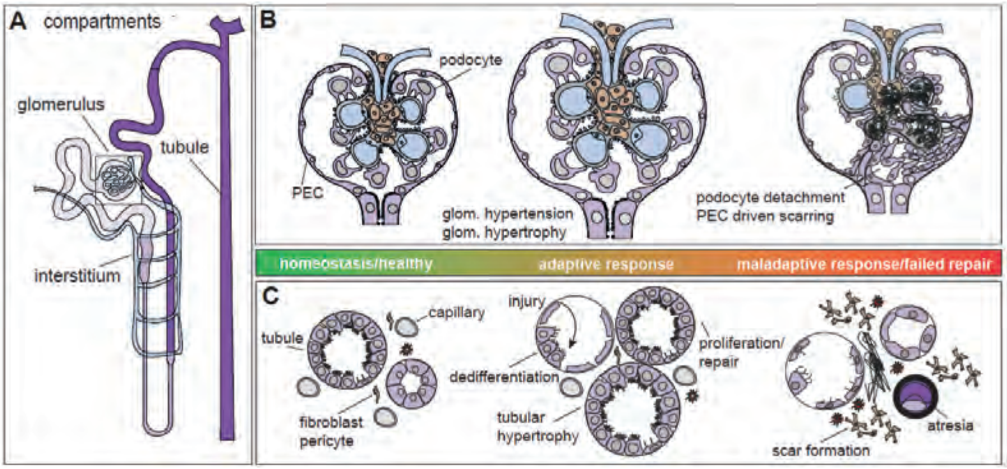


Fig. 1 Kidney compartment specific CKD adaptive and maladaptive mechanisms.

A. The functional units of mammalian kidneys can be divided into different compartments: glomerular compartment, tubular compartment, and interstitial compartment. **B.** Upon glomerular injury adaptive changes occur due to glomerular hypertension which leads to glomerular hypertrophy. If a certain threshold of podocyte stress and podocyte loss is reached, parietal epithelia cells become activated and invade the glomerular capillaries causing irreversible scarring and nephron loss. **C.** Upon tubular injury proteinuria and misdirected filtration leads to tubular stress and activation as well as to secretion of cytokines and chemokines that promote interstitial inflammation and activation of fibroblasts. Compensatory hypertrophy of tubules of the remnant nephrons occurs.

process does not proceed to complete regeneration of renal tissue, resulting in the formation of an interstitial scar (IF) that lacks regenerative capacity, leading to irreversible loss of renal tissue (Venkatachalam et al. 2015). The cellular origin of fibrogenesis in both acute and chronic forms of tubulointerstitial fibrosis is not yet known. It was postulated by Frank et al. 1993 that damaged tubular epithelial cells secrete cytokines that activate surrounding fibroblasts (Frank et al. 1993). In addition, Hewitson et al. demonstrated that the number of intrarenal fibroblasts increases exponentially in the renal interstitium after injury (Hewitson 2009). In addition to orthotropic fibroblasts, other cell types have been discussed as cells of origin, e.g., renal tubule cells that differentiate into myofibroblasts by means of an epithelial-mesenchymal transition (EMT) (Zeisberg et al. 2003). In addition, pericytes, circulating bone marrow mesenchymal stem cells, or macrophages or endothelial cells have been discussed as progenitor cells involved in fibrogenesis (Kramann, DiRocco, and Humphreys 2013). However, that epithelial cells contribute to the myofibroblast pool through EMT has been refuted in recent years (Duffield 2014). Without doubt, epithelial cells, especially proximal tubule cells, play an important role in the development and progression of renal fibrosis through a release of second messengers such as cytokines. In this process, termed partial EMT, epithelial cells do not leave the tubular compartment (Lovisa, Zeisberg, and Kalluri 2016). Also discussed were endothelial cells as progenitors of myofibroblasts as the so-called endothelial-mesenchymal transition (EndoMT) (Piera-Velazquez, Li, and Jimenez 2011). However, recent data show that the studies describing that phenomenon have used Cre-Driver lines that are not specific for endothelial cells. For example, a Tie2Cre mouse was used, which is also expressed by hematopoietic cells and pericytes. So-called fibrocytes are CD45+ cells of the bone marrow that are also thought to contribute to fibrosis. After organ injury, these cells migrate into the injured organ via the blood circulation and differentiate into myofibroblasts. Another population that has been discussed is MSCs (Mesenchymal Stem Cells) from bone marrow, but there is little evidence for this to date. In recent years, there is a consensus that resident mesenchymal cells, such as pericytes and resident fibroblasts, are the most likely cells from which the majority of myofibroblasts arise. Humphreys et al. demonstrated that FoxD1+ interstitial pericytes and fibroblasts give rise to nearly 100% of all renal myofibroblasts (Humphreys et al. 2010). However, FoxD1 is a relatively non-specific marker that is also expressed by other cells that do not differentiate into myofibroblasts, such as: vascular smooth muscle cells, mesangial cells, and few tubular epithelial cells. Interestingly, it was shown more than a hundred years ago that fibrosis very often has a vascular origin. Moreover, cell markers such as alpha-smooth muscle actin (aSMA) have been frequently used in various experimental approaches, without really knowing whether this cellular marker is exclusively expressed in renal myofibroblasts. Which cell types are involved in the development of human tubulointerstitial fibrosis has not yet been clarified. Accurate molecular analysis and characterization of the activation processes in these cells could not only improve understanding and end a decade-long debate about the cell of origin of renal fibrosis, but also provide potential therapeutic targets or signaling pathways.

1.3. Myocardial infarction and heart failure

Myocardial infarction (MI) and heart failure (HF) are major public health concerns worldwide, with high morbidity and mortality rates(Jayaraj et al. 2019). Myocardial infarction, also known as a heart attack, occurs when the blood flow to a part of the heart is blocked, typically by a build-up of plaque in the coronary arteries (Frangogiannis 2015). This results in damage or death of heart muscle cells, which can lead to a range of complications, including heart failure. Heart failure is a condition in which the heart is unable to pump enough blood to meet the body's needs. This can be caused by a variety of factors, including damage from a heart attack, hypertension, diabetes, and other underlying heart conditions. Heart failure is characterized by a range of symptoms, including shortness of breath, fatigue, and swelling in the legs and feet. The incidence of myocardial infarction and heart failure is increasing globally, due in part to the aging population and the rising prevalence of risk factors such as obesity, diabetes, and hypertension. Despite advances in treatment, myocardial infarction and heart failure continue to be major causes of morbidity and mortality worldwide. The pathophysiology of myocardial infarction and heart failure is complex and involves a range of cellular and molecular mechanisms (Bergmann 2010). The acute phase of myocardial infarction is characterized by the formation of an infarct zone, which is the area of the heart muscle that is damaged or killed due to the lack of blood flow. This is followed by a series of remodeling processes that occur in the weeks and months following the infarction, including inflammation, fibrosis, and changes in the structure and function of the heart muscle. Heart failure is characterized by a progressive decline in cardiac function, which can be caused by a range of underlying mechanisms, including myocardial infarction, hypertension, and other underlying heart conditions. These mechanisms can lead to structural changes in the heart, such as hypertrophy and fibrosis, as well as changes in the function of the heart muscle, such as decreased contractility and relaxation. The management of myocardial infarction and heart failure involves a range of interventions, including lifestyle changes, medications, and surgical and non-surgical procedures. Lifestyle changes, such as quitting smoking, eating a healthy diet, and exercising regularly, can reduce the risk of myocardial infarction and heart failure. Medications, such as ACE inhibitors, beta-blockers, and diuretics, can be used to manage the symptoms of heart failure and reduce the risk of complications. Surgical and non-surgical procedures, such as coronary artery bypass surgery and angioplasty, can be used to treat myocardial infarction. In addition, heart failure can be treated by implantation of devices such as cardiac resynchronization therapy (CRT) or left ventricular assist devices (LVADs) which help to improve the function of the heart. In conclusion, myocardial infarction and heart failure are major public health concerns worldwide, with high morbidity and mortality rates. Despite advances in treatment, myocardial infarction and heart failure continue to be major causes of morbidity and mortality worldwide. The pathophysiology of myocardial infarction and heart failure is complex and involves a range of cellular and molecular mechanisms. The management of myocardial infarction and heart failure involves a range of interventions, including lifestyle changes, medications, and surgical and non-surgical procedures. It is important to continue research to improve our understanding of the underlying mechanisms of these conditions and to develop new and more effective treatments.

1.4. Outline of the thesis

The research described in this thesis aims to unravel mechanisms underlying progression of chronic kidney and ischemic heart disease. We characterized human kidney and heart cells from healthy and diseased human samples using single cell multiomics assays and spatial transcriptomics with the specific focus to unravel cellular and molecular heterogeneity of disease driving cell states. Moreover, we studied for the first time the spatial molecular changes following acute myocardial infarction in humans.

Chapter 2 provides an overview of the emerging single cell genomics field and covers both wet-lab aspects of the technology as well as the different computational approaches in order to analyse the data. The article discusses the use of experimental and computational technologies to study the kidney at the single-cell level. These technologies include single-cell RNA sequencing, imaging, and computational modeling. These methods allow for a more detailed and accurate understanding of the complex cellular and molecular processes that occur in the kidney. The article also highlights the potential benefits of using these technologies in understanding kidney disease and developing new treatments.

In **Chapter 3** we aimed to understand the origin of myofibroblasts, a type of cell that plays a key role in the development of fibrosis in the kidney. We used single-cell RNA sequencing and computational analysis to identify and track the origin and progression of myofibroblasts in patients with kidney fibrosis. The results of the study revealed that myofibroblasts can originate from multiple cell types, including fibroblasts, endothelial cells, and pericytes, and that the origin of these cells may vary depending on the specific type of kidney fibrosis. We wish to investigate that the myofibroblasts that originate from different cell types have distinct gene expression profiles and functional characteristics. Our approach may result in the identification of specific novel myofibroblast antigens which can be utilized as promising strategy for treating kidney fibrosis.

In **Chapter 4** we describe a computational method called scOpen that is used to analyze data from single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) experiments. The scOpen method enables the estimation of chromatin accessibility at the single-cell level, which provides insights into the regulation of gene expression and cellular differentiation. We aim to show that the scOpen method outperforms existing methods in terms of accuracy and robustness when applied to various scATAC-seq datasets from different cell types and organisms. The scOpen method enables the identification of cell-type specific cis-regulatory elements and the study of chromatin accessibility dynamics across cell states and developmental processes. We aim to suggest that scOpen is a powerful tool for understanding the mechanisms of cellular differentiation and gene regulation in different cell types and organisms.

In **Chapter 5** we describe a computational method to integrate multi-omics data with prior knowledge to generate mechanistic hypotheses about the underlying biology.
Integration of multiple data types and prior knowledge can improve the understanding of the relationships between different molecules and processes. The method uses a causal inference approach to identify the key regulatory molecules and interactions that are likely to be involved in the biological processes of interest, by taking into account the direction of the causal effects between variables. The authors demonstrated the usefulness of the method by applying it to the analysis of multi-omics data in cancer and demonstrated the ability of the method to generate testable hypotheses about the underlying biology. The study will describe the method as a powerful tool to integrate multi-omics data with prior knowledge to generate testable hypotheses about the underlying biology.

In **Chapter 6** we aimed to generate a detailed map of the human cardiac remodeling process following myocardial infarction (MI) using single-cell gene expression, chromatin accessibility, and spatial transcriptomic profiling. We will use multimodal data integration of physiological zones at distinct time points in myocardium from patients with MI and controls. The data integration enables researchers to evaluate cardiac cell-type compositions at an increased resolution and yielded insights into changes in the cardiac transcriptome and epigenome through the identification of distinct tissue structures of injury, repair, and remodeling. The study will identify and validated disease-specific cardiac cell states of major cell types, and analyzed them in their spatial context. The study will provide an integrative molecular map of human myocardial infarction, represents an essential reference for the field, and paves the way for advanced mechanistic and therapeutic studies of cardiac disease.

In **Chapter 7** we describe a study that aimed to investigate the origin of adult human kidney organoids and their potential as a model for polycystic kidney disease. In this study we will use adult human kidney organoids originating from CD24+ cells, which are a population of adult renal progenitor cells. The organoids will be analysed for gene expression profile to that of polycystic kidney disease. The study aims to reveal that organoids derived from polycystic kidney disease patients have different gene expression profiles compared to those derived from healthy individuals, providing new insights into the disease. Overall, we aim to demonstrate that adult human kidney organoids represent an advanced model for studying adult polycystic kidney disease and can be used to gain a better understanding of the disease and to identify potential therapeutic targets.

Finally, in **Chapter 8**, the main conclusions of this thesis will be discussed considering the broader context and future perspectives are provided.

References

Basile, David P., Melissa D. Anderson, and Timothy A. Sutton. 2012. “Pathophysiology of Acute Kidney Injury.” *Comprehensive Physiology* 2 (2): 1303–53.

Bergmann, Martin W. 2010. “WNT Signaling in Adult Cardiac Hypertrophy and Remodeling: Lessons Learned from Cardiac Development.” *Circulation Research* 107 (10): 1198–1208.

Coelho, Silvia, Guadalupe Cabral, José A. Lopes, and António Jacinto. 2018. “Renal Regeneration after Acute Kidney Injury.” *Nephrology* 23 (9): 805–14.

Duffield, Jeremy S. 2014. “Cellular and Molecular Mechanisms in Kidney Fibrosis.” *The Journal of Clinical Investigation* 124 (6): 2299–2306.

Frangogiannis, Nikolaos G. 2015. “Pathophysiology of Myocardial Infarction.” *Comprehensive Physiology* 5 (4): 1841–75.

Frank, R. S., T. S. Frank, G. B. Zelenock, and L. G. D’Alecy. 1993. “Ischemia with Intermittent Reperfusion Reduces Functional and Morphologic Damage Following Renal Ischemia in the Rat.” *Annals of Vascular Surgery* 7 (2): 150–55.

GBD 2019 Diseases and Injuries Collaborators. 2020. “Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019.” *The Lancet* 396 (10258): 1204–22.

Hewitson, Tim D. 2009. “Renal Tubulointerstitial Fibrosis: Common but Never Simple.” *American Journal of Physiology. Renal Physiology* 296 (6): F1239–44.

Humphreys, Benjamin D., Shuei-Liong Lin, Akio Kobayashi, Thomas E. Hudson, Brian T. Nowlin, Joseph V. Bonventre, M. Todd Valerius, Andrew P. McMahon, and Jeremy S. Duffield. 2010. “Fate Tracing Reveals the Pericyte and Not Epithelial Origin of Myofibroblasts in Kidney Fibrosis.” *The American Journal of Pathology* 176 (1): 85–97.

Jayaraj, Joshua Chadwick, Karapet Davatyan, S. S. Subramanian, and Jemmi Priya. 2019. “Epidemiology of Myocardial Infarction.” *Myocardial Infarction* 10. https://books.google.com/books?hl=en&lr=&id=HG-QDwAAQBAJ&oi=fnd&pg=PA9&dq=myocardial+infarction+heart+failure+epidemiology+worldwide&ots=6SpUUcQ0GQ&sig=v2uiM8ulG300CrT-3h5wOc7peC8.

Kramann, Rafael, Derek P. DiRocco, and Benjamin D. Humphreys. 2013. “Understanding the Origin, Activation and Regulation of Matrix-Producing Myofibroblasts for Treatment of Fibrotic Disease.” *The Journal of Pathology* 231 (3): 273–89.

Kriz, W., I. Hartmann, and H. Hosser. 2001. “Tracer Studies in the Rat Demonstrate Misdirected Filtration and Peritubular Filtrate Spreading in Nephrons with Segmental Glomerulosclerosis.” *Journal of the*. https://jasn.asnjournals.org/content/12/3/496.short.

Levey, Andrew S., and Josef Coresh. 2012. “Chronic Kidney Disease.” *The Lancet* 379 (9811): 165–80.

Lovisa, Sara, Michael Zeisberg, and Raghu Kalluri. 2016. “Partial Epithelial-to-Mesenchymal Transition and Other New Mechanisms of Kidney Fibrosis.” *Trends in Endocrinology and Metabolism: TEM* 27 (10): 681–95.

Piera-Velazquez, Sonsoles, Zhaodong Li, and Sergio A. Jimenez. 2011. “Role of Endothelial-Mesenchymal Transition (EndoMT) in the Pathogenesis of Fibrotic Disorders.” *The American Journal of Pathology* 179 (3): 1074–80.

Romagnani, Paola, Giuseppe Remuzzi, Richard Glassock, Adeera Levin, Kitty J. Jager, Marcello Tonelli, Ziad Massy, Christoph Wanner, and Hans-Joachim Anders. 2017. “Chronic Kidney Disease.” *Nature Reviews. Disease Primers* 3 (November): 17088.

Ruggenenti, Piero, Paolo Cravedi, and Giuseppe Remuzzi. 2012. “Mechanisms and Treatment of CKD.” *Journal of the American Society of Nephrology: JASN* 23 (12): 1917–28.

Venkatachalam, Manjeri A., Joel M. Weinberg, Wilhelm Kriz, and Anil K. Bidani. 2015. “Failed Tubule Recovery, AKI-CKD Transition, and Kidney Disease Progression.” *Journal of the American Society of Nephrology: JASN* 26 (8): 1765–76.

Zeisberg, Michael, Jun-Ichi Hanai, Hikaru Sugimoto, Tadanori Mammoto, David Charytan, Frank Strutz, and Raghu Kalluri. 2003. “BMP-7 Counteracts TGF-beta1-Induced Epithelial-to-Mesenchymal Transition and Reverses Chronic Renal Injury.” *Nature Medicine* 9 (7): 964–68.

Zeisberg, Michael, and Eric G. Neilson. 2010. “Mechanisms of Tubulointerstitial Fibrosis.” *Journal of the American Society of Nephrology: JASN* 21 (11): 1819–34.

Zoccali, Carmine, Raymond Vanholder, Ziad A. Massy, Alberto Ortiz, Pantelis Sarafidis, Friedo W. Dekker, Danilo Fliser, et al. 2017. “The Systemic Nature of CKD.” *Nature Reviews. Nephrology* 13 (6): 344–58.

Affiliations

[1]Division of Nephrology, RWTH Aachen University, Aachen, Germany,
[2]Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands,
[3]Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University, Heidelberg, Germany,
[4]Joint Research Center for Computational Biomedicine, RWTH Aachen University Hospital, Aachen, Germany and
[5]Molecular Medicine Partnership Unit, European Molecular Biology Laboratory and Heidelberg University, Heidelberg, Germany

2

Experimental and computational technologies to dissect the kidney at the single-cell level

Christoph Kuppe[1,2], Javier Perales-Patón[1,3,4], Julio Saez-Rodriguez [3,4,5] and Rafael Kramann[1,2]

Abstract

The field of single-cell technologies, in particular single-cell genomics with transcriptomics and epigenomics, and most recently single-cell proteomics, is rapidly growing and holds promise to advance our understanding of organ homoeostasis and disease and facilitate the identification of novel therapeutic targets and biomarkers. This review offers an introduction to these technologies. In addition, as the size and complexity of the data require sophisticated computational methods for analysis and interpretation, we will also provide an overview of these methods and summarize the single-cell literature specifically pertaining to the kidney.

Introduction

Next-generation sequencing (NGS) has enabled highthroughput measurements of DNA and RNA, and revolutionized biomedical research over the last decade. Performing DNA or RNA sequencing from the entire sample, often referred to as bulk sequencing, is an approach that has prompted various major breakthroughs in understanding disease mechanisms, identifying novel diagnostics and promoting target discovery.

However, bulk sequencing techniques have two major drawbacks. Since they measure biomolecules across an entire sample, they contain data originating from many different cells likely belonging to a variety of cell types. Another major drawback is that cellular resolution is entirely lost in these bulk genomic measurements, especially in tissues where cells, even those right next to each other, can have a distinct transcriptomic signature. Thus, some genes might be expressed at a very high level only in certain cells, e.g. inflammatory cells that home to a tissue after injury. In the bulk measurements, these genes are detected as upregulated without the ability to further determine from which cell this information originated.

NGS-based technologies for genomics, transcriptomics and epigenomics have been developed recently on the single-cell level, allowing us to study thousands of genes in any given single cell. Furthermore, spatial genomic technologies have emerged that allow us to study gene expression without losing spatial information within the tissue. All these technologies require specific computational pipelines and complex algorithms to mine the huge datasets they produce. International efforts that bring together different scientific disciplines, like the Human Cell Atlas, aim to create a comprehensive map with a cellular resolution of all human cells, including those in the human kidney[1]. The mapping of all different cell types in the human kidney, and specifically where they are located in health, will help us acquire a better understanding of kidney homoeostasis. Ultimately, the goal is to identify the complex cellular interactions and networks in disease that will enable the development of novel targeted therapies.

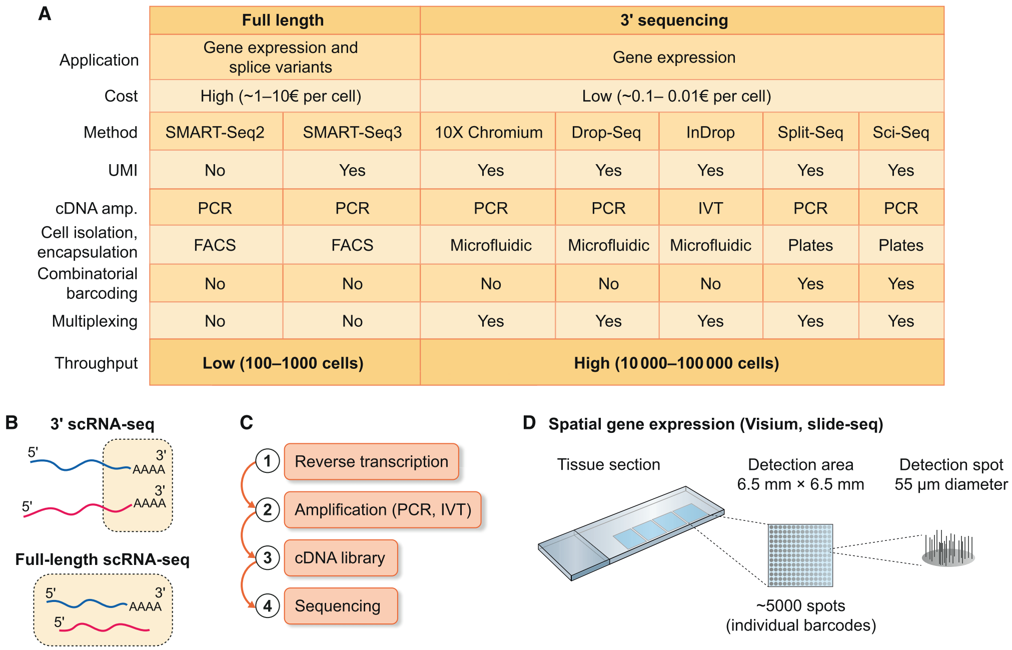
These advances have generated a critical rethinking of the concept of cell-type definition. Canonical cell types, like proximal tubular cells or podocytes in the kidney, were originally defined by their function in the tissue and characteristics such as their unique morphology. However, investigations using single-cell RNA-sequencing (scRNA-seq) data have demonstrated considerable variation in gene expression within defined cell types, thus blurring the lines between cell-type distinctions. A first recent example in the kidney domain is the discovery of a transitional cell state between principal cells and intercalating cells in the collecting duct of mice[2].

Various reviews have recently covered scRNA-seq in the kidney, offering a broad introduction to the field and presenting the most common approaches[3-11]. To complement these excellent reviews, we now focus on topics that have not yet been extensively discussed in the recent kidney literature, in particular by addressing computational approaches, spatial transcriptomics and measurement of DNA accessibility. We also discuss their promising potential for kidney research to identify disease-driving mechanisms, biomarkers and novel therapeutic targets.

Introduduction to sc-RNA-seq

Methods that allow the analysis of the genome-wide transcriptome of individual single cells are referred to as scRNA-seq methods. scRNA-seq is an approach that maps cell states of heterogeneous samples and has become increasingly popular across many biological and biomedical fields. Depending on the scRNA-seq platform used, the costs may range from cents per cell to euros per cell, mostly determined by throughput of the assays and subsequent sequencing costs. The first critical step in all scRNA-seq protocols is the isolation of single cells, which still represents a challenge in complex solid tissues like the kidney. Single nuclei can also be utilized for the so-called single nucleus RNA-sequencing (snRNA-seq), with the advantage that frozen tissue can be used. Depending on the starting material (fresh versus frozen tissue), protocols use enzymatic digestion combined with mechanical disruption or isolation of nuclei from fresh or frozen tissue by bounce homogenization similar to the generation of nuclear lysates for western blots. Both methods have advantages and disadvantages, as pointed out recently[12,13], including potential activation of stress responses using heat-activated enzymes in fresh tissue and the lack of immune cell heterogeneity in single nuclei preparations. Enzymatic digestion protocols using cold-active enzymes have been developed and have shown a decreased activation of stress response-related genes[14,15]. However, for fibrotic tissue or less accessible regions (e.g. glomeruli), this approach might have reached its limitations. Enrichment for distinct cell types by antibody staining or genetic tagging in mouse tissue followed by sorting fluorescence-activated cell sorting (FACS) from fresh tissue can help to remove other cell types and thus obtain a higher resolution for the cells of interest. However, this enrichment is more challenging in nuclei, yet recently developed strategies like Probe-Seq[16] allow the isolation of distinct nuclei from fresh or frozen tissue using *in situ* hybridization for cell type-specific mRNAs followed by FACS. scRNA-seq data from isolated nuclei tend to have a higher background signal due to ambient RNA, and the RNA contained in the nucleus is mostly unspliced. Therefore, the methods for alignment have to be adjusted to include exonic and intronic reads.

The first step of most scRNA-seq protocols (3′) starts with polyA-mRNA capturing using oligo-dT primers and subsequent first-strand cDNA synthesis using an Moloney Murine Leukemia Virus reverse transcriptase (see schematic in Figure 1). Following the second-strand synthesis, which mostly relies on a template-switching oligo reaction,



A

	Full length		3' sequencing				
Application	Gene expression and splice variants		Gene expression				
Cost	High (~1–10€ per cell)		Low (~0.1– 0.01€ per cell)				
Method	SMART-Seq2	SMART-Seq3	10X Chromium	Drop-Seq	InDrop	Split-Seq	Sci-Seq
UMI	No	Yes	Yes	Yes	Yes	Yes	Yes
cDNA amp.	PCR	PCR	PCR	PCR	IVT	PCR	PCR
Cell isolation, encapsulation	FACS	FACS	Microfluidic	Microfluidic	Microfluidic	Plates	Plates
Combinatorial barcoding	No	No	No	No	No	Yes	Yes
Multiplexing	No	No	Yes	Yes	Yes	Yes	Yes
Throughput	**Low (100–1000 cells)**		**High (10 000–100 000 cells)**				

Figure 1: Common scRNA-seq methods.
(**A**) Overview of scRNA-seq methods and characteristics. (**B**) Schematic illustrating the difference between 3′ versus full-length scRNA-seq methods. (**C**) Main protocol steps used for scRNA-seq. (**D**) Schematic illustrating principles of spatial gene expression. In brief, a tissue section is placed on a detection area. The detection area comprises about 5000 detection spots that are individually DNA-barcoded for polyA-mRNA capture. The transcripts detected can be later assigned to their corresponding detection spot on the slide, thus resolving their spatial organization.

the cDNA is amplified using polymerase chain reaction. Using standard NGS library preparation, sequence-ready individually indexed libraries are synthesized. Recent novel protocols that combine full-length transcriptome coverage with a 5′-unique molecular identifier significantly increase the sensitivity approaching the sensitivity of single-molecular RNA-fluorescence *in situ* hybridization (FISH) and will help to overcome the data sparsity (increasing the transcript detection per cell) in scRNA-seq data in the near future.

Single-cell epigenome/chromatin organization (ATAC-Seq)

Chromatin accessibility is crucial for gene expression regulation and cell fate in development, homoeostasis and disease. Regulation of gene expression is a dynamic interaction between chromatin structure regulating chromatin accessibility and recruitment of transcription factors to promoter regions, enhancers and activator sequences[17]. Accessible chromatin regions can be determined by DNase-seq or Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq)[18]. ATAC-seq uses a mutated Tn5 transposase that identifies open chromatin regions and inserts sequencing adapters in the probed DNA region.

Recently, ATAC-seq has been developed on a single-cell level[19,20]. The investigation of the epigenomic landscape by scATAC-seq holds great promise to uncover the heterogeneity in gene regulatory programmes between cells[21]. In a landmark study, Cusanovich *et al*. performed scATAC-seq in 13 adult mouse tissues, including from the kidney, providing a first atlas of chromatin accessibility in a variety of identified renal cell types[22].

Importantly, scATAC data are even more sparse than scRNA- seq data; therefore, the analysis methods need to take this into account and measure accessibility across groups of cells or across sets of genomic features. Thus, novel computational tools are needed that address the issue of extreme sparsity. scOpen might be such a tool that uses positive-unlabelled learning of matrices and estimates the probability that a region is open in a given cell[23]. We have recently performed scATAC sequencing of mouse kidney in homoeostasis and at different stages of fibrosis using the unilateral ureteral obstruction model[23]. Using scOpen, we could report gene regulatory programmes associated with all major murine kidney cell types in homoeostasis and fibrosis[23]. Interestingly, we detected various shared regulatory programmes in different kidney cell types such as myofibroblasts and tubule epithelium that might drive their expansion and or de/differentiation[23]. Several studies that apply scATAC-seq in human and mouse kidneys are ongoing and will certainly shed light on the heterogeneity of chromatin accessibility and gene-regulatory programmes in renal cell types during homoeostasis and disease.

Spatial Genomic Technologies

Single-cell RNA-seq and ATAC-seq data sets are generated after dissociation of the tissue and thus spatial and morphological information of the cells within the tissue environment is lost. This information can be critical to understanding the role of different cells in tissue functioning, in particular for a complex organ as spatially structured as the kidney.

The field of spatial transcriptomics has recently emerged utilizing technologies that can identify the location of several dozens to thousands of mRNAs in intact tissue slices. The detection f mRNA in tissues by FISH is widely used but has been limited to only a few mRNAs in parallel in the past. However, there have been several recent advances to increase the number of mRNAs that can be imaged in a given tissue by stripping and rehybridization and/or by combinatorial barcoding[24-26]. multiplexed error-robust fluorescence *in situ* hybridization was able to image up to several hundreds of transcripts[27] and more recently seqFISHþ achieved super-resolution imaging with detection of up to 10 000 genes in a single cell[28]. Other approaches include technologies that encode RNA-seq with barcoded oligonucleotide capture arrays on a slide such as the so-called spatial transcriptomics[29], which has been commercialized and improved in the Visium platform (10x genomics). Visium allows the capture of several thousands of genes (up to 10 000) in a given tissue area using barcoded spots that are 55 mm wide. Five thousand such spots are arranged in an imaging area of 6.5 x 6.5 mm. The technology is relatively easy to use in virtually all cryopreserved tissue but does not reach the single-cell level given the diameter of the individual spots as well as the distance to the next spot (100 mm centre to centre). Slide-seq is a similar technique that uses uniquely barcoded microbeads that are 10 mm wide and bound to a rubber-coated glass coverslip in a monolayer reaching a single-cell resolution in >60% of the beads applied to brain tissue[30]. One current issue with increasing the resolution is that this comes with a decreased depth of gained information, for example, number of genes detected. To increase the resolution, the options are either to reduce the size of the barcoded spots/beads or to increase the size of the tissue, which could potentially be achieved chemically while preserving the mRNA. Alternatively, one can use scRNA and known *in situ* information of certain marker genes to computationally and virtually assign cells to a location[31-34].

So far, no high-throughput spatial transcriptomic work has been published on the kidney that holds promise to give novel insight into kidney homoeostasis and disease, especially when used in combination with some of the single-cell technologies described here. Integration could potentially be used to increase virtually the resolution to the real single-cell level.

Single-cell Proteomics

Beyond genomics, single-cell proteomics technologies have also developed rapidly. Mass spectrometry technologies enable a high (yet not complete) coverage of the proteome[35]. The proteome can be a very informative omic between the transcriptome and the phenotype[36]. Obtaining single-cell resolution is technically challenging. Among other reasons, there is no method that allows amplification of proteins, in contrast to DNA and RNA. However, the amount of material required by mass spectrometry is continuously decreasing, such that we can envision that it will soon be possible to measure high-coverage proteomics in single cells[37]. Complementarily, antibodybased methods can measure single cells, but measured proteins are limited to those for which antibodies are available and with limited multiplexing capabilities. Flow cytometry has been available for quite some time, and recent technologies such as mass cytometry allow us to increase the multiplexing to several dozens of proteins at a time. This is achieved by marking the antibodies with rare metals and subsequently ionizing the sample and measuring the metals in a mass spectrometer[38]. Antibody-based strategies like CITE-Seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) can also be combined with sequencing-based technologies, enabling the simultaneous measurement of proteins and transcripts at the single- cell resolution[39]. Furthermore, antibody-based methods can be applied to obtain spatially resolved data. This is done mainly by using two approaches: by adapting the mass cytometry approach to tissues[40-41] or by using microscopes. For the latter, repeating cycles of staining and 'washing' of the antibodies enables an increase of coverage[42-43] akin to the processes for RNA described above. The mass cytometry-based approach was recently applied to characterize different cell types in healthy and diseased human kidneys[44]. These single-cell proteomic technologies have only very recently been developed, and we expect them to be more broadly applied to the kidney in the near future, providing complementary information to other single-cell data.

Data Analysis

All of the new technologies discussed above provide large and complex data sets. In contrast to molecular data derived from classical high-throughput technologies on bulk samples, single- cell genomics data present novel sources of biological and technical variability[45]. The low amount of starting material from a single-cell hampers the sensitivity for the genome-wide profiling, resulting in low gene coverage and sporadic drop out measurements (data sparsity)[46]. Moreover, the cell-to-cell heterogeneity of individualized cell profiling at large scale provides a landscape of biological heterogeneity without precedent in molecular data analysis. Cellular states and transition phases are embedded in a continuous space with the main variability that stems from the differentiated cell types[47]. These aspects have motivated the development of computational methods specific to single-cell analysis.

Computational methods for single-cell genomics aim to delineate the relevant biological information from the aforementioned confounding factors. As these data types are very recent and technologies still in development, the corresponding computational tools are under very active development. The ratio that single-cell-specific methods are being developed is roughly proportional to the number of studies generating these new data types and is particularly prominent in the field of scRNA-seq[48]. The fact that the bioinformatics community embraces open-source release of code and preprints boosts the development and adoption of these methods. At the same time, it poses challenges in keeping up with the most suitable methods of each family at the

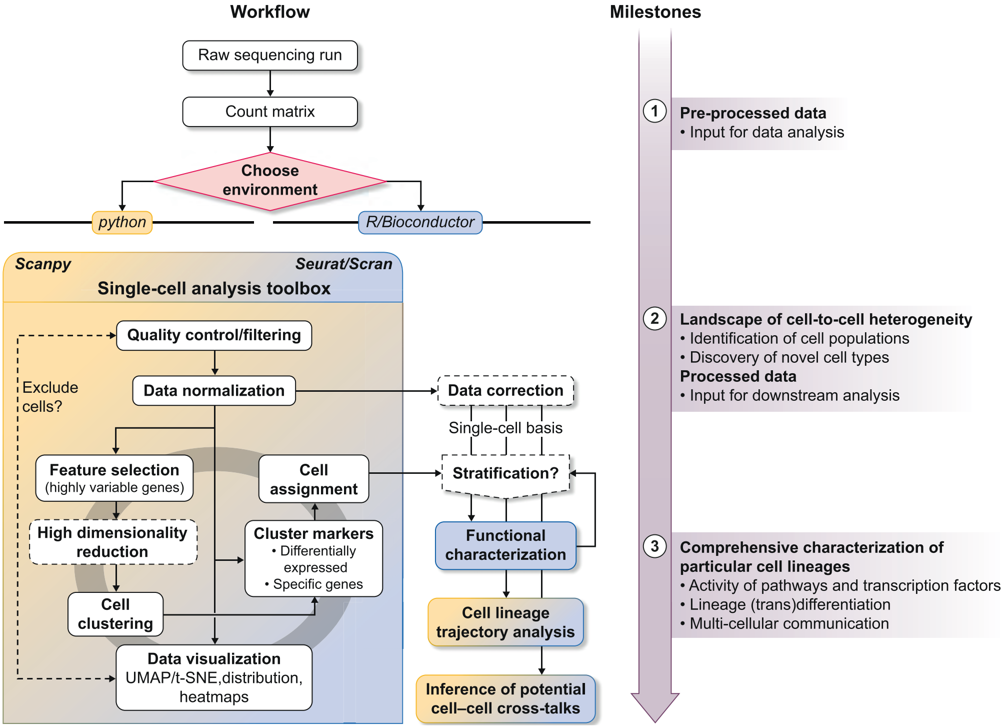

Workflow
Milestones
Raw sequencing run
Count matrix
Choose environment
python
R/Bioconductor
Scanpy
Seurat/Scran
Single-cell analysis toolbox
Quality control/filtering
Exclude cells?
Data normalization
Data correction
Single-cell basis
Feature selection (highly variable genes)
Cell assignment
Stratification?
High dimensionality reduction
Cluster markers
• Differentially expressed
• Specific genes
Functional characterization
Cell clustering
Cell lineage trajectory analysis
Data visualization
UMAP/t-SNE,distribution, heatmaps
Inference of potential cell–cell cross-talks
1
Pre-processed data
• Input for data analysis
2
Landscape of cell-to-cell heterogeneity
• Identification of cell populations
• Discovery of novel cell types
Processed data
• Input for downstream analysis
3
Comprehensive characterization of particular cell lineages
• Activity of pathways and transcription factors
• Lineage (trans)differentiation
• Multi-cellular communication

Figure 2: Workflow and milestones of the single-cell data RNA-seq analysis.
Pre-processing raw sequencing data are the input required for the analysis (Milestone 1). The choice of the analytical programing environment will determine the available methods (see Table 1 for a detailed list). The use of a well-established toolbox for single-cell analysis supports the performance of a conventional standard analysis (Milestone 2). The analysis is driven by the biological interpretation, and it might become laborious with several refining loops prior to downstream analysis. Depending on initial and data-driven hypotheses, a comprehensive characterization focussed on particular cell types might complete the full characterization of the cell populations present in the sample (Milestone 3). Discontinuous lines indicate optional steps.

time these are being released. Special issues and articles dedicated to benchmarking the different families of methods will help researchers to navigate and choose between the alternatives. These benchmark exercises are important to understanding the applicability and limitations of the methods, either adapted from bulk data or bespoke to single cells[49,50]. A method can 'run' on the data but this does not mean that the results produced are reliable or meaningful.

The majority of these tools are embedded in the most popular analytical programming environment in data science, Python and R. Well-established bioinformatics frameworks include Scanpy (Python[51]), Seurat (R[52]) and bioconductor packages such as Scran and Scater (R[53]). These toolboxes propose an integrative methodology following a pipeline of analyses that covers a low-level conventional single-cell data analysis, including the transcriptomics (scRNA-seq), epigenomics (ATAC-seq) and spatial data.

As an illustration, we here describe a methodology for a standard analysis of scRNA-seq data (Figure 2 and Table 1). It would be similarly applicable to other single-cell omics with quantitative data, although the preprocessing of the raw sequencing data might differ significantly. The analysis involves data pre-processing (quantification of gene expression in single cells, quality control, data normalization and data correction), cell clustering and cell identity assignment, and visualization of the distinct cell populations driven by the use of the single-cell toolbox[51-53]. Depending on the biological hypothesis, downstream analysis might focus on further characterization of certain cell populations. If applicable, the trajectory inference of cell lineages with deregulated genes along pseudotime could be investigated. Current best practices are described in Luecken and Theis[54]. Downstream analyses might include the functional characterization of the cell population in terms of the activity of pathways and transcription factors[49]—or other transcriptome programmes of interest (e.g. cellular states related to a biological condition)—and the inference of potential cell–cell communication crosstalk between cell populations[55,56].

General Advice

- For reproducibility and transparency, it is important that the raw and processed data be publicly available, along with the code used to process it.
- The manuscript should explain the different methods and software used, the values chosen for the corresponding parameters and the assumptions made.
- Data are typically noisy and incomplete, and findings obtained from downstream analysis such as identification of key pathways should be considered in general hypotheses to be validated rather than the findings by themselves.
- Doublets are inherent to droplet-generated scRNA-seq data sets and beyond computational doublet detection tools additional experimental validation (e.g. *in situ* multiplex mRNA hybridization) is needed.
- Batch effects have to be carefully taken into account when analysing integrated multiple scRNA-seq data sets, and again, findings need to be further validated. Multiplexing of several conditions or samples could be used to reduce batch effects *a priori*.
- Confirmation experiments of scRNA-seq data can include a second technology to identify cell types/states, e.g. scATAC sequencing and also spatial information such as *in situ* hybridization or spatial gene expression data. Furthermore, protein expression

data by multiplexed immunostaining can be used. To validate mechanisms, cell-culture overexpression and knockdown/knockout experiments in either monolayer, co-culture or organoids or *in vivo* experiments in rodents are state-of-the-art approaches.

Single cell studies in the kidney

Although the single-cell field is quite young, there is already a large number of studies that have used single-cell technologies (all mostly restricted to scRNA-seq) in the human and mouse kidney as well as in kidney organoids (Table 2). Discussing all these studies would be beyond the scope of this review, but we would like to highlight the results of a few. The first landmark whole mouse kidney scRNA-seq dataset was generated by the Susztak lab[2]. They identified overall 21 major kidney cell types and demonstrated that Mendelian disease genes show cell specificity. Interestingly, they also discovered a new transitional cell type in the collecting duct of adult mice and validated these data by genetic fate tracing and further demonstrated a role of the Notch signaling pathway in collecting duct cell plasticity[2]. The Humphrey's lab published several important articles including scRNA-seq of a human kidney allograft biopsy[71], comparison of scRNA-seq with single-nucleus snRNA-seq in the kidney[12], first human diabetic nephropathy data from tumour nephrectomy samples[72] and a comparison of induced pluripotent stem cell (iPSC)-derived kidney organoids with human kidney data[73]. This important work has shed light on a heterogeneous immune response in mixed rejection[71]. It has also demonstrated that snRNA-seq can be utilized to analyse kidney tissue and might have some advantages, including the use of biobanked frozen specimens[12]. The diabetic nephropathy data suggest that epithelial cells of the thick ascending limb, late distal convoluted tubule and collecting duct adopt a gene expression profile that is consistent with increased potassium secretion, among various other interesting findings[72].

Several other labs have also reported very interesting single-cell datasets on either human or mouse kidney specimens. Karaiskos *et al.* reported a single-cell atlas of the mouse glomerulus using fixed cells[74]. Young *et al.* have sequenced single cells from different human kidney cancer types as well as adult and foetal healthy kidney tissue[75]. Their data give important insight into human kidney cancer heterogeneity and novel cell types that might be the origin of the cancer, including aberrant foetal cells as the origin of Wilms tumour and a subtype of the proximal convoluted tubule in clear cell renal cell carcinoma[75]. Lake *et al.* provide a map of the human kidney using snRNA-seq from tumour nephrectomies and discarded deceased donor specimens describing 30 distinct cell populations[76]. The McMahon lab has performed scRNA-seq of the human nephrogenic niche and has given novel insights into the early patterning processes and

Table1. Selected computational tools for scRNA-seqanalysis

	Data pre-processing	
🟩	CellRanger	Private software for the pre-processing of the popular 10× genomics chromium raw scRNA-seq data for the quantification of UMI read counts
🟩	Kallisto BUS	Fast pseudo-alignment quantification of UMI read counts for scRNA-seq [57]
	Single-cell toolbox	
🟦	Seurat	Bioinformatics toolbox for the quality control, normalization and exploration of single-cell data [52]
🟨	Scanpy	Bioinformatics toolbox for the scalable quality control, normalization and exploration oflarge single-cell data sets [51]
🟦	Bioconductor packages[a]	Bioconductor packages such as scran and scater for the low-level standard scRNA-seq analysis. See [53] for a review
	Cell marker extraction	
	Statistical contrasts[a]	Test for differential gene expression such as Wilcoxon rank-sum test, t-test, MAST [58] or DESeq2 [59] as stand-alone or their wrappers in single-cell toolbox
🟦	GenesorteR	Tool to rank marker genes by their specificity scores and conditional expression probabilities from each cell cluster [60]
	Data visualization	
	Visualization tools[a]	High dimensionality reduction methods to embed cells into two dimensions to investigate biological patterns in the local and global structure of the data such as t-Distributed Stochastic Neighbour Embedding and Uniform Manifold Approximation and Projection. Violin plot and heat map to show cell-to-cell heterogeneity of selected features
🟦	SWNE (Similarity Weighted Nonnegative Embedding)	Method for the visualization of cells and main biological factors in a 2-dimensional space that captures discrete cell types and continuous lineage trajectories. SWNE combines non-negative matrix factorization for dimensionality reduction, with shared nearest neighbour networks to smooth the matrix decomposition of cells by their similarity in the high-dimensional space [61]
	Data correction	
🟦	Combat	Correction method based on a linear model to regress out covariates (e.g. batch effect) from the gene expression taking into account both the mean and the variance [62]
	Imputation tools[a]	Family of methods under development for the imputation of gene expression over drop outs, which are under evaluation in [63]
	Functional characterization	
🟦	DoRothEA	Knowledge-based transcription factor regulons for the inference of their activities [64]
🟦	PROGENy	Footprint-based method for the inference of pathway activities [65]
🟦	SCENIC	Data-driven workflow for Gene Regulatory Network reconstruction and inference of transcription factor activities [56]
🟦	AUCell	Gene set enrichment method specific to single-cell data [66]
	Lineage trajectory	
🟦	Monocle	Toolkit for time-series pseudotime single-cell analysis: differential expression, clustering and trajectory with cell fate branch analyses [67]
🟨	PAGA	Partition-based graph abstraction tool for complex data sets [68]
🟦	Slingshot	Tree-based method for the lineage trajectory inference [69]
	Cell–cell communication	
🟨	CellPhoneDB	Database of curated ligand–receptor interaction coupled with a simple algorithm to select potential cell–cell interactions based on the actual expression ofligand–receptor pairs [70]
🟦	NicheNetR	Trained model ofligand–receptor relationships linked to downstream signaling interactions to estimate the most likely communications between two cell populations (sender and receiver cells) based on the actual expression ofligand–receptor pairs and downstream footprint of said interaction [56]

[a]Denotes broad categories of methods that are implemented in the single-cell toolbox. UMI, unique molecular identifier.

The name of the tool with a short description is shown. Colours on the left indicate the environment the tool is wrapped in: Python¼ yellow, R¼ blue and bioinformatics pipeline¼ green.

developmental trajectories of nephron progenitor cells in forming a human nephron[77]. Additional work from the same group has provided the most detailed single-cell atlas of the mouse kidney using specimens from cortex, medulla and papilla providing novel data on sex differences as well as distinct cell composition of nephrons dependent on the time of nephron specification and lineage convergence[14]. Dumas *et al.* recently provided a mouse atlas of endothelial kidney cells describing 24 different subpopulations[78]. They further reported that endothelial cells upregulate genes involved in hypoxia response and oxidative phosphorylation in response to dehydration and hypertonicity[78]. A recent article from the Clatworthy group has performed scRNA-seq of adult and foetal human kidneys with a special focus on spatiotemporal immune topology[79]. They have identified anatomically defined expression patterns of immune genes within epithelial compartments and describe expression of antimicrobial peptide transcripts in the pelvic

Table 2. Overview of single-cell studies in kidney and or organoids to date [2, 12, 14, 71-106]

	References	Date	Number of cells or nuclei	Human	Mouse	Organoid	sc-seq
1	Conway [103]	May 2020	> 30 000		🟦		🟪
2	Menon [104]	May 2020	> 100 000	🟧			🟩
3	Kalucka et al. [83]	February 2020	> 30 000		🟦		🟩
4	do Valle Duraeset al. [84]	February 2020	> 50 000		🟦		🟩
5	Liao et al. [85]	January 2020	23 366	🟧			🟩
6	Dumas et al. [78]	January 2020	> 40 000		🟦		🟩
7	Barry et al. [86]	December 2019	5936		🟦		▨
8	Ransick et al. [14]	November 2019	31 265		🟦		🟩
9	Subramanianet al. [105]	November 2019	450 118			🟨	🟩
10	Wilson et al. [72]	September 2019	23 980	🟧			🟩
11	Stewartet al. [79]	September 2019	27 203	🟧			🟩
12	Lowet al. [87]	September 2019	62 506			🟨	🟩
13	Der et al. [80]	July 2019	6041	🟧			🟪
14	Lakeet al. [76]	June 2019	17 659	🟧			⬛
15	Arazi et al. [89]	June 2019	2881	🟧			🟨
16	Combeset al. [90]	June 2019	6732		🟦		
17	Fu et al. [91]	April 2019	644		🟦		🟪
18	Schutgenset al. [92]	March 2019	192			🟨	🟨
19	Harder et al. [93]	January 2019	12 000^a			🟨	⬛
20	Combeset al. [81]	January 2019	> 8000	🟧		🟨	🟩
21	Wu et al. [10]	January 2019	11 391	🟧		🟨	🟩
22	Wu et al. [71]	December 2018	83 130	🟧		🟨	🟩
23	Tabula Muris Consortiumet al. [94]	October 2018	800^a		🟦		🟨
24	Wanget al. [95]	September 2018	3543	🟧			
25	Cao et al. [96]	September 2018	11 296		🟦		
26	Younget al. [75]	August 2018	72 051	🟧			🟩
27	Karaiskos et al. [74]	August 2018	12 954		🟦		⬛
28	Wu et al. [55]	August 2018	8746	🟧			🟥
29	Menonet al. [97]	August 2018	6141	🟧	🟦		
30	Gillies et al. [98]	August 2018	4743	🟧			🟩
31	Lindström et al. [77]	June 2018	3367	🟧		🟨	🟩
32	Czerniecki et al. [82]	June 2018	10 535		🟦		⬛
33	Park et al. [2]	May 2018	57 979		🟦		🟩
34	Kramann et al. [99]	May 2018	194	🟧	🟦		🟪
35	Lindström et al. [100]	March 2018	2800				🟩
36	Sivakamasundariet al. [105]	December 2017	22 469				🟩
37	Chenet al. [102]	November 2017	218	🟧	🟦		🟪
38	Pode-Shakkedet al. [106]	June 2017	80	🟧			🟪
39	Der et al. [88]	May 2017	899	🟧	🟦	🟨	🟪

^{a}Estimated number of cells.

🟩	🟪	🟨	⬛	🟥	🟧	▨	⬜
10X Genomics	SMART-Seq1	Cel-Seq2	Drop-Seq	InDrop	STRT-Seq	ddSeq Biorad	Sci-CAR

epithelium in adult but not in foetal kidneys. These data give novel insight into immune cell heterogeneity of the human kidney at unprecedented resolution[79].

Der *et al.* has performed scRNA-seq of renal biopsies and skin specimens of lupus nephritis patients[80]. Their analysis suggests that a high interferon Type I response signature in keratinocytes or tubular epithelium distinguishes lupus nephritis patients from healthy control patients. Interestingly, a high interferon Type I response signature was also associated with a failure to respond to treatment[80].

Several groups have performed scRNA-seq experiments in kidney organoids and partly compared this data to adult or developmental human kidney specimens. Wu *et al.* have compared iPSC and human embryonic stem cell-derived kidney organoids to adult and foetal human kidney by scRNA-seq[72]. Their analysis indicates that the organoid cell types in the protocols used correlate better with foetal human kidneys as compared with adult human kidneys and, thus, represent developmental kidney stages. Furthermore, the data suggest that the differentiation protocols still lead to around 10% or more non-renal, off-target cell types[72]. Combes *et al.* have also performed scRNA-seq of iPSC-derived kidney organoids demonstrating similarities to foetal human kidneys[81]. In general, the differentiation protocol used highly influences the cell-type composition in the kidney organoid (e.g. distal nephron versus glomerular development) and thus should be taken into account.

Czernicki *et al.* have demonstrated a high-throughput screening platform for kidney organoid screening and differentiation and performed scRNA-seq from some organoids demonstrating an effect of vascular endothelial growth factor on endothelial cell differentiation and abundance[82]. Subramanian *et al.* performed scRNA-seq of four human iPSC cell line-derived kidney organoids with a total of more than 450 000 cells sequenced[105]. Interestingly, they report that transplantation of the organoids under the renal capsule diminishes off-target cell types[105].

In summary, several groups have already started to use state-of-the-art single-cell genomics tools to study kidney organoids, development, homoeostasis and disease, while many more studies with larger sample sizes and improved technologies are ongoing. These studies will certainly offer novel insight into kidney function, cell heterogeneity and crosstalk in homoeostasis and disease. This most promising era of novel genomic tools is the perfect time to join these efforts to map kidney physiology and pathophysiology using single-cell genomic approaches.

Future Developments

The rapidly developing field of single-cell genomics allows insights into kidney homoeostasis and disease at unprecedented resolution. Future studies will likely help identify novel disease mechanisms, and therapeutic targets as well as diagnostic and prognostic biomarkers. Single-cell sequencing of urinary cells might serve as a liquid biopsy to help diagnose specific kidney diseases non-invasively. The multiomic analysis of single compartments in the kidney, like single glomeruli or single nephrons, might also add further insight into physiological changes in gene expression and in diseases. Single-cell multiomics are developing rapidly[107], such as the combined measurement of single-cell RNA expression with protein expression and epigenetic changes. These methods provide multiple molecular angles on the cells and will likely accelerate the understanding of cellular pathophysiology and result in the discovery of novel drug targets.

References

1. Regev A, Teichmann SA, Lander ES et al. The human cell atlas. eLife [Internet] 2017; 6: e27041
2. Park J, Shrestha R, Qiu C et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science 2018;360: 758–763
3. Rao DA, Arazi A, Wofsy D et al. Design and application of single-cell RNA sequencing to study kidney immune cells in lupus nephritis. Nat Rev Nephrol 2020;16:238–250
4. Stewart BJ, Ferdinand JR, Clatworthy MR. Using single-cell technologies to map the human immune system—implications for nephrology. Nat Rev Nephrol 2020; 16: 112–128
5. Park J, Chang LL, Kim J et al. Understanding the kidney one cell at a time. Kidney Int 2019; 96: 862–870
6. Wilson PC, Humphreys BD. Single-cell genomics and gene editing: implications for nephrology. Nat Rev Nephrol 2019; 15: 63–64
7. Wu H, Humphreys BD. The promise of single-cell RNA sequencing for kidney disease investigation. Kidney Int 2017; 92: 1334–1342
8. Wilson PC, Humphreys BD. Kidney and organoid single-cell transcriptomics: the end of the beginning. Pediatr Nephrol 2020; 35: 191–197
9. Saez-Rodriguez J, Rinschen MM, Floege J et al. Big science and big data in nephrology. Kidney Int 2019; 95: 1326–1337
10. Malone AF, Wu H, Humphreys BD. Bringing renal biopsy interpretation into the molecular age with single-cell RNA sequencing. Semin Nephrol 2018; 38: 31–39
11. Clark AR, Greka A. The power of one: advances in single-cell genomics in the kidney. Nat Rev Nephrol 2020; 16: 73–74
12. Wu H, Kirita Y, Donnelly EL et al. 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 2019; 30: 23–32
13. O'Sullivan ED, Mylonas KJ, Hughes J et al. Complementary roles for single-nucleus and single-cell RNA sequencing in kidney disease research. J Am Soc Nephrol 2019; 30: 712–713
14. Ransick A, Lindstro¨m NO, Liu J et al. Single-cell profiling reveals sex, lineage, and regional diversity in the mouse kidney. Dev Cell 2019; 51: 399–413.e7
15. Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development 2017; 144: 3625–3632
16. Amamoto R, Garcia MD, West ER et al. Probe-Seq enables transcriptional profiling of specific cell types from heterogeneous tissue by RNA-based isolation. eLife 2019; 8
17. Gottesfeld JM, Carey MF. Introduction to the thematic minireview series: chromatin and transcription. J BiolChem2018; 293: 13775–13777
18. Buenrostro JD, Giresi PG, Zaba LC et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013; 10: 1213–1218
19. Buenrostro JD, Wu B, Litzenburger UM et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015; 523: 486–490
20. Cusanovich DA, Daza R, Adey A et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 2015; 348: 910–914
21. Chen H, Lareau C, Andreani T et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol 2019; 20: 241
22. *Cusanovich DA, Hill AJ, Aghamirzaie D et al. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell 2018; 174: 1309–1324.e18*
23. Li Z, Kuppe C, Cheng M et al. scOpen: chromatin-accessibility estimation of single-cell ATAC data.bioRxiv preprint; doi: 10.1101/865931
24. La Manno G, Gyllborg D, Codeluppi S et al. Molecular diversity of midbrain development in mouse, human, and stem cells. Cell 2016; 167: 566–580.e19

25. Codeluppi S, Borm LE, Zeisel A et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods 2018; 15: 932–935

26. Lubeck E, Coskun AF, Zhiyentayev T et al. Single-cell in situ RNA profiling by sequential hybridization. NatMethods 2014; 11: 360–361

27. Chen KH, Boettiger AN, Moffitt JR et al. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015; 348:aaa6090

28. Eng C-HL, Lawson M, Zhu Q et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature 2019; 568: 235–239

29. Ståhl PL, Salme´n F, Vickovic S et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016; 353: 78–82

30. Rodriques SG, Stickels RR, Goeva A et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019; 363: 1463–1467

31. Achim K, Pettit J-B, Saraiva LR et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol 2015; 33: 503–509

32. Satija R, Farrell JA, Gennert D et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015; 33: 495–502

33. Tanevski J, Nguyen T, Truong B et al. Predicting cellular position in the drosophila embryo from single-cell transcriptomics data. bioRxiv 2019. https://www.biorxiv.org/content/10.1101/796029v1.abstract (10 January 2020, date last accessed)

34. Nitzan M, Karaiskos N, Friedman N et al. Gene expression cartography. Nature 2019; 576: 132–137

35. Hohner M, Frese CK, Grahammer F et al. Single-nephron proteomes connect morphology and function in proteinuric kidney disease. Kidney Int 2018; 93: 1308–1319

36. Rinschen M, Saez-Rodriguez J. The tissue proteome in the multi-omic landscape of kidney disease. Nat Rev Nephrol 2020; doi: 10.1038/s41581-020-00348-5

37. Slavov N. Unpicking the proteome in single cells. Science 2020; 367: 512–513

38. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. Cell 2016; 165: 780–791

39. Stoeckius M, Hafemeister C, Stephenson W et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 2017; 14: 865–868

40. Angelo M, Bendall SC, Finck R et al. Multiplexed ion beam imaging of human breast tumors. NatMed 2014; 20: 436–442

41. Giesen C,Wang HAO, Schapiro D et al. Highlymultiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods 2014; 11: 417–422

42. Rashid R, Gaglia G, Chen Y-A et al. Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer. Sci Data 2019; 6: 323

43. Gut G, Herrmann MD, Pelkmans L. Multiplexed proteinmaps link subcellular organization to cellular states. Science 2018; 361: eaar7042

44. Singh N, Avigan ZM, Kliegel JA et al. Development of a 2-dimensional atlas of the human kidney with imaging mass cytometry. JCI Insight 2019; 4: e129477

45. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015; 16: 133–145

46. Hicks SC, Townes FW, TengMet al. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 2018; 19: 562–578

47. Rostom R, Svensson V, Teichmann SA et al. Computational approaches for interpreting scRNA-seq data. FEBS Lett 2017; 591: 2213–2225

48. Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 2018; 14: e1006245

49. Holland CH, Tanevski J, Perales-Pato´n J et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol 2020; 21: 36

50. Pratapa A, Jalihal AP, Law JN et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods 2020; 17: 147–154

51. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018; 19: 15

52. Butler A, Hoffman P, Smibert P et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018; 36: 411–420

53. Amezquita RA, Lun ATL, Becht E et al. Orchestrating single-cell analysis with Bioconductor. NatMethods 2020; 17: 137–145

54. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 2019; 15: e8746

55. Vento-Tormo R, Efremova M, Botting RA et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature 2018; 563: 347–353

56. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods 2020; 17: 159–162

57. Melsted P, Ntranos V, Pachter L. The barcode, UMI, set format and BUStools. Bioinformatics 2019; 35: 4472–4473

58. Finak G, McDavid A, Yajima M et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015; 16: 278

59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15: 550

60. Ibrahim MM, Kramann R. genesorteR: feature ranking in clustered single cell data. biorixiv preprint 2019; doi: https://doi.org/10.1101/676379 (10 January 2020, date last accessed)

61. Wu Y, Tamayo P, Zhang K. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. Cell Systems 2018; 7: 656–666.e4

62. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118–127

63. Hou W, Ji, JZ Hicks, HSC. A systematic evaluation of single-cell RNA-sequencing imputation methods. bioRxiv preprint 2020; https://doi:.org/10.1101/2020.01.29.925974(2March 2020, date last accessed)

64. Garcia-Alonso L, Holland CH, IbrahimMMet al. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res 2019; 29: 1363–1375

65. Schubert M, Klinger B, Klu¨nemann M et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 2018;9: 20

66. Aibar S, Gonza´lez-Blas CB, Moerman T et al. SCENIC: single-cell regulatory network inference and clustering. NatMethods 2017; 14: 1083–1086

67. Trapnell C, Cacchiarelli D, Grimsby J et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014; 32: 381–386

68. Wolf FA, Hamey FK, Plass M et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol 2019; 20: 59

69. Street K, Risso D, Fletcher RB et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 2018; 19: 477

70. Efremova M, Vento-Tormo M, Teichmann SA et al. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 2020; 15: 1484–1506

71. Wu H, Malone AF, Donnelly EL et al. Single-cell transcriptomics of a human kidney allograft biopsy specimen defines a diverse inflammatory response. J Am Soc Nephrol 2018; 29: 2069–2080

72. Wilson PC,Wu H, Kirita Y et al. The single-cell transcriptomic landscape of early human diabetic nephropathy. Proc Natl Acad Sci USA 2019; 116: 19619–19625

73. Wu H, Uchimura K, Donnelly EL et al. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with singlecell transcriptomics. Cell Stem Cell 2018; 23: 869–881.e8

74. Karaiskos N, Rahmatollahi M, Boltengagen A et al. A single-cell transcriptome atlas of the mouse glomerulus. J Am Soc Nephrol 2018; 29: 2060–2068

75. Young MD, Mitchell TJ, Vieira Braga FA et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 2018; 361: 594–599

76. Lake BB, Chen S, Hoshi M et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. Nat Commun 2019; 10: 2832

77. Lindstro¨m NO, De Sena Brandine G, Tran T et al. Progressive recruitment of mesenchymal progenitors reveals a time-dependent process of cell fate acquisition in mouse and human nephrogenesis. Dev Cell 2018; 45: 651–660.e4

78. Dumas SJ, Meta E, Borri M et al. Single-cell RNA sequencing reveals renal endothelium heterogeneity and metabolic adaptation to water deprivation. J Am Soc Nephrol 2020; 31: 118–138

79. Stewart BJ, Ferdinand JR, Young MD et al. Spatiotemporal immune zonation of the human kidney. Science 2019; 365: 1461–1466

80. Der E, Suryawanshi H, Morozov P et al.; the Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium. Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways. Nat Immunol 2019; 20: 915–927

81. Combes AN, Zappia L, Er PX et al. Single-cell analysis reveals congruence between kidney organoids and human fetal kidney. Genome Med 2019; 11: 3

82. Czerniecki SM, Cruz NM, Harder JL et al. High-throughput screening enhances kidney organoid differentiation from human pluripotent stem cells and enables automated multidimensional phenotyping. Cell Stem Cell 2018; 22: 929–940.e4

83. Kalucka J, de Rooij LPMH, Goveia J et al. Single-cell transcriptome atlas of murine endothelial cells. Cell 2020; 180: 764–779.e20

84. Duraes F do V, do Valle Duraes F, Lafont A et al. Immune cell landscaping reveals a protective role for regulatory T cells during kidney injury and fibrosis. JCI Insight 2020; 5: e130651

85. Liao J, Yu Z, Chen Y et al. Single-cell RNA sequencing of human kidney. Sci Data 2020; 7: 4

86. Barry DM, McMillan EA, Kunar B et al. Molecular determinants of nephron vascular specialization in the kidney. Nat Commun 2019; 10:5705

87. Low JH, Li P, Chew EGY et al. Generation of human PSC-derived kidney organoids with patterned nephron segments and a de novo vascular network. Cell Stem Cell 2019; 25: 373–387.e9

88. Der E, Ranabothu S, Suryawanshi H et al. Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis. JCI Insight 2017; 2: e93009

89. Arazi A, Rao DA, Berthier CC et al.; the Accelerating Medicines Partnership in SLE network. The immune cell landscape in kidneys of patients with lupus nephritis. Nat Immunol 2019; 20: 902–914

90. Combes AN, Phipson B, Lawlor KT et al. Single cell analysis of the developing mouse kidney provides deeper insight into marker gene expression and ligand-receptor crosstalk. Development [Internet] 2019; 146: dev178673

91. Fu J, Akat KM, Sun Z et al. Single-cell RNA profiling of glomerular cells shows dynamic changes in experimental diabetic kidney disease. J AmSoc Nephrol 2019; 30: 533–545

92. Schutgens F, Rookmaaker MB, Margaritis T et al. Tubuloids derived from human adult kidney and urine for personalized disease modeling. Nat Biotechnol 2019; 37: 303–313

93. Harder JL,Menon R, Otto EA et al. Organoid single cell profiling identifies a transcriptional signature of glomerular disease. JCI Insight 2019; 4: e122697

94. Tabula Muris Consortium, Overall coordination, Logistical coordination et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 2018; 562: 367–372

95. Wang P, Chen Y, Yong J et al. Dissecting the global dynamic molecular profiles of human fetal kidney development by single-cell RNA sequencing. Cell Rep 2018; 24: 3554–3567.e3

96. Cao J, Cusanovich DA, Ramani V et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 2018; 361: 1380–1385

97. Menon R, Otto EA, Kokoruda A et al. Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. Development 2018; 145: dev164038

98. GilliesCE, Putler R, Menon R et al. An eQTL landscape of kidney tissue in human nephrotic syndrome. Am J Hum Genet 2018; 103: 232–244

99. Kramann R, Machado F, Wu H et al. Parabiosis and single-cell RNA sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. JCI Insight 2018; 3: e99561

100. Lindstro¨m NO, Guo J, Kim AD et al. Conserved and divergent features of mesenchymal progenitor cell types within the cortical nephrogenic niche of the human and mouse kidney. J Am Soc Nephrol 2018; 29: 806–824

101. Sivakamasundari V, BolisettyM, Sivajothi S et al. Comprehensive cell type specific transcriptomics of the human kidney. bioRxiv preprint 2017; doi: 10.1101/238063

102. Chen L, Lee JW, Chou C-L et al. Transcriptomes of major renal collecting duct cell types in mouse identified by single-cell RNA-seq. Proc Natl Acad Sci USA 2017; 114: E9989–E9998

103. Conway BR. Kidney single-cell atlas reveals myeloid heterogeneity in progression and regression of kidney disease. bioRxiv 2020; 10.1101/ 2020.05.14.095166

104. Menon R. SARS-CoV-2 receptor networks in diabetic kidney disease, BKvirus nephropathy and COVID-19 associated acute kidney injury. medRxiv 2020; 10.1101/2020.05.09.20096511

105. Subramanian A, Sidhom E-H, Emani M et al. Single cell census of human kidney organoids shows reproducibility and diminished off-target cells after transplantation. Nat Commun 2019; 10: 5462

106. Pode-Shakked N, Gershon R, Tam G et al. Evidence of in vitro preservation of human nephrogenesis at the single-cell level. Stem Cell Reports 2017; 9: 279–291

107. Method of the year 2019: single-cell multimodal omics. Nat Methods 2020; 17: 1. Doi: 10.1038/s41592-019-0703-5

Affiliations

1 Division of Nephrology and Clinical Immunology, RWTH Aachen University, Aachen, Germany.

2 Institute of Experimental Medicine and Systems Biology, RWTH Aachen University, Aachen, Germany.

3 Department of Urology and Paediatric Urology, St Antonius Hospital, Eschweiler, Germany. 4Department of Urology, Kidney Transplantation Centre, Martin-Luther-University, Halle, Germany.

5 Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University and Heidelberg University Hospital, BioQuant, Heidelberg, Germany.

6 Joint Research Center for Computational Biomedicine, RWTH Aachen University Hospital, Aachen, Germany.

7 Department of Pathology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.

8 Department of Pediatric Nephrology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Amalia Children's Hospital, Nijmegen, The Netherlands. 9Department of Cell Biology, Institute for Biomedical Technologies, RWTH Aachen University, Aachen, Germany.

10 Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK.

11 III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

12 Department of Anatomy and Developmental Biology, Monash Biomedical Discovery Institute, Monash University, Melbourne, Victoria, Australia.

13 Department of Pathology, RWTH Aachen University, Aachen, Germany.

14 Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands. 15Molecular Medicine Partnership Unit, European Molecular Biology Laboratory, Heidelberg University, Heidelberg, Germany.

16 MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK.

17 Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands.

18 Present address: Bayer Pharma AG, Berlin, Germany.

19 These authors contributed equally: Christoph Kuppe, Mahmoud M. Ibrahim.

20 These authors jointly supervised this work: Neil C. Henderson, Rafael Kramann. e-mail: rkramann@gmx.net

3

Decoding myofibroblasts origins in human kidney fibrosis

Christoph Kuppe[1,2,19], Mahmoud M. Ibrahim[1,2,18,19], Jennifer Kranz[2,3,4], Xiaoting Zhang[2], Susanne Ziegler[2], Javier Perales-Patón[2,5,6], Jitske Jansen[2,7,8], Katharina C. Reimer[1,2,9], James R. Smith[10], Ross Dobie[10], John R. Wilson-Kanamori[10], Maurice Halder[1,2], Yaoxian Xu[2], Nazanin Kabgani[2], Nadine Kaesler[1,2], Martin Klaus[11], Lukas Gernhold[11], Victor G. Puelles[11,12], Tobias B. Huber[11], Peter Boor[1,13], Sylvia Menzel[2], Remco M. Hoogenboezem[14], Eric M. J. Bindels[14], Joachim Steffens[3], Jürgen Floege[1], Rebekka K. Schneider[9,14], Julio Saez-Rodriguez[5,6,15], Neil C. Henderson[10,16,20] & Rafael Kramann[1,2,17,20]

Abstract

Kidney fibrosis is the hallmark of chronic kidney disease progression; however, at present no antifibrotic therapies exist[1–3]. The origin, functional heterogeneity and regulation of scar forming cells that occur during human kidney fibrosis remain poorly understood[1,2,4]. Here, using single-cell RNA sequencing, we profiled the transcriptomes of cells from the proximal and non-proximal tubules of healthy and fibrotic human kidneys to map the entire human kidney. This analysis enabled us to map all matrix-producing cells at high resolution, and to identify distinct subpopulations of pericytes and fibroblasts as the main cellular sources of scar-forming myofibroblasts during human kidney fibrosis. We used genetic fate-tracing, time-course single-cell RNA sequencing and ATAC–seq (assay for transposase-accessible chromatin using sequencing) experiments in mice, and spatial transcriptomics in human kidney fibrosis, to shed light on the cellular origins and differentiation of human kidney myofibroblasts and their precursors at high resolution. Finally, we used this strategy to detect potential therapeutic targets, and identified *NKD2* as a myofibroblast-specific target in human kidney fibrosis.

Main Text

Chronic kidney disease (CKD) affects more than 10% of the world population. The final common pathway of kidney injury is fibrosis and its extent is inextricably linked to clinical outcomes[1,4]. No approved therapies exist at present, and the cellular origin, functional heterogeneity and regulation of scar-producing cells in the human kidney continues to be debated[1,4]. Using single-cell RNA sequencing (scRNA-seq), we profiled approximately 135,000 human and mouse kidney cells during homeostasis and fibrosis, allowing us to determine the heterogeneity of cells that produce extracellular matrix (ECM) at high resolution. We identified several subpopulations of mesenchymal cells as the main contributors to human kidney fibrosis, whereas injured tubular epithelia, endothelium and monocytes only exhibited minor ECM expression. Genetic fate-tracing and time-course scRNA-seq and ATAC–seq experiments in mice, and spatial transcriptomics in human kidney fibrosis, supported these findings, and shed light on the origin and regulation of human kidney myofibroblasts. This approach also identified candidate therapeutic targets, such as the myofibroblast-specific naked cuticle homologue 2 (*NKD2*) gene.

Single-cell atlas of human chronic kidney disease

To understand which resident human renal cell types secrete ECM during homeostasis and CKD, we generated a single-cell map of the kidneys, with a focus on the tubulointerstitium. More than 80% of renal cortical cells are proximal tubular epithelial cells and have thus dominated previous single-cell maps, masking other populations[5]. We therefore sorted for viable, non-proximal tubular cells (CD10$^-$) and CD10$^+$ proximal tubular cells to map the entire kidney (Extended Data Fig. 1a, b). Although CD10 is also expressed by other cell types, this strategy allows an enrichment or depletion of proximal tubular cells. We performed scRNA-seq analysis of both CD10$^+$ and CD10$^-$ fractions from 13 patients with CKD due to hypertensive nephrosclerosis or control patients without CKD (n = 7; estimated glomerular filtration rate (eGFR) > 60 and n = 6; eGFR < 60) (Extended Data Fig. 1a–i, Supplementary Table 1). We profiled 53,672 CD10$^-$ cells from 11 patients (n = 7 eGFR > 60; n = 4 eGFR < 60) (Supplementary Table 1). To integrate the data across patients, we used an unsupervised graph-based clustering method and identified 50 different CD10$^-$ cell clusters that were represented in both eGFR groups (Fig. 1a–d). This strategy enabled us to determine the heterogeneity of the renal interstitium including the identification of rare cell types such as Schwann cells (Fig. 1a–d, Extended Data Figs. 1j–u, 2a–d). Next, we profiled 33,690 CD10$^+$ proximal tubular cells (5 patients with eGFR > 60 and 3 with eGFR < 60) and arranged these into 7 clusters (Fig. 1e, Extended Data Fig. 2e–j). Cell-cycle analysis indicated increased cycling in CKD, probably reflecting epithelial repair (Extended Data Fig. 2k, l). Analysis of KEGG pathway and Gene Ontology (GO) terms in CD10$^+$ cells suggested increased fatty acid metabolism in CKD (Fig. 1f, Extended Data Fig.

2m, n). Notably, dysregulated fatty acid metabolism has been shown to cause tubular de-differentiation and fibrosis[6].

Origin of extracellular matrix in human CKD

To identify the cell types that contribute to the production of ECM in kidney fibrosis, we established a single-cell ECM expression score that included collagens, glycoproteins and proteoglycans[7], and confirmed an increased score in published CKD data[8] (Extended Data Fig. 2o–u). ECM scores demonstrated a clear shift towards high ECM-expressing cells in CKD (Fig. 1g). Mesenchymal cells exhibited the highest ECM expression and this increased further in CKD (Fig. 1h, i, Extended Data Fig. 2q–u). All fibroblasts and myofibroblasts expanded in CKD (Fig. 1j). Although *ACTA2* has been used as a myofibroblast marker previously, we defined myofibroblasts as cells that express the most ECM genes. To assess the putative myofibroblast differentiation processes, we generated a uniform manifold approximation and projection (UMAP) embedding of (myo)fibroblasts and pericytes (Extended Data Fig. 3a–c). This embedding supported our unsupervised graph clustering (Fig. 1b), highlighting the heterogeneity of the renal mesenchyme. Myofibroblasts were identified as cells that express periostin (*POSTN*) (Extended Data Fig. 3b). Diffusion mapping of high ECM-expressing mesenchymal cells suggested that myofibroblasts arise from both pericytes and fibroblasts (Fig. 1k, Extended Data Fig. 3d).

Minor upregulation of ECM genes occurred in epithelial cells (Fig. 1h), which suggests a minor contribution of the long-debated epithelial-to-mesenchymal transition[1,9,10]. Injured proximal tubules showed the highest expression of ECM genes among CD10$^-$ epithelium, with various expressed genes and GO terms suggesting de-differentiation (Extended Data Fig. 3e–j). In CD10$^+$ proximal tubules, ECM expression increased slightly in CKD (Extended Data Fig. 3k-n). Injured cells were defined by expression of *SOX9*, *CD24A* and *CD133* (also known as *PROM1*) for proximal tubules and *VCAM1* and *ACKR1* for endothelium[11–13].

Thus, the vast majority of ECM in human kidney fibrosis originates from mesenchymal cells, with a minor contribution from de-differentiated proximal tubule cells.

Cellular source of myofibroblasts

Our CD10$^-$ scRNA-seq data identified most *COL1A1*-expressing cells as PDGFRβ$^+$ (Extended Data Fig. 3o). Unsupervised clustering of 37,380 PDGFRβ$^+$ cells sorted from human kidneys ($n = 4$; eGFR > 60 and $n = 4$; eGFR < 60) (Supplementary Table 1) identified mesenchymal populations and some epithelial, endothelial and immune cells, which were annotated by correlation with the CD10$^-$ populations (Fig. 2a, b, Extended Data Fig. 4a–e). ECM gene expression again dominated in pericyte, fibroblast and myofibroblast clusters (Extended Data Fig. 4f–i). Some macrophage, monocyte, endothelial and injured epithelial populations also expressed COL1α1 and PDGFRβ, but at much lower levels (Fig.

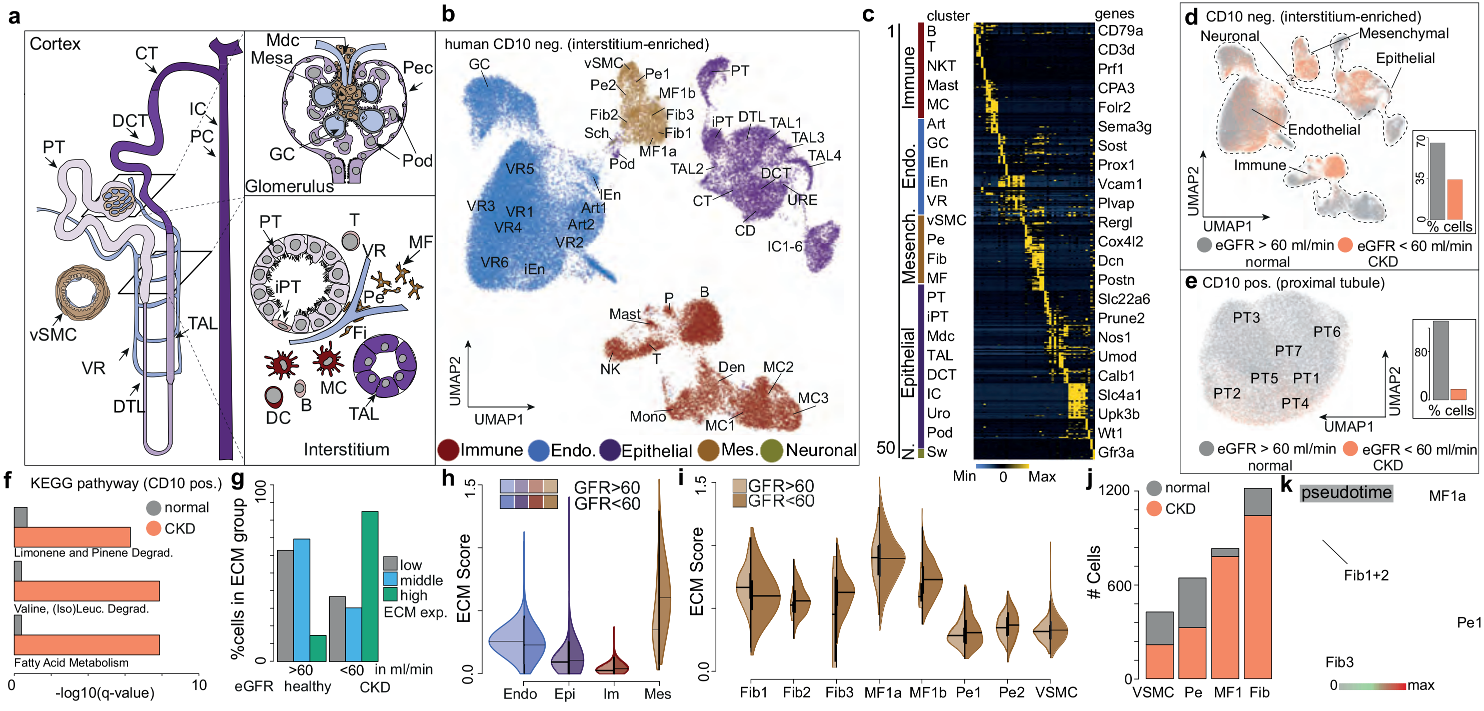


Fig. 1 Single cell atlas of human CKD

a, Scheme of the kidney. CT, connecting tubule; DC, dendritic cell; DCT, distal convoluted tubule; DTL, descending thin limb; GC, glomerular capillary; IC, intercalated cell; Mdc, macula densa; Mast, mast cell; Mesa, mesangial cells; MC, macrophage; MF, myofibroblast; PC, peritubular capillary; Pe, pericyte; Pec, parietal epithelial cells; Pod, podocyte; PT, proximal tubule; Sch, Schwann cell; TAL, thick ascending limb; VR, vasa recta; vSMC, vascular smooth muscle cell. **b**, UMAP embedding of 51,849 CD10$^-$ single cells from 15 human kidneys. Labels refer to 50 clusters identified (see Supplementary Data 1). Endo., endothelial; Epi., epithelial; Mes, mesenchymal; Neuro., neuronal. **c**, Scaled gene expression of the top 10 specific genes in each cluster (see Supplementary Data 2 for detailed information). Each column is the average expression of all cells in a cluster. **d**, Stratification of cells by eGFR. **e**, UMAP embedding of 31,875 CD10$^+$ single cells stratified by eGFR. **f**, KEGG pathway enrichment for CD10$^+$ cells. **g**, CD10$^-$ clustering by ECM score stratified by eGFR (see Extended Data Fig. 2p). **h**, ECM score stratified by cell type and eGFR, mesenchymal ($P \approx 0$), immune ($P \approx 0$), epithelial ($P \approx 0$), endothelial ($P \approx 0$). **i**, Single-cell ECM score for mesenchymal cells, stratified by the main cell types and by eGFR. Bonferroni-corrected *P* values based on two-sided *t*-test for differences in eGFR categories: fibroblast 1 (Fib1) ($P = 0.00015$), Fib2 ($P = 1$), Fib3 ($P = 0.54$), myofibroblast 1a (MF1a) ($P = 1$), MF1b ($P = 0.59$), pericyte 1 (Pe1) ($P = 0.096$), Pe2 ($P = 1$), smooth muscle cell (SMC) ($P = 0.162$). **j**, Number of cells per mesenchymal cell type and clinical parameter. Adjusted *P* values for hypergeometric test for fibroblast and myofibroblast = 0; for pericyte and vSMCs = 1. **k**, Diffusion mapping of mesenchymal cells. Pseudotime indicates cell ordering along putative differentiation processes. For details on statistics and reproducibility, see Methods.

2a, b, Extended Data Fig. 4f–i). Doublet-likelihood scores were low for endothelial and injured epithelial cells, but slightly increased in macrophages (Extended Data Fig. 4j). We verified *COL1A1* mRNA expression in these cells by in situ hybridization (ISH) (Extended Data Fig. 4k–m). These data partially explain the controversy regarding the contributions of non-mesenchymal lineages to fibrosis[1,14], because we indeed observed minor ECM gene expression in these cells, whereas most ECM is derived from mesenchymal cells.

Pseudotime trajectory and diffusion map analysis of major ECM-expressing cells from the PDGFRβ$^+$ populations indicated three main sources of myofibroblasts in human kidneys: (1) NOTCH3$^+$ RGS5$^+$PDGFRα$^-$ pericytes; (2) MEG3$^+$PDGFRα$^+$ fibroblasts; and (3) COLEC11$^+$CXCL12$^+$ fibroblasts (Fig. 2c, Extended Data Fig. 5a). Diffusion mapping places non-CKD cells within populations of low ECM-expressing pericytes and fibroblasts, indicating a differentiation trajectory from low-ECM, non-CKD pericytes and fibroblasts to high-ECM CKD myofibroblasts (Fig. 2c, Extended Data Fig. 5a–i). We confirmed this directionality and also the main lineages of the diffusion map, consisting of NOTCH3$^+$ pericytes (lineage 1) and MEG3$^+$ fibroblasts (lineage 2), using ISH in human kidneys (Fig. 2d, Extended Data Fig. 5j–m). We observed a potential intermediate stage of cells that co-expressed *NOTCH3*, *MEG3* and *POSTN*, possibly representing differentiating cells in the centre of the diffusion map (Fig. 2d, Extended Data Fig. 5k–m). Distinct spatial tissue locations could be identified for the myofibroblast 1 population (POSTN$^+$), which increased in fibrosis, and for the myofibroblast 3 population (CCL19$^+$CCL21$^+$), which were enriched around glomeruli (Extended Data Fig. 5n–r). The gene expression program of pericyte-to-myofibroblast differentiation (lineage 1) demonstrated changes in the cell cycle consistent with differentiation and expansion (Fig. 2e). Ordering their pathway enrichment along pseudotime yielded early canonical WNT and activator protein-1 (AP-1), intermediate *ATF2*, *PDGFRA* and late integrin, ECM receptor interaction and TGFβ signalling among other pathways (Fig. 2e, bottom, Extended Data Fig. 6a). Cessation of the cell cycle also characterized fibroblast-tomyofibroblast differentiation, followed by increased proliferation (lineages 2 and 3) (Extended Data Fig. 6b, c) with early AP-1, and inflammatory pathways, followed by integrin and ECM interaction pathways (Extended Data Fig. 6d–g). Late TGFβ signalling was prevalent in the analysis of lineages 1 and 3 (Fig. 2e, Extended Data Fig. 6a, g). A comparison of ligand and receptor expression within this pathway suggested a mechanism in which myofibroblasts promote the differentiation of fibroblasts and pericytes by TGFβ signalling (Extended Data Fig. 6h–k).

Many of the above pathways are known regulators of fibrosis, including integrin[15] and AP-1 signalling[16]. To further understand transcriptional regulation of mesenchymal populations, we performed motif enrichment analysis of transcription factor sequences in promoters and distal regions of marker genes. This highlighted a potential key regulatory role of AP-1 in the differentiation of fibroblasts to myofibroblasts (Extended

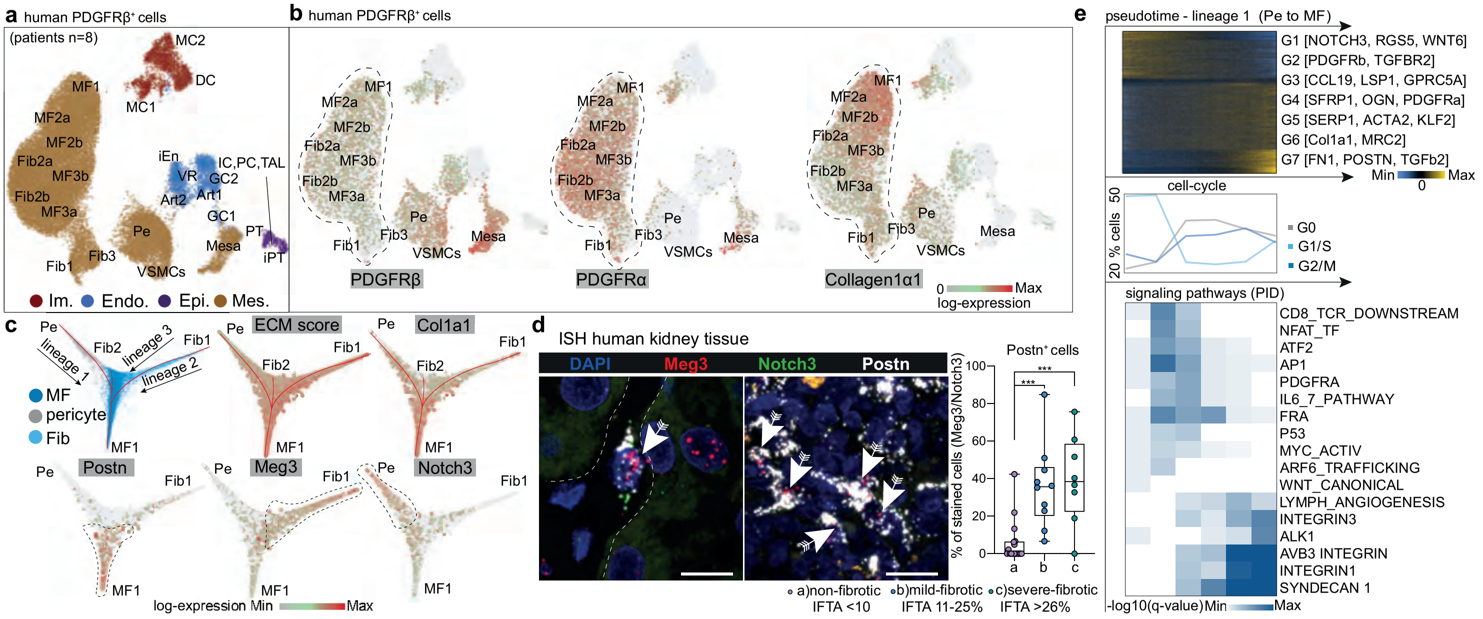


Fig. 2 Origin of myofibroblasts in the human kidney

a, UMAP embedding of 37,800 PDGFRβ$^+$ single cells from 8 human kidneys. Labels refer to identified cell types by unsupervised clustering (see Supplementary Data 1). b, Expression of selected genes on the embedding in a. c, Diffusion map embedding of PDGFRβ$^+$ fibroblasts (Fib), myofibroblasts (MF) and pericytes (*n* = 23,883) and the expression of selected genes on the same embedding. Red lines correspond to the three lineage trajectories (lineage 1 (L1), L2 and L3) predicted by Slingshot given the diffusion map coordinates and the clusters depicted in Extended Data Fig. 5b. d, Representative images and quantification of RNA ISH for *MEG3*, *NOTCH3*, *POSTN* in 35 human kidneys (patient data in Supplementary Table 2). IFTA, interstitial fibrosis tubular atrophy score. *n* = 17 (1), 10 (2), 8(3). ***$P < 0.001$, one-way ANOVA followed by Bonferroni's correction. Tukey box and whisker plot. Scale bars, 10 μm (left), 25 μm (right). e, Top, gene expression dynamics along pseudotime for lineage 1 from pericytes to myofibroblasts (see c, Methods). Middle, cell cycle stage as a percentage of each of the 2,000 cells along pseudotime. Bottom, enrichment of signalling pathways along pseudotime based on the pathway interaction database (PID). For details on statistics and reproducibility, see Methods.

Data Fig. 6l). To validate this finding functionally, we generated a human PDGFRβ$^+$ kidney cell line (Extended Data Fig. 6m). Inhibition of AP-1 significantly decreased cell proliferation and expression of *OGN*, whereas *POSTN* expression was increased, which suggests myofibroblast differentiation (Extended Data Fig. 6n).

In the human PDGFRβ data, *OGN* marked fibroblast 1+3 whereas *POSTN* marked myofibroblasts 1 (Extended Data Fig. 6o). Consistent with this, expression of AP-1 transcription factors negatively correlated with average collagen expression, whereas the expression of putative AP-1 target genes positively correlated with average collagen expression (Extended Data Fig. 6p), indicating that AP-1 may have a repressor role. However, the role of AP-1 is likely to be multifunctional and it may have additional roles that could also promote fibrosis.

We next studied which cells signal towards the key ECM-expressing cells (Extended Data Fig. 6q). Lowest signalling came from healthy proximal tubule, whereas injured proximal tubule cells were among the top signalling partners, suggesting tubule-interstitial signalling as a hallmark of fibrosis[17] (Extended Data Fig. 6q). This interaction involves Notch, TGFβ, WNT and PDGFα signalling (Extended Data Fig. 6r).

PDGFRα$^+$PDGFRβ$^+$ marks ECM-expressing cells

In genetic fate-tracing experiments in kidneys from *Pdgfrb-creER-tdTomato* mice, ISH and immunostaining confirmed that almost all myofibroblasts are derived from the PDGFRβ lineage (Fig. 3a–c, Extended Data Fig. 7a–c). A Smart-Seq2 time-course study in *Pdgfrb-eGFP* mice demonstrated that the abundance of smooth muscle cells and pericytes decreased after unilateral ureteral obstruction (UUO), whereas mesangial cells and COL1α1$^+$PDGFRα$^+$ matrix-producing cells increased (Fig. 3d–f, Extended Data Fig. 7d, e). Similar to the human kidney, the main ECM-expressing cell population exhibited expression of *Pdgfra*, *Pdgfrb* and *Postn* (Fig. 3g, Extended Data Fig. 7e–g). Other cells showed significantly lower ECM expression than the PDGFRα$^+$PDGFRβ$^+$ population (Extended Data Fig. 7f, g).

Immunostaining and ISH in mice confirmed double positivity for PDGFRα$^+$ and tdTomato in *Col1a1*-expressing cells, which confirms that cells that express PDGFRα and PDGFRβ are the main source of ECM (Extended Data Fig. 7h, i). This was confirmed via multiplex ISH in a cohort of 62 patients (Extended Data Fig. 7j, k). Diffusion map embedding of matrix-producing cells and pericytes was also consistent with our human PDGFRβ data, and suggested that pericytes (PDGFRβ$^+$PDGFRα$^-$NOTCH3$^+$) are one origin of the main ECM-producing cells (PDGFRβ$^+$PDGFRα$^+$COL1α1$^+$POSTN$^+$) (Extended Data Fig. 7l–p).

Combined, our data demonstrate that PDGFRα$^+$PDGFRβ$^+$ dual-positive mesenchymal cells, including all fibroblast and myofibroblast populations but not non-activated PDGFRα$^-$PDGFRβ$^+$ pericytes (low ECM-expressing pericytes), represent most ECM-expressing cells.

Heterogeneity of the mesenchyme

We next generated scRNA-seq data from 7,245 PDGFRα^{+}PDGFRβ^{+} cells in mouse kidney fibrosis experiments (Fig. 3h). These cells expanded approximately 140-fold after injury and UMAP embedding revealed four major, distinct populations that correspond to mesenchyme, epithelial, endothelial and immune cells (Fig. 3i–k, Extended Data Fig. 7q–r), all of which have been described as origins of kidney fibrosis[1,14,18]. We did not detect undifferentiated pericytes in this data, because pericytes are PDGFRα^{-} in humans and mice (Fig. 2c, Extended Data Fig. 7e).

Non-mesenchymal cells expressed markedly lower ECM and collagen levels than mesenchymal cells (Fig. 3k, Extended Data Figs. 7r, s, 8a), supporting our human data that non-mesenchymal cells contribute little to scarring (Figs. 1, 2). Doublet scores were low in these clusters (Extended Data Fig. 8b).

Unsupervised clustering revealed two key classes within mesenchymal cells in this dataset: (1) fibroblast 1 marked by *Scara5* and *Meg3* expression; and (2) myofibroblasts consisting of various myofibroblast subpopulations (Fig. 3j, k, Extended Data Fig. 8a). In our human data, the myofibroblast 1 subset corresponded to terminally differentiated myofibroblasts with the highest ECM expression preceded in differentiation pseudotime by myofibroblast 2 (OGN^{+}), whereas the fibroblast 1 cells appeared as a ‘progenitor’ non-activated fibroblast population (Fig. 2c). Fibroblast 1 cells differed from myofibroblasts in the PDGFRα^{+}PDGFRβ^{+} data by three main features: first, *Col15a1*, a mouse myofibroblast-specific collagen (Extended Data Fig. 7e), was expressed at lower levels in fibroblast 1 cells than in myofibroblasts (Extended Data Fig. 8c); second, although *Meg3* was expressed in some other cells (Extended Data Fig. 8d, e), it was confined to fibroblast 1 cells within the mesenchyme (Fig. 3k), as validated by ISH in human kidney (Extended Data Fig. 8d–f); (3) the fibroblast 1 population is SCARA5^{+} but FRZB^{-} (Extended Data Fig. 8g), again demonstrating that they are distinct from myofibroblasts.

Having established fibroblast 1 cells as a distinct population, we generated UMAP and diffusion map embeddings and performed pseudotime analyses of all PDGFRα^{+}PDGFRβ^{+} mesenchymal cells to gain insight into their lineage relationships (Fig. 3l–n). This analysis suggested fibroblast 1 (MEG3^{+}SCARA5^{+}) and myofibroblast 2 (COL14A1^{+}OGN^{+}) as early states, myofibroblast 3a as an intermediate state, and myofibroblast 1a (NRP3^{+}NKD2^{+}), myofibroblast 1b (GREM2^{+}) and myofibroblast 3b (FRZb^{+}) as terminal states (Fig. 3l–n). Thus, fibroblast 1 and myofibroblast 2 cells are the main source of myofibroblasts in mouse kidney fibrosis. The myofibroblast 2 subset (OGN^{+}COL14A1^{+}) might exist in healthy mouse kidneys or may arise as an intermediate state via pericyte-to-myofibroblast differentiation (Fig. 2c, human data). Expression of *Agtr1a* in these cells points towards a pericyte origin (Fig. 3n).

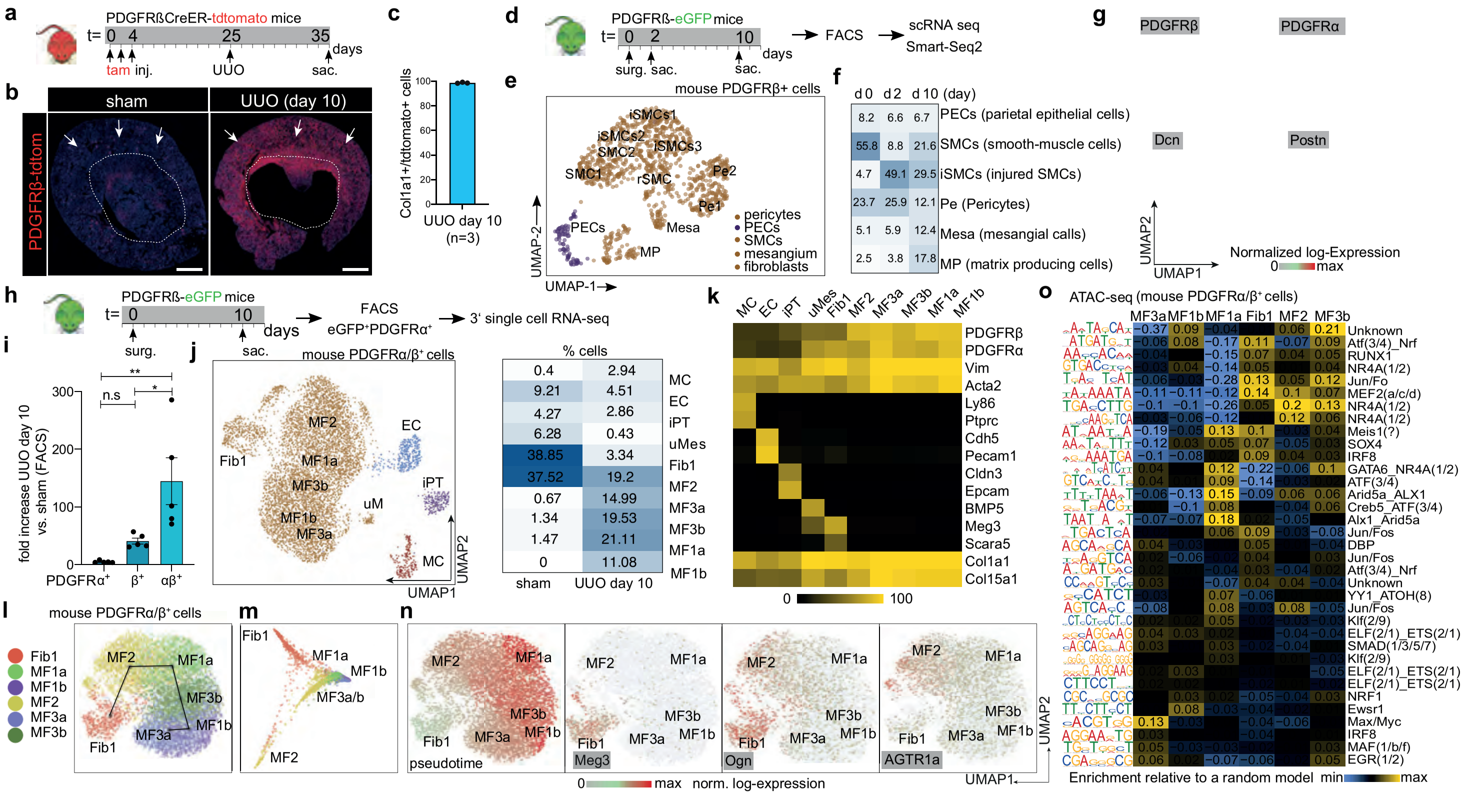


Fig. 3 Origin of myofibroblast in mice

a, Design of fate-tracing experiments. **b**, Representative images of the kidney from a *Pdgfrb-creER;tdTomato* mouse (sham versus UUO). Scale bars, 1,000 μm. **c**, Percentage of cells expressing *Col1a1* mRNA that co-express tdTomato at day 10 after UUO (n = 3; data are mean values). **d**, Time-course UUO experiment design. **e**, UMAP embedding of the mouse PDGFRβ$^+$ cells. Labels refer to cell types identified. PECs, parietal epithelial cells. **f**, Percentage of cells per cell type and time point. D, day; iSMCs, injured SMCs; MP, matrix-producing cells. **g**, Expression of selected genes on the UMAP embedding from **e**. **h**, Scheme of the PDGFRα/PDGFRβ isolation UUO experiment. **i**, Quantification of PDGFRα$^+$ and PDGFRβ$^+$ cells by flow cytometry (n = 5 per group). *P < 0.05, **P < 0.01, one-way ANOVA with post hoc Bonferroni correction. Data are mean ± s.e.m. **j**, Left, UMAP embedding of the PDGFRα$^+$ PDGFRβ$^+$ cells. Right, percentage of cells per cluster. EC, endothelial cells; uMes, unknown mesenchymal. **k**, Expression of selected genes in each of the cell clusters in **j**. **l**, **m**, UMAP (**l**) and diffusion map (**m**) embedding of fibroblasts and myofibroblasts. **n**, Computational cell ordering (pseudotime) and expression of selected genes on the embedding in **n**. **o**, Enrichment of the occurrence of transcription factor motifs in fibroblasts and myofibroblasts. Transcription factor motifs were identified from PDGFRα$^+$PDGFRβ$^+$ day 10 UUO ATAC–seq data (Methods). For details on statistics and reproducibility, see Methods.

Supervised classification of the mouse PDGFRα$^+$PDGFRβ$^+$ single-cell data based on our human PDGFRβ$^+$ cells confirmed the distinctness of fibroblast 1 cells and myofibroblasts in both species (Extended Data Fig. 9a, b).

Our data suggest a model in which PDGFRβ$^+$PDGFRα$^+$POSTN$^+$ high-ECM expressing myofibroblasts (here termed myofibroblast 1) arise from PDGFRβ$^+$PDGFRα$^-$NOTCH3$^+$ pericytes, PDGFRβ$^+$PDGFRα$^+$SCARA5$^+$ fibroblasts (fibroblast 1) and PDGFRβ$^+$PDGFRα$^+$CXCL12$^+$ fibroblasts (fibroblast 2) (Extended Data Fig. 9c). Pericytes potentially differentiate through an intermediate ECM-expressing PDGFRβ$^+$PDGFRα$^+$OGN$^+$ COL14A1$^+$ (myofibroblast 2) state into myofibroblast 1 cells (Extended Data Fig. 9c).

Distinct fibroblast and myofibroblast cell states

We next sought to determine whether the above fibroblast and myofibroblast cell states represent distinct cell types with distinct gene regulatory profiles[19]. We generated bulk ATAC–seq[20] data from PDGFRα$^+$PDGFRβ$^+$ cells isolated from mouse kidneys after UUO and deconvoluted the open chromatin region signatures based on proximity to marker genes identified in the scRNA-seq clusters (Fig. 3o). Fibroblast 1 and myofibroblast 2 cells were distinct from each other and from other myofibroblasts. Myofibroblast 1a cells were distinct from myofibroblast 1b and featured enrichment of ATF factors. The myofibroblast 2 and 3b subclusters showed enrichment of the orphan receptor *Nr4a1*, previously reported as a regulator of TGFβ signalling and fibrosis[21]. Fibroblast 1 cells showed enrichment of AP-1 motifs (Fig. 3o), in line with the human data (Extended Data Fig. 6l). RNA expression of the factors identified by ATAC–seq (Extended Data Fig. 9d–g) confirmed the sequence motif enrichment (Fig. 3o), indicating divergent transcriptional regulation in these populations.

Consistent with our ATAC–seq data, analysis of signalling pathways based on our scRNA-seq data indicated that fibroblast 1 and myofibroblasts are distinct populations with different regulatory programs (Extended Data Fig. 9h).

NKD2 as potential therapeutic target

We analysed our data to identify potential therapeutic targets for kidney fibrosis. *Nkd2* is specifically expressed in mouse PDGFRα$^+$PDGFRβ$^+$ terminally differentiated myofibroblasts (Fig. 4a, Extended Data Fig. 9i), and cells positive for both NKD2 and PDGFRα constituted more than 40% of all COL1α1$^+$ cells (Fig. 4b). In human PDGFRβ$^+$ cells, *NKD2* marks high ECM myofibroblasts, its expression correlates positively with *POSTN* and ECM and negatively with genes associated with pericytes and fibroblasts (Fig. 4c, Extended Data Fig. 10a, b). NKD2$^+$ myofibroblasts exhibited increased TGFβ, WNT and TNF pathway activity compared with NKD2$^-$ cells (Extended Data Fig. 10c), using the PROGENy method[22]. Multiplex ISH in 36 patients confirmed that a subpopulation of PDGFRα$^+$PDGFRβ$^+$ cells expresses *NKD2* and expands in fibrosis (Fig. 4d, e). *NKD2* is a

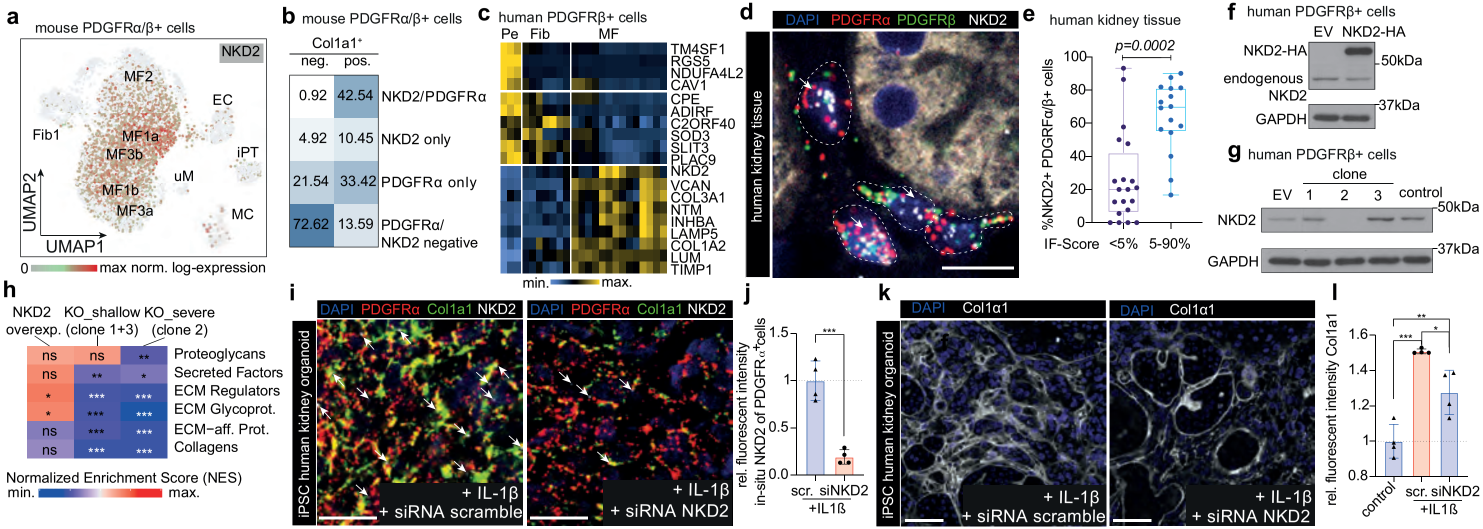


Fig. 4 NKD2 as a therapeutic target

a, Expression of *Nkd2* in mouse PDGFRα$^+$PDGFRβ$^+$ cells visualized on the UMAP embedding from Fig. 3j. **b,** Percentage of *Col1a1*$^{+/-}$ cells in mouse PDGFRα$^+$PDGFRβ$^+$ cells (Fig. 3j, stratified by PDGFRα and NKD2 expression). **c,** Scaled gene expression of *NKD2* correlating or anti-correlating genes in human PDGFRβ$^+$ cells (Fig. 2). **d, e,** RNA ISH of *PDGFRA*, *PDGFRB* and *NKD2* in human kidneys (d) and quantification of triple-positive cells (e) (n = 36, patient data in Supplementary Table 2). n = 20 and 16 (for box plots in e). *P* values determined by two-tailed Mann–Whitney test. Tukey box and whisker plot. Scale bar, 10 µm. **f, g,** Representative western blots of *NKD2* overexpression (f) and knockout (g) in human PDGFRβ cells. EV, empty vector. For gel source data, see Extended Data Fig. 10e. **h,** Gene set enrichment analysis (GSEA) of ECM genes in NKD2-perturbed PDGFRβ$^-$ kidney cells. NES, normalized enrichment score; NS, not significant; OE, overexpression. n = 3 each. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, fast GSEA-multilevel method after adjusting *P* values for multiple testing correction (Benjamini and Hochberg). **i,** ISH of *PDGFRA*, *PDGFRB* and *NKD2* in human iPS cell-derived kidney organoids after transfection of *NKD2* siRNA (siNKD2) or scrambled (Scr.) control. **j,** Quantification of *NKD2* RNA expression in PDGFRα$^+$ organoids. n = 4 each. *P* values determined by two-tailed unpaired *t*-test. **k, l,** Immunofluorescence staining (k) and quantification (l) of COL1α1 in kidney organoids derived from human iPS cells. n = 4 each. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, one-way ANOVA followed by Bonferroni's correction. Scale bars, 50 µm. Data are mean ± s.d. For details on statistics and reproducibility, see Methods.

modulator of TNF and the WNT pathway[23,24]. To study the role of *NKD2* in kidney fibrosis, we used our human PDGFRβ$^+$ data to predict a gene-regulatory network focused on genes that correlated with *NKD2*, using the GRNboost2 framework[25] (Extended Data Fig. 10d–f). This analysis suggests that *NKD2* regulates *ETV1* and affects paracrine signalling through *LAMP5* (Extended Data Fig. 10f, g).

Lentiviral overexpression of *NKD2* in our human PDGFRβ cell line induced expression of ECM molecules in response to TGFβ, whereas knockout of *NKD2* markedly reduced the expression of *COL1A1*, *FN1* and *ACTA2* in the presence or absence of TGFβ (Fig. 4f, g, Extended Data Fig. 10h–j). RNA-seq analysis from cells overexpressing *NKD2* demonstrated upregulated ECM regulators and glycoproteins, whereas *NKD2*-knockout cells exhibited loss of ECM regulators, glycoproteins and collagens (Fig. 4h). Pathway and GO analysis placed *NKD2* in ECM expression programs and suggested an interaction with AP-1 and integrin signalling (Extended Data Fig. 10k, l). We further observed strong changes in the expression of WNT receptors and ligands after *NKD2* knockout (Extended Data Fig. 10m).

To validate *NKD2* as a therapeutic target, we generated induced pluripotent stem (iPS) cell-derived kidney organoids that contained all major compartments of the human kidney (Extended Data Fig. 10n–p). IL-1β can induce fibrosis in iPS cell-derived kidney organoids[26], and short interfering RNA (siRNA)-mediated knockdown of *NKD2* inhibited IL-1β-induced *COL1A1* expression (Fig. 4i–l). Thus, *NKD2* marks myofibroblasts in kidney fibrosis, is required for collagen expression, and represents a potential therapeutic target. However, because these organoids do not contain immune cells, further in vivo data will be required to fully verify this finding.

Discussion

Myofibroblasts represent the main source of ECM during kidney fibrosis, but their cellular origin was controversial[1,9]. scRNA-seq analysis allows the dissection of the cellular heterogeneity of complex tissues and disease processes, and generates insights into disease mechanisms at unprecedented resolution[13,27,28].

Genetic fate-tracing data in mice and histology analyses of human tissue have suggested that epithelial, endothelial, haematopoietic cells and resident mesenchymal cells all contribute to fibrosis[1]. Here we provide a comprehensive cell atlas of human and mouse kidney fibrosis and show that most scar tissue originates from dual-positive PDGFRα$^+$PDGFRβ$^+$ fibroblasts and myofibroblasts. In both humans and mice, these myofibroblasts predominantly derive from pericytes and fibroblasts. Our scRNA-seq strategy pointed to new disease mechanisms and potential therapeutic targets, such as myofibroblast-expressed *NKD2*. Although *NKD2* has been reported as a WNT inhibitor, our data indicate that it may also act as an activator of some aspects of WNT signalling.

Our work highlights the intricate cell differentiation mechanisms involved in fibrosis and provides a resource for future clinical research in kidney disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2941-1.

Methods

Ethics

The local ethics committee of the University Hospital RWTH Aachen approved all human tissue protocols (EK-016/17). Kidney tissue was collected from the Urology Department of the Hospital Eschweiler from patients undergoing (partial) nephrectomy due to kidney cancer. All patients provided informed consent and the study was performed in accordance with the Declaration of Helsinki.

Processing of human tissue

The tissue was snap-frozen on dry-ice or placed in prechilled in University of Wisconsin solution (BTLBUW, Bridge to Life). Tissues were sliced into approximately 0.5–1 mm3 pieces and transferred to a C-tube (Miltenyi Biotec) and processed on a gentle-MACS (Miltenyi Biotec) using the program spleen 4. The tissue was digested for 30 min at 37 °C with agitation at 300 rpm in a digestion solution containing 25 μg ml−1 liberase TL (Roche) and 50 μg ml−1 DNase (Sigma) in RPMI (Gibco). After incubation, samples were processed again on a gentle-MACS (Miltenyi Biotec) using the same program. The resulting suspension was passed through a 70-μm cells strainer (Falcon), washed with 45 ml cold PBS and centrifuged for 5 min at 500g at 4 °C. Live, single cells were enriched by FACS-sorting and gating on DAPI negative cells with further enrichment of epithelial cells by CD10 staining or PDGFRβ staining for fibroblasts.

Mice

Pdgfrb-creER^{t2} (that is, *B6-Cg-Gt(Pdgfrb-cre/ERT2)*$^{6096Rha/J}$, JAX Stock 029684) and *Rosa26tdTomato* (that is, *B6-Cg-Gt(ROSA) 26Sort*$^{tm(CAG\text{-}tdTomato)Hze}$*/J*, JAX Stock 007909) were purchased from Jackson Laboratories. *Pdgfrb-BAC-eGFP* reporter mice were developed by N. Heintz (The Rockefeller University) for the GENSAT project. UUO was performed as previously described using male and female mice[29]. Animal experiment protocols were approved by the LANUV-NRW, Germany, and by the UK Home Office Regulations. For Smart-Seq2, *Pdgfrb-eGFP* male mice born within 10 days of each other were used, and between 9 and 11 weeks old at the time of surgery. For inducible fate-tracing experiments, *Pdgfrb-creER;tdTomato* mice (8 weeks of age) received tamoxifen (10 mg orally) 3 times via gavage followed by a washout period of 21 days and then subjected to UUO surgery or sham (as above) and killed at 10 days after surgery. Mice were housed at two to five mice per cage with a 12-h light–dark cycle (lights on from 0700 to 1900) at sustained temperature (20 °C ± 0.5 °C) and humidity (approximately 50% ± 10%) with ad libitum access to food and water.

Single-cell isolation in mouse

Euthanized mice were perfused via the left heart with 20 ml NaCl 0.9% to remove blood residues from the vasculature. To isolate single kidney cells, a combination of enzymatic and mechanical disruption was used as described above for human single cell isolation. Overall, the viability was more than 80% using this method.

FACS

Cells were labelled with the following antibodies: anti-CD10 human (clone HI10a, Biolegend, 1:100), anti-PDGFRβ mouse (clone PR7212, R&D, 1:100), anti-PDGFRα mouse (clone APA5, Biolegend, 1:100), anti-CD45 mouse (clone 30_F11). Isolated cells were resuspended in 1% FBS in PBS on ice at a final concentration of 1 × 10^7 cells per ml. Cells were pre-incubated with Fc-Block (TruStainFx human, TruStainFx mouse Clone 91, Biolegend) and then incubated with the above antibodies for 30 min on ice protected from light diluted 1:100 in 2% FBS in PBS. For human anti-PDGFRβ staining, goat anti-mouse Dyelight 405 (poly24091, Biolegend, 1:100) was used as a secondary antibody.

All compensation was performed at the time of acquisition using single colour staining and negative staining and fluorescence minus one controls. The cells were sorted in the semi-purity mode targeting an efficiency of more than 80% with the SONY SH800 sorter (Sony Biotechnology; 100-µm nozzle sorting chip Sony). For plate-based sorting for SMART-Seq, cell sorting was performed on a FACS Aria II machine (Becton Dickinson) using BD FACSDiva software. FACS data analysis was performed using FlowJo.

Single-cell assays including Smart-Seq2 and 10X Genomics 3' scRNA-seq (V2 and V3)

For Smart-Seq2 single cells were processed by SciLifeLab Eukaryotic Single Cell Genomic Facility (Karolinska Institute). Before shipping, single cells were sorted into wells of a 384-well plate containing pre-prepared lysis buffer. Libraries were sequenced on Illumina HiSeq4000. The single-cell solution of primary human kidney cells was run on a

Chromium Single Cell Chip kit and libraries were performed using Chromium Single Cell 3′ library kit V2/3 and i7 Multiplex kit (PN-120236, PN-120237, PN-120262, 10x Genomics) according to the manufacturer’s protocol. The library quality was determined using D1000 ScreenTape on a 2200 TapeStation system (Agilent Technologies). Libraries were sequenced on a Illumina Novaseq targeting a read depth as suggested by 10X Genomics 3‘ single-cell RNA kits V2/3.

Human kidney fibrosis evaluation

Periodic acid–Schiff (PAS)-stained sections of the kidneys were analysed and scored in a blinded fashion. The extent of interstitial fibrosis and tubular atrophy were assessed as two separate parameters as the percentage of affected cortical area. For collagen I and III immunohistochemistry (collagen I (Southern Biotech) 1310-01; collagen III (Southern Biotech) 1330-01), sections of formalin-fixed and paraffin-embedded renal tissues were processed for indirect immunoperoxidase staining as previously described29. Using a whole slide scanner (NanoZoomer HT, Hamamatsu Photonics), fully digitalized images of immunohistochemically stained slides were further processed and analysed using the viewing software NDP.view (Hamamatsu Photonics) and ImageJ (National Institutes of Health). The percentage of positively stained area was analysed in the kidney cortex in a blinded fashion.

Antibodies and immunofluorescence staining

Kidney tissues were fixed in 4% formalin for 2 h at room temperature and frozen in OCT after dehydration in 30% sucrose overnight. Using 5–10-μm cryosections, slides were blocked in 5% donkey serum followed by 1-h incubation of the primary antibody, washing three times for 5 min in PBS and subsequent incubation of the secondary antibodies for 45 min. After 4′,6-diamidino-2-phenylindole (DAPI) staining (Roche, 1:10,000), the slides were mounted with ProLong Gold (Invitrogen, P10144). The following antibodies were used: anti-mouse PDGFRα (AF1062, 1:100, R&D), anti-CD10 human (clone HI10a, 1:100, Biolegend), anti-HNF4α (clone C11F12, 1:100, Cell Signaling), anti-Pan-Cytokeratin TypeI/II (Invitrogen, MA1-82041), anti-DACH1 (Sigma, HPA012672, 1:100), anti-COL1α1 (Abcam, ab34710, 1:100), anti-ERG (abcam, ab92513, 1:100), anti-CXCL12/SDF-1 (R&D, MAB350, 1:100), AF488 donkey antigoat (1:200, Jackson Immuno Research), and AF647 donkey anti-rabbit (1:200, Jackson Immuno Research)

Confocal imaging

Images were acquired using a Nikon A1R confocal microscope using 40× and 60× objectives (Nikon). Raw imaging data were processed using Nikon Software, ImageJ, Adobe Photoshop and Adobe Illustrator.

Human kidney tissue microarray

Paraffin-embedded, formalin-fixed kidney specimens from 98 non-tumorous human kidney samples of the Eschweiler/Aachen biobank were selected based on a previously performed PAS staining. Areas were randomly selected per sample and one 2-mm core was taken from each kidney sample using the TMArrayer (Pathology Devices, Beecher Instruments). Each core was arrayed into a recipient block in a 2-mm-spaced grid covering approximately 2.5 cm^2, and 5-μm thick sections were cut and processed using standard histological techniques.

RNA ISH

ISH was performed using formalin-fixed paraffin-embedded tissue samples and the RNAScope Multiplex Detection KIT V2 (RNAScope, 323100) following the manufacturer's protocol with minor modifications. The antigen retrieval was performed for 22 min at 96 °C instead of 15 min at 99 °C in a water bath. Then, 3–5 drops of pre-treatment 1 solution were incubated at room temperature for 10 min after performing antigen retrieval. The washing steps were performed three times for 5 min. The following probes were used for the RNAscope assay: Hs-PDGFRβ 548991-C1, Hs-PDGFRα 604481-C3, Hs-COL1α1 401891, Hs-COL1α1 401891-C2, Hs-MEG3 400821, Hs-NKD2 581951-C2 (targeting 236-1694 of NM_033120.3), Hs-POSTN 409181-C2 and 409181-C3, Hs-PECAM1 487381-C2, Hs-Ccl19 474361-C3, Hs-CCL21 474371-C2, Hs-NOTCH3 558991-C2, Mm-COL1α1 319371, Mm-PDGFRα 480661-C2, and Mm-PDGFRβ 411381-C3.

ISH image analysis

Systematic random sampling was applied to subsample at least three representative tubulo-interstitial areas per image. Next, every fluorescent dot (transcript) was manually annotated using the cell counting tool from Fiji (Max Planck Institute of Molecular Cell Biology and Genetics). Single nuclei were then isolated using an in-house made tool (https://gitlab.com/mklaus/segment_cells_register_marker) based on watershed (limits: 0.1–0.4) to identify neighbouring nuclei, edge detection for incompleteobjects andobject sizeselection(limits: 12–180 μm2). The total number of individual dots was then retrieved for every isolated nucleus. Dots located outside of nuclei were not included in this analysis. For MEG3 and NKD2 analysis of PDGFRα and PDGFRβ cells, images were analysed using QuPath after segmenting the nuclei and counting cells based on >1 positive spot per imaging channel. For quantification of COL1α1 immunofluorescence and NKD2 ISH, images were split in RGB channels and the integrated fluorescent density was determined per image using ImageJ.

qRT–PCR

Cell pellets were collected and washed with PBS followed by RNA extraction according to the manufacturer's instructions using the RNeasy Mini Kit (Qiagen). Total RNA (200 ng) was reverse transcribed with High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and qRT–PCR was carried out further as previously described29. Data were analysed using the 2-Ct method. The primers used are listed in Supplementary Table 3.

Generation of a human PDGFRβ+ cell line

PDGFRβ+ cells were isolated from healthy human kidney cortex of a nephrectomy specimen (71-year-old male individual) by generating a single-cell suspension (as above). For the isolation, cells were stained in two steps using a specific PDGFRβ antibody (R&D MAB1263 antibody, dilution 1:100) followed by Anti-Mouse IgG1-MicroBeads solution (Miltenyi, 130-047-102). After magnetic-activated cell sorting (MACS), cells were cultured in DMEM medium (Thermo Fisher 31885) for 14 days and immortalized using SV40LT and HTERT. Retroviral particles were produced by transient transfection of HEK293T cells using TransIT-LT (Mirus). Two types of amphotropic particles were generated by co-transfection of plasmids pBABE-puro-SV40-LT (Addgene 13970) or xlox-dNGFR-TERT (Addgene 69805) in combination with a packaging plasmid pUMVC (Addgene 8449) and a pseudotyping plasmid pMD2.G (Addgene 12259). Retroviral particles were concentrated 100 times using Retro-X concentrator (Clontech) 48 h after transfection. Cell transduction was performed by incubating the target cells with serial dilutions of the retroviral supernatants (1:1 mix of concentrated particles containing SV40-LT or hTERT) for 48 h. Next, the infected PDGFRβ+ cells were selected with 2 μg ml−1 puromycin at 72 h after transfection for 7 days.

Culturing human iPS cell-derived kidney organoids

Human iPS cell C-15 clone 0001 was received from the Stem Cell Facility of the Radboud University Center. Human iPS cells were grown on Geltrex-coated plates using E8 medium (Life Technologies). After 70–80% confluency, iPS cells were detached using 0.5 mM EDTA and cell aggregates were reseeded by splitting 1:3. iPS cells were differentiated using a modified published protocol30 and seeded at a density of 18,000 cells per cm2 on Geltrex-coated plates (Greiner). Differentiation towards intermediate mesoderm was initiated using CHIR99021 (6 μM, Tocris) in Essential 6 (E6) medium (Life Technologies) for 3 and 5 days, followed by FGF9 (200 ng ml−1, RD Systems) and heparin (1 μg ml−1, Sigma Aldrich) supplementation in E6 medium up to day 7. After 7 days of differentiation, cell aggregates (300,000 cells per organoid, mixture of 3 and 5 day CHIR-differentiated cells) were cultured on Costar Transwell inserts to stimulate self-organizing nephrogenesis using E6 differentiation medium. On days 7 + 18 (2D culture + 3D culture conditions), the kidney organoids were used for siRNA knockdown experiments as described below.

siRNA knockdown of NKD2 in human iPS cell-derived kidney organoids

NKD2 siRNA knockdown was carried out according to the manufacturer protocol (DharmaFECT transfection reagent and NKD2-specific smartpool siRNA, both Horizon Discovery). The transfection master mix and scrambled controls were prepared in E6 medium (Gibco) and added to the organoids. After an initial incubation of 24 h, the transfection master mixes were refreshed, and IL-1β (Sigma-Aldrich) was added at a concentration of 100 ng ml–1 to induce fibrosis. The addition of IL-1β together with fresh transfection master mix was repeated every 24 h for two days. Then, 96 h after the initiation of transfection, the organoids were collected and processed for paraffin sectioning. Fluorescence in-situ hybridization (FISH) and immunofluorescence staining was performed as described above.

TGFβ treatment experiments

TGFβ (100-21-10UG, Peprotech) at 10 ng ml–1 in PBS was added to 75% confluent PDGFRβ cells for 24 h after 24 h of serum starvation with 0.5% FCS containing medium. For inhibitor experiments, the T-5224 inhibitor (or vehicle) was added to the culture wells 1 h before the addition of TGFβ. All experiments were performed in triplicates.

AP-1 inhibitor treatment

T-5224 (c-Fos/AP-1 inhibitor, Cayman Chemicals, 22904) was dissolved in DMSO and stored at –80 °C. DMSO was always added in the same proportions to control wells.

Cell proliferation (WST-1 assay)

WST-1 assay with PDGFRβ cells was performed in 96-wells as recommended by the manufacturer (Roche Applied Science). In brief, 1 × 104 PDGFRβ cells were seeded into each well of 96-well plates and the cells were treated with T-5224 or vehicle (DMSO) with the indicated concentrations in triplicates. Cells were incubated with WST-1 reagent for 2 h before collecting at the indicated time points. Absorbance at both450nm and 650 nm (as a reference) was measured.

CRISPR–Cas9 vector construction, virus production and transduction

The NKD2-specific guide RNA (forward 5′-CACCGACTCCAG TGCGATGTCTCGG-3′; reverse5′-AAACCCGAGACATCGCACTGGAGTC-3′) were cloned into pL-CRISPR.EFS.GFP (Addgene 57818) using BsmBI restriction digestion. Lentiviral particles were produced by transient co transfection of HEK293T cells with lentiviral transfer plasmid, packaging plasmid psPAX2 (Addgene 12260) and VSVG packaging plasmid pMD2.G (Addgene 12259) using TransIT-LT (Mirus). Viral supernatants were collected 48–72 h after transfection, clarified by centrifugation, supplemented with 10% FCS and Polybrene (Sigma-Aldrich, final concentration of 8 μg ml−1) and 0.45-μm filtered (Millipore; SLHP033RS). Cell transduction was performed by incubating the PDGFRβ cells with viral supernatants for 48 h. eGFP-expressing cells were single-cell sorted into 96-well plates. Expanded colonies were assessed for mutations with mismatch detection assay: gDNA spanning the CRISPR target site was PCR amplified and analysed by T7EI digest (T7 Endonuclease, NEB M0302S). To determine specific mutation events on both alleles within the clones grown, the PCR product was subcloned into the pCR 4Blunt-TOPO vector (Thermo Scientific K287520). Minimum 6 colonies per CRISPR-clone were grown and sent for sanger sequencing (clone C2: 30 colonies have been sequenced).

Western blot

Cell lysates were prepared by RIPA buffer with protease inhibitor cocktail (Roche). The protein concentrations of the lysates were quantified using BCA assay (23225, Pierce, ThermoScientific). The protein lysates were heated for 5 min at 95 °C in 4× SDS sample loading buffer (BioRad) and loaded into 10% SDS–PAGE gels. Afterwards samples were transferred onto PVDF membranes and the blots were probed with primary antibody in 5% Blotto (Thermo Fisher): (1:3,000 rabbit anti-human NKD2 polyclonal antibody, Invitrogen PA5-61979) for 2 h, followed by incubation with secondary antibody for 1 h after washing (1:5,000 horseradish-peroxidase (HRP)-conjugated anti rabbit, Vector Laboratories) and developed using Pierce ECL Western Blotting Substrate A and B. Mouse monoclonal anti-GAPDH antibody (Novus Biologicals NB300-320; 1:1,000) followed by HRP-conjugated anti-mouse secondary antibody (Vector Laboratories) was used to stain GAPDH as a loading control.

Lentiviral overexpression of NKD2

The human cDNA of NKD2 was PCR amplified using the primer sequences 5′-ATGGGGAAACTGCAGTCGAAG-3′ and 5′-CTAGGACGGGTGGAAGTGGT-3′. Restriction sites and N-terminal 1xHA-Tag have been introduced into the PCR product using the primer 5′-CACTCG AGGCCACCATGTACCCATACGATGTTCCAGATTACGCTGGGAAACTGCAGTCGAAG-3′ and 5′-ACGGAATTCCTAGGACGGGTGGAAGTG-3′.

Subsequently, the PCR product was digested with XhoI and EcoRI and cloned into pMIG (pMIG was a gift from W. Hahn (Addgene plasmid 9044; http://n2t.net/addgene:9044; RRID:Addgene_9044). Retroviral particles were produced by transient transfection in combination with packaging plasmid pUMVC (pUMVC was a gift from B. Weinberg; Addgene plasmid 8449) and pseudotyping plasmid pMD2.G (pMD2.G was a gift from D. Trono; Addgene plasmid 12259; http://n2t.net/addgene:12259; RRID:Addgene_12259) using TransIT-LT (Mirus). Viral supernatants were collected 48–72 h after transfection, clarified by centrifugation, supplemented with 10% FCS and Polybrene (Sigma-Aldrich, final concentration of 8 µg ml−1) and passed through a 0.45-µm filter (Millipore; SLHP033RS). Cell transduction was performed by incubating the PDGFRβ cells with viral supernatants for 48 h. eGFP-expressing cells were single-cell sorted.

Bulk RNA sequencing

RNA was extracted according to the manufacturer's instructions using the RNeasy Mini Kit (QIAGEN). For rRNA-depleted RNA-seq using 1 and 10 ng of diluted total RNA, sequencing libraries were prepared with KAPA RNA HyperPrep Kit with RiboErase (Kapa Biosystems) according to the manufacturer's protocol.

ATAC–seq preparation

PDGFRα+PDGFRβ+ cells were sorted by FACS from freshly isolated UUO kidneys as described above, washed twice with cold PBS and centrifuged at500gfor5min. Cellpelletswerelysedin50µlice-coldlysisbuffer(10mM Tris-HCl, pH 7.5; 10mM NaCl, 3mM MgCl2, 0.08% NP40 substitute (74385, Sigma), 0.01% Digitonin (G9441, Promega)), and immediately centrifuged at 500g for 9 min. Pellets were resuspended in 50 µl of a transposase reaction mix as previously described31. Transposed DNA was amplified by PCR using NEBNext 2x Master mix (M0541S; New England Biolabs) with custom Nextera PCR primers. The first PCR was performed with 50 µl volume and 6 cycles using NEBNext 2x Master mix and 1.25 µM custom primers; the second RT–PCR was performed with 15 µl volume for 20 cycles using 5 µl (10%) of the pre-amplified mixture plus 0.125 µM primers to determine the number of additional cycles needed as described previously[31]. The amplified DNA library was purified using MinElute PCR Purification kit (28004, Qiagen) and eluted in 20 µl of 10 mM Tris-HCl (pH 8) for subsequent sequencing.

Smart-Seq2 data processing

The initial single-cell transcriptomic data was processed at the Eukaryotic Single-Cell Genomics Facility at the Science for Life Laboratory in Stockholm, Sweden. Obtained reads were mapped to the mm10 build of the mouse genome (concatenated with transcripts for eGFP and the ERCC spike-in set) to yield a count for each endogenous gene, spike-in, and eGFP transcript per cell. Ribosomal RNA genes, ribosomal proteins and ribosomal pseudo-genes were filtered out. We noticed that cells that did not feature any alignments assigned to either eGFP or PDGFRβ clustered into a single cluster after unsupervised cell clustering (see 'Mouse Smart-Seq2 single-cell data integration strategy'). Therefore, we opted to remove those cells, and performed all analysis and clustering without considering those cells (17 cells).

10x single cell RNA-seq data processing

Fastq files were processed using Alevin[32] and Salmon (Alevin parameters -l ISR, Salmon version 0.13.1)[33], Gencode v29 human transcriptome, and Gencode vM20 mouse transcriptome as reference transcriptomes[34.] The Alevin expected cells parameter was set according to thrice the number of cells estimated according to the knee-method applied to the read counts per cell barcodes distribution. Therefore, the unique molecular identifier (UMI) count matrix produced by Alevin resulted in a large number of putative cells that we could filter later.

10x scRNA-seq cell filtering

We removed ribosomal RNA genes (0–1% on average of detected RNA content per cell) and mitochondrially encoded genes (0–80% on average of detected RNA content per cell) from the main gene expression matrix. Mitochondrially encoded genes were removed to avoid introducing unwanted variation between cells that might be solely dependent on changes in mitochondrial content[35]. The log10(total UMI counts per cell) distribution from the count matrix produced by Alevin (see above) typically showed a bimodal distribution, therefore log10(total UMI counts per cell) were clustered into two clusters using mclust R package v5.4.3 setting modelNames to 'E'[36]. Cells that belong to the cluster with the higher counts were kept. Then cells were filtered based on mitochondrial RNA content and bias towards highly expressed genes as follows:

(1) cells were clustered into two clusters using a bivariate Gaussian mixture with two components learned on log10(total UMI counts per cell) and the percentage of mitochondrial UMI per cell. Clustering was performed using the R package Mclust setting modelNames to 'EII'. Cells falling into the cluster with higher mitochondrial content cells were excluded. This filtering step was followed only for libraries that showed a clear bimodal distribution of mitochondrial content (only three 10x libraries in this study). (2) The total number of UMIs per cell should correlate with the total number of unique detected genes. Cells that do not follow this relationship (outliers) were filtered by clustering nuclei using a bivariate Gaussian mixture model on log10(total UMI counts) and log10(total unique detected genes) using the mclust R package setting model-Names to "VEV","VEE". (3) Cells whose percentage of total counts in the top 500 genes represented more than 5 times absolute median deviation for all cells were removed. (4) Finally, to exclude cells that consisted mainly of ribosomal proteins and pseudogenes, we removed cells with a percentage of ribosomal protein and pseudo-gene expression that was more than 5 absolute median deviations of all other cells. Mitochondrial-based filtering was not performed for CD10+ libraries because libraries from proximal tubule epithelial cells are expected to result in a high number of mitochondrial reads. Note that not all filtering steps were performed for all libraries as this depends on the quality of each library and UMI-cell-gene distribution. The script for quality control, cell filtering is

available at: https://github.com/mahmoudibrahim/KidneyMap/blob/master/templates/process_scRNA.r

Human 10x single-cell data integration strategy

Upon initial analysis of our data, we noted several points: (1) Cell types are not guaranteed to be equally represented across patients and across conditions (healthy or CKD). This is because the cell types captured in any single 10x Chromium run are determined by random sampling of cells. (2) Samples from both healthy controls and patients with CKD consisted of cells in healthy and disease states, because this categorization is based on clinical parameters and not on molecular data or a controlled in vitro experiment. We would expect mainly a change in the proportion of healthy and disease cell states between healthy and diseased patient samples. (3) Samples from different patients were processed and prepared on different days as dictated by the surgery schedule at the Eschweiler Hospital. Therefore, potential technical (batch) effects could not be controlled on the experimental side. (4)The ability to discover highly resolved cell clusters in underrepresented cell types might be affected by class imbalance as certain cell types may be markedly more abundant than others, and the size of the dataset (number of cells) that affects clustering results using unsupervised modularity-based graph clustering algorithms[37].

Our experimental strategy involved obtaining separate libraries from CD10+ and CD10− cell fractions (see main text), which was designed to mitigate class imbalance on the level of cell type capturing frequency by the 10x Chromium protocol. To further mitigate the points discussed above we aimed to (1) cluster the data on a local level while keeping global information on the relation between cell types intact, and (2) correct for potential technical (batch) differences between samples while retaining important differences, such as different cell types or different states of cell types due to disease. To do so, we followed a strategy comprised of the following steps.

Step one: after quality control and cell filtering (see above), cells in each 10x library were clustered separately and each cell cluster was assigned to one of six main cell types: CD10+ epithelial, CD10− epithelial, immune, endothelial, mesenchymal and neuronal cells.

Step two: for each one of the six main cell types, cells from all 10x libraries were integrated together. Variability between cells due to technical reasons was corrected and cells were clustered using unsupervised graph clustering. This process resulted in six separate endothelial, CD10+ epithelial, CD10− epithelial, mesenchymal, immune and neuronal maps. Each map consisted of cells from multiple 10x libraries.

Step three: we integrated three single cell maps for: (1) CD10+ cells (proximal tubule; Fig. 1), (2) CD10− cells (proximal tubule-depleted; Fig. 1) and (3) PDGFRβ+ cells (mesenchymal; Fig. 2), by combining single-cell expression (UMI counts) and clustering information from all main cell type individual maps of each data set from step two. All plots presented in this Article are reproducible from those three integrated maps.

This approach accomplished local clustering and technical variability removal, and allowed for high-resolution discovery of cell states consisted of 24 cells, whereas the largest cluster consisted of 5,355 cells. Relative to ‘a high-level clustering followed by sub-clustering’ approach, our approach produces highly resolved clusters in a data-driven unbiased manner, while avoiding the question of which clusters to subcluster altogether. We note that a similar data integration approach was previously described[38].

Step one details

Cell clustering

After cell filtering and quality control (see above), we used marker genes previously described[39,40] and BioGPS41 (for neuronal genes) as a priori defined highly variable gene list. Two lists were constructed for human and mouse based on gene symbol conversion according to the biomaRt database[42,43]. We followed a graph-clustering approach to determine cell clusters, similar to that of Seurat[44] and previously described[45,46]. The clustering approach consisted of dimensionality reduction of the normalized expression matrix (restricted to the highly variable gene list) using singular value decomposition as a first step. The left singular vectors are Eigengenes that describe gene expression programs across single cells[47]. The top n left singular vectors were selected based on the knee of the singular values curve, and used to construct a k-nearest neighbour graph, in which the average k per cell was defined as the square root of the number of cells. The function nn2 from the R package RANN was first used to define the k-nearest neighbours (https://CRAN.R-project.org/package=RANN) and the final graph was constructed based on the top n nearest neighbours by similarity in which n = k × number of cells. Cells were clustered on the graph using the Infomap graph clustering method48 as implemented in the iGraph R package (https://igraph.org). Infomap is a state-of-the-art graph community detection method that we selected for this step as we noticed it tends to produce higher resolved clusters than other graph-clustering methods. At this step, we also calculated a single-cell doublet score for all cells using the doublet score function in the Scran R package that implements the doublet score method previously described[49]. This score is aggregated per cluster and reported for each integrated map (Extended Data Figs. 4i, 8e), but not used to exclude cells.

Assigning cell clusters to five main types

We obtained a ranking for each gene in each cluster according to whether it is unique to a cluster and also highly expressed in this cluster using the function sortGenes in the genesorteR R package[50], setting binarize Method to 'naive'. We intersected the top 50 genes in each cluster with the a priori highly variable gene list (see above) and used this intersection to determine which of the five main cell types (epithelial, endothelial, immune, neuronal or mesenchymal) the cell cluster belongs to.

Scripts and metadata

The a priori putative variable gene list is provided here: https://github.com/mahmoudibrahim/KidneyMap/blob/master/assets/public/all_markers_Human_MMI_Apr2020.txt and https://github.com/mahmoudibrahim/KidneyMap/blob/master/assets/public/all_markers_Mouse_MMI_Mar2020.txt. The script for quality control, cell filtering, clustering and cell type assignment is provided here: https://github.com/mahmoudibrahim/KidneyMap/blob/master/templates/process_scRNA.r

Step two details

Combining data. We combined all cells belonging to each main cell type from all samples and patients (all 10x libraries) as well as their clustering information obtained via graph clustering in step one. Then, for each main cell type the following steps were followed.

Data integration and iterative clustering.

We have previously observed that marker genes or differentially expressed genes identified after cell clustering can often differ from those used as a feature set input o the clustering procedure[50]. It is also generally established that clustering results will vary depending on the input feature set. Therefore, we followed an iterative clustering approach that cyclically refines the variable gene set that is input to the clustering procedure, the technical effect mitigation parameters and the cell cluster assignments. In detail, the algorithm consists of the following steps: (1) Given the clustering obtained from step one, we define highly variable features based on gene specificity ranking per cluster using the sortGenes function in the genesorteR R package setting binarizeMethod to 'naive' (see above). We use the combined set of the top 500 genes in each cluster as highly variable genes. (2) Technical effects were removed using the mutual nearest neighbour (MNN) method[51] as implemented in the fastMNN function of the batchelor R package[35,51], setting the number of dimensions to 30 and auto.order to TRUE. This method removes technical differences while retaining differences due to cell types and returns reduced dimensions directly. (3) Cells were clustered based on the reduced dimensions returned by fastMNN. The clustering approach is similar to that followed for clustering in step one except that we use the Louvain algorithm, a widely used algorithm for community detection on graphs and for single-cell clustering[52]. To control the resolution at which the clustering occurs, we define the average number of k-nearest neighbours used to construct the graph as r.squareroot(n) and vary r between 1 and 0.01. We select the r that returns the most informative clustering as determined using the getClassAUC function from the genesorteR R package[50]. This function defines clustering quality by an internal evaluation procedure, and expresses clustering quality as a function of the specificity of the marker genes in each cell cluster. The number of nearest neighbours that produces the clustering with highest average class AUC is selected. (4) Raw gene expression counts (UMI counts) were normalized using the deconvolution strategy for scaling normalization[53] as implemented in the computeSumFactors function in the Scran R package[35], setting the clusters argument to the cluster labels obtained from (3). We repeated steps (1)–(4) until there was no longer any appreciable increase in agreement in cell cluster assignments between consecutive iteration, quantified by the slope of change of the adjusted rand index[54]. We noticed that this algorithm resulted in a progressive increase in the rand index (between cluster assignments in the i-th iteration and those in the i-1-iteration) and increase in class AUC value measured by the genesorteR getClassAUC function. Typically, no more than three iterations were needed. An approach to refine the variable gene list and cell clustering was previously proposed[38,55].

Cluster quality control

We determine low-quality cell clusters as those with no differentially expressed genes at a P value cut-off of 0.05, as determined by the getPValues function from genesorteR R package[50], or those whose differentially expressed genes are dominated by ribosomal proteins or genes typically known as housekeeping genes (such as B2M, GAPDH). We also controlled for potential doublet clusters based on marker gene expression. For example, if a cell cluster expresses both EPCAM (epithelial marker) and PTPRC (encoding CD45, immune marker) at high levels simultaneously, we assume it may represent an epithelial cell or immune cell doublet. This is a similar approach to that previously described56. We repeated the clustering procedure again after this cell removal.

Scripts and metadata. Scripts for data integration and clustering is provided here: https://github.com/mahmoudibrahim/KidneyMap/ blob/master/templates/clusterCells.r

Step three details

Integrated maps. Integrated maps were generated by combining the clustering results (step two), patient or mouse metadata and cell expression (UMI count) information as detailed below. For the whole kidney CD10+ or CD10– data, we generated two maps accordingly.

The CD10^{-} map contained all epithelial, immune, endothelial, mesenchymal and neuronal cells, whereas the CD10^{+} combined all epithelial CD10^{+} sorted cells. PDGFRβ^{+} data were analysed separately from CD10^{+} or CD10 data. We generated one integrated map comprising all cells from all PDGFRβ^{+} libraries.

Cluster merging and filtering. We first removed genes that were detected in less than 0.1% of all cells (that is, at least in 1 out of every 1,000 cells) given the full integrated map, and used the remaining genes to produce gene specificity ranking per cell cluster using the sortGenes function from the genesorteR R package setting binarizeMethod to 'naive'. Clusters that shared more than 80 out of the top 100 specific genes were merged. We have experimented with different ways to merge similar clusters, and this was our choice as a conservative method that tended to maintain different cell states and merge only very highly similar clusters. Despite our efforts to remove low-quality droplets during cell filtering and low-quality clusters in step two, we still noticed the possibility of observing low-quality clusters given the entire integrated map. Therefore, having merged the cell clusters, we checked cell clusters for differential expression using the getPValues function in the genesorteR R package setting numPerm to 20 and removed cell clusters with no differentially expressed genes. Those were consistently low-quality cells with lower transcript capture rate overall. For the PDGFRβ^{+} data, we also removed cell clusters in which PDGFRβ was detected in less than 1 median absolute deviation of its expression in all cell clusters (calculated cut-off was: 4% of cells in the cluster); those were immune and epithelial cell clusters. After removing those cell clusters, we reformed an expression matrix containing all possible genes and performed gene filtering again (see above). We normalized gene expression over the full integrated map using the computeSumFactor function from the Scran R package[35] using the clustering information from step two.

Scripts and metadata

Scripts for combining data into full integrated maps and producing all subsequent plots are available here: https:// github.com/mahmoudibrahim/KidneyMap/tree/master/make_intergrated_maps. Details for various analyses are described below. Overall, this approach was biologically informed, and allowed us to correct for potential technical effects during cell clustering such that almost all cell clusters contained cells from more than one patient/ library, while preserving interesting differences between patients such as diseased cell states (for example injured proximal tubule cells), differences in (myo) fibroblast states and differences in ECM expression.

Mouse 10x single-cell data integration strategy

Mouse 10x data were analysed and integrated as described for human data. The script used to produce the integrated map is available here: https://raw.githubusercontent.com/mahmoudibrahim/KidneyMap/master/make_intergrated_maps/mouse_PDGFRABpositive.r

Mouse Smart-Seq2 single-cell data integration strategy

Because single-cell plate sorting was performed such that cells from all three time points were equally represented in all plates, no further batch effect mitigation was performed during the analysis. Variable genes were determined using the Scran R package decomposeVar function, after running the trendVar function on the ERCC transcripts[35]. Genes with a false discovery rate (FDR) value < 0.01 and biological variance component > 1 were kept as highly variable genes. Using those variable genes, we followed the same clustering approach as described for the 10x Chromium data, but we ran only two clustering iteration and did not vary the number of nearest neighbours. Script used for analysis of mouse Smart-Seq2 data are available here: https://github.com/ mahmoudibrahim/KidneyMap/blob/master/make_intergrated_maps/mouse_PDGFRBpositive.r.

Cluster annotation

A gene ranking per cluster was produced using the sortGenes function in the genesorteR R package[50] setting binarizeMethod to 'adaptiveMedian' (Smart-Seq2 data) or to 'naive' (10x data). We then annotated our highly resolved cell clusters manually based on previous knowledge and information from literature. We refer to this annotation as 'level 3 annotation' in the Supplementary Information. There were 50 such clusters in CD10$^-$ data, 7 clusters in CD10$^+$ data, 26 clusters in PDGFRβ$^+$ human data, 10 clusters in mouse Smart-Seq2 data and 10 clusters in mouse PDGFRα$^+$PDGFRβ$^+$ data. At that highly resolved level (level 3), a cell cluster can represent either a bona fide cell type or a different cell state. Thus, we also grouped those highly resolved cell clusters into canonical cell types based on our annotation. This resulted in 29 cell types in the CD10$^-$ map, 1 cell type in the CD10$^+$ map, 16 cell types in the PDGFRβ$^+$ map, 5 cell types in the mouse PDGFRα$^+$PDGFRβ$^+$ map and 6 cell types in the Smart-Seq2 mouse PDGFRβ$^+$ map. We refer to this cell grouping as 'level 2 annotation' in the Supplementary Information. We then further annotated the cell clusters as epithelial, endothelial, mesenchymal, immune or neuronal for plot and figure annotation to allow easier data interpretation.

UMAPs and diffusion maps

Integrated full-map UMAP[57] projections (Figs. 1–5) were generated via the UMAP Python package (https://github.com/lmcinnes/umap) on the reduced corrected dimensions returned from fastMNN setting min_dist to 0.6 and the number of neighbours to square root the number of cells. Local UMAP projections (Figs. 1, 4, Extended Data Fig. 5) were produced setting min_dist to 1, as those parameters tend to produce more geometrically accurate embeddings (see https://umap-learn. readthedocs.io/en/latest/). Diffusion maps were produced using the destiny R package (https://github.com/theislab/destiny) also using the reduced dimensions returned from fastMNN as input and setting the number of neighbours to square root the number of cells. We tested various randomization seeds for UMAP and diffusion map and various diffusion map distance metrics (as recommended in the destiny R package manual) and confirmed that no qualitative difference occurs in the resulting single cell projections.

Lineage trees or trajectories and pseudotime

The Slingshot R package[58] was used for lineage tree inference and pseudotime cell ordering inference based on the UMAP/diffusion map projection. The cell clustering (see 'Step two details') was used as input cell clusters. Start and end clusters were chosen based on reasonable expectation given our prior knowledge as discussed and described previously[58] (for example, myofibroblast is the end cluster in a pericyte/ fibroblast/ myofibroblast map).

Gene dynamics along pseudotime

Genes with expression that varied with cell ordering were defined as those in which normalized expression correlated with cell ordering as quantified by the Spearman correlation coefficient at a Bonferroni–Hochberg corrected *P* value cutoff of 0.001. Gene clusters and expression heat maps (for example, Fig. 2f, top) were produced by ordering cells along the pseudotime predicted by SlingShot and using the genesorteR function plotMarkerHeat. This function clusters genes using the *k*-means algorithm, and we set the plot and clustering to average every 10 cells along pseudotime. Pathway enrichment and cell cycle analyses were calculated by grouping every 2,000 cells along pseudotime.

Pathway enrichment and GO analysis

For the single-cell data, we used KEGG pathway and PID pathway data downloaded in November 2019 from MSigDB 3[59,60] as '.gmt' files. Pathway enrichment analysis was performed using the clusterProfiler R package[61] using the top 100 genes for each cell cluster/group as defined by the sortGenes function from the genesorteR package. The enricher function was used setting minGSSize to 10 and maxGSize to 200. The top five terms by *q* value for each cell cluster or group were plotted as heat maps of $-\log_{10}(q$ value). GO biological process[62] analysis was performed on the top 200 genes via the same method. The enricher function was used setting minGSSize to 100 and maxGSize to 500. To compare pathway activity between NKD2$^+$ and NKD2$^-$ mesenchymal cells, we used PROGENy to estimate the activity of 14 pathways in a single-cell basis[22,63], using the top 500 most responsive genes from the model as it is recommended from a benchmark study[63].

Cell cycle analysis

Cell cycle analysis was done following the previously described method[64] and explained in the tutorial by P.-Y. Tung (https://jdblischak.github.io/singleCellSeq/analysis/cell-cycle.html), using normalized gene expression as input and setting the gene correlation value to 0.1. We used cell cycle gene sets previously provided[65]. To quantify enrichment/depletion of single-cell cycle assignments (Fig. 1g), we plot the $\log_2$-transformed fold-change of those frequencies relative to the average frequency obtained by randomizing the true frequency matrix 1,000 times while keeping row and column sums constant. Randomization was performed using the R package Vegan (https://CRAN.R-project.org/package=vegan). Positive numbers indicate enrichment relative to what would be expected by chance, negative numbers indicate depletion.

ECM and collagen score

The expression of core matrisome genes previously described[7] were summarized based on normalized gene expression data using the same method used for cell cycle analysis. Also see Extended Data Fig. 2g–u.

Gene expression heat maps

Scaled gene expression heat maps such as those in Fig. 2d were produced using the plotMarkerHeat and plotTopMarkerHeat functions in the genesorteR R package[50]. The fraction of expressed cells heat maps such as Fig. 3d were produced using plotBinaryHeat function from the genesorteR R package. All other heat maps were produced using ComplexHeatmap R package (v.2.4.2)[66].

ATAC–seq analysis

Illumina Tn5 adaptor sequences were trimmed from ATAC–seq reads using bbduk command from BBmap suite (v.38.32, settings: trimq = 18, k = 20, mink = 5, hdist = 2, hdist2 = 0)[67]. STAR (v.2.7.0e) was used to map ATAC–seq reads to the mm10 genome assembly retaining only uniquely mapped pairs (settings: alignEndsType EndToEnd, alignIntronMax 1, alignMatesGapMax 2000, alignEndsProtrude 100 ConcordantPair, outFilterMultimapNmax 1, outFilterScoreMinOverLread 0.9, outFilterMatchNminOverLread 0.9)[68]. The Picard MarkDuplicates command (v.2.18.27) was used to remove sequence duplicates (settings: remove_duplicates = TRUE, http://broadinstitute.github.io/picard/). Non-concordant read pairs were then removed from the BAM file using Samtools (v.1.3.1)[69]. bedtools (v.2.17.0) was used to convert BAM files to BED files and to extend each read to 15 bp upstream and 22 bp downstream from the read 5′-end in a stranded manner[70], to account for steric hindrance of Tn5-DNA contacts[71]. JAMM (v.1.0.7rev5) was used to identify open regions from the final BED files keeping the two replicates separate, retaining peaks that were at least 50 bp in width in the all list for further analysis (parameters: -r peak,

-f 38,38, -e auto, -b 100)[72]. ATAC–seq signal bigwig files were produced using JAMM SignalGenerator pipeline (settings: -f 38,38 -n depth). To deconvolute ATAC–seq signal from bulk ATAC–seq data according to scRNA-seq clustering, we followed the following strategy. To deconvolute the ATAC–seq signal three main steps in the data analysis were taken: (1) each open chromatin peak (where transcription factors are expected to bind DNA) was first assigned to a specific gene.

(2) These genes were ranked per scRNA-seq cluster (Fib, MF1/2 etc) depending on their expression in the scRNA-seq dataset. (3) The top 2,000 ATAC peaks were used to identify enriched transcription factor motif sequences. In more detail, each open chromatin ATAC–seq peak was assigned to a gene according to its closest annotated transcription start site using the bedtools closest function, setting 100 kb as the maximum possible assignment distance. ATAC–seq peak ranking per scRNA-seq cluster was obtained by ranking the peaks according to the ranking of their assigned gene in the single cell RNA-seq cluster. The top 2,000 ATAC– seq peaks for each scRNA-seq cluster were selected and XXmotif[73] was used for de novo motif finding for each scRNA-seq cluster open chromatin regions separately (settings:–revcomp–merge-motif-threshold MEDIUM). We kept only motifs whose occurrence was more than 5%, as defined by XXmotif, for further analysis. Motif occurrence from all motifs from all 4 scRNA-seq clusters were quantified using FIMO[74] with default parameters (MEME v.5.0.1) in the peaks assigned to the top 200 genes in each scRNA-seq cluster. This produced a frequency matrix of motif occurrence in scRNA-seq clusters. To quantify enrich-ment/depletion of motif occurrence in scRNA-seq clusters we plot the $\log_2$-transformed fold change of those frequencies relative to the average frequency obtained by randomizing the true frequency matrix 1,000 times while keeping row and column sums constant. Randomization was performed using the R package Vegan (https://CRAN.R-project.org/package=vegan). Positive numbers indicate enrichment relative to what would be expected by chance, negative numbers indicate depletion (see Fig. 4k). We selected *Irf8*, *Nrf1*, *Creb5*, *Atf3*, *Elf* or *Ets* transcription factor family and *Klf2* or *Klf5* for further investigation. We plotted the signal from all peaks that contained those motifs using DeepTools v.3.3.1[75], using the bigwig file generated by JAMM as input (see above and Supplementary Fig. 11). We visualized the same bigwig file and motif occurrence in the Integrative Genomics Viewer[76] (v.2.4.10, Supplementary Fig. 11).

Other visualization analysis

Heat maps that do not quantify gene expression were produced using the heatmap2 function in the gplots R package (https://CRAN.R-project. org/package=gplots). Violin plots were produced using the vioplot R package (https://CRAN.R-project.org/package=vioplot).

Quantification and statistical analysis used outside of the single-cell sequencing data

Data are presented as mean ± s.e.m. if not specified otherwise. Comparison of two groups was performed using unpaired *t*-test. For multiple group comparison, one-way ANOVA with Bonferroni's multiple comparison test was applied or two-way ANOVA with Sidak's multiple comparisons test. Statistical analyses were performed using GraphPad Prism 8 (GraphPad Software). $P < 0.05$ was considered significant.

Gene regulatory network analysis

Gene expression was l1-scaled per gene and the Pearson correlation coefficient was calculated between *Nkd2* and all other genes along pericyte, fibroblast and myofibroblast single cells. The top 100 correlating and top 100 anti-correlating genes were selected for pathway enrichment analysis. Furthermore, the expression of those 200 genes along single cells was used as input to GRNboost2+ python package to predict putative regulatory links between genes. The output network was filtered by removing connections with strength ≤ 10. The resulting network was plotted as an undirected network (because regulators are not known beforehand) using ggraph package (https:// cran.r-project.org/web/packages/ggraph/index.html) and clustered into 4 modules using the Louvain algorithm as implemented in the igraph package.

Transcription factor predictions from single-cell data

To obtain transcription factor scores in distal and proximal regions, we used the top 200 marker genes for fibroblast, pericyte and myofibroblast cell clusters as input gene lists to RCisTarget[77]. We followed the RCisTarget Vignette to perform the analysis with default parameters (available at https://bioconductor.org/packages/release/bioc/ vignettes/RcisTarget/inst/doc/RcisTarget.html). To quantify AP-1 expression, we used all *Jun* and *Fos* genes as a geneset and applied the same method to obtain an AP-1 score as we did for ECM score. To quantify AP-1 activity (defined as the expression of putative target genes[78,79], we defined AP-1 target genes according to the Dorothea regulon database[63,80] and applied the same method as ECM score to obtain a single cell AP-1 activity score.

Mouse supervised cell classification

We classified single cells in the mouse PDGFRα$^+$PDGFRβ$^+$ dataset using the human PDGFRβ$^+$ dataset as a reference using the CHETAH algorithm with default parameters[81]. Human gene symbols were converted to mouse gene symbols using the biomaRt database[43].

CellphoneDB analysis

CellPhoneDB (v.2.1.1) was used to estimate cell–cell interactions among the cell types found in the human CD10$^-$ fraction using the version 2.0.0 of the database[82], and the normalized gene expression as input, with default parameters (10% of cells expressing the ligand/ receptor). Interactions with $P < 0.05$ were considered significant. We consider only ligand–receptor interactions based on the annotation from the database, for which only and at least one partner of the interacting pair was a receptor, thus discarding receptor–receptor and other interactions without a clear receptor. Ligand–receptor interactions from pathways involved in kidney fibrosis were selected using the membership from KEGG database for Hedgehog, Notch, TGFβ and WNT signalling, and REACTOME database for EGFR signalling from MSigDB 3[59,60], and manual curation for PDGF signalling.

Bulk RNA-seq data analysis

Gene expression was quantified at the transcript level using Salmon v1.1.0, with the–validatMappings and–gcBias parameters switched on, to the human Gencode v29 transcriptome. Transcript level counts were aggregated to gene level counts using the import in tximport R package, setting countsFromAbundance to 'lengthScaledTPM'[83]. Limma R package (v.3.44.1) was used to test for differential gene expression between *NKD2*-perturbed human kidney PDGFRβ$^+$ as compared to controls using the empirical Bayes method after voom transformation[84]. We found that two out of the three clones of CRISPR–Cas9 *NKD2* knockout group together in the principal component analysis and exhibited a shallow phenotype, whereas the third clone grouped independently and presented a more severe phenotype. Thus, we grouped the first two clone knockouts, to have two independent knockout conditions for the statistical contrasts. Differentially expressed genes were ranked by the moderated *t*-statistic from the statistical test for pathway and GO analysis. *P* values were adjusted for multiple testing using Benjamini and Hochberg method. Genes and pathways with FDR < 0.05 were considered significant.

For pathway and GO analysis, we also used clusterProfiler R package with KEGG and PID pathways using genes with adjusted $P < 0.01$ in the *NKD2*-perturbed cells as compared to the control and absolute log-transformed fold change higher than 1 for knockout comparison (higher than 0 for overexpression comparison) with a maximum of 200 genes, ranked by the adjusted *P* value. We used GSEA-preranked to test for an enrichment of ECM genes in the phenotypes using fgsea R package (v.1.14.0)[85], with MatrisomeDB gene set collection[7].

Statistics and reproducibility

Data are presented as mean ± s.e.m. if not specified otherwise in the legends. Unless otherwise stated, statistical significance was assessed by a two-tailed Student's *t*-test or one ANOVA with Bonferroni's multiple comparison with $P < 0.05$ being considered statistically significant. Statistical analyses were performed using GraphPad Prism 8 (GraphPad Software) or as described in above. Results are presented in dot plots, with dots representing individual values, violin plots (horizontal line indicates the median, the box indicates the span of the 25% to the 75% percentiles, whiskers extend to maximum 1.5× this interquartile range) and Tukey box and whisker plots (horizontal line indicates the median, the box indicates the span of the 25% to the 75% percentiles, whiskers extend to maximum and minimum values). The number of samples for each group was chosen on the basis of the expected levels of variation and consistency. The depicted RNAscope, immunofluorescence micrographs and western blot micrographs are representative. All studies were performed at least twice, and all repeats were successful.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Processed data for all human and mouse RNA-seq and ATAC–seq libraries produced in this study are available at the Zenodo data archive (https://zenodo.org/record/4059315, https://doi.org/10.5281/ zenodo.4059315). Processed and raw mouse data are available via the Gene Expression Omnibus (GEO) under the accessions GSE145173 (for mouse PDGFRαβ scRNA-seq and ATAC–seq) and GSE144528 (for mouse PDGFRβ Smart-Seq.). Source data are provided with this paper.

Code availability

Custom scripts used in single cell and bulk RNA-seq data analysis are available at: https://github.com/mahmoudibrahim/KidneyMap.
Scripts used for imaging in-situ hybridization data quantification are available at: https://gitlab.com/mklaus/segment_cells_register_marker.

References

1. Duffield, J. S. Cellular and molecular mechanisms in kidney fibrosis. J. Clin. Invest. 124, 2299–2306 (2014).
2. Kramann, R., DiRocco, D. P. & Humphreys, B. D. Understanding the origin, activation and regulation of matrix-producing myofibroblasts for treatment of fibrotic disease. J. Pathol. 231, 273–289 (2013).
3. Friedman, S. L., Sheppard, D., Duffield, J. S. & Violette, S. Therapy for fibrotic diseases: nearing the starting line. Sci. Transl. Med. 5, 167sr1 (2013).
4. Falke, L. L., Gholizadeh, S., Goldschmeding, R., Kok, R. J. & Nguyen, T. Q. Diverse origins of the myofibroblast—implications for kidney fibrosis. Nat. Rev. Nephrol. 11, 233–244 (2015).
5. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 361, 594–599 (2018).
6. Kang, H. M. et al. Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. Nat. Med. 21, 37–46 (2015).
7. Naba, A. et al. The extracellular matrix: Tools and insights for the "omics" era. Matrix Biol. 49, 10–24 (2016).
8. Fan, Y. et al. Comparison of kidney transcriptomic profiles of early and advanced diabetic nephropathy reveals potential new mechanisms for disease progression. Diabetes 68, 2301–2314 (2019).
9. Kriz, W., Kaissling, B. & Le Hir, M. Epithelial-mesenchymal transition (EMT) in kidney fibrosis: fact or fantasy? J. Clin. Invest. 121, 468–474 (2011).
10. Huang, S. & Susztak, K. Epithelial plasticity versus EMT in kidney fibrosis. Trends Mol. Med. 22, 4–6 (2016).
11. Elices, M. J. et al. VCAM-1 on activated endothelium interacts with the leukocyte integrin VLA-4 at a site distinct from the VLA-4/fibronectin binding site. Cell 60, 577–584 (1990).
12. Kang, H. M. et al. Sox9-positive progenitor cells play a key role in renal tubule epithelial regeneration in mice. Cell Rep. 14, 861–871 (2016).
13. Ramachandran, P. et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. Nature 575, 512–518 (2019).
14. Wang, Y.-Y. et al. Macrophage-to-myofibroblast transition contributes to interstitial fibrosis in chronic renal allograft injury. J. Am. Soc. Nephrol. 28, 2053–2067 (2017).
15. Henderson, N. C. et al. Targeting of αv integrin identifies a core molecular pathway that regulates fibrosis in several organs. Nat. Med. 19, 1617–1624 (2013).
16. Wernig, G. et al. Unifying mechanism for different fibrotic diseases. Proc. Natl Acad. Sci. USA 114, 4757–4762 (2017).
17. Venkatachalam, M. A., Weinberg, J. M., Kriz, W. & Bidani, A. K. Failed tubule recovery, AKI-CKD transition, and kidney disease progression. J. Am. Soc. Nephrol. 26, 1765–1776 (2015).
18. Kramann, R. et al. Parabiosis and single-cell RNA sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. JCI Insight 3, e99561 (2018).
19. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–100 (2012).
20. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
21. Palumbo-Zerr, K. et al. Orphan nuclear receptor NR4A1 regulates transforming growth factor-β signaling and fibrosis. Nat. Med. 21, 150–158 (2015).
22. Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. 9, 20 (2018).
23. Zhao, S. et al. NKD2, a negative regulator of Wnt signaling, suppresses tumor growth and metastasis in osteosarcoma. Oncogene 34, 5069–5079 (2015).
24. Li, C. et al. Myristoylated Naked2 escorts transforming growth factor α to the basolateral plasma membrane of polarized epithelial cells. Proc. Natl Acad. Sci. USA 101, 5571–5576 (2004).
25. Moerman, T. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics 35, 2159–2161 (2019).

26. Lemos, D. R. et al. Interleukin-1β activates a MYC-dependent metabolic switch in kidney stromal cells necessary for progressive tubulointerstitial fibrosis. J. Am. Soc. Nephrol. 29, 1690–1705 (2018).

27. Tsukui, T. et al. Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis. Nat. Commun. 11, 1920 (2020).

28. Adams, T. S. et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci. Adv. 6, eaba1983 (2020).

29. Kramann, R. et al. Perivascular Gli1+ progenitors are key contributors to injury-induced organ fibrosis. Cell Stem Cell 16, 51–66 (2015).

30. Takasoto, M. et al. Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. Nature 526, 564–568 (2015).

31. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962 (2017).

32. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol. 20, 65 (2019).

33. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).

34. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47 (D1), D766–D773 (2019).

35. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000 Res. 5, 2122 (2016).

36. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J. 8, 289–317 (2016).

37. Fortunato, S. & Barthélemy, M. Resolution limit in community detection. Proc. Natl Acad. Sci. USA 104, 36–41 (2007).

38. Zeisel, A. et al. Molecular architecture of the mouse nervous system. Cell 174, 999–1014. e22 (2018).

39. 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).

40. 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).

41. Wu, C. et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 10, R130 (2009).

42. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21, 3439–3440 (2005).

43. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191 (2009).

44. Stuart, T. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21 (2019).

45. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 31, 1974–1980 (2015).

46. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).

47. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl Acad. Sci. USA 97, 10101–10106 (2000).

48. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. Proc. Natl Acad. Sci. USA 105, 1118–1123 (2008).

49. Dahlin, J. S. et al. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. Blood 131, e1–e11 (2018).

50. Ibrahim, M. M. & Kramann, R. genesorteR: feature ranking in clustered single cell data. Preprint at https://doi.org/10.1101/676379 (2019).

51. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421–427 (2018).

52. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

53. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 17, 75 (2016).

54. Rand, W. M. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66, 846–850 (1971).

55. Yang, L., Liu, J., Lu, Q., Riggs, A. D. & Wu, X. SAIC: an iterative clustering approach for analysis of single cell RNA-seq data. BMC Genomics 18 (Suppl. 6), 689 (2017).

56. Karaiskos, N. et al. The Drosophila embryo at single-cell transcriptome resolution. Science 358, 194–199 (2017).

57. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).

58. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 19, 477 (2018).

59. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).

60. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740 (2011).

61. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–287 (2012).

62. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).

63. Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 21, 36 (2020).

64. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).

65. Yang, J. et al. Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors. Nucleic Acids Res. 45, 1281–1296 (2017).

66. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016).

67. Bushnell, B. BBMap: a Fast, Accurate, Splice-Aware Aligner https://www.osti.gov/ biblio/1241166 (2014).

68. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

69. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

70. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.1–11.12.34 (2014).

71. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11, R119 (2010).

72. Ibrahim, M. M., Lacadie, S. A. & Ohler, U. JAMM: a peak finder for joint analysis of NGS replicates. Bioinformatics 31, 48–55 (2015).

73. Luehr, S., Hartmann, H. & Söding, J. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. Nucleic Acids Res. 40, W104–W109 (2012).

74. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).

75. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44 (W1), W160–W165 (2016).

76. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).

77. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086 (2017).

78. Schacht, T., Oswald, M., Eils, R., Eichmüller, S. B. & König, R. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics 30, i401–i407 (2014).

79. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. 48, 838–847 (2016).

80. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 29, 1363–1375 (2019).

81. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res. 47, e95 (2019).

82. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand– receptor complexes. Nat. Protoc. 15, 1484–1506 (2020).

83. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000 Res. 4, 1521 (2015).

84. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).

85. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. Preprint at https://doi.org/10.1101/060012 (2019).

Affiliations

[1] Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, 52074, Aachen, Germany.

[2] Institute of Experimental Medicine and Systems Biology, RWTH Aachen University Medical School, 52074 Aachen, Germany.

[3] Division of Nephrology and Clinical Immunology, RWTH Aachen University, 52074 Aachen, Germany.

[4] Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, 52074 Aachen, Germany.

[5] Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Aachen, Germany.

[6] Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, 3015GD Rotterdam, The Netherlands.

[7] These authors contributed equally: Zhijian Li, Christoph Kuppe.

4

Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen

Zhijian Li[1,7], **Christoph Kuppe**[2,3,7], Susanne Ziegler[2], Mingbo Cheng[1], Nazanin Kabgani[2], Sylvia Menzel[2], Martin Zenke[4,5], Rafael Kramann [2,3,6]& Ivan G. Costa[1]

Abstract

A major drawback of single-cell ATAC-seq (scATAC-seq) is its sparsity, i.e., open chromatin regions with no reads due to loss of DNA material during the scATAC-seq protocol. Here, we propose scOpen, a computational method based on regularized non-negative matrix factorization for imputing and quantifying the open chromatin status of regulatory regions from sparse scATAC-seq experiments. We show that scOpen improves crucial downstream analysis steps of scATAC-seq data as clustering, visualization, *cis*-regulatory DNA interactions, and delineation of regulatory features. We demonstrate the power of scOpen to dissect regulatory changes in the development of fibrosis in the kidney. This identifies a role of Runx1 and target genes by promoting fibroblast to myofibroblast differentiation driving kidney fibrosis.

Introduction

The simplicity and low cell number requirements of the assay for transposase-accessible chromatin using sequencing (ATAC-seq)[1] made it the standard method for detection of open chromatin (OC), enabling the first study of OC of cancer cohorts[2]. Moreover, careful consideration of digestion events by the enzyme Tn5 allowed insights on regulatory elements such as positions of nucleosomes[1-3], transcription factor (TF) binding sites, and the activity level of TFs[4]. The combination of ATAC-seq with single-cell sequencing (scATAC-seq)[5] further expanded ATAC-seq applications by measuring the OC status of thousands of single cells from healthy[6-7] and diseased tissues[8]. Computational tasks for analysis of scATAC-seq include detection of cell types with clustering (scABC[9], cisTopic[10], SnapATAC[11]); identification of TF regulating individual cells (chromVAR[12]); and prediction of co-accessible DNA regions in groups of cells (Cicero[13]).

Usually, the first step for analysis of scATAC-seq data is the detection of OC regions by calling peaks on the scATAC-seq library by ignoring cell information. Next, a matrix is built by counting the number of digestion events per cell in each of the previously detected regions. This matrix usually has a very high dimension (up to >10^6 regions) and a maximum of two digestion events are expected for a region per cell. As with scRNA-seq[14-16], scATAC-seq is affected by dropout events due to the loss of DNA material during library preparation. These characteristics render the scATAC-seq count matrix sparse, i.e. 3% of non-zero entries. In contrast, scRNA-seq have less severe sparsity (>10% of non-zeros) than scATAC-seq due to smaller dimension (< 20,000 genes for mammalian genomes) and lower dropout rates for genes with high or moderate expression levels. This sparsity poses challenges in the identification of cell-specific OC regions and is likely to affect downstream analysis as clustering and detection of regulatory features. Although several computational methods have been developed to address this issue for scRNA-seq data (e.g., MAGIC[14], scImpute[17], DCA[18], and SAVER[19]), these methods were not designed to deal with the sparse and low count nature of scATAC-seq data. Until date, there are only two approaches for imputation methods for scATAC-seq data e.g., SCALE[20] and cisTopic[10]. SCALE, which is based on deep learning, requires a graphics processing unit (GPU) for training. The usual small size of GPU memory limits the number of cells to be analyzed. cisTopic is a Bayesian-based method, which was reported to have an exponential increase of the running time for an increasing number of reads[21]. Therefore, both approaches are likely to have scalability issues with large data sets.

We here present scOpen, an unsupervised learning model for scATAC-seq data imputation. It estimates accessibility scores to indicate if a region is open in a particular cell. scOpen is based on a non-negative matrix factorization (NMF), which makes no assumption on the data distribution as SCALE or cisTopic. It also includes a regularization, which makes it less prone to overfitting. To speed up the learning, we make use of a cyclic coordinate descent (CCD) algorithm. Moreover, we adopt an elbow detection approach to automatically determine the number of dimensions of the input data. The imputed matrix can be used as input for usual computational methods of scATAC-seq data as clustering, visualization, and prediction of DNA interactions (Fig. 1a). We demonstrate the power of scOpen on a comprehensive benchmarking analysis using publicly available scATAC-seq data with true labels. Moreover, we use scOpen together with HINT-ATAC[4] footprinting analysis to infer regulatory networks driving the development of fibrosis with a scATAC-seq time-course dataset of 31,000 cells in murine kidney fibrosis, identifying Runx1 as a regulator of myofibroblast differentiation.

Results

OC estimation with scOpen

scOpen performs imputation and denoising of a scATAC-seq matrix via a regularized NMF based on a binarized scATAC-seq cell count matrix, where features represent OC regions which are obtained by peak calling based on aggregated scATAC-seq profiles. This matrix is transformed using the term frequency-inverse document frequency (TF-IDF), which weighs the importance of an OC region to a cell. Next, it applies a regularized NMF using a coordinate descent algorithm[22]. In addition, it provides a computational approach to optimize the dataset-specific rank k of the NMF approach based on a knee detection method[23]. scOpen provides as results imputed and reduced dimension matrices, which can be used for distinct downstream analysis as visualization, clustering, inference of regulatory players, and *cis*-regulatory DNA interactions (Fig. 1a).

First, we made use of simulated scATAC-seq similar as in ref. 21 to evaluate the parameterization of two hyper-parameters of scOpen, i.e., the rank k and the regularization term λ (see the “Methods” section; Supplementary Fig. 1a–d). Results indicate that the scOpen automatic procedure for rank selection obtains close to optimal results, i.e. selected rank had similar accuracy than best ranks for both imputation and clustering problems. Regarding λ, a value of 1 is optimal in the imputation problem, where values in the range [0, 1] were optimal for the clustering problem. This indicates the importance of the regularization parameter in scATAC-seq data imputation. The $\lambda = 1$ and the rank selection strategy are used as default by scOpen.

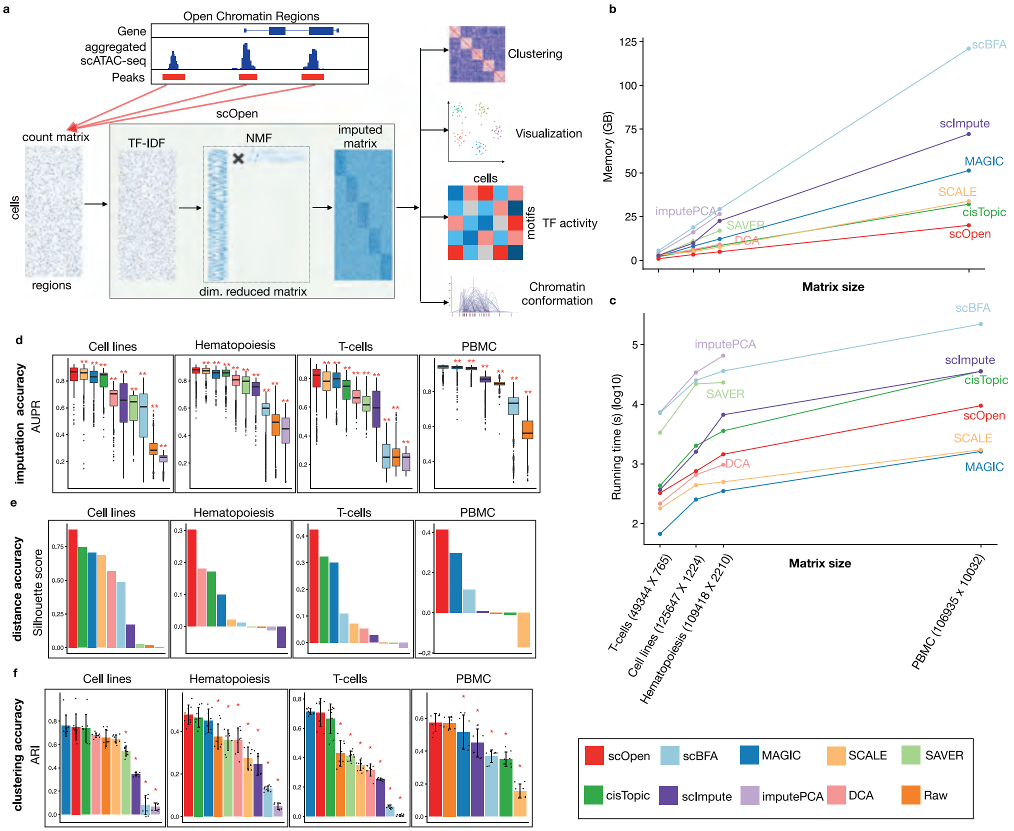


Fig.1 a scOpen receives as input a sparse peak by cell count matrix.

After matrix binarization, scOpen performs TF–IDF transformation followed by NMF for dimension reduction and matrix imputation. The imputed or reduced matrix can then be given as input for scATAC-seq methods for clustering, visualization, and interpretation of regulatory features. **b** Memory requirements of imputation/denoising methods on benchmarking datasets. The *x*-axis represents the number of elements of the input matrix (number of OC regions by cells). **c** Same as **b** for running time requirements. **d** Boxplot showing the evaluation of imputation/denoising methods for recovering true peaks. The *y*-axis indicates the area under the precision-recall curve (AUPR). Methods are ranked by the mean AUPR. The asterisk and the two asterisks mean that the method is outperformed by the top-ranked method (scOpen) with significance levels of 0.05 and 0.01 at a confidence level of 0.95 (Wilcoxon Rank Sum test, paired, two-sided), respectively (*n* = 1224 cells for Cell lines, *n* = 2210 cells for Hematopoiesis, *n* = 765 cells for T-cells, and *n* = 10,032 for PBMC). The box plot represents the median (central line), first and third quartiles (box bounds). The whiskers present the 1.5 interquartile range (IQR) and external dots represent outliers (data greater than or smaller than 1.5IQR). **e** Barplots showing silhouette score (*y*-axis) for benchmarking datasets. **f** Barplots showing the clustering accuracy for distinct imputation methods. The *y*-axis indicates the mean adjusted Rand Index (ARI). Dots represent individual ARI values of distinct clustering methods. Error bars represent the standard deviation (SD) of ARI. Data are represented as mean ± SD. The asterisk and the two asterisks mean that the method is outperformed by the top-ranked method with significance levels of 0.05 and 0.01 at a confidence level of 0.95 (*n* = 8 independent clustering experiments, Wilcoxon Rank Sum test, paired, two-sided), respectively. Source data for Fig. 1 are provided as a Source Data file.

Benchmarking of scOpen for imputation of scATAC-seq

For benchmarking, we made use of four public scATAC-seq data sets: cell lines[5], human hematopoiesis composing of eight cell types[6], four sub-types of T cells[8], and a multi-omics RNA-ATAC from peripheral blood mononuclear cells (PBMCs) with 14 cell types (see the "Methods" section). These datasets were selected due to the presence of external labels, which were defined independently of the scATAC-seq at hand. After processing, we generated a count matrix for each dataset and detected 50k to 120k OC regions with 3–7% of non-zero entries, confirming the sparsity of scATAC-seq data (Supplementary Table 1). For comparison, we selected top-performing imputation/denoising methods[24] proposed for scRNA-seq (MAGIC[14], SAVER[19], scImpute[17], DCA[18], and scBFA[25]); two scATAC-seq imputation methods (cisTopic[10] and SCALE[20]); a PCA-based imputation method (imputePCA[26]); and the raw count matrix (Supplementary Fig. 2a).

We first evaluated the time and memory requirement of imputation methods (see the "Methods" section). scOpen had the overall lowest memory requirements, i.e it required at least 2 fold less memory as compared to cisTopic, MAGIC, or SCALE (Fig. 1b) and had a maximum requirement of 16 GB on the PBMC dataset (Supplementary Data 1). Regarding computing time, MAGIC was the fastest followed by SCALE and scOpen. These were the only methods performing the imputation of the large PBMCs dataset (10k cells vs. 100k peaks) in < 3 h (Fig. 1c), while imputePCA, SAVER, and DCA failed to execute at the PBMCs dataset.

We next tested if imputation methods can improve the recovery of true OC regions. For this, we created true and negative OC labels for each cell type by peak calling of bulk ATAC-seq profiles. Next, we evaluated the correspondence between imputed scATAC-seq values and peaks of the corresponding cell type with the area under precision-recall curve (AUPR) (see the "Methods" section). scOpen significantly outperformed all competing methods by presenting the highest mean AUPR (Fig. 1 d). The combined ranking indicates SCALE and MAGIC as runner-up methods (Supplementary Fig. 2b). Next, we evaluated the influence on the number of cells per cluster in the AUPR. Despite an overall decrease in AUPR with sample size, we observed that top performing methods (scOpen, SCALE, and MAGIC) were less sensible to cell numbers (Supplementary Fig. 2c).

We also investigated the impact of imputation on the estimation of distances between cells and the impact on standard clustering methods. Distance between cells was evaluated with the silhouette score, while clustering accuracy was evaluated with adjusted Rand index (ARI)[27] both regarding the agreement with known cell labels. scOpen was the best performer in all data sets regarding the silhouette score (Fig. 1e). The combined ranking demonstrated that scOpen had significantly better results than competing methods, while cisTopic and MAGIC were runner-up methods (Supplementary Fig. 2d). Regarding clustering, scOpen was best in the hematopoiesis and multi-omics PBMCs datasets and second-best for cell lines and T cell datasets (Fig. 1f). When considering the combined ranking, scOpen performed best followed by cisTopic and MAGIC (Supplementary Fig. 2e). Visual representations with UMAP[28] projections of these datasets and methods are provided in Supplementary Fig. 3. Altogether, these results support that scOpen outperforms state-of-the-art imputation methods while providing the lowest memory footprint and above-average time performance.

Benchmarking of scATAC-seq clustering methods

Another relevant question was to compare scOpen with top-performing state-of-the-art scATAC-seq pipelines: cisTopic, SnapATAC and Cusanovich2018[21] (see the "Methods" section; Supplementary Fig. 4a). Here, pipelines were evaluated with the default clustering methods, i.e graph-based clustering for SnapATAC[11] and density-based clustering for other methods[10]. We also evaluated the use of both reduced and imputed matrices for scOpen and cisTopic, as these methods provide both types of representations.

The evaluation of distance matrices with the silhouette score indicated that both imputed or low dimension scOpen matrices presented the highest score in all data-sets (Fig. 2a) and both scOpen matrix representations tied as first in the combined rank (Supplementary Fig. 4b). cisTopic, which was the runner-up method, performed well in cell lines, hematopoiesis, and T-cells but poorly for multi-omics PBMCs. Next, we evaluated the clustering performance of competing pipelines. Again, scOpen performed best on cell lines and hematopoiesis data sets and ranked first/second in the combined rank (Supplementary Fig. 4c). Overall, this analysis indicates that both reduced dimension and imputed scOpen matrices obtain the best overall results for distance and clustering representations on evaluated datasets. Of note, the low-dimensional matrix reduces the memory footprint on clustering by more than 1000-fold in comparison to using full imputed matrices and serves as an alternative for cluster analysis of large dimensional data sets.

Improving scATAC-seq downstream analysis using scOpen estimated matrix

Next, we tested the benefit of using scOpen estimated matrices as input for scATAC-seq computational pipelines, which have as objective the identification of regulatory features associated with single cells (chromVAR[12]), estimation of gene activity scores and DNA-interactions (Cicero[13]), or a clustering method tailored for scATAC-seq data (scABC[9]) (Supplementary Fig. 4d). Both chromVAR and Cicero first transform the scATAC-seq matrix to either TFs and genes feature spaces respectively. Clustering was then performed using the standard pipelines from each approach. We compared the clustering accuracy (ARI) and distance (silhouette score) of these methods with either raw or scOpen estimated matrices. In all combinations of methods and datasets, we observed a higher or equal ARI/

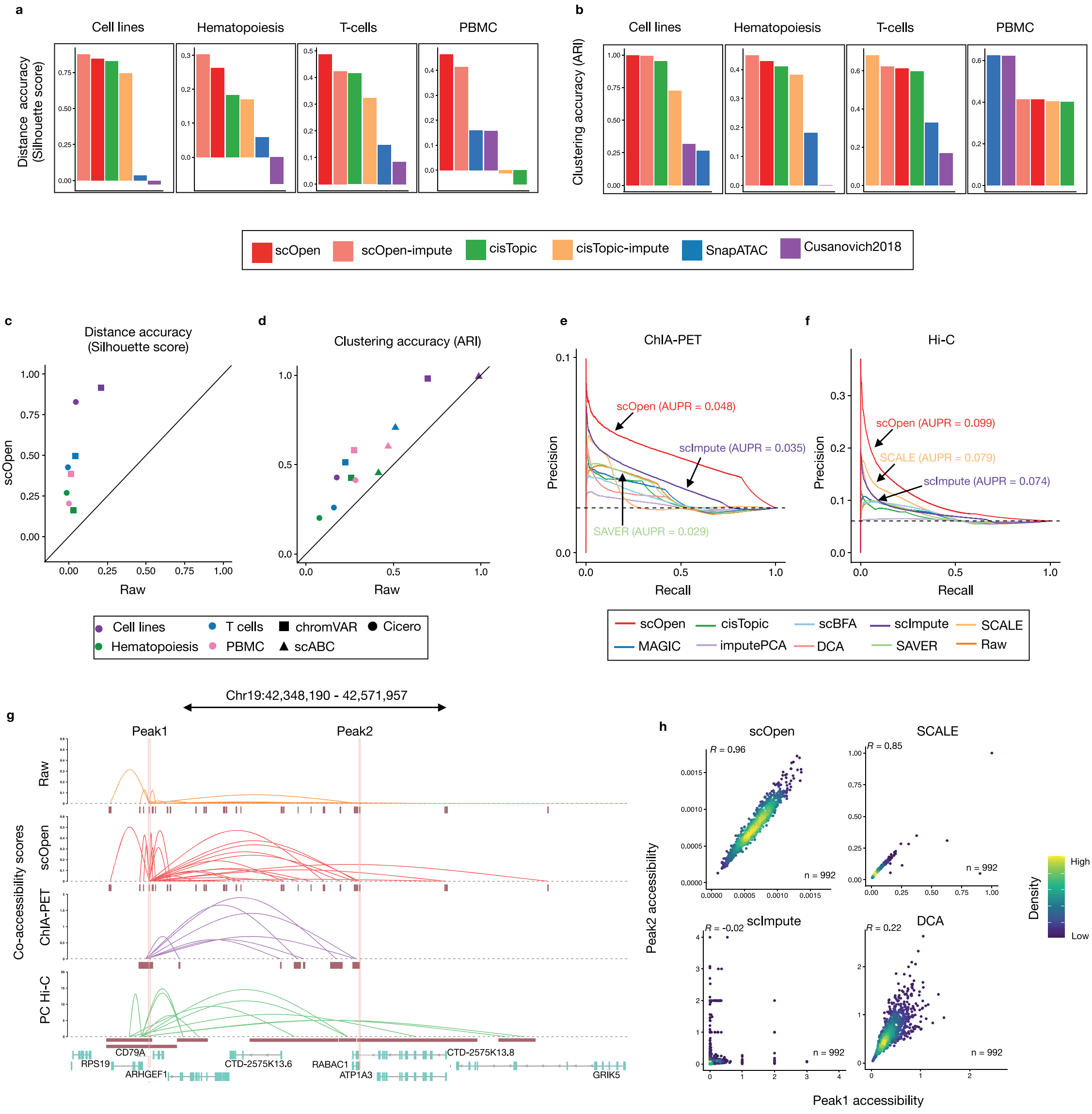


Fig. 2

a Bar plot showing an evaluation of distances estimated on distinct scATAC-seq representations with a silhouette score. **b** Bar plots showing the clustering accuracy (ARI) for distinct clustering pipelines. **c**Scatter plot comparing silhouette score of datasets by providing raw (*x*-axis) and scOpen estimated matrices (*y*-axis) as input for Cicero and chromVAR. Colors represent datasets and shapes represent methods. scABC is not evaluated as it does not provide a space transformation. **d** Same as **c** for clustering results (ARI) of Cicero, chromVAR, and scABC. **e** Precision-recall curves showing the evaluation of the predicted links on GM12878 cells using the raw and imputed matrix as input. We used data from pol-II ChIA-PET as true labels. Colors refer to methods. We reported the AUPR for the top 3 methods. **f** Same as **e** by using Hi-C data as true labels. **g** Visualization of co-accessibility scores (*y*-axis) of Cicero predicted with raw and scOpen estimated matrices contrasted with scores based on RNA pol-II ChIA-PET (purple) and promoter capture Hi-C (green) around the CD79A locus (*x*-axis). For ChIA-PET, the log-transformed frequencies of each interaction PET cluster represent co-accessibility scores, while the negative log-transformed *p*-values from the CHiCAGO software indicate Hi-C scores. **h** Scatter plot showing single-cell accessibility scores estimated by top-performing imputation methods (according to **f**) for the link between peak 1 and peak 2 (supported by Hi-C data). Each dot represents a cell and color refers to density. Pearson correlation is shown on the left-upper corner. Source data for Fig. 2 are provided as a Source Data file.

silhouette whenever a scOpen matrix was provided as input (Fig. 2c, d). Results can be inspected with UMAP visualization with and without scOpen imputation (Supplementary Fig. 5).

Prior to estimating gene-centric OC scores, Cicero first predicts co-accessible pairs of DNA regions in groups of cells, which potentially form *cis*-regulatory interactions. We compared Cicero predicted interactions on human lymphoblastoid cells (GM12878) by using Hi-C and ChIA-PET from this cell type as true labels for all imputation methods with data as provided in ref. 13. Both AUPR values and odds ratios indicated that the scOpen matrix improves the detection of GM12878 interactions globally (Fig. 2 e, f; Supplementary Fig. 6a, b). To evaluate the impact on the number of cell on these predictions, we have down-sampled the data to only consider 50% or 25% of cells. We observed a residual decrease in the AUPR of scOpen for 25% of cells (Supplementary Fig. 6 c). This supports that chromatin conformation prediction works well even for cell types with low abundance. The power of scOpen imputation was clear when checking the individual locus (Fig. 2g), as previously described by Cicero[13]. This is evident when contrasting accessibility scores between pairs of peak-to-peak links supported by Hi-C predictions (Fig. 2h; Supplementary Fig. 6d–h). scOpen obtained highly correlated accessibility scores, while other imputation methods showed quite diverse association patterns. Together, these results indicated that the use of scOpen estimated matrices improves downstream analysis of state-of-the-art scATAC-seq methods.

Applying scOpen to scATAC-seq of fibrosis driving cells

Next, we evaluated scOpen in its power to improve the detection of cells in a complex disease dataset. For this, we performed whole mouse kidney scATAC-seq in C57Bl6/WT mice in homeostasis (day 0) and at two-time points after injury with fibrosis: 2 and 10 days after unilateral ureteral obstruction (UUO)[29-30]. Experiments recovered a total of 30,129 high-quality cells after quality control with an average of 13,933 fragments per cell, a fraction of reads in promoters of 0.46, and high reproducibility ($R > 0.99$) between biological duplicates (Supplementary Fig. 7a, b; Supplementary Table 1). After data aggregation, 150,593 peaks were detected, resulting in a highly dimensional and sparse scATAC-seq matrix (4.2% of non-zeros). Next, we performed data integration for batch effect removal using Harmony[31]. For comparison, we used a dimension reduced matrix from either LSI (Cusanovich2018), cisTopic, SnapATAC, or scOpen. We annotated the

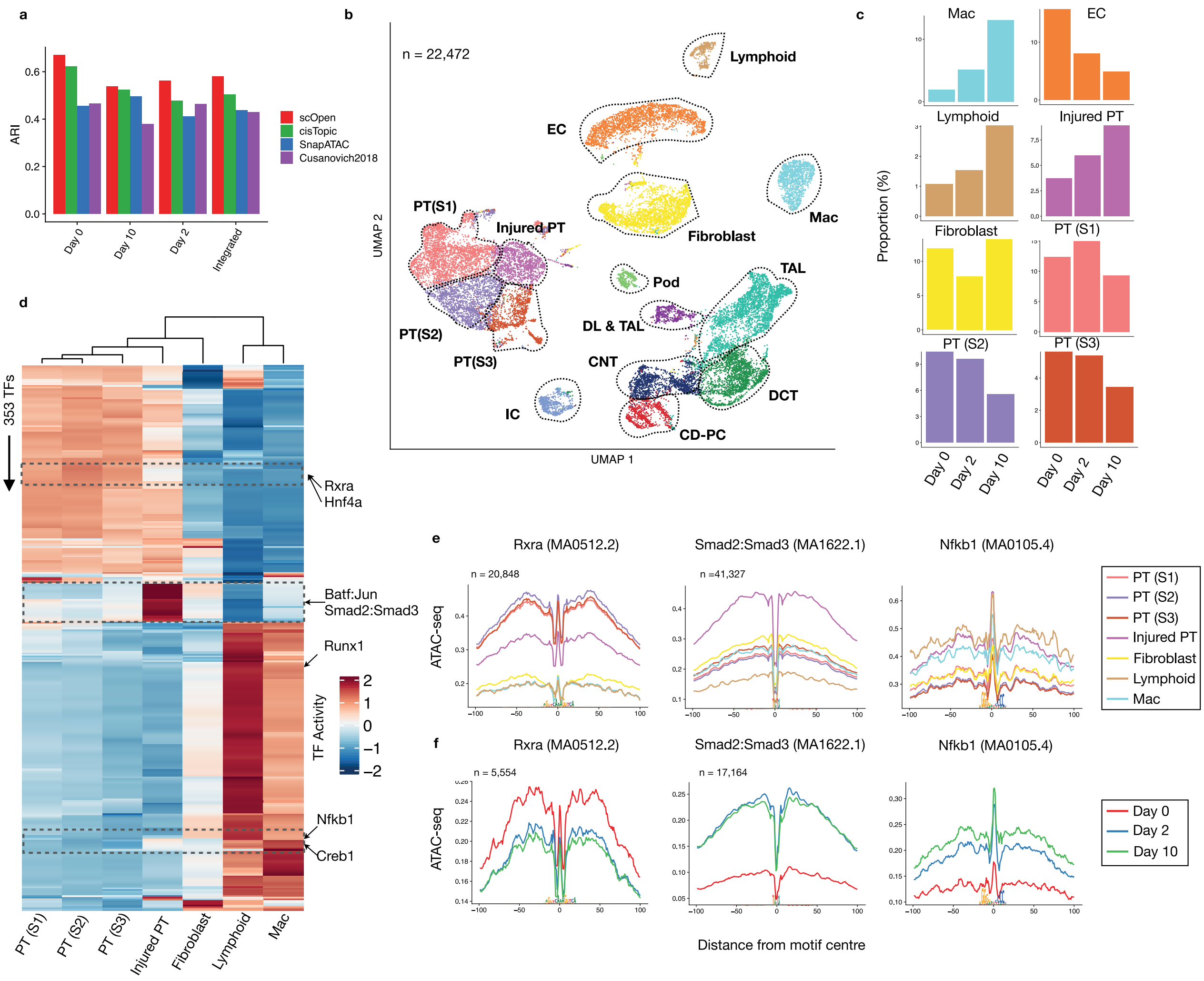

a
ARI
0.0
0.2
0.4
0.6
Day 0
Day 10
Day 2
Integrated
scOpen
cisTopic
SnapATAC
Cusanovich2018
b
n = 22,472
UMAP 2
UMAP 1
Lymphoid
EC
Mac
Fibroblast
PT(S1)
Injured PT
Pod
TAL
PT(S2)
PT(S3)
DL & TAL
CNT
DCT
IC
CD-PC
c
Proportion (%)
Mac
EC
Lymphoid
Injured PT
Fibroblast
PT (S1)
PT (S2)
PT (S3)
Day 0
Day 2
Day 10
d
353 TFs
Rxra
Hnf4a
Batf:Jun
Smad2:Smad3
Runx1
TF Activity
2
1
0
−1
−2
Nfkb1
Creb1
PT (S1)
PT (S2)
PT (S3)
Injured PT
Fibroblast
Lymphoid
Mac
e
Rxra (MA0512.2)
Smad2:Smad3 (MA1622.1)
Nfkb1 (MA0105.4)
n = 20,848
n =41,327
ATAC-seq
PT (S1)
PT (S2)
PT (S3)
Injured PT
Fibroblast
Lymphoid
Mac
f
Rxra (MA0512.2)
Smad2:Smad3 (MA1622.1)
Nfkb1 (MA0105.4)
n = 5,554
n = 17,164
ATAC-seq
Day 0
Day 2
Day 10
−100
−50
0
50
100
Distance from motif centre

Fig. 3

a ARI values (*y*-axis) contrasting clustering results and transferred labels using distinct dimensional reduction methods for scATAC-seq. Clustering was performed by only considering UUO kidney cells on day 0 (WT), day 2, or day 10 or the integrated data set (all days). **b** UMAP of the integrated UUO scATAC-seq after doublet removal with major kidney cell types: fibroblasts, descending loop of Henle and thin ascending loop of Henle (DL & TAL); macrophages (Mac), Lymphoid (T and B cells), endothelial cells (EC), thick ascending loop of Henle (TAL), distal convoluted tubule (DCT), collecting duct-principal cells (CD-PC), intercalated cells (IC), podocytes (Pod) and proximal tubule cells (PT S1; PT S2; PT S3; Injured PT). **c** Proportion of cells of selected clusters on either day 0, day 2 or day 10 experiments. **d** Heatmap with TF activity score (*z*-transformed) for TFs (*y*-axis) and selected clusters (*x*-axis). We highlight TFs with the decrease in activity scores in injured PTs (Rxra and Hnf4a), with high TF activity scores in injured PTs (Batf:Jun; Smad2:Smad3) and immune cells (Creb1; Nfkb1). **e**Transcription factor footprints (average ATAC-seq around predicted binding sites) of Rxra, Smad2::Smad3 and Nfkb1 for selected cell types. The logo of underlying sequences is shown below and the number of binding sites is shown top-left corner. **f** Transcription factor footprints of Rxra, Smad2::Smad3, and Nfkb1 for injured PT cells in day 0, day 2, and day 10. Source data for Fig. 3 are provided as a Source Data file.

scATAC-seq profiles using single nuclei RNA-seq (snRNA-seq) data of the same kidney fibrosis model from an independent study[32] via label transfer[33] to serve as cell labels. We then evaluated the batch correction results using silhouette score and clustering. We observed that clusters based on scOpen were more similar to the transferred labels (higher ARI) than clusters based on competing methods (Fig. 3a). Furthermore, scOpen also provided better distance metrics and visualization than competing methods (Supplementary Figs. 7c–e and 8). These results support the discriminative power of scOpen in this large and complex dataset.

Next, we annotated the clusters of scOpen by using known marker genes and transferred labels after removing doublets with ArchR[34]. We identified all major kidney cell types including PT cells, distal/connecting tubular cells, collecting duct and loop of Henle, endothelial cells (ECs), fibroblasts as well as the rare populations of podocytes and lymphocytes (Fig. 3b; Supplementary Fig. 9a). Lymphocytes were not described in the previously scRNA-seq study[32], which supports the importance of annotation of scATAC-seq clusters independently of scRNA-seq label transfer. Of particular interest were cell types with population changes during the progression of fibrosis (Fig. 3c; Supplementary Fig. 9b–d). We observed an overall decrease of normal proximal tubular (PT), glomerular and ECs and an increase of immune cells as expected in this fibrosis model with tubule injury, the influx of inflammatory cells, and capillary loss[35-36]. Importantly, we detected an increased PT sub-population, which we characterized as injured PT by increased accessibility around the PT injury markers *Vcam1* and *Kim1* (*Havrc1*)[37] (Supplementary Fig. 9a).

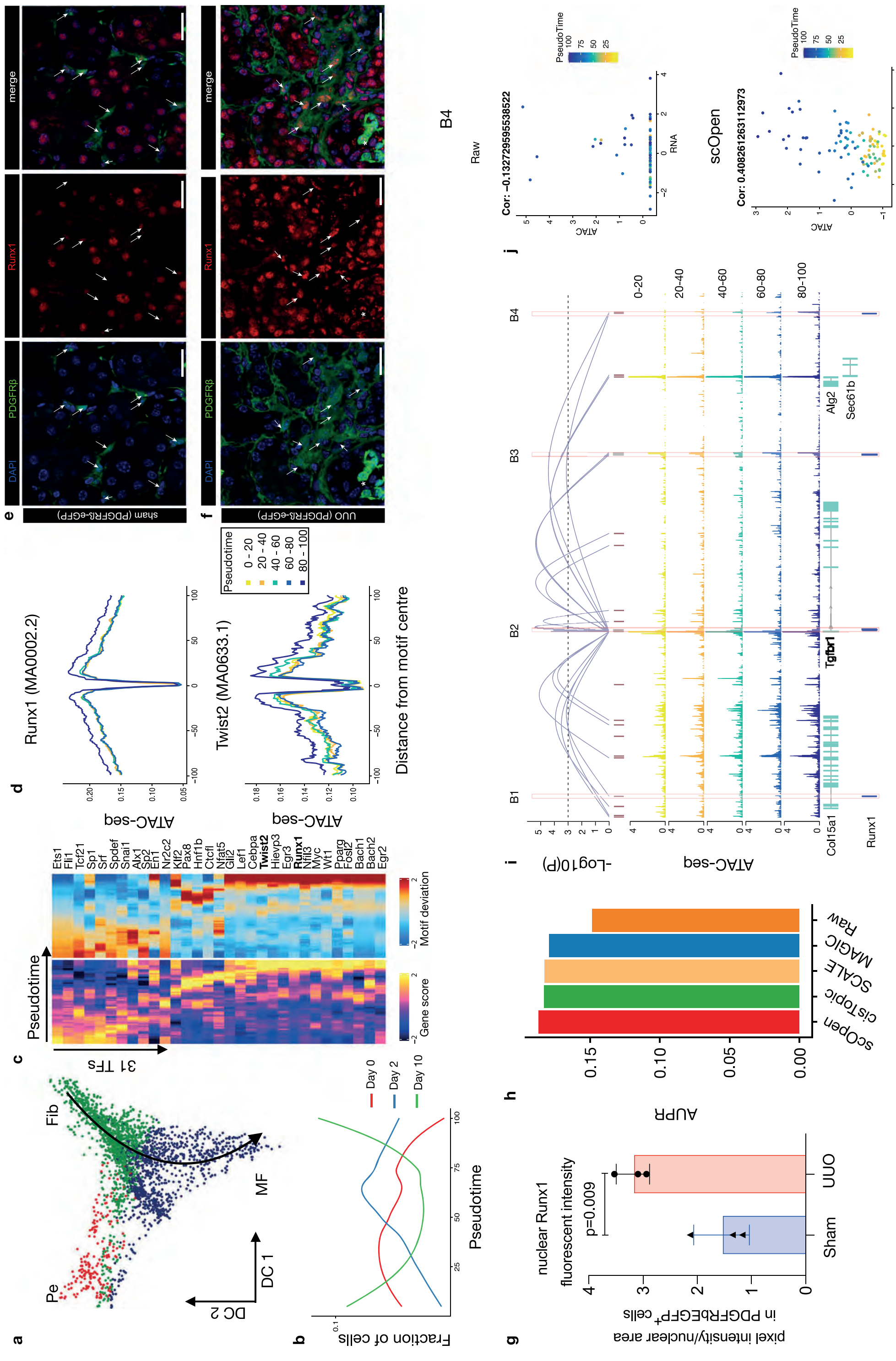

a
Pe
Fib
MF
DC 1
DC 2
b
Fraction of cells
0.1
Pseudotime
25
50
75
100
Day 0
Day 2
Day 10
c
Pseudotime
31 TFs
Ets1
Fli1
Tcf21
Sp1
Srf
Spdef
Snai1
Alx1
Sp2
En1
Nr2c2
Klf2
Pax8
Hnf1b
Ctcfl
Nfat5
Gli2
Lef1
Cebpa
Twist2
Hlevp3
Egr3
Runx1
Nfil3
Myc
Wt1
Pparg
Fosl2
Bach1
Bach2
Egr2
Gene score
Motif deviation
d
Runx1 (MA0002.2)
Twist2 (MA0633.1)
ATAC-seq
Distance from motif centre
Pseudotime
0 - 20
20 - 40
40 - 60
60 -80
80 - 100
e
sham (PDGFRβ-eGFP)
DAPI
PDGFRβ
Runx1
merge
f
UUO (PDGFRβ-eGFP)
g
nuclear Runx1 fluorescent intensity
pixel intensity/nuclear area in PDGFRβEGFP+ cells
p=0.009
Sham
UUO
h
AUPR
0.00
0.05
0.10
0.15
scOpen
cisTopic
SCALE
MAGIC
Raw
i
-Log10(P)
ATAC-seq
B1
B2
B3
B4
0-20
20-40
40-60
60-80
80-100
Col15a1
Tgfbr1
Alg2
Sec61b
Runx1
j
B4
Raw
Cor: −0.132729559538522
scOpen
Cor: 0.408261263112973
ATAC
RNA
PseudoTime
100
75
50
25

Fig. 4

a Diffusion map showing sub-clustering of fibroblasts. Colors refer to sub-cell-types and arrow represents differentiation trajectory from fibroblast to myofibroblast. Pe pericyte, Fib fibroblast, MF myofibroblast. **b** Line plots showing cell proportion from the day after UUO along the trajectory. **c** Pseudotime heatmap showing gene activity (left) and TF motif activity (right) along the trajectory. **d** Footprinting profiles of Runx1 and Twist2 binding sites along the trajectory. **e** Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney. In sham-operated mice, Runx1 staining shows a reduced intensity in PDGFRb-eGFP+ cells compared to remaining kidney cells (arrows). **f** Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney at 10 days after UUO as compared to sham. Arrows indicate Runx1 staining in expanding PDGFRb-eGFP+ myofibroblasts. **g** Quantification of Runx1 nuclear intensity in PDGFRb-eGFP+ cells in sham vs. UUO mice. Error bars represent the SD of the intensity. Data are presented as mean ± SD. Statistical significance was assessed by a two-tailed Student's *t*-test with $p < 0.05$ being considered statistically significant ($n = 3$ mice). **h** Performance of top-performing imputation methods on the prediction of Runx1 target genes measured with AUPR. **i** Peak-to-Gene links (top) predicted on scOpen matrix and associated to Tgfbr1 in fibroblast cells. The height of links represents its significance. Dash line represents the threshold of significance (FDR = 0.001). ATAC-seq tracks (below) were generated from pseudo-bulk profiles of fibroblast/myofibroblast cells with increasing pseudo time (0–20, 20–40, 40–60, 60–80, and 80–100). Binding sites of Runx1 (B1–B4) supported by ATAC-seq footprints and overlapping to peaks are highlighted on the bottom. **j** Scatter plot showing gene activity of Tgfbr1 and normalized peak accessibility from raw (upper) or scOpen imputed matrix (lower) for peak-to-gene link B4. Each dot represents cells in a given pseudotime and the overall correlation is shown in the left-upper corner. Scale bars in **e** and **f** represent 50 µm. For details on statistics and reproducibility, see the "Methods" section. Source data for Fig. 4 are provided as a Source Data file.

Dissecting cell-specific regulatory changes in fibrosis

Next, we adapted HINT-ATAC[4] to dissect regulatory changes in scATAC-seq clusters. For each cluster, we created a pseudo-bulk ATAC-seq library by combining reads from single cells in the cluster. We then performed footprinting analysis and estimated TF activity scores for all footprint-supported motifs. We only kept TFs with changes (high variance) in TF activity scores among clusters. We focused here on clusters associated with PT cells, fibroblasts, and immune cells, as these represent key players in kidney remodeling and fibrosis after injury. As shown in Fig. 3d, the TF activity scores capture regulatory programs associated with these three major cell populations (Supplementary Data 2). Injured PTs have overall lower TF activity scores than all TFs of the PT cluster. TFs with a high decrease in activity in injured PTs include Rxra, which is important for the regulation of calcium homeostasis in tubular cells[38], and Hnf4a, which is important in PT development[39] (Fig. 3d, e). Footprint profile of Rxra in injured PTs display a gradual loss of TF activity over time indicating that injured PT acquires a de-differentiated phenotype during fibrosis progression and tubular dilatation (Fig. 3f). A group of TFs with high activity scores in injured PTs also have increased TF activity scores in fibroblasts (Smad2:Smad3 and Batf:Jun) indicating shared regulatory programs in these cells. Smad proteins are downstream mediators of TGFβ signaling, which is a known key player of fibroblast to myofibroblast differentiation and fibrosis[40]. The high activity of Smad2:Smad3 also indicates a role of TGFβ in the de-differentiation of injured PTs. Also, both Smad2:Smad3 reach a peak in TF activity level at day 2 after UUO in injured PTs (Fig. 3f), which indicates these TFs are activated post-transcriptionally. We also detect the high activity of Nfkb1 in injured PTs (and lymphocytes), which fits with the known role of Nfkb1 in injured and failed repair PTs[41-42]. Moreover, our analysis also shows a gradual TF activity increase over time in injured PT (Fig. 3f), suggesting that Nfkb1 plays an important role in sustaining the injured PT phenotype.

scOpen reveals TF driving myofibroblast differentiation

A key process in kidney injury is fibrosis, which is caused by the differentiation of fibroblasts and pericytes to matrix secreting myofibroblasts[43]. To dissect potential differentiation trajectories, we performed a diffusion map embedding of the fibroblasts (Fig. 4a), which revealed the presence of three major branches formed by fibroblasts, pericytes, and myofibroblasts, as supported by the accessibility of *Scara5*, *Ng2* (*Cspg4*), *Postn* and *Col1a1*(Supplementary Fig. 10)[43,44].

We next created a cellular trajectory across the differentiation from fibroblasts to myofibroblasts using ArchR (Fig. 4a; Supplementary Fig. 10c). We observed that there is an increase in cells after injury (Day 2 and Day 10) along the trajectory (Fig. 4b). We next characterized TFs by correlating their gene activity with TF activity along the trajectory (Fig. 4c) and ranked these by their correlation (Supplementary Fig. 10d). The correlation of Runx1, which has a well-known function in blood cells[45], stood out, besides showing a steady increase in activity in myofibroblasts. Another TF with high correlation and similar myofibroblast specific activity was Twist2, which has a known role in epithelial to mesenchymal transition in kidney fibrosis[46] (Fig. 4d).

To validate the yet uncharacteristic role of Runx1 in myofibroblasts, we performed immunostaining and quantification of Runx1 signal intensity in transgenic PDGFRb-eGFP mice that genetically tag fibroblasts and myofibroblasts[43,47]. Runx1 staining in control mice (sham) revealed positive nuclei in tubular epithelial cells and rarely in PDGFRb-eGFP+ mesenchymal cells (Fig. 4e). In kidney fibrosis after UUO surgery (day 10), Runx1 staining intensity increased significantly in PDGFRb+ myofibroblasts (Fig. 4f,g). Next, we performed retroviral overexpression experiments and RNA-sequencing in a human kidney PDGFRb+ fibroblast cell-line that we have generated[43] to ask whether Runx1 might be functionally involved in myofibroblast differentiation in humans (Supplementary Fig. 11a, b). Runx1 overexpression led to reduced proliferation (Supplementary Fig. 11c) and strong gene expression changes (Supplementary Fig. 11d). Gene ontology (GO) and pathway enrichment analysis indicated enrichment of cell adhesion, cell differentiation, and TGFB signaling following Runx1 overexpression (Supplementary Fig. 11e). Various extracellular matrix genes (*Fn1*, *Col13A1*), as well as a TGFB receptor

(Tgfbr1) and Twist2, were up-regulated following Runx1 overexpression (Supplementary Fig. 11d). Furthermore, we observed increased expression of the myofibroblast marker gene *Postn* after Runx1 overexpression. Altogether, this suggests that Runx1 might directly drive myofibroblast differentiation of human kidney fibroblasts since overexpression reduced cell proliferation and induced expression of various myofibroblast genes.

Identification of Runx1 target genes

Another important application of scATAC-seq is the prediction of *cis*-regulatory DNA interactions (peak-to-gene links) by measuring the correlation between gene activity and reads counts in proximal peaks. To compare the impact of imputation on this task, we predicted peak-to-gene links in fibroblasts on distinct scATAC-seq matrices using ArchR[34] after imputation with top-performing imputation methods. The use of imputation methods led to improved signals on peak-to-gene links predictions as indicated by higher correlation values after imputation (Supplementary Fig. 12a, b). We considered all genes with at least one link, where the peak has a footprint supported Runx1-binding site, as Runx1 targets. We then compared the predicted Runx1 targets from distinct scATAC-seq imputed matrices with differentially expressed genes after Runx1 over-expression (true labels). All imputation methods obtained higher AUPR values than the use of a raw matrix, while scOpen obtained the highest AUPR (Fig. 4h; Supplementary Fig. 12c). Among others, scOpen predicted *Tgfbr1* and *Twist2* as prominent Runx1 target genes (Fig. 4i; Supplementary Fig. 12 d). We observed several peaks with high peak-to-gene correlation, increasing accessibility upon myofibroblast differentiation and presence of Runx1-binding sites. The positive impact of imputation was clear when observing scatter plots contrasting gene activity and peak accessibility of these peak-to-gene links (Fig. 4j; Supplementary Fig. 12e–i).

Another interesting question is the association of the predicted link with distinct regulatory features. While we observed no clear association of the correlation of predicted links with the size of the link (Supplementary Fig. 13a), our analysis suggested that links associated with active kidney enhancers have a higher correlation than other active regulatory regions. This further supports the functional relevance of predicted links. These results suggest that Runx1 is an important regulator of myofibroblast differentiation by regulating the EMT-related TF Twist2 and by amplifying TGFB signaling by increasing the expression of a TGFB receptor 1 and affecting the expression of extracellular matrix genes. Altogether, these results uncover a complex cascade of regulatory events across cells during the progression of fibrosis and reveal a yet unknown function of Runx1 in myofibroblast differentiation in kidney fibrosis.

Discussion

In ATAC-seq, Tn5 generates a maximum of 2 fragments per cell in a small (~200 bp) OC region. Subsequent steps of the ATAC-seq protocol cause loss of a large proportion of these fragments. For example, only DNA fragments with the two distinct Tn5 adapters, which are only present in 50% of the fragments, are amplified in the PCR step[48]. Further DNA material losses occur during single-cell isolation, liquid handling, sequencing, or by simple financial restrictions of sequencing depth. Assuming that 25% of accessible DNA can be successfully sequenced, we expect that 56% of accessible chromatin sites will not have a single digestion event causing the so-called dropout events, assuming that digestion events follow a binomial distribution. Despite this major signal loss, imputation and denoising have been widely ignored in the scATAC-seq literature[5,6,8,9,12,13] and common scATAC-seq pipelines (e.g., Signac[49] and ArchR[34]).

We demonstrated here that scOpen estimated matrices have a higher recovery of dropout events and also improved distance and clustering results when compared to imputation methods for scRNA-seq[14,17,18,19,25] and the few available imputations methods tailored for scATAC-seq (cisTopic-impute[10], SCALE[20]). scOpen also presented very good scalability with the lowest memory requirements and tractable computational time on large data sets. From a methodological perspective, scOpen is the only method performing regularization of estimated models to prevent over-fitting. This is in line with a previous study, which indicated over-fitting as one of the largest issues on scRNA-seq imputation[50]. Moreover, it is also possible to use the scOpen factorized matrix as a dimension reduction. We have shown that both dimensions reduced and imputed matrices from scOpen displayed the best performance on distance representation and clustering when compared to diverse state-of-the-art scATAC-seq dimension reduction/clustering pipelines (cisTopic, SnapATAC, and Cusanovich et al. 2018). Of note, the ArchR pipeline is equivalent to Cusanovich et al. 2018 and based on the same dimension reduction method (LSI). It is worth noting that LIGER[51] is another method of employing NMF for single-cell ATAC-seq analysis. It uses integrative NMF to extract shared factors with the objective of multi-modal data integration of scATAC-seq and scRNA-seq. Moreover, denoising in bulk ATAC-seq has also been approached with the use of deep learning methods[52]. These closely related approaches, however, have distinct applications than scOpen and are therefore not evaluated here.

Finally, we have demonstrated that the use of scOpen-corrected matrices improves the accuracy of existing state-of-art scATAC-seq methods (cisTopic[10], chromVAR[12], Cicero[13]). Particularly positive results were obtained in the prediction of chromatin conformation with Cicero, where all methods perform better than raw matrices. Cicero works by measuring the correlation between pairs of proximal links. Due to the fact that dropout events are independent for two regions, it is not surprising that imputation has strong benefits. This is equivalent to observations from van Dijk et al. 2018[14] in the context of scRNA-seq, where the prediction of gene–gene interactions after MAGIC imputation was significantly improved. Altogether, these results support the importance of dropout event correction with scOpen in any computational analysis of scATAC-seq. Of note, a sparsity similar to scATAC-seq is also expected in single-cell protocols based on DNA enrichment such as scChIP-seq[53,54], scCUT&Tag[55], or scBisulfite-seq[56]. Denoising and imputation of count matrices from these protocols represent a future challenge.

Moreover, we used scOpen to characterize complex cascades of regulatory changes associated with kidney injury and fibrosis. Our analyses demonstrated that a major expanding population of cells, i.e. injured PTs, myofibroblasts, and immune cells, share regulatory programs, which are associated with cell de-/differentiation and proliferation. Of all methods evaluated, scOpen obtained the best clustering results in the kidney cell repertoire using a scRNA-seq on the same kidney injury model as a reference. Trajectory analysis identified Runx1 as the major TF driving myofibroblast differentiation, which was validated by Runx1 staining in the mouse model and by retroviral over-expression studies in human PDGFRb+ kidney cells. Computational prediction with peak-to-gene links combined with footprint-supported Runx1-binding sites indicated the role of Runx1 in the regulation of *Tgfbr1* and *Twist2*. These were validated on over-expression experiments in human fibroblasts. Altogether, results suggest that Runx1 makes fibroblasts more sensitive to TGFB signaling via increasing expression of the TGFB receptors.

Runx1 has recently been reported as a potential inducer of EMT in PT cells[57]. Furthermore, in vitro data of mesenchymal stem cells (MSCs) isolated from bone marrow or prostate gland points towards a potential myofibroblast differentiation role of Runx1[58]. In vivo evidence for a functional role of Runx1 in regulating fibrogenesis has been demonstrated in zebrafish[59]. Single-cell RNA-seq data from zebrafish heart after cryo-injury suggests that endocardial cells and thrombocytes up-regulate Runx1 while Runx1 mutant zebrafish demonstrated enhanced cardiac regeneration after cryoinjury with an ameliorated fibrotic response. Here we show for the first time in vivo and in vitro evidence that Runx1 in myofibroblasts regulates scar formation following a fibrogenic kidney injury in mice. Runx1 deficiency caused reduced myofibroblast formation and enhanced recovery. To this end, inhibiting Runx1 could lead to reduced myofibroblast differentiation and increased endogenous repair after fibrogenic organ injuries in the kidney and heart. Our results shed light on mechanisms of myofibroblasts differentiation driving kidney fibrosis and chronic kidney disease (CKD). Altogether, this demonstrates how scOpen can be used to dissect complex regulatory processes by footprinting analysis combined with peak-to-gene link predictions.

Methods

UUO data pre-processing.

We used Cell-Ranger ATAC (v1.1.0) pipeline toperform low-level data processing (https://support.10xgenomics.com/single-cellatac/ software/pipelines/latest/algorithms/overview). We first demultiplexed raw base call files using cellranger-atac mkfastq with its default setting to generate FASTQ files for each flowcell. Next, cellranger-atac count was applied to perform read trimming, filtering, and alignment. We then estimated the transcription start site (TSS) enrichment score using the obtained fragment files and filtered lowquality cells using a TSS score of 8 and a number of unique fragments of 1000 as thresholds. The obtained barcodes are considered valid cells for the following analysis.

UUO data dimension reduction, data integration, and clustering. We next performed peak calling using MACS2 for each sample and merged the peaks to generate a union peak set, which was used to create a peak by cell matrix. For comparison, we applied distinct methods, i.e., scOpen, cisTopic, SnapATAC, and LSI/Cusanovich2018, to the matrix and used the dimension reduced matrix for data integration, clustering, and visualization. Next, we used Harmony31 to integrate the scATAC-seq profiles from different conditions (day 0, day 2, and day 10) using either LSI/Cusanovich2018, cisTopic, scOpen, or SnapATAC dimension reduced matrix as input. Specifically, we created a Seurat object for each of the lowdimension matrices and ran the Harmony algorithm with the function RunHarmony. We then used k-medoids to cluster the cells taking batch-corrected lowdimension matrix as input. The number of clusters was set to 17 given that the single-nucleus RNA-seq that we used as a reference for annotation identified 17 unique cell types (see below).

UUO label transfer.

To evaluate and annotate the clusters obtained from data integration, we downloaded a publicly available snRNA-seq dataset of the same fibrosis model (GSE119531) and performed label transfer using Seurat333. This dataset contains 6147 single-nucleus transcriptomes with 17 unique cell types32. For label transfer, we used the gene activity score matrix estimated by ArchR and transferred the cell types from the snRNA-seq dataset to the integrated scATACseq dataset by using the function FindTransferAnchors and TransferData in Seurat333. For benchmarking purposes, the predicted labels were used as true labels to compute ARI for evaluation of the clustering results and silhouette score for evaluation distances after using different dimension reduction methods as input for data integration (Supplementary Fig. 7c–e). We also performed the same analysis for each sample separately and evaluated the results (Fig. 3a).

UUO cluster annotation.

For the biological interpretation, we estimated doublet scores using ArchR34 and removed cells with a doublet score higher than 2.5. Next, we named the cluster by assigning the label with the highest proportion of cells to the cluster and checking marker genes (Supplementary Fig. 9a). In total, we recovered 16 unique cell types from the 17 labels, as two clusters (2 and 17) were annotated as TAL cells. Specifically, we denoted clusters 6, 1, 3 as proximal tubule (PT) S1, S2, and S3 cells. We annotated cluster 2 as thick ascending limb (TAL), cluster 5 as distal convoluted tubule (DCT), cluster 7 as collecting duct-principal cell (CD-PC), cluster 8 as an EC, cluster 9 as connecting tubule (CNT), cluster 10 as an intercalated cell (IC), cluster 11 as fibroblast, cluster 12 as descending limb + thin ascending limb (DL and TAL), cluster 13 as macrophage (MAC), cluster 16 as podocytes (Pod). Cluster 14 was identified as injured PT, which was not described in ref. 32, given the increased accessibility of marker Vcam1 and Havcr1 (Supplementary Fig. 9a). We also renamed the cells of cluster 15, which were label as Mac2 in ref. 32, as lymphoid cells given that these cells express B and T cell markers Ltb and Cd1d, but not macrophage markers C1qa and C1qb. Finally, cluster 4 was removed based on the doublet analysis.

UUO Cell-type-specific footprinting with HINT-ATAC.

We have adapted the footprinting-based differential TF activity analysis from HINT-ATAC for scATACseq. In short, we created pseudo bulk atac-seq libraries by combining reads of cells for each cell type and performed footprinting with HINT-ATAC. Next, we predicted TF binding sites by motif analysis (FDR = 0.0001) inside footprint sequences using RGT (v0.12.3; https://github.com/CostaLab/reg-gen). Motifs were obtained from JASPAR Version 202073. We measured the average digestion profiles around all binding sites of a given TF for each pseudo-bulk ATAC-seq library. We used then the protection score4, which measures the cell-specific activity of a factor by considering the number of digestion events around the binding sites and depth of the footprint. Higher protection scores indicate higher activity (binding) of that factor. Finally, we only considered TFs with more than 1000 binding sites and variance in activity score higher than 0.3. We also performed smoothing for visualization of average footprint profiles. In short, we performed a trimmed mean smoothing (5 bps window) and ignored cleavage values in the top 97.5% quantile for each average profile.

Identifying trajectory from fibroblast to myofibroblast.

We performed further sub-clustering of fibroblast cells on batch-corrected low-dimension scOpen matrix. In total, 3 clusters were obtained and annotated as pericyte (cluster 1), myofibroblast (cluster 2), and Scara5+ fibroblast (cluster 3) using known marker genes

(Supplementary Fig. 10a), respectively. For visualization, a diffusion map 2D embedding was generated using R package density74. Next, a trajectory from Scara5+ fibroblast to myofibroblast was created using function addTrajectory and visualized using function plotTrajectory from ArchR (Supplementary Fig. 10c).

Identifying key TF drivers of myofibroblast differentiation.

To identify TFs that drive this process, we first performed peak calling based on all fibroblasts using MACS2 to obtain specific peaks and then estimated motif deviation per cell using chromVAR. The deviation scores were normalized to allow for comparison between TFs. Next, we selected the TFs with high variance of deviation and gene activity score along the trajectory and calculated the correlation of TF activity and gene activity. This was done by using the function correlateTrajectories from ArchR. We only consider the 31 TFs with significant correlation (FDR < 0.1) (Fig. 4c). We then sorted the TFs by correlation, which identifies Runx1 as the most relevant TF for the differentiation (Supplementary Fig. 10d).

Prediction of peak-to-gene links.

We obtained TSS from annotation BSgenome. Mmusculus.UCSC.mm10 for each gene and extended it by 250 kbps for both directions. Then, we overlapped the peaks from fibroblasts and the TSS regions using function findOverlaps to identify putative peak-to-gene links. We next created 100 pseudo-bulk ATAC-seq profiles by assigning each cell to an interval along the trajectory of myofibroblast differentiation. The gene score matrix and peak matrix were aggregated according to the assignment to generate two pseudo-bulk data matrices. For each putative peak-to-gene link, we calculated the correlation between peak accessibility and gene activity. The p-values are computed using. distribution and corrected by Benjamini–Hochberg method. For comparison, we also performed matrix imputation using the four top methods, i.e., scOpen, SCALE, MAGIC, and cisTopic, as evaluated by peaks recovering (Supplementary Fig. 2b) and computed the correlation based on the imputed matrix. To compare the scOpen predicted peak-to-gene correlation from different types of peaks, we used the annotation generated by R package ArchR34 and classified the peaks as distal, exonic, intronic, and promoter. We also tested if the correlation is different between activate enhancers and nonenhancers. For this, We obtained H3K27ac (ENCSR000CDG) and H3K4me1 (ENCSR436FYE) ChIP-seq peaks of mouse kidneys from ENCODE. The peaks were classified as active enhancers if they are overlapping with H3K27Ac and H3K4me1, and other active regions if they are only overlapping with H3K27Ac.

Prediction and evaluation of Runx1 target genes.

With each peak being associated with genes, we next sought to link Runx1 to its target genes. For this, we first performed a footprinting pseudo-bulk ATAC-seq profile to identify TF footprints inside peaks linked to genes in the previous peak-to-gene analysis. Next, we identified Runx1-binding sites using a motif-matching approach. We defined the genes that have at least one footprint-support binding site of Runx1 in their associated peaks as Runx1 target genes. We then used the peak-to-gene correlation as a prediction between Runx1 and the target genes. This procedure was for the peak to gene links predicted by distinct imputation approaches, thus generating various predictions. To evaluate the results, we used the DE genes obtained from RNA-seq of Runx1 overexpression as true labels (see below), and computed the AUPR (Fig. 4h).

Affiliations

[1] Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg University, Heidelberg, Germany

[2] Faculty of Medicine, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Aachen, Germany

[3] Faculty of Medicine, Institute of Experimental Medicine and Systems Biology, RWTH Aachen University, Aachen, Germany

[4] Division of Nephrology and Clinical Immunology, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

[5] Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands

[6] MRC Cancer Unit, Hutchison/MRC Research Centre, University of Cambridge, Cambridge, UK

[7] Faculty of Health and Medical Sciences, Proteomics Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

[8] Centre of Biological Engineering, University of Minho - Campus de Gualtar, Braga, Portugal

[9] Department of Urology and Pediatric Urology, St. Antonius Hospital Eschweiler, Academic Teaching Hospital of RWTH Aachen, Eschweiler, Germany

[10] Department of Urology and Kidney Transplantation, Martin Luther University, Halle (Saale), Germany

[11] Department of Hematology, Erasmus MC, Rotterdam, The Netherlands

[12] Institute for Research and Innovation in Health (i3s), Porto, Portugal

[13] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

[14] Molecular Medicine Partnership Unit, European Molecular Biology Laboratory, Heidelberg University, Heidelberg, Germany

5

Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses

Aurelien Dugourd[1,2,3,4], **Christoph Kuppe**[3,4,5], Marco Sciacovelli[6], Enio Gjerga[1,2], Attila Gabor[1], Kristina B. Emdal[7], Vitor Vieira[7], Dorte B. Bekker-Jensen[7], Jennifer Kranz[3,9,10], Eric.M.J. Bindels[11], Ana S.H. Costa[6,15], Abel Sousa[12,13], Pedro Beltrao[13], Miguel Rocha[8], Jesper V. Olsen[7], Christian Frezza[6], Rafael Kramann[3,4,5] and Julio Saez-Rodriguez[1,2,14]

Abstract

Multi-omics datasets can provide molecular insights beyond the sum of individual omics. Various tools have been recently developed to integrate such datasets, but there are limited strategies to systematically extract mechanistic hypotheses from them. Here, we present COSMOS (Causal Oriented Search of Multi-Omics Space), a method that integrates phosphoproteomics, transcriptomics, and metabolomics datasets. COSMOS combines extensive prior knowledge of signaling, metabolic, and gene regulatory networks with computational methods to estimate activities of transcription factors and kinases as well as network-level causal reasoning. COSMOS provides mechanistic hypotheses for experimental observations across multi-omics datasets. We applied COSMOS to a dataset comprising transcriptomics, phosphoproteomics, and metabolomics data from healthy and cancerous tissue from eleven clear cell renal cell carcinoma (ccRCC) patients. COSMOS was able to capture relevant crosstalks within and between multiple omics layers, such as known ccRCC drug targets. We expect that our freely available method will be broadly useful to extract mechanistic insights from multi-omics studies.

SYNOPSIS

A new approach integrates multi-omics datasets with a prior knowledge network spanning signaling, metabolism and allosteric regulations. Application to a kidney cancer patient cohort captures relevant cross-talks among deregulated processes.

- A causal multi-omics network is built by integrating multiple ressources spanning signaling, metabolism and allosteric regulations.
- Transcriptomics, phosphoproteomics and metabolomics data are integrated in a set of coherent mechanistic hypotheses using CARNIVAL, a tool contextualizing causal networks.
- This set of coherent mechanistic hypotheses can be mined to identify disease mechanisms and therapeutic targets.
- A network built for a cohort of kidney cancer patients shows coherence with other studies and known therapeutic targets.

Introduction

“Omics” technologies measure at the same time thousands of molecules in biological samples, from DNA, RNA, and proteins to metabolites. Omics datasets are an essential component of systems biology and are made possible by the popularization of analytical methods such as next-generation sequencing or mass spectrometry. Omics data have enabled the unbiased characterization of the molecular features of multiple human diseases, particularly in cancer (preprint: Jelinek & Wu, 2012; Iorio *et al*, 2016; Subramanian *et al*, 2017). It is becoming increasingly common to characterize multiple omics layers in parallel, with so-called “trans-omics analysis”, to gain biological insights spanning multiple types of cellular processes (Sciacovelli *et al*, 2016; Kawata *et al*, 2018; Vitrinel *et al*, 2019). Consequently, many tools are developed to analyze such data (Tenenhaus *et al*, 2014; Argelaguet *et al*, 2018; Sharifi-Noghabi *et al*, 2019; Singh *et al*, 2019; Liu *et al*, 2019c), mainly by adapting and combining existing “single omics” methodologies to multiple parallel datasets. These methods identify groups of measurements and derive integrated statistics to describe them, effectively reducing the dimensionality of the datasets. These methods are useful to provide a global view of the data, but additional processing is required to extract mechanistic insights from them.

To extract mechanistic insights from datasets, some methods (such as pathway enrichment analysis) use prior knowledge about the players of the process being investigated. For instance, differential changes in the expression of the genes that constitute a pathway can be used to infer the activity of that pathway. Methods that a priori define groups of measurements based on known regulated targets (that we call footprints (Dugourd & Saez-Rodriguez, 2019)) of transcription factors (TFs; Alvarez *et al*, 2016; Garcia-Alonso *et al*, 2019), kinases/phosphatases (Wiredja *et al*, 2017), and pathway perturbations (Schubert *et al*, 2018) provide integrated statistics that can be interpreted as a proxy of the activity of a molecule or process. These methods seem to estimate more accurately the status of processes than classic pathway methods (Cantini *et al*, 2018; Schubert *et al*, 2018; Dugourd & Saez-Rodriguez, 2019). Since each of these types of footprint methods works with a certain type of omics data, finding links between them could help to interpret them collectively in a mechanistic manner. For example, one can use a network diffusion algorithm, such as TieDIE (Paull *et al*, 2013), to connect different omics footprints together (Drake *et al*, 2016). This approach provides valuable insights, but diffusion (or random walk) based algorithms do not typically take into account causal information (such as activation/inhibition) that is available and are essential to extract mechanistic information. TieDIE partially addressed this problem by focusing the diffusion process on causally coherent subparts of a network of interest, but it is thus limited to local causality.

Recently, we proposed the CARNIVAL tool (Liu *et al*, 2019b) to systematically generate mechanistic hypotheses connecting TFs through global causal reasoning supported by Integer Linear Programming. CARNIVAL connects activity perturbed nodes such as drug targets with deregulated TFs activities by contextualizing a signed and directed Prior Knowledge Network (PKN). We had hypothesized how such a method could potentially be used to connect footprint-based activity estimates across multiple omics layers (Dugourd & Saez-Rodriguez, 2019).

In this study, we introduce COSMOS (Causal Oriented Search of Multi-Omics Space). This approach connects TF and kinase/phosphatases activities (estimated with footprint-based methods) as well as metabolite abundances with a novel PKN spanning across multiple omics layers (Fig 1). COSMOS uses CARNIVAL's Integer Linear Programming (ILP) optimization strategy to find the smallest coherent subnetwork causally connecting as many deregulated TFs, kinases/phosphatases, and metabolites as possible. The subnetwork is extracted from a novel integrated PKN spanning signaling, transcriptional regulation, and metabolism of > 117,000 edges. CARNIVAL's ILP formulation effectively allows to evaluate the global network's causal coherence given a set of known TF, kinases/phosphatases activities and metabolite abundances. While we showcase this method using transcriptomics, phosphoproteomics, and metabolomics inputs, COSMOS can theoretically be used with any other additional inputs, as long as they can be linked to functional insights (for example, a set of deleterious mutations). As a case study, we generated transcriptomics, phosphoproteomics, and metabolomics datasets from kidney tumor tissue and corresponding healthy tissue out of nine clear cell renal cell carcinoma (ccRCC) patients. We estimated changes of activities of TFs and kinase/phosphatases as well as metabolite abundance differences between tumor and healthy tissue. We integrated multiple curated resources of interactions between proteins, transcripts, and metabolites together to build a meta PKN. Next, we contextualized the meta PKN to a specific experiment. To do so, we identified causal pathways from our prior knowledge that connect the observed changes in activities of TFs, kinases, phosphatases, and metabolite abundances between tumor and healthy tissue. These causal pathways can be used as hypothesis generation tools to better understand the molecular phenotype of kidney cancer. We also refactored all functions to run the COSMOS approach into an R package.

Results

Building the multi-omics dataset

To build a multi-omics dataset of renal cancer, we performed transcriptomics, phosphoproteomics, and metabolomics analyses of renal nephrectomies and adjacent normal tissues of 11 renal cancer patients (for details on the patients see Dataset EV1).

First, we processed the different omics datasets to prepare for the analysis. For the transcriptomics dataset, 15,919 transcripts with average counts > 50 were kept for subsequent analysis. In the phosphoproteomics dataset, 14,243 phosphosites detected in at least four samples were kept. In the metabolomics dataset, 107 metabolites detected across 16 samples were kept. Principal component analysis (PCA) of each omics dataset independently showed a clear separation of healthy and tumor tissues on the first component (transcriptomics: 40% of explained variance (EV), phosphoproteomics: 26% of EV, metabolomics: 28% of EV, Fig EV1), suggesting that tumor sample displayed molecular deregulations spanning across signaling, transcription and metabolism. Each omics dataset was independently submitted to differential (tumor vs. healthy tissue) analysis using LIMMA (Ritchie *et al*, 2015). Consistently with the PCA, a volcano plot overlapping the results of the differential analysis of each omics showed that the transcriptomics dataset led to larger differences and smaller *P*-values than phosphoproteomics and metabolomics extracted from the same samples (Fig EV2). This is further apparent by the number of hits under a given false discovery rate (FDR, Benjamini & Hochberg, 1995) threshold. We obtained 6,699 transcripts and 21 metabolites significantly regulated with FDR < 0.05. While only 11 phosphosites were found under 0.05 FDR, 447 phosphosites had an FDR < 0.2. This result confirmed the deep molecular deregulations of tumors spanning across signaling, transcription, and metabolism. Then, the differential statistics for all tested (not just the ones under the FDR threshold) transcripts, phosphopeptides, and metabolites were used for further downstream analysis, as explained below.

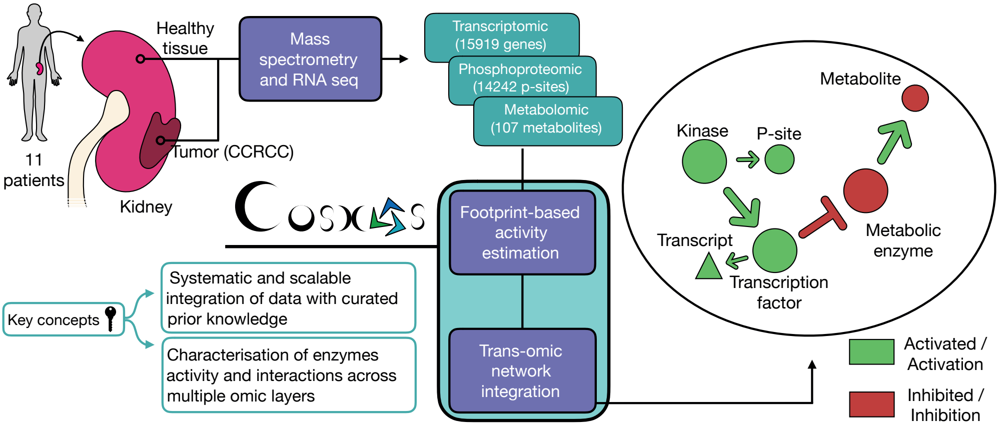

Healthy tissue
11 patients
Kidney
Tumor (CCRCC)
Mass spectrometry and RNA seq
Transcriptomic (15919 genes)
Phosphoproteomic (14242 p-sites)
Metabolomic (107 metabolites)
COSMOS
Footprint-based activity estimation
Trans-omic network integration
Key concepts
Systematic and scalable integration of data with curated prior knowledge
Characterisation of enzymes activity and interactions across multiple omic layers
Metabolite
Kinase
P-site
Transcript
Transcription factor
Metabolic enzyme
🟩 Activated / Activation
🟥 Inhibited / Inhibition

Figure 1. Overview of analysis pipeline.

From left to right: We sampled and processed 11 patient tumors and healthy kidney tissues from the same kidney through RNA-sequencing and 9 of those same patients through mass spectrometry to characterize their transcriptomics, phosphoproteomics, and metabolomics profiles. We calculated differential abundance for each detected gene, phosphopeptide, and metabolite. We estimated kinase and transcription factor activities using the differential analysis statistics and footprint-based methods. We used the estimated activities alongside the differential metabolite abundances to contextualize (i.e., extract the subnetwork that better explains the phenotype of interest) a generic trans-omics causal prior knowledge network (meta PKN).

Footprint-based transcription factor, kinase, and phosphatase activity estimation

We then performed computational footprint analysis to estimate the activity of proteins responsible for changes observed in specific omics datasets. By the term "activity", we refer to a quantifiable proxy of the function of a protein, estimated based on the footprint left by said activity. This definition can apply, but is not limited to, an enzyme's catalytic activity. Footprint-based activity estimation (Dugourd & Saez-Rodriguez, 2019) relies on the concept that the measured abundances of molecules (such as phosphopeptides or transcripts) can be used as a proxy of upstream (direct or indirect) regulator activities responsible for those changes (Rhodes *et al*, 2005; Casado *et al*, 2013; Ochoa *et al*, 2016). In the case of TF activity estimation, this means that measured changes in the abundances of transcripts give us information about the changes of activities of the transcription factors that regulate their abundance. An activity estimation only depends on the changes of the abundances measured in its target transcripts, not its own transcript abundance. In this study, we used the VIPER algorithm (Alvarez *et al*, 2016) to estimate the activity of transcription factors and kinases based on transcript and phosphopeptide abundances changes, respectively. For transcriptomics and phosphoproteomics data, this analysis estimates transcription factor and kinases/phosphatase activity, respectively. 24,347 transcription factors (TFs) to target interactions (i.e., transcript under the direct regulation of a transcription factor) were obtained from DoRothEA (Garcia-Alonso *et al*, 2019), a meta-resource of TF-target interactions. Those TF-target interactions span over 365 unique transcription factors. In parallel, 33,616 interactions of kinase/phosphosphate and their phosphosite targets (i.e., phosphopeptides directly (de)phosphorylated by specific kinases (phosphatases)) were obtained from OmniPath (Türei *et al*, 2016) kinase substrate network, a meta-resource focused on curated information on signaling processes. Only TFs and kinases/phosphatases with at least 10 and 5 detected substrates, respectively, were included. This led to the activity estimation of 328 TFs and 174 kinases. In line with the results of the differential analysis, where fewer phosphosites were deregulated than transcripts, TF activities displayed a stronger deregulation than kinases. TF activity scores reached a maximum of 8.7 standard deviations (sd) for Transcription Factor Spi-1 Proto-Oncogene (*SPI1*) (compared to the null score distribution; sd compared to null is also referred to as a normalized enrichment score, NES), while kinase activity scores reached a maximum of 4.6 NES for Casein Kinase 2 Alpha 1 (*CSNK2A1*). In total, 102 TFs and kinases/phosphatase had an absolute score over 1.7 NES ($P < 0.05$) and were considered significantly deregulated in kidney tumor samples (Fig 2A). The presence of several known signatures of ccRCC corroborated the validity of our analysis. For instance, hypoxia (*HIF1A*), inflammation (*STAT2,*Fig 2B), and oncogenic (*MYC*, Cyclin Dependent Kinase 2 and 7 (*CDK2/7,* (Fig 2C)) markers were up-regulated in tumors compared to healthy tissues (Zeng *et al*, 2014; Schödel *et al*, 2016; Clark *et al*, 2020). Furthermore, among suppressed TFs we identified, the *HNF4A* gene has been previously associated with ccRCC (Lucas *et al*, 2005).

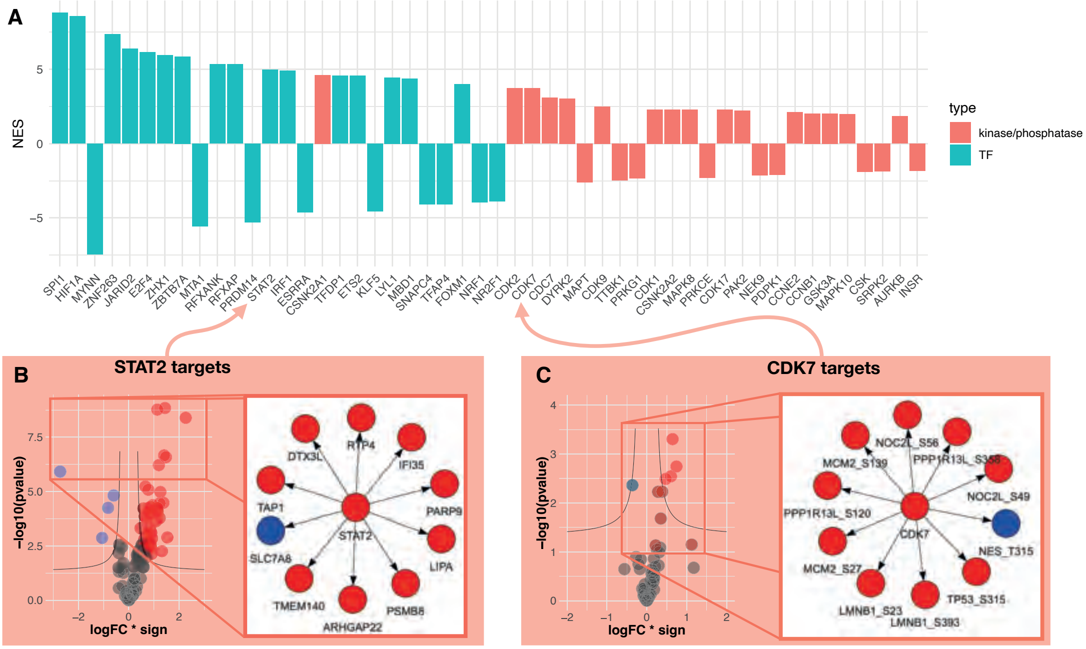


Figure 2. Differentially regulated transcription factor, kinase, and phosphatase activities cancer vs. healthy tissue.

A Bar plot displaying the normalized enrichment score (NES, proxy of activity change) of the 25 up- or down-regulated TF and top 25 up- or down-regulated kinase and phosphatases activities between kidney tumor and adjacent healthy tissue. **B** Right panel shows the 10 most changing RNA abundances of the STAT2 regulated transcripts . Left panel shows the change of abundances of all STAT2 regulated transcripts that were used to estimate its activity change. X-axis represents log fold change of regulated transcripts multiplied by the sign of regulation (1 for inhibition and 1 for activation of transcription). Y-axis represents the significance of the log fold change (log10 of P-value, LIMMA moderated unpaired t-test Pvalues). The black line is defined by the following function when fold change is negative : y = abs(hAss 1 + x/(x + vAss)); and y = abs(hAss 1 + x/(x vAss)) when fold change is positive. abs() is the absolute value, hAss is the horizontal asymptote (hAss = 1.3) and vAss is the vertical asymptote (vAss = 0.3). C Right panel shows the 10 most changing phosphopeptide abundances of the CDK7 regulated phosphopeptides. Left panel shows the change of abundances of all CDK7 regulated phosphopeptides that were used to estimate its activity change. X-axis represents log fold change of regulated transcripts multiplied by the sign of regulation (1 for inhibition and 1 for activation of transcription). Y-axis represents the significance of the log fold change (log10 of P-value, LIMMA moderated unpaired t-test P-values). The black line is defined by the following function when fold change is negative : y = abs(hAss 1 + x/(x + vAss)); and y = abs (hAss 1 + x/(x vAss)) when fold change is positive. Where abs() is the absolute value, hAss is the horizontal asymptote (hAss = 1.3) and vAss is the vertical asymptote (vAss = 0.3).

Causal network analysis

We set out to find potential causal mechanistic pathways that could explain the changes we observed in TF, kinases/phosphatase activities, and metabolic abundances. Thus, we developed a systematic approach to search in public databases, such as OmniPath, for plausible causal links between significantly deregulated TFs, kinases/phosphatases and metabolites. In brief, we investigated if changes in TF, kinase/phosphatase activities, and metabolite abundance can explain each other with the support of literature-curated molecular interactions. An example of such a mechanism can be the activation of the transcription of *MYC* gene by *NFKB1*. Since both *NFKB1* and *MYC* display increased activities in tumors, and there is evidence in the literature that *NFKB1* can regulate *MYC* transcription (FANTOM4 database), it may indicate that this mechanism is responsible for this observation.

First, we needed to map the deregulated TFs, kinases, and metabolites on a causal prior knowledge network spanning over signaling pathways, gene regulation, and metabolic networks. Hence, we combined multiple sources of experimentally curated causal links together to build a meta causal prior knowledge network. This meta PKN must include direct causal links between proteins (kinase to kinase, TF to kinase, TF to metabolic enzymes, etc…), between proteins and metabolites (reactants to metabolic enzymes and metabolic enzymes to products) and between metabolites and proteins (allosteric regulations). High confidence (≥ 900 combined score) allosteric regulations of the STITCH database (Szklarczyk *et al*, 2016) were used as the source of causal links between metabolites and enzymes (Fig 3A). The directed signed interactions of the OmniPath database were used as a source of causal links between proteins (Fig 3B). The human metabolic network Recon3D (Brunk *et al*, 2018) (without cofactors and hyper-promiscuous metabolites, see Material and Methods) was converted to a causal network and used as the source of causal links between metabolites and metabolic enzymes (Fig 3C). The resulting meta PKN consists of 117,065 interactions and contains causal paths linking TFs/kinases/phosphatases with metabolites and vice versa in a machine readable format. This network is available in the COSMOS R package.

We then used the meta PKN to systematically search causal paths between the deregulated TFs, kinases/phosphatases, and metabolites using an ILP optimization approach (see Material and Methods, Meta PKN contextualization). Here, we used CARNIVAL with our meta PKN to find the smallest sign-coherent subnetwork connecting as many deregulated TFs, kinases/phosphatases, and metabolites as possible. First, we filtered out all interactions that do not involve genes expressed in our samples. Then, we removed nodes beyond a given number of steps downstream of inputs. We also removed any edge that leads to an incoherence between a TF activity score and the transcript abundance change of its targets (Appendix Fig S1A). We then performed a CARNIVAL run from TFs/kinases/phosphatases to metabolites to estimate the activity of TFs in the COSMOS solution network. These activities are used to filter out incoherent transcriptional regulation events from the meta PKN. Then, CARNIVAL is used to find causal paths going from TFs/kinases/phosphatases to the metabolites (the "forward network"). Finally, CARNIVAL is used to go from metabolites to TFs/Kinases/phosphatases ("backward network"). The choice of TFs/Kinases/phosphatases and metabolites to be included is

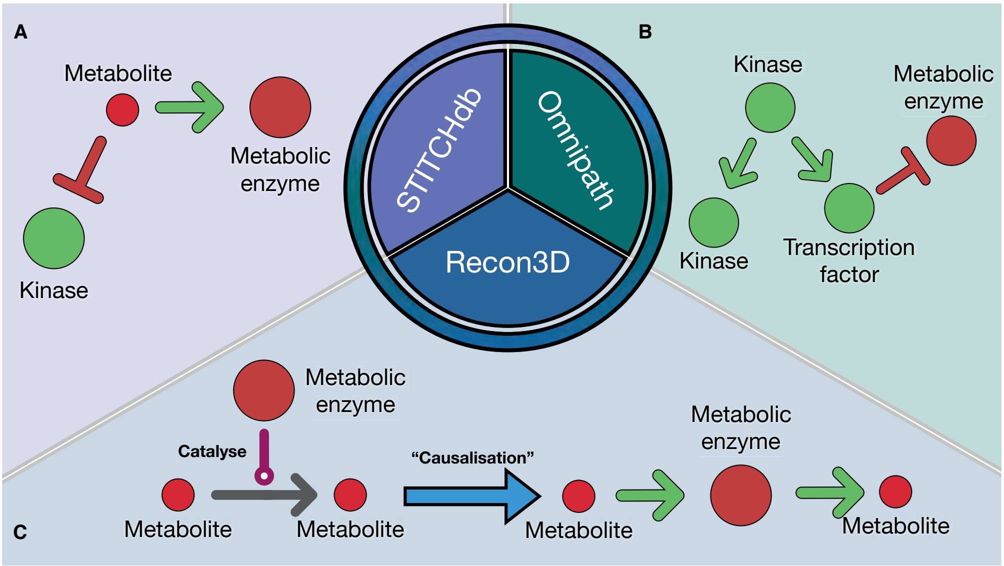


Figure 3. Graphical explanation of meta PKN sources.

A–C Schematic representation of the meta generic network (meta PKN) created combining STITCH, OmniPath and Recon3D. (**A**) STITCH provides information on inhibition/activation of enzyme activities mediated by metabolites. (**B**) OmniPath provides information inhibition/activation of enzyme activities mediated by other enzymes based mainly on curated resources. (**C**) Recon3D provides information on reactants and products associated with metabolic enzymes. To make this information compatible with the causal edges from OmniPath and STITCH, the interactions of recon3D are converted so that reactants "activate" their metabolic enzymes, which themselves "activate" their products.

detailed in Appendix Note 1. We combined the two networks (making union of the two sets edges and the union of the two sets of node attributes) to obtain a network with 449 unique edges (Appendix Fig S2, Dataset EV5). CARNIVAL finds a direct path connecting downstream measurements with upstream nodes, and thus, the solution networks do not contain loops. Loops can however appear in the final merged network when nodes are overlapping between "forward" and "backward" runs.

We then used our network to investigate the regulation of relevant signaling cascades and metabolic reactions in ccRCC. An over-representation analysis of the network solution nodes (with the hallmark genesets of MSigDB) displayed the interferon gamma (IFNg) response as the top significant pathway in our COSMOS network. Hence, we focused on the interaction members of this pathway (such as *NFKB1*, *HIF1A*, and *PNP*) and their crosstalks with metabolic deregulations to assess the relevance of the mechanistic hypotheses generated by COSMOS. We found that *NFKB1*, a central actor of the IFNg pathway is activated in ccRCC, consistently with other reports (Zhang *et al*, 2018; Rodrigues *et al*, 2018) where it was also demonstrated to be regulated by the *PI3K/AKT* pathway (An & Rettig, 2007). Interestingly, COSMOS also proposed the activation of *BCAT1*, one of the key enzymes of the branched-chain amino acid metabolism, orchestrated by *HIF1A* and *MYC* (Gordan *et al*, 2008; Ananieva & Wilkinson, 2018). Both mechanisms are shown in Fig 4 (1) and (2).

Of note, COSMOS provided deeper insights into these molecular mechanisms by linking *MYC* activation to *NFKB1*. The COSMOS model suggests that *MYC* up-regulates the expression of the metabolic enzyme *BCAT1*, potentially leading to the observed higher levels of glutamate, glutamine and reduced glutathione in ccRCC (marked as (2) in Fig 4). A strong role of *MYC* and glutamine metabolism in ccRCC development is known (Shroff *et al*, 2015). Consistently with what was hypothesized in a recent proteogenomics ccRCC study (Clark *et al*, 2020), we were able to capture crosstalks between members of the interferon gamma pathway (such as *JUN*), *YY1* and metabolic down-regulation observed in our data ((3) in Fig 4). COSMOS highlighted how *YY1* inhibition might be connected with the depletion of adenine, hypoxanthine, and inosine through regulation of the *ADA* and *PNP* metabolic enzyme (Popławski et al, 2017). The low levels of adenosine predicted by COSMOS might also be potentially linked to the down-regulation of *AKT3* and

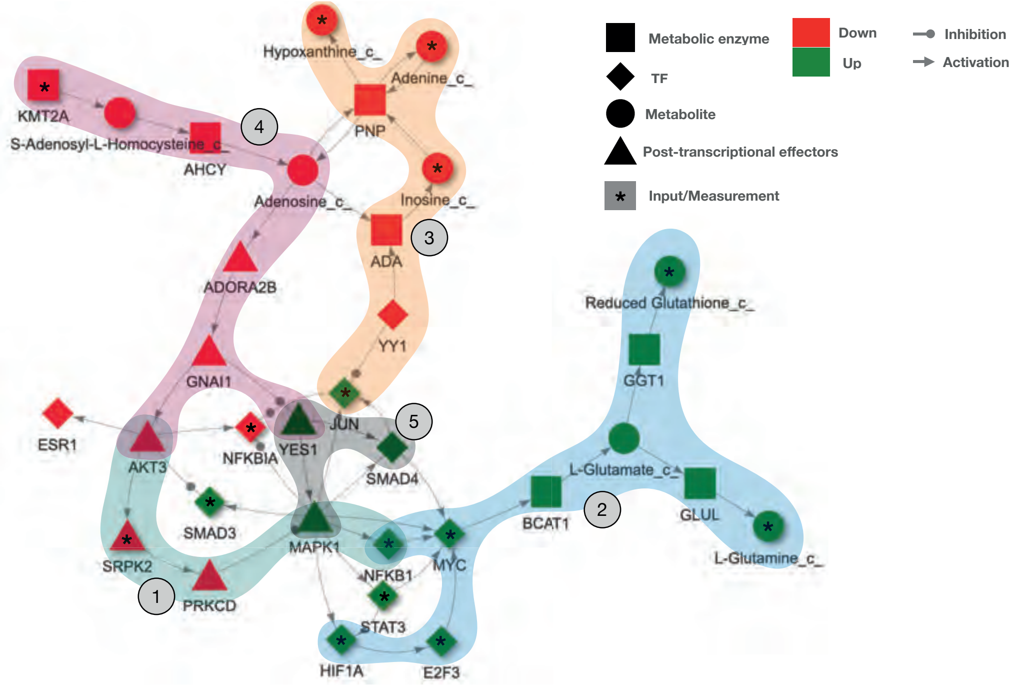


Figure 4. COSMOS subnetwork centered on the interferon gamma response pathway.

The figure includes the main members of the interferon gamma response pathway, the most enriched cancer hallmark in the full COSMOS network. We also display the metabolic enzymes that were hypothesized to be influenced downstream of this pathway, such as BCAT1 and PNP. The numbered mechanisms are discussed in the main text.

up-regulation of *YES1*, through a cascade which involves both *ADORA2B* and *GNAI1*, downstream of s-Adenosyl-L-homocysteine and inhibition of *KMT2A* ((4) in Fig 4). Finally, the COSMOS model showed a significant activation of *MAPK1* and *SMAD4* downstream of *YES1* (a member of the SRC family) ((5) in Fig 4).

Consistency, robustness, and flexibility

Due to the combined effect of experimental noise and incompleteness of prior knowledge (kinase/substrate interactions, TF/targets interactions and meta PKN), it is critical to assess the performance of the pipeline presented above.

One way to estimate the performance is to check if the COSMOS mechanistic hypotheses correspond to correlations observed in tumor tissues (Appendix Fig S1B). Thus, on the one hand, a topology-driven co-regulation network was generated from the COSMOS network. The assumption behind this network is that direct downstream targets of the same enzymes should be co-regulated. On the other hand, a data-driven correlation network of TFs, kinases, and phosphatases was generated from tumor tissues alone. Assuming thresholds of absolute values of correlation ranging between 0 and 1 to define true positive co-regulations, the comparison between the topology-driven co-regulation network and the data-driven correlation network yielded a TPR ranging between 0.55 and 0 (*n* = 269 pairs of predicted/measured co-regulations) for the predictions (Appendix Fig S3). It performed consistently better than a random baseline (see Material and Methods) over the considered range of correlation coefficient thresholds. We also compared the results with network solutions obtained hiding either TFs or kinases/phosphatases. When TFs were hidden, COSMOS performed consistently better than the random baseline and reached a maximum TPR of 0.62. Of note, this curve was estimated from only *n* = 21 co-regulation events. When kinases and phosphatases were hidden, COSMOS performed again consistently better than the random baseline and reached a maximum TPR of 0.58 (*n* = 228). In both cases, the performance of COSMOS was slightly larger than the full COSMOS performance (TPR = 0.55). This could be due to a lack of consistency across the omics data, although due to the low number of comparisons we could not make a conclusive statement. These results suggest that COSMOS' performance is relatively robust to removing either the phosphoproteomics or transcriptomics layers when trying to find connection between signaling and metabolism. However, using the three omics layers together yielded a larger network (367 edges (full) vs. 294 (hidden kinases) and 135 edges (hidden TFs)) and denser (1.67 edge/node ratio vs. 1.54 and 1.19 edge/node ratio, respectively) than when one omics layer was removed (Dataset EV3). Hence, using all layers yield a greater number of mechanistic hypotheses, even if not necessarily of higher quality.

To study the robustness of COSMOS to changes in the PKN, we generated a series of partially degraded PKN by randomly shuffling an increasing number of edges in the original PKN (2, 10, 20, 30, 40, 50% of all edges shuffled completely randomly). We ran COSMOS with each version of the PKN. We first compared the results of the "forward" COSMOS runs (connecting TFs and kinases with downstream metabolites). We calculated the absolute difference between the edge weight of the results (see Materials and Methods, meta PKN contextualization) obtained from each shuffled PKN with the result obtained from the original PKN. The edge weight represents the frequency of appearance (in %) of an edge across all the networks in the pool of network solutions. This showed that for the 2% shuffled network, the differences were relatively small (median of the absolute weight difference = 10), with 4% of edges flipped (i.e., 0 weight in shuffled network and 100 weight in original network, or vice versa). As expected, the differences were higher regarding the other shuffled networks, with medians weight differences of 10, 14, 23, 50, 28, and 35 for the 2, 10, 20, 30, 40, and 50% shuffled PKN, respectively (Appendix Fig S4A).

We then compared the results of the "backward" COSMOS runs (connecting metabolites with downstream TFs and kinases). Here, the comparison was far less quantitative because the optimization reported only a single solution for all runs except in the case of the 20% shuffled PKN (Appendix Fig S4B). 61, 31, 18, 12, 8, and 8% of edge weight differences were equal to 0 for the 2, 10, 20, 30, 40, and 50% shuffled PKN, respectively.

In both "forward" and "backward" runs, the network results had a relatively similar number of edges from the original and shuffled PKNs (min = 142, max = 342, mean = 263, SD = 63). The optimization thus consistently excluded a common set of edges covering the vast majority of the network, that contains over 56,000 edges.

We also compared the results we obtained from our samples with results obtained using another independent ccRCC dataset. We obtained the transcriptomics and phosphoproteomics dataset of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) ccRCC patient cohort (Clark *et al*, 2020). Following the same approach as with our patient samples, we performed the differential analysis between tumor and healthy tissue for both omics datasets and estimated TFs and kinase/phosphatase activities. Then, we ran COSMOS to find mechanistic hypotheses explaining the connections between deregulated transcription factors and kinases/phosphatases. The resulting COSMOS network was coherent with the results shown in the original publication and also provided additional information on the crosstalks between deregulated kinases and transcription factors. In particular, COSMOS captured the signaling crosstalks between *EGF*, *VEGF*, *AKT*, *MAPK*, *MTOR*, *NFKB,* and *MYC* (Dataset EV4). Finally, we compared which biological processes were captured in the COSMOS network generated from the data of our patient samples and the COSMOS network generated from the CPTAC ccRCC patient cohort. As shown in Fig EV3, the top over-represented pathways were very consistent between the two studies. Notably, *PI3K-AKT-MTOR* signaling and G2M checkpoint (Clark *et al*, 2020), *TNFA* signaling via *NFKB* (Al-Lamki *et al*, 2010), interferon gamma response (Thapa *et al*, 2013), *WNT* beta catenin signaling (Xu *et al*, 2016), and *IL6 JAK STAT3* signaling pathway were all significantly over-represented (*P* < 0.02).

Finally, we applied COSMOS to a public breast cancer dataset including transcriptomics and fluxomics measurements (Katzir *et al*, 2019) to connect signaling directly with metabolic flux estimation, instead of metabolite abundance measurements as done in the previous cases. We performed a differential analysis of transcript abundance and flux values between tumor cells cultured with and without glutamine. We then looked for mechanistic hypotheses connecting TF activity deregulations and changes in flux values. Coherently with the original study, almost all metabolites of the TCA cycle, glycolysis and pentose phosphate pathway were predicted to be down-regulated by COSMOS (Appendix Fig S5). Interestingly, COSMOS finds *HIF1A* as a master regulator of glycolysis through his effect on *HK1/2*, *GAPDH*, *GCK*, *ENO1,* and *LDHA* transcription. This is consistent with the known role of *HIF1A* in breast cancer (Samanta *et al*, 2014; Masoud & Li, 2015; Zhang *et al*, 2015; Singh *et al*, 2017). The down-regulation of *MYC* is also in line with the decreased activity of *HK2* and *LDHA* and *GLS1* enzymes which are important in aerobic glycolysis and glutamine catabolism (Dong *et al*, 2020).

Discussion

In this paper, we present COSMOS, an analysis pipeline to systematically generate mechanistic hypotheses by integrating multi-omics datasets with a broad range of curated resources of interactions between protein, transcripts, and metabolites.

We first showed how TF, kinase, and phosphatase activities could be coherently estimated from transcriptomics and phosphoproteomics datasets using footprint-based analysis. This is a critical step before further mechanistic exploration. Indeed, transcript and phosphosite usually offer limited functional insights by themselves as their relationship with corresponding protein activity is usually not well characterized. Yet, they can provide information on the activity of the upstream proteins regulating their abundances. Thus, the functional state of kinases, phosphatases, and TFs is estimated from the observed abundance change of their known targets, i.e., their molecular footprint. Thanks to this approach, we could simultaneously characterize protein functional states in tumors at the level of signaling pathway and transcriptional regulation. Key actors of hypoxia response, inflammation pathway, and oncogenic genes were found to have especially strong alteration of their functional states, such as *HIF1A* , *EPAS1*, *STAT1/2*, *MYC,* and *CDK2*. Loss of *VHL* is a hallmark of ccRCC and is directly linked to the stability of the *HIF* (*HIF1A* and *EPAS1*) proteins found deregulated by our analysis (Maxwell *et al*, 1999; Ivan *et al*, 2001; Jaakkola *et al*, 2001). Finding these established signatures of ccRCC to be deregulated in our analysis is a confirmation of the validity of this approach.

We then applied COSMOS with a novel meta causal Prior Knowledge Network spanning signaling, transcription, and metabolism to systematically find potential mechanisms linking deregulated protein activities and metabolite concentrations. To the best of our knowledge, this is the first attempt to integrate these three omics layers together in a systematic manner using causal reasoning. Previous methods studying signaling pathways with multi-omics quantitative datasets (Drake *et al*, 2016) connected TFs with kinases but they were limited by the preselected locally coherent subnetwork of the TieDIE algorithm. Introducing global causality along with metabolomics data allows us to obtain a direct mechanistic interpretation of links between proteins at different regulatory levels and metabolites. The goal of our approach is to find a coherent set of such mechanisms connecting as many of the observed deregulated protein activities and metabolite concentrations as possible. Using COSMOS is particularly interesting as all the proposed mechanisms between pairs of molecules (proteins and metabolites) have to be plausible not only in the context of their own pairwise interaction but also with respect to all other molecules that we wish to include in the model. For example, the proposed activation of *MYC* by *NFKB1* and *MAPK1* is further supported by *STAT3* activation, because *MAPK1* is also known to activate *STAT3*. Thus, we developed COSMOS to scale this type of reasoning up to the entire PKN with all significantly deregulated protein activities and metabolites. We relied on an ILP optimization through the CARNIVAL R package (Liu *et al*, 2019b) to contextualize this PKN with our data. We refined the optimization procedure to handle this very large PKN and built an R package to facilitate others to use it with their own data. Given a set of deregulated TFs, kinases/phosphatases, or metabolites, COSMOS provides the users with a set of coherent mechanistic hypotheses to explain changes observed in a given omics layer with upstream regulators from other omics layers. Thus, its aim is to integrate measured data with prior knowledge in a consistent and systematic manner, not to explicitly predict the outcome of new experiments.

Since the interferon gamma response pathway was the most over-represented cancer hallmark in the COSMOS network solution, we investigated further the relevance of the mechanistic hypothesis connecting members of this pathway. The network showed that the crosstalks between *MAPK1*, *NFKB1*, *MYC*, *HIF1A,* and *YY1* could explain the deregulation in glutamine and reduced glutathione metabolism, as well as inosine, hypoxanthine, and adenine. These were particularly relevant as they were important interactions in ccRCC. *MYC* and glutamine metabolism appear to be an interesting therapeutic target of ccRCC (Shroff *et al*, 2020). *YY1* is a known indirect inhibitor of *MYC* involved in cancer development (Austen *et al*, 1998). The COSMOS network showed *YY1* could also potentially have a role in the down-regulation of the *ADA* and *PNP* metabolic enzyme activities. Coherently, *PNP* has been shown to be non-essential in ccRCC cell lines, which is expected from down-regulated metabolic enzymes (Gatto *et al*, 2015). In addition, the link shown by COSMOS between *NFKB1* and *MYC* can have implications for the treatment of ccRCC, due to its pivotal role in arsenite (a drug used in chemotherapy) treatment of cancer (Huang *et al*, 2014). Furthermore, the activation of the *NFKB1-MYC* link in *FBW7*-deficient cells seems to sensitize them to Sorafinib (a *MEK-Raf* inhibitor), a drug used in treatment of primary kidney cancer (Huang *et al*, 2014). In addition, *NFKB1* and *MYC* are both promising ccRCC treatment targets (Peri *et al*, 2013; Bailey *et al*, 2017). The link shown by COSMOS between *KMT2A* and adenosine is interesting, because *KMT2A* mutations have been reported in a number of ccRCC patients (Yan *et al*, 2019), suggesting that this enzyme might play a functional role in ccRCC development. Moreover, it has been proposed, at least in vitro, that ccRCC cell lines with low basal levels of phospho-*AKT* were sensitive to treatment with an adenosine analog (Kearney *et al*, 2015). The link betwe en *YES1*, *MAPK1,* and *SMAD4* in the COSMOS network is especially relevant considering that *YES1* is a known targetable oncogene (Hamanaka *et al*, 2019). These examples illustrate the ability of COSMOS to extract mechanistic hypotheses to understand and potentially improve treatment of cancer by integration of multiple omics data and prior knowledge.

However, it is important to mention that COSMOS is only aimed at providing hypotheses to further explore experimentally. COSMOS does not aim at recapitulating all the molecular interactions that may be happening in a given context. Currently, COSMOS simply provides a large set of coherent mechanistic hypotheses, given the data and prior knowledge available. We argue that this facilitates the interpretation of a complex multi-omics dataset and guides the exploration of biological questions.

We assessed the performances and robustness of our approach. We computed a tumor specific correlation network of TF and kinase activities and compared it to the co-regulation predicted by COSMOS. This yielded encouraging results, though imperfect, underscoring the fact that the mechanisms proposed by COSMOS—like those by any similar tool—are hypotheses. It also highlighted that adding more omics data to integrate allows to generate more hypotheses and connect them together, but does not necessarily improve their predictive performances.

There are three main known limits to the predictions of COSMOS. First, the input data are incomplete. Only a limited fraction of all potential phosphosites and metabolites are detected by mass spectrometry. This means that we have no information on a significant part of the PKN; part of the unmeasured network is kept in the analyses and the values are estimated as intermediate “hidden values”. Second, not all regulatory events between TFs, kinase, and phosphatases and their targets are known, and activity estimation is based only on the known regulatory relationships. Thus, many TFs, kinase, and phosphatases are not included because they have no curated regulatory interactions or no detected substrates in the data. Third, and conversely, COSMOS will find putative explanations within the existing prior knowledge that may not be the true mechanism.

These problems mainly originate from the importance that is given to prior knowledge in this method. Since prior knowledge is by essence incomplete, the next steps of improvement could consist of finding ways to extract more knowledge from the observed data to weight in the contribution of prior knowledge. For instance, one could use the correlations between transcripts, phosphosites and metabolites to quantify the interactions available in databases such as OmniPath. Importantly, any other omics that relate to active molecules (such as miRNAs or metabolic enzyme fluxes) can be used to estimate protein activities through footprint approaches (such as DNA accessibility or PTMs other than phosphorylation) can be seamlessly integrated (as we showed with the fluxomic of the breast cancer dataset). Moreover, COSMOS was designed to work with bulk omics datasets, and it will be very exciting to find ways of applying this approach to single cell datasets. Encouragingly, the footprint methods that bring data into COSMOS seem fairly robust to the characteristics of single-cell RNA data such as dropouts (Holland *et al*, 2020). Related to the importance of prior knowledge, the PKN can also depend on how we interpret the information we have about molecular interactions. In particular, we converted the reaction network of Recon3D into a causal network where metabolite reactants “activate” metabolic enzymes, and metabolic enzymes “activate” metabolite products. This first approximation assumes that metabolite abundances are only driven by their production rates. We plan to refine this in the future to include that metabolite abundances can change as a result of consumption as well. Finally, we expect that in the future, data generation technologies will increase coverage and our prior knowledge will become more complete, reducing the mentioned limitations. In the meantime, we believe that COSMOS is already a useful tool to extract causal mechanistic insights from multi-omics studies.

Methods and Protocols

Sample collection and processing

We included a total of 22 samples from 11 renal cancer patients (6 men, age 65.0 ± 14.31, 5 women, age 65.2 ± 9.257 (mean ± SD)) for transcriptomics. Phosphoproteomics was also measured in a subset of 18 samples from 9 of these patients (6 men, age 65 ± 14.31; 3 women, age 63.33 ± 11.06 (mean ± SD)), and metabolomics was also measured in 16 samples from 8 out of these 9 patients (5 men, age 62 ± 13.23; 3 women, age 63.33 ± 9.89 (mean ± SD), Fig EV4, Dataset EV1). Patients underwent nephrectomy due to renal cancer. We processed tissue from within the cancer and a distant unaffected area of the same kidney. The tissue was snap-frozen immediately after nephrectomy within the operation room. The clinical data of the included patients is outlined in Dataset EV1. Histological evaluation showed clear renal cell carcinoma in all patients.

Ethics

The local ethics committee of the University Hospital RWTH Aachen approved all human tissue protocols for this study (EK-016/17). The study was performed according to the declaration of Helsinki. Kidney tissues were collected from the Urology Department of the University Hospital Eschweiler from patients undergoing partial/- or nephrectomy due to kidney cancer. All patients gave informed consent.

Human tissue processing

Kidney tissues were sampled by the surgeon from normal and tumor regions. The tissue was snap-frozen on dry-ice or placed in prechilled University of Wisconsin solution (#BTLBUW, Bridge to Life Ltd., Columbia, U.S.) and transported to our laboratory on ice.

RNA Isolation, library preparation, NGS sequencing

RNA was extracted according to the manufacturer 's instructions using the RNeasy Mini Kit (QIAGEN). For rRNA-depleted RNA-seq using 1 and 10 ng of diluted total RNA, sequencing libraries were prepared with KAPA RNA HyperPrep Kit with RiboErase (Kapa Biosystems) according to the manufacturer's protocol. Sequencing libraries were quantified using quantitative PCR (New England Biolabs, Ipswich, USA). Equimolar pooling of the libraries was normalized to 1,4 nM, denatured using 0.2 N NaOH and neutralized with 400 nM Tris pH 8.0 prior to sequencing. Final sequencing was performed on a Novaseq6000 platform (Illumina) according to the manufacturer's protocols (Illumina, CA, USA).

Metabolomics

Snap-frozen tissue specimens were cut and weighed into Precellys tubes prefilled with ceramic beads (Stretton Scientific Ltd., Derbyshire, UK). An exact volume of extraction solution (30% acetonitrile, 50% methanol, and 20% water) was added to obtain 40 mg specimen per mL of extraction solution. Tissue samples were lysed using a Precellys 24 homogenizer (Stretton Scientific Ltd., Derbyshire, UK). The suspension was mixed and incubated for 15 min at 4°C in a Thermomixer (Eppendorf, Germany), followed by centrifugation (16,000 *g*, 15 min at 4°C). The supernatant was collected and transferred into autosampler glass vials, which were stored at −80°C until further analysis.

Samples were randomized to avoid bias due to machine drift and processed blindly. LC-MS analysis was performed using a Q Exactive mass spectrometer coupled to a Dionex U3000 UHPLC system (both Thermo Fisher Scientific). The liquid chromatography system was fitted with a Sequant ZIC-pHILIC column (150 mm × 2.1 mm) and guard column (20 mm × 2.1 mm) from Merck Millipore (Germany) and temperature maintained at 45 °C. The mobile phase was composed of 20 mM ammonium carbonate and 0.1% ammonium hydroxide in water (solvent A), and acetonitrile (solvent B). The flow rate was set at 200 µL/min with the gradient described previously (Mackay *et al*, 2015). The mass spectrometer was operated in full MS and polarity switching mode. The acquired spectra were analyzed using XCalibur Qual Browser and XCalibur Quan Browser software (Thermo Scientific).

Phosphoproteomics

Snap-frozen tissues were heat inactivated (Denator T1 Heat Stabilizor, Denator) and transferred to a GndCl solution (6 M GndCl, 25 mM Tris, pH 8.5, Roche Complete Protease Inhibitor tablets (Roche)) and homogenized by ceramic beads using 2 steps of 20 s at 5,500 rpm (Precellys 24, Bertin Technologies). The tissues were heated for 10 min at 95°C followed by micro tip sonication on ice and clarified by centrifugation (20 min, 16,000 *g*, 4°C). Samples were reduced and alkylated by adding 5 mM tris(2-carboxyethyl)phosphine and 10 mM chloroacetamide for 20 min at room temperature.

Lysates were digested by Lys-C (Wako) in an enzyme/protein ratio of 1:100 (w/w) for 1 h, followed by a dilution with 25 mM tris buffer (pH 8.5), to 2 M guanidine-HCl and further digested overnight with trypsin (Sigma-Aldrich; 1:100, w/w). Protease activity was quenched by acidification with TFA, and the resulting peptide mixture was concentrated on C18 Sep-Pak Cartridges (Waters). Peptides were eluted with 40% ACN followed by 60% ACN. The combined eluate was reduced by SpeedVac, and the final peptide concentration was estimated by measuring absorbance at *A*280 on a NanoDrop (Thermo Fisher Scientific). Peptide (300 µg) from each sample was labeled with 1 of 11 different TMT reagents according to the manufacturer's protocol (Thermo Fisher Scientific) for a total of four TMT sets. Each set comprised 10 samples and a common internal reference (composed of equal amounts of digested material from all samples).

After labeling, the samples were mixed and phosphopeptides were further enriched using titanium dioxide beads (5 µm Titansphere, GL Sciences, Japan). TiO2 beads were pre-incubated in 2,5-dihydroxybenzoic acid (20 mg/ml) in 80% ACN and 1% TFA (5 µl/mg of beads) for 20 min. Samples were brought to 80% ACN and 5% TFA. 1.5 mg beads (in 5 µl of DHB solution) were added to each sample, which was then incubated for 20 min while rotating. After incubation, the beads were pelleted and fresh TiO2 beads were added to the supernatant for a second enrichment step. Beads were washed with five different buffers: (i) 80% ACN and 6% TFA, (ii) 10% ACN and 6% TFA, (iii) 80% ACN and 1% TFA, (iv) 50% ACN and 1% TFA, (v) 10% ACN and 1% TFA. The final washing step was performed on a C8 stage tip, from which the phosphopeptides were with 20 µl 5% NH4OH followed by 20 µl 10% NH4OH with 25% ACN. Eluted peptides were fractionated using a reversed-phase Acquity CSH C18 1.7 µm 1 × 150 mm column (Waters, Milford, MA) on an UltiMate 3000 high-pressure liquid chromatography (HPLC) system (Dionex, Sunnyvale, CA) operating at 30 µl/min. Buffer A (5 mM ammonium bicarbonate) and buffer B (100% ACN) were used. Peptides were separated by a linear gradient from 5% B to 35% B in 55 min, followed by a linear increase to 70% B in 8 min and 12 fractions were collected in a concatenated manner.

The peptide solution was adjusted in volume to an appropriate concentration and kept in loading buffer (5% ACN and 0.1% TFA) prior to autosampling. An in-house packed 15 cm, 75 µm ID capillary column with 1.9 µm Reprosil-Pur C18 beads (Dr. Maisch, Ammerbuch, Germany) was used with an EASY-nLC 1200 system (Thermo Fisher Scientific, San Jose, CA). The column temperature was maintained at 40°C using an integrated column oven (PRSO-V1, Sonation, Biberach, Germany) and interfaced online with a Q Exactive HF-X mass spectrometer. Formic acid (FA) 0.1% was used to buffer the pH in the two running buffers used. The gradients went from 8 to 24% acetonitrile (ACN) in 50 min, followed by 24–36% in 10 min. This was followed by a washout by a 1/2 min increase to 64% ACN, which was kept for 4.5 min. Flow rate was kept at 250 nL/min. Re-equilibration was done in parallel with sample pickup and prior to loading with a minimum requirement of 0.5 µl of 0.1% FA buffer at a pressure of 600 bar.

The mass spectrometer was running in data-dependent acquisition mode with the spray voltage set to 2 kV, funnel RF level at 40, and heated capillary at 275°C. Full MS resolutions were set to 60,000 at m/z 200 and full MS AGC target was 3E6 with an IT of 25 ms. Mass range was set to 375–1500. AGC target value for fragment spectra was set at 1E5, and intensity threshold was kept at 2E5. Isolation width was set at 0.8 m/z and a fixed first mass of 100 m/z was used. Normalized collision energy was set at 33%. Peptide match was set to off, and isotope exclusion was on.

Raw MS files were analyzed by MaxQuant software version 1.6.0.17 using the Andromeda search engine. Proteins were identified by searching the higher-energy collisional dissociation (HCD)–MS/MS peak lists against a target/decoy version of the human UniProt protein database (release April 2017) using default settings. Carbamidomethylation of cysteine was specified as fixed modification, and protein N-terminal acetylation, oxidation of methionine, pyro-glutamate formation from glutamine, and phosphorylation of serine, threonine, and tyrosine residues were considered as variable modifications. The “maximum peptide mass” was set to 7,500 Da, and the “modified peptide minimum score” and “modified maximum peptide score” were set to 25. Everything else was set to default values. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository.

Data normalization and differential analysis

In the phosphoproteomics dataset, 19285 unique phosphosites were detected across 18 samples. Visual inspection of the raw data PCA first 2 components indicated two major batches of samples (1st batch : “38KI”, “38TU”, “15KI”, “15TU”, “29KI”, “29TU”, “16KI”, “16TU”, “32KI”, “32TU”, “35KI”, “35TU”; 2nd batch : “40KI”, “40TU”, “24KI”, “24TU”, “11KI”, “11TU”). Thus, each batch was first normalized using the VSN R package (Huber *et al*, 2002; Välikangas *et al*, 2018). We removed p-sites that were detected in < 4 samples, leaving 14,243 unique p-site to analyze. Visual inspection of the PCA first two components of the normalized data revealed that the first batch of samples could itself be separated in 3 batches (4 batches across all samples). Thus, we used the removeBatchEffect function of LIMMA to remove the linear effect of the 4 batches. Differential analysis was performed using the standard sequence of lmFit, contrasts.fit and eBayes functions of LIMMA, with FDR correction.

For the transcriptomics data, counts were extracted from fast.q files using the RsubRead R package and GRCh37 (hg19) reference genome. Technical replicates were averaged, and genes with average counts under 50 across samples were excluded, leaving 15919 genes measured across 22 samples. To allow for logarithmic transformation, 0 count values were scaled up to 0.5 (similar to the voom function of LIMMA). Counts were then normalized using the VSN R package function and differential analysis was performed with LIMMA package, in the same way as the phosphoproteomics data.

For the metabolomics data, 107 metabolites were detected in 16 samples. Intensities were normalized using the VSN package. Differential analysis was done using LIMMA in the same manner as for phosphoproteomics and transcriptomics. All data are available at: https://github.com/saezlab/COSMOS.

Footprint-based analysis

TF-target collection was obtained from DoRothEA A,B,C and D interaction confidence levels from the DoRothEA R package (version 1.1.0). For the enrichment analysis, the viper algorithm (Alvarez *et al*, 2016) was used with the LIMMA moderated t-value as gene level statistic (Zyla *et al*, 2017). The eset.filter parameter was set to FALSE. Only TFs with at least 25 measured transcripts were included.

Kinase substrate collection was obtained using the default resource collection of OmniPath, with the URL “http://omnipathdb.org/ptms?fields=sources,references&genesymbols=1” (version of 2020 Feb 05). For the enrichment analysis, the viper algorithm was used with the LIMMA moderated *t*-value as phosphosite level statistic. The eset.filter parameter was set to FALSE. Only TFs with at least 5 measured transcripts were included. All data are available at https://github.com/saezlab/COSMOS_MSB/tree/main/data.

Meta PKN construction

To propose mechanistic hypotheses spanning through signaling, transcription and metabolic reaction networks, multiple types of interactions have to be combined together in a single network. Thus, we built a meta Prior Knowledge Network (PKN) from three online resources, to incorporate three main types of interactions. The three types of interactions are protein–protein interactions, metabolite-protein allosteric interactions, and metabolite-protein interactions in the context of a metabolic reaction network. Protein–protein interaction was imported from OmniPath with the URL http://omnipathdb.org/interactions?types=post_translational,transcriptional&datasets=omnipath,pathwayextra,dorothea&fields=sources,references,curation_effort,dorothea_level,type&genesymbols=yes (version of 2020 July 17), and only signed directed interactions were included (is_stimulation or is_inhibition columns equal to 1). Metabolic-protein allosteric interactions were imported from the STITCH database (version of 2019 November 06), with combined confidence score ≥ 900 after exclusion of interactions relying mainly on text mining.

For metabolic-protein interactions in the context of metabolic reaction network, Recon3D was downloaded from https://www.vmh.life/#downloadview (version of 2019 Feb 19). Then, the gene rules ("AND" and "OR") of the metabolic reaction network were used to associate reactants and products with the corresponding enzymes of each reaction. When multiple enzymes were associated with a reaction with an "AND" rule, they were combined together as a single entity representing an enzymatic complex. Then, reactants were connected to corresponding enzymatic complexes or enzymes by writing them as rows of simple interaction format (SIF) table of the following form: reactant;1;enzyme. In a similar manner, products were connected to corresponding enzymatic complexes or enzymes by writing them as rows of simple interaction format (SIF) table of the following form: enzyme;1;product. Thus, each row of the SIF table represents either an activation of the enzyme by the reactant (i.e., the necessity of the presence of the reactant for the enzyme to catalyze it's reaction) or an activation of the product by an enzyme (i.e., the product presence is dependent on the activity of its corresponding enzyme). Most metabolite–protein interactions in metabolic reaction networks are not exclusive, thus measures have to be taken in order to preserve the coherence of the reaction network when converted to the SIF format. First, metabolites that are identified as "Coenzymes" in the Medical Subject Heading Classification (as referenced in the PubChem online database) were excluded. Then, we looked at the number of connections of each metabolite and searched the minimum interaction number threshold that would avoid excluding main central carbon metabolites. Glutamic acid has 338 interactions in our Recon3D SIF network and is the most connected central carbon metabolite, thus any metabolites that had more than 338 interactions was excluded. An extensive list of Recon3D metabolites (PubChem CID) with their corresponding number of connections is available in Dataset EV2. Metabolic enzymes catalyzing multiple reactions were uniquely identified for each reaction to avoid cross-links between reactants and products of different reactions. Finally, exchange reactions were further uniquely identified according to the relevant exchanged metabolites, as to avoid confusion between transformation of metabolites and simply exchanging them between compartments.

Finally, each network (protein–protein, allosteric metabolite–protein, and reaction network metabolite–protein) was combined into a single SIF table. This network is available in the COSMOS R package.

Meta PKN contextualization

COSMOS uses the CARNIVAL R package to perform the network optimization via an ILP algorithm. In brief, we try to minimize the value of an objective function that depends on two main factors: (i) the mismatch between the simulated values of kinases, TFs, and metabolites for a given causal network and the corresponding available values estimated from the measurements and (ii) the size of the solution network. For each run, given the prior knowledge network and the input and measurements, a set of constraints are generated to define the solution space (based on the objective function) that the ILP solver (IBM CPLEX in our study) explores to find an optimal solution (Melas *et al*, 2015; preprint: Liu *et al*, 2019a). After a given amount of time (decided by the user), the search is stopped and the best solution at this point is returned by CPLEX. The solution is typically a pool (or family) of networks that are all equally optimal with respect to the objective function. Thus, CARNIVAL reports the solution as a set of edges with an associated weight that represent their frequency of appearance in the current network pool. CARNIVAL needs a set of starting and end nodes to look for paths in between. TFs, kinases, and phosphatases absolute normalized enrichment scores greater than 1.7 standard deviation were considered deregulated. Coherently, metabolites with uncorrected *P*-values smaller than 0.05 were considered deregulated. We give more information on the rational to choose an appropriate threshold in the Appendix Note 1. This yielded a set of 98 TFs, 25 kinases/phosphatase, and 41 metabolites to be used as input and measurements for COSMOS.

Then, the PKN is pre-processed in three steps to make it easier for CARNIVAL to find a solution network, as detailed below.

Filtering

The generic meta PKN contained 117,065 edges. We first filtered the meta PKN to keep only genes that are expressed. With the main dataset presented in this paper, we considered the 15,919 genes that remained after removing the lowly expressed genes (defined as those with average count under 50 across the 22 samples, based on the count distribution) as expressed. This reduced the size of the meta PKN from 117,065 edges to 66,749 edges.

Reduction

At this stage, the meta PKN may contain independent network modules that do not include any of the actual input nodes (the significant TF, kinase/phosphatase activities, and metabolites). Thus, we filter out any gene that cannot be connected to any input node. We define a maximum given number of steps to avoid excessively long causal paths that would be un-plausible and thus have unclear biological relevance. We chose 8 steps downstream of signaling inputs for the "forward" run (signaling to metabolism) and 7 steps downstream of metabolic inputs for the "backward" run (metabolism to signaling) as > 90% of the PKN could be captured in that number of steps.

Correction

We use the transcriptomics data differential gene expression analysis results to directly remove any edge that leads to an incoherence between a TF activity and its target transcript abundance change (which is a wrongly predicted transcriptional regulation event). This is done once before running CARNIVAL, using TF activities predicted with DoRothEA. Then, we do a pre-run of CARNIVAL (TFs/kinases/phosphatases -> metabolites) to generate a first solution network. We can subsequently use TF activities predicted by CARNIVAL to filter out any wrongly predicted transcriptional regulation event from the meta PKN (Appendix Fig S3A).

Then, we first set the deregulated kinases, phosphatases, and TFs as starting points and deregulated metabolites as end points ("forward" run). This direction represents regulations first going through the signaling and transcriptional part of the cellular network and stops at deregulated metabolites in the metabolic reaction network. However, since metabolite concentration can also influence the activity of kinases and TFs through allosteric regulations, we also ran CARNIVAL by setting deregulated metabolites as starting points and deregulated TFs, kinases, and phosphatases as end points ("backward" run). The "forward" run was performed with a time limit of 7,200 s and yielded a network of 162 edges. The "backward" run was performed with a time limit of 21,800 s and yielded a network of 302 edges.

There was a single incoherence in the predicted sign of *ARNT2* transcription factor (–1 in "forward" run, 1 in "backward" run) between the common part of the two resulting networks. We made the union of the two networks, resulting in a combined network of 449 unique edges, while preserving the incoherent sign of *ARNT2* in the corresponding node attributes of the network (Dataset EV5).

Coherence between COSMOS mechanistic hypotheses and omics measurements

To assess the robustness of COSMOS predictions, we compared co-regulations predicted by the COSMOS solution network with co-regulations estimated from correlation between kinase, phosphatase, and TF activities. When multiple nodes are co-regulated by a common parent node in the COSMOS network, we can assume that the activity of the co-regulated nodes should be correlated. Thus, we created a correlation network with the TF and kinase/phosphatase activities estimated at a single sample level. To estimate the single sample level activities, normalized RNA counts and phosphosite intensities were scaled (minus mean over standard deviation) across samples. Thus, the value of each gene and phosphosite is now a z-score relative to an empirical distribution generated from the measurements across all samples. We used these z-scores as input for the viper algorithm to estimate kinase/phosphatases and TF activities at single sample level. Thus, the resulting activity scores in a sample are relative to all the other samples. Then, a correlation network was built using only tumor samples. Thus, the correlation calculated this way represents co-regulations that are supported by the available data in tumor. We defined the ground truth for co-regulations as over a range of absolute correlation coefficients between 0 and 1 with a 0.01 step. Thus, a True Positive here is a co-regulation predicted from the topology of the COSMOS network that also has a corresponding absolute correlation coefficient in tumor samples above the given threshold. Since defining a ground truth in such a manner can yield many false positives (a correlation can often be spurious), the TPR of COSMOS was always compared to a random baseline. This approach was repeated for COSMOS solution networks obtained after hiding either kinase and phosphatases or TFs.

Robustness analysis

We generated a series of subsets of the original meta PKN where increasing amounts of interactions are shuffled randomly. Starting from the full meta PKN, we shuffled 2, 10, 20, 30, 40, and 50% of interactions. Each shuffling is independent from the others (the missing interactions are all selected randomly at each percentage case). Then, COSMOS was run for each meta PKN subset with the same parameter as the original run.

CPTAC ccRCC data analysis

The CPTAC ccRCC transcriptomics and phosphoproteomics datasets of the proteogenomics study of ccRCC (Clark *et al*, 2020) were obtained from the CPTAC data portal. We kept 20,284 phosphosites that were detected across at least 10% of the 185 patient and healthy samples (110 and 75, respectively).

We filtered out lowly expressed genes (RPKM (Reads Per Kilobase of transcript, per Million mapped reads) < 170, based on the inflexion point observed in the RPKM distribution) from the transcriptomics dataset, keeping 14,921 genes for further analysis.

LIMMA was used for both phosphoproteomics and transcriptomics to perform a differential analysis between healthy and tumor samples.

Kinase and transcription factor activities were performed with the same parameters as with our own ccRCC patient samples (see footprint-based analysis). 57 kinases and 97 TFs with absolute NES > 1.7 were used as input and measurements in the COSMOS pipeline. The meta PKN was reduced to keep only nodes with a maximum distance of 8 steps downstream of input kinases and TFs. The kinase to TF CARNIVAL run was performed with a time limit of 7,200 s. The TF to kinase run was performed with a time limit of 21,800 s. The union of the “forward” and “backward” run networks resulted in a final COSMOS network of 480 edges.

Breast cancer data analysis

Multi-omics experimental data for breast cancer cell lines was obtained from (Katzir *et al*, 2019). The authors performed experimental measurements on the MCF7 cell line under normal growth conditions, glutamine deprivation, and oligomycin supplementation.

We obtained mRNA expression quantification of 1,905 metabolic genes and filtered those whose mean across all conditions was at least 0.1% of the maximum observed expression value. The experiments were split in 2 batches, leading us to regress this effect out. We then fit a linear model using the LIMMA package, from which we obtained t-statistic values at the gene level for a given comparison pair. Finally, TF activity scores were estimated using regulon confidence A, B, and C with a minimum of 25 targets per TF with the VIPER package, using the pleiotropy correction.

Fluxomics measurements estimated from ^{13}C-assisted metabolomics were available for 44 metabolic reactions included in the Recon 3D genome-scale metabolic model. We computed the $\log_2$ fold change between each pair of conditions to be analyzed.

COSMOS was then used to generate context-specific sub-networks using the transcription factor NES and the fluxomics $\log_2$ fold change as inputs and measurements. It was run without using the correction and reduction step, with a time limit of 7,200 s on the “forward” and “backward” runs.

Acknowledgements

A.D. and E.G. were Marie-Curie Early Stage Researchers supported by the European Union’s Horizon 2020 research and innovation program (675585 Marie-Curie ITN “SymBioSys”) to J.S.R. A.D. was funded by German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung BMBF) MSCoreSys research initiative research core SMART-CARE (031L0212A). This work was further supported by the JRC for Computational Biomedicine which was partially funded by Bayer AG, and the Medical Research Council (MC_UU_12022/6 to C.F. and M.S.). The Novo Nordisk

Foundation Center for Protein Research is supported by Novo Nordisk Foundation grant number NNF14CC0001. J.V.O. was funded by a grant from Danish Council for Independent Research (8020-00100B) to partly support K.B.E. who was also supported in part by the Lundbeck Foundation (R193-2015-243). R.K. was supported by grants of the German Research Foundation (DFG: SFBTRR57, P30; SFBTRR219 C05, CRU344, P1), by a Grant of the European Research Council (ERC-StG 677448), a Grant of the State of North Rhine-Westphalia (Return to NRW), the BMBF eMed Consortia Fibromap, the ERA-CVD Consortia MEND-AGE, the Else Kroener Fresenius Foundation (EKFS) and the Interdisciplinary Centre for Clinical Research (IZKF) within the faculty of Medicine at the RWTH Aachen University (O3-11). C.K. was supported by the German Society of Internal Medicine (DGIM). Thanks to Hyojin Kim for her contribution to the original COSMOS logo design. Thanks to Denes Turei for his help with putting the meta PKN online. We thank E. Ruppin and R. Katzir for helping us with the breast cancer dataset from Katzir *et al* (2019). Open Access funding enabled and organized by ProjektDEAL.

Authors contributions

AD and JS-R designed the method. AD coded the pipeline and ran the analysis. AD with input from CK, MS, CF and JS-R. interpreted results. CK collected the patient samples and generated RNA-Seq libraries. EG developed and adapted CARNIVAL to the pipeline. KBE, DBB-J and JVO generated the phosphoproteomics dataset. EMJB performed final RNA-sequencing on Novaseq6000 platform. MS and ASHC performed the liquid chromatography mass spectrometry-based metabolomics analyses and processed the data. JK collected patient consents and samples. CF, RK and JS-R supervised the project. AD wrote the manuscript with help from JS-R. AS and PB processed the CPTAC data. VV and MR analyzed the breast cancer dataset. AG has developed the R package version of the method.

Conflict of interest

The authors declare that they have no conflict of interest. C.F. is a member of the scientific advisory board of Owlstone and scientific adviser of Istesso. JSR has received funding from GSK and expects to receive funding from Sanofi and consultant fees from Travere Therapeutics.

References

- Al-Lamki RS, Sadler TJ, Wang J, Reid MJ, Warren AY, Movassagh M, Lu W, Mills IG, Neal DE, Burge J ***et al*** (2010) Tumor necrosis factor receptor expression and signaling in renal cell carcinoma. ***Am J Pathol* 177**: 943–954
- Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. ***Nat Genet* 48**: 838–847
- An J, Rettig MB (2007) Epidermal growth factor receptor inhibition sensitizes renal cell carcinoma cells to the cytotoxic effects of bortezomib. ***Mol Cancer Ther* 6**: 61–69
- Ananieva EA, Wilkinson AC (2018) Branched-chain amino acid metabolism in cancer. ***Curr Opin Clin Nutr Metab Care* 21**: 64–70
- Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. ***Mol Syst Biol* 14**: e8124
- Austen M, Cerni C, Lüscher-Firzlaff JM, Lüscher B (1998) YY1 can inhibit c-Myc function through a mechanism requiring DNA binding of YY1 but neither its transactivation domain nor direct interaction with c-Myc. ***Oncogene* 17**: 511–520
- Bailey ST, Smith AM, Kardos J, Wobker SE, Wilson HL, Krishnan B, Saito R, Lee HJ, Zhang J, Eaton SC ***et al*** (2017) MYC activation cooperates with Vhl and Ink4a/Arf loss to induce clear cell renal cell carcinoma. ***Nat Commun* 8**: 15770
- Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. ***J R Stat Soc Series B Stat Methodol* 57**: 289–300
- Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, Gatto F, Nilsson A, Preciat Gonzalez GA, Aurich MK ***et al*** (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. ***Nat Biotechnol* 36**: 272–281
- Cantini L, Calzone L, Martignetti L, Rydenfelt M, Blüthgen N, Barillot E, Zinovyev A (2018) Classification of gene signatures for their information value and functional redundancy. ***NPJ Syst Biol Appl* 4**: 2
- Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, Vanhaesebroeck B, Joel S, Cutillas PR (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. ***Sci Signal* 6**: rs6
- Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, da Veiga LF, Reva B, Lih T-SM, Chang H-Y ***et al*** (2020) Integrated proteogenomic characterization of clear cell renal cell carcinoma. ***Cell* 180**: 207
- Dong Y, Tu R, Liu H, Qing G (2020) Regulation of cancer cell metabolism: oncogenic MYC in the driver's seat. ***Signal Transduct Target Ther* 5**: 124
- Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, Stoyanova T, Faltermeier CM, Uzunangelov V, Carlin DE ***et al*** (2016) Phosphoproteome integration reveals patient-specific networks in prostate cancer. ***Cell* 166**: 1041–1054
- Dugourd A, Saez-Rodriguez J (2019) Footprint-based functional analysis of multiomic data. ***Curr Opinion Syst Biol* 15**: 82–90
- Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. ***Genome Res* 29**: 1363–1375
- Gatto F, Miess H, Schulze A, Nielsen J (2015) Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. ***Sci Rep* 5**: 10738
- Gordan JD, Lal P, Dondeti VR, Letrero R, Parekh KN, Oquendo CE, Greenberg RA, Flaherty KT, Rathmell WK, Keith B ***et al*** (2008) HIF-alpha effects on c-Myc distinguish two subtypes of sporadic VHL-deficient clear cell renal carcinoma. ***Cancer Cell* 14**: 435–446
- Hamanaka N, Nakanishi Y, Mizuno T, Horiguchi-Takei K, Akiyama N, Tanimura H, Hasegawa M, Satoh Y, Tachibana Y, Fujii T ***et al*** (2019) YES1 is a targetable oncogene in cancers harboring YES1 gene amplification. ***Cancer Res* 79**: 5734–5745
- Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H ***et al*** (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. ***Genome Biol*21**: 36

- Huang H, Ma L, Li J, Yu Y, Zhang D, Wei J, Jin H, Xu D, Gao J, Huang C (2014) NF-κB1 inhibits c-Myc protein degradation through suppression of FBW7 expression. ***Oncotarget* 5**: 493–505
- Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. ***Bioinformatics* 18**(Suppl 1): S96–104
- Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H ***et al*** (2016) A landscape of pharmacogenomic interactions in cancer. ***Cell* 166**: 740–754
- Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS, Kaelin WG (2001) HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. ***Science* 292**: 464–468
- Jaakkola P, Mole DR, Tian Y-M, Wilson MI, Gielbert J, Gaskell SJ, Kriegsheim AV, Hebestreit HF, Mukherji M, Schofield CJ ***et al*** (2001) Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. ***Science* 292**: 468–472.
- Jelinek D, Wu X (2012) Faculty of 1000 evaluation for The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. F1000 - Post-publication peer review of the biomedical literature https://doi.org/10.3410/f.14264142.15777309 [PREPRINT].
- Katzir R, Polat IH, Harel M, Katz S, Foguet C, Selivanov VA, Sabatier P, Cascante M, Geiger T, Ruppin E (2019) The landscape of tiered regulation of breast cancer cell metabolism. ***Sci Rep* 9**: 1–12
- Kawata K, Hatano A, Yugi K, Kubota H, Sano T, Fujii M, Tomizawa Y, Kokaji T, Tanaka KY, Uda S ***et al*** (2018) Trans-omic Analysis Reveals Selective Responses to Induced and Basal Insulin across Signaling, Transcriptional, and Metabolic Networks. ***iScience* 7**: 212–229
- Kearney AY, Fan Y-H, Giri U, Saigal B, Gandhi V, Heymach JV, Zurita AJ (2015) 8-Chloroadenosine Sensitivity in Renal Cell Carcinoma Is Associated with AMPK Activation and mTOR Pathway Inhibition. ***PLoS One* 10**: e0135962
- Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J (2019a) From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. ***bioRxiv***: 541888 [PREPRINT]
- Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J (2019b) From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. ***NPJ Syst Biol Appl* 5**: 40
- Liu W, Payne SH, Ma S, Fenyö D (2019c) Extracting Pathway-level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis. ***Mol Cell Proteomics* 18**: S169–S182
- Lucas B, Grigo K, Erdmann S, Lausen J, Klein-Hitpass L, Ryffel GU (2005) HNF4 α reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. ***Oncogene* 24**: 6418–6431
- Mackay GM, Zheng L, van den Broek NJF, Gottlieb E (2015) Analysis of Cell Metabolism Using LC-MS and Isotope Tracers. ***Methods Enzymol* 561**: 171–196
- Masoud GN, Li W (2015) HIF-1α pathway: role, regulation and intervention for cancer therapy. ***Acta Pharm Sin B* 5**: 378–389
- Maxwell PH, Wiesener MS, Chang GW, Clifford SC, Vaux EC, Cockman ME, Wykoff CC, Pugh CW, Maher ER, Ratcliffe PJ (1999) The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. ***Nature* 399**: 271–275
- Melas IN, Sakellaropoulos T, Iorio F, Alexopoulos LG, Loh W-Y, Lauffenburger DA, Saez-Rodriguez J, Bai JPF (2015) Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. ***Integr Biol* 7**: 904–920
- Ochoa D, Jonikas M, Lawrence RT, El Debs B, Selkrig J, Typas A, Villén J, Santos SD, Beltrao P (2016) An atlas of human kinase regulation. ***Mol Syst Biol* 12**: 888
- Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). ***Bioinformatics* 29**: 2757–2764
- Peri S, Devarajan K, Yang D-H, Knudson AG, Balachandran S (2013) Meta-analysis identifies NF-κB as a therapeutic target in renal cancer. ***PLoS One* 8**: e76746
- Popławski P, Tohge T, Bogusławska J, Rybicka B, Tański Z, Treviño V, Fernie AR, Piekiełko-Witkowska A (2017) Integrated transcriptomic and metabolomic analysis shows that disturbances in metabolism of tumor cells contribute to poor survival of RCC patients. **Biochim Biophys Acta Mol Basis Dis 1863**(3): 744–752.
- R Core Team (2020) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
- Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM (2005) Mining for regulatory programs in the cancer transcriptome. ***Nat Genet* 37**: 579–583

- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. ***Nucleic Acids Res* 43**: e47
- Rodrigues P, Patel SA, Harewood L, Olan I, Vojtasova E, Syafruddin SE, Zaini MN, Richardson EK, Burge J, Warren AY ***et al*** (2018) NF-κB-Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis. ***Cancer Discov* 8**: 850–865
- Samanta D, Gilkes DM, Chaturvedi P, Xiang L, Semenza GL (2014) Hypoxia-inducible factors are required for chemotherapy resistance of breast cancer stem cells. ***Proc Natl Acad Sci USA* 111**: E5429–E5438
- Schödel J, Grampp S, Maher ER, Moch H, Ratcliffe PJ, Russo P, Mole DR (2016) Hypoxia, hypoxia-inducible transcription factors, and renal cancer. ***Eur Urol* 69**: 646–657
- Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N, Saez-Rodriguez J (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. ***Nat Commun* 9**: 20
- Sciacovelli M, Gonçalves E, Johnson TI, Zecchini VR, da Costa ASH, Gaude E, Drubbel AV, Theobald SJ, Abbo SR, Tran MGB ***et al*** (2016) Fumarate is an epigenetic modifier that elicits epithelial-to-mesenchymal transition. ***Nature* 537**: 544–547
- Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M (2019) MOLI: multi-omics late integration with deep neural networks for drug response prediction. ***Bioinformatics* 35**: i501–i509
- Shroff EH, Eberlin LS, Dang VM, Gouw AM, Gabay M, Adam SJ, Bellovin DI, Tran PT, Philbrick WM, Garcia-Ocana A ***et al*** (2015) MYC oncogene overexpression drives renal cell carcinoma in a mouse model through glutamine metabolism. ***Proc Natl Acad Sci USA* 112**: 6539–6544
- Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Cao K-AL (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. ***Bioinformatics* 35**: 3055–3062.
- Singh D, Arora R, Kaur P, Singh B, Mannan R, Arora S (2017) Overexpression of hypoxia-inducible factor and metabolic pathways: possible targets of cancer. ***Cell Biosci* 7**: 62
- Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK ***et al*** (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. ***Cell* 171**: 1437–1452.e17
- Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M (2016) STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. ***Nucleic Acids Res* 44**: D380–D384
- Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V (2014) Variable selection for generalized canonical correlation analysis. ***Biostatistics* 15**: 569–583
- Thapa RJ, Chen P, Cheung M, Nogusa S, Pei J, Peri S, Testa JR, Balachandran S (2013) NF-κB inhibition by bortezomib permits IFN-γ-activated RIP1 kinase-dependent necrosis in renal cell carcinoma. ***Mol Cancer Ther* 12**: 1568–1578
- Türei D, Korcsmáros T, Saez-Rodriguez J (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. ***Nat Methods* 13**: 966–967
- Välikangas T, Suomi T, Elo LL (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. ***Brief Bioinform* 19**: 1–11
- Vitrinel B, Koh HWL, Kar FM, Maity S, Rendleman J, Choi H, Vogel C (2019) Exploiting inter-data relationships in next-generation proteomics analysis. ***Mol Cell Proteomics* 18**(8 suppl 1): S5–S14
- Wiredja DD, Koyutürk M, Chance MR (2017) The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. ***Bioinformatics* 33**: 3489–3491
- Xu Q, Krause M, Samoylenko A, Vainio S (2016) Wnt signaling in renal cell carcinoma. ***Cancers* 8**: 57
- Yan L, Zhang Y, Ding B, Zhou H, Yao W, Xu H (2019) Genetic alteration of histone lysine methyltransferases and their significance in renal cell carcinoma. ***PeerJ* 7**: e6396
- Zeng Z, Que T, Zhang J, Hu Y (2014) A study exploring critical pathways in clear cell renal cell carcinoma. ***Exp Ther Med* 7**: 121–130
- Zhang J, Wu T, Simon J, Takada M, Saito R, Fan C, Liu X-D, Jonasch E, Xie L, Chen X ***et al***(2018) VHL substrate transcription factor ZHX2 as an oncogenic driver in clear cell renal cell carcinoma. ***Science* 361**: 290–295
- Zhang J-Y, Zhang F, Hong C-Q, Giuliano AE, Cui X-J, Zhou G-J, Zhang G-J, Cui Y-K (2015) Critical protein GAPDH and its regulatory mechanisms in cancer cells. ***Cancer Biol Med* 12**: 10–22
- Zyla J, Marczyk M, Weiner J, Polanska J (2017) Ranking metrics in gene set enrichment analysis: do they matter? ***BMC Bioinformatics* 18**: 256

Affiliations

[1] Institute of Experimental Medicine and Systems Biology, RWTH Aachen University, Medical Faculty, Aachen, Germany.

[2] Division of Nephrology and Clinical Immunology, RWTH Aachen University, Medical Faculty, Aachen, Germany.

[3] Institute for Computational Biomedicine, Heidelberg University, Faculty of Medicine, Heidelberg University Hospital, Bioquant, Heidelberg, Germany.

[4] Informatics for Life, Heidelberg, Germany.

[5] Institute for Computational Genomics, RWTH Aachen University, Medical Faculty, Aachen, Germany.

[6] Joint Research Center for Computational Biomedicine, RWTH Aachen University Hospital, Aachen, Germany.

[7] Department of General Internal Medicine and Psychosomatics, Heidelberg University Hospital, Heidelberg, Germany.

[8] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia.

[9] Erich and Hanna Klessmann Institute for Cardiovascular Research and Development, Clinic for Thoracic and Cardiovascular Surgery, Heart and Diabetes Center NRW, Bad Oeynhausen, Germany.

[10] Heart and Diabetes Center, North Rhine-Westphalia, Bad Oeynhausen, Germany.

[11] Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [12]Institute for Molecular Cardiovascular Research IMCAR, RWTH Aachen University, Medical Faculty, Aachen, Germany.

[13] Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, The Netherlands.

[14] Department of Pathology, RWTH Aachen University, Aachen, Germany.

[15] Cardiopathology, Institute for Pathology and Neuropathology, University Hospital Tübingen, Tübingen, Germany.

[16] Department of Cardiology, Regenerative Medicine Center and Circulatory Health Lab, University Medical Center Utrecht, Utrecht, The Netherlands.

[17] Department of Pathology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

[18] Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands. [19]National Heart and Lung Institute, Imperial College London, London, UK.

[20] Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany.

[21] Institute of Cell and Tumor Biology, RWTH Aachen University, Medical Faculty, Aachen, Germany.

[22] Oncode Institute, Erasmus Medical Center, Rotterdam, The Netherlands.

[23] Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands.

[24] These authors contributed equally

[25] These authors jointly supervised this work

6

Spatial multi-omic map of human myocardial infarction

Christoph Kuppe[1,2,24], Ricardo O. Ramirez Flores[3,4,24], Zhijian Li[5,6,24], Sikander Hayat[1], Rebecca T. Levinson[3,4,7], Xian Liao[1], Monica T. Hannani[1,3], Jovan Tanevski[3,8], Florian Wünnemann[3], James S. Nagai[5,6], Maurice Halder[1], David Schumacher[1], Sylvia Menzel[1], Gideon Schäfer[1], Konrad Hoeft[1], Mingbo Cheng[5,6], Susanne Ziegler[1], Xiaoting Zhang[1], Fabian Peisker[1], Nadine Kaesler[1,2], Turgay Saritas[1,2], Yaoxian Xu[1], Astrid Kassner[9], Jan Gummert[10], Michiel Morshuis[10], Junedh Amrute[11], Rogier J. A. Veltrop[12,13], Peter Boor[2,14], Karin Klingel[15], Linda W. Van Laake[16], Aryan Vink[17], Remco M. Hoogenboezem[18], Eric M. J. Bindels[18], Leon Schurgers[1,13], Susanne Sattler[19], Denis Schapiro[3,20], Rebekka K. Schneider[21,22], Kory Lavine[11], Hendrik Milting[9,25], Ivan G. Costa[5,6,25], Julio Saez-Rodriguez[3,4,25] & Rafael Kramann[1,2,23,25]

Abstract

Myocardial infarction is a leading cause of death worldwide[1]. Although advances have been made in acute treatment, an incomplete understanding of remodelling processes has limited the effectiveness of therapies to reduce late-stage mortality[2]. Here we generate an integrative high-resolution map of human cardiac remodelling after myocardial infarction using single-cell gene expression, chromatin accessibility and spatial transcriptomic profiling of multiple physiological zones at distinct time points in myocardium from patients with myocardial infarction and controls. Multi-modal data integration enabled us to evaluate cardiac cell-type compositions at increased resolution, yielding insights into changes of the cardiac transcriptome and epigenome through the identification of distinct tissue structures of injury, repair and remodelling. We identified and validated disease-specific cardiac cell states of major cell types and analysed them in their spatial context, evaluating their dependency on other cell types. Our data elucidate the molecular principles of human myocardial tissue organization, recapitulating a gradual cardiomyocyte and myeloid continuum following ischaemic injury. In sum, our study provides an integrative molecular map of human myocardial infarction, represents an essential reference for the field and paves the way for advanced mechanistic and therapeutic studies of cardiac disease.

Main

Coronary heart disease driving acute myocardial infarction is the largest contributor to cardiovascular mortality, which in turn is the leading cause of death worldwide1. Substantial progress has been made in the acute therapy of myocardial infarction, focusing primarily on percutaneous coronary intervention resulting in decreased acute mortality. However, the morbidity and mortality caused by left ventricular cardiac remodelling after myocardial infarction remain unacceptably high[2]. Cardiac remodelling after myocardial infarction involves immune cell recruitment and demarcation of the infarcted area followed by resorption of necrotic tissue, phagocytosis, myofibroblast activation, scar formation and neovascularization[3]. Understanding the exact cellular and molecular mechanisms of cardiac remodelling processes from the acute ischaemic event to the chronic cardiac scar formation in their spatial context will be key to developing novel therapeutics.

Here we used a combination of single-cell gene expression, chromatin accessibility and spatially resolved transcriptomics to study the events of cardiac tissue reorganization and to characterize the cell-type-specific changes in gene regulation, providing an integrated spatial multi-omic map of cardiac remodelling after myocardial infarction. Our multi-omic data-driven approach, including spatial context, enables us to understand how a given cell state changes based on the cells' neighbourhood and how this relates to transcriptional and regulatory variations. By deconvoluting spatial transcriptomics spots into cell-type abundances, we characterized cell niches occurring in different stages following acute myocardial infarction. We identified different cell states of cardiomyocytes, endothelial cells, myeloid cells and fibroblasts that are associated with disease progression on the basis of the integrated single-cell multi-omics data. Moreover, we inferred the gene-regulatory networks differentiating these cell states and projected this information onto specific tissue locations, thus mapping putative regulators controlling gene expression on specific myocardial tissue zones and disease stages. This enabled us to gain novel insights into the gene-regulatory programmes driving injury of cardiomyocytes, activated phagocytic macrophages and their relation to myofibroblast differentiation in cardiac tissue remodelling. Our results provide a comprehensive spatially resolved characterization of gene regulation of the human heart in homeostasis and after myocardial infarction. We have released our spatial multi-omics data through publicly available platforms to enable users to interactively explore the dataset. We anticipate that this data will be a reference map for future studies and ultimately for the development of novel therapeutics.

Multi-omic map of myocardial infarction

We applied an integrative single-cell genomics strategy with single-nucleus RNA sequencing (snRNA-seq) and single-nucleus assay for transposase-accessible chromatin sequencing (snATAC-seq) together with spatial transcriptomics from the same tissue mapping human cardiac cells in homeostasis and after myocardial infarction at unprecedented spatial and molecular resolution (Fig. 1a-c and Supplementary Table 1). We profiled a total of 31 samples from 23 individuals, including four non-transplanted donor hearts as controls, and samples from tissues with necrotic areas (ischaemic zone and border zone) and the unaffected left ventricular myocardium (remote zone) of patients with acute myocardial infarction (Fig. 1a). These acute myocardial infarction specimens were collected from heart tissues obtained at different time points after the onset of clinical symptoms (chest pain), before the patients received an artificial heart or a left-ventricular assist device because of cardiogenic shock and as a bridge to transplantation (Supplementary Fig. 1a-c). We also analysed nine human heart specimens at later stages after myocardial infarction (fibrotic zone; Fig. 1b) that exhibited ischaemic heart disease and were available from heart transplantation recipients at the time of orthotopic heart transplantation.

For each cardiac sample, we obtained 10-μm cryo-sections and isolated nuclei from the remaining tissue directly adjacent to the cryo-section with subsequent fluorescence-activated nuclei sorting (FANS) for snRNA-seq and snATAC-seq (Fig. 1c). After filtering out low-quality nuclei, we obtained a total of 191,795 nuclei from all samples for snRNA-seq, with an average of 2,020 genes per nucleus, together with chromatin accessibility data from 46,086 nuclei overall with an average of 28,066 fragments per nucleus (Supplementary Fig. 2a,b and Supplementary Tables 2–5). After controlling for data quality, the spatial transcriptomics datasets contained a total of 91,517 spots (average of 3,389 spots per specimen and 2,001 genes per spot) (Supplementary Figs. 2c,e–g and 3a,b). Quantification based on histology revealed an average of four nuclei per spatial transcriptomic spot from all slides (Supplementary Fig. 2c and Supplementary Table 6). Samples from the ischaemic zone had the lowest abundance of nuclei and an enriched expression of genes associated with cell death and the regulated necrosis pathway, suggesting increased necrotic cell death (Supplementary Fig. 2d). This integrated dataset represents, to our knowledge, the largest and most comprehensive multi-modal profiling of human myocardial infarction tissue including spatial information and samples at distinct disease progression stages. We devised an integrative data analysis approach spanning all three modalities of our single-cell experiments to study cardiac cell-specific information and cell-specific interactions in their spatial and disease progression context (Extended Data Fig. 1a).

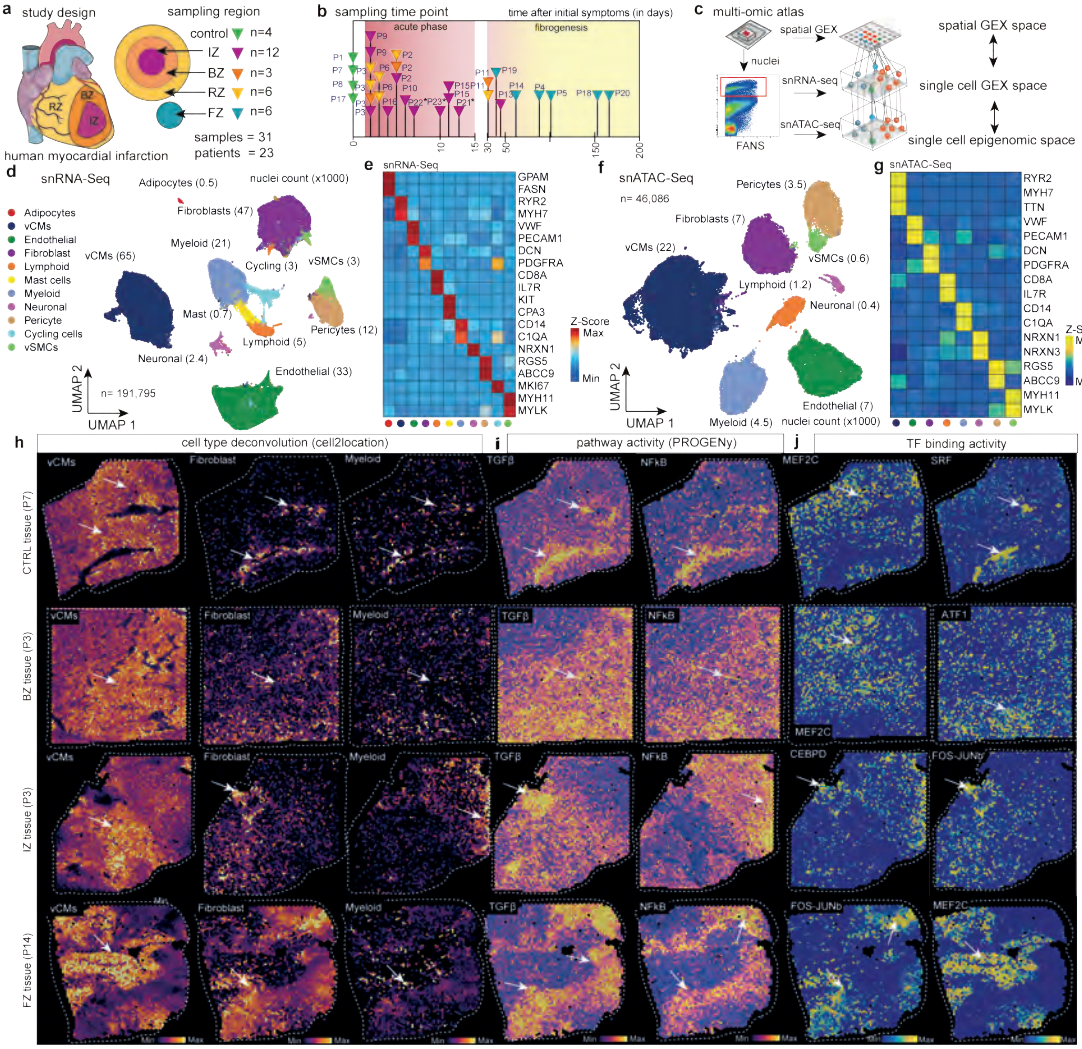


Fig. 1

a, Study schematic. RZ, remote zone; BZ, border zone; IZ, ischaemic zone; FZ, fibrotic zone. **b**, Sampling time points. P indicates patient number. Asterisks indicate snRNA-seq samples that were used for validation only (P21–P23). **c**, Data modalities. GEX, gene expression. **d**, UMAP of snRNA-seq data from all samples ($n = 191{,}795$). vCMs, ventricular cardiomyocytes. vSMCs, vascular smooth muscle cells. **e**, Average marker gene expression after *z*-score transformation. Colours along the bottom correspond to the cell types in **d. f**, Uniform manifold approximation and projection (UMAP) of snATAC-seq data for all samples ($n = 46{,}068$). **g**, Chromatin accessibility of marker genes after *z*-score transformation. Colours along the bottom correspond to the cell types in **d. h–j**, Characterization of spatial transcriptomics data using cell-type deconvolution (**h**), pathway activity (**i**) and transcription factor (TF) binding activity (**j**) for control (Ctrl), border zone, ischaemic zone and fibrotic tissue samples. Max, maximum; min, minimum.

We established a map of major human heart cell types using the snRNA-seq and snATAC-seq datasets independently. First, we clustered cells on the basis of the integrated snRNA-seq data from all samples after batch correction (Extended Data Fig. 1b). Clusters were annotated with curated marker genes from the literature[4,5,6] and ten major cardiac cell types were identified (Fig. 1d,e). We also found an additional cluster with enriched expression of the cell-cycle marker gene MKI67, which showed a high score of cell-cycle G2/M and S phases and was mainly recovered in ischaemic zone samples (Extended Data Fig. 1c,d). To validate the annotations, we compared the data with a recent study on healthy human hearts[4] and an independent novel dataset of ischaemic heart samples (n = 3, generated during this study) and observed a high agreement and correlation in terms of molecular profiles and cellular composition (Extended Data Fig. 1e–g). Of note, the cycling cells were also captured in the independent ischaemic dataset (Extended Data Fig. 1f).

We next integrated and clustered the snATAC-seq data from all samples (Extended Data Fig. 2a). These clusters were annotated on the basis of gene chromatin accessibility with the same markers as for snRNA-seq. This approach identified eight major cell types, matching all cell types from snRNA-seq data with the exception of two rare cell types (that is, mast cells and adipocytes) (Fig. 1f,g). Label transfer from snRNA-seq to snATAC-seq indicated that the annotations between these two modalities were consistent (Extended Data Fig. 2b,c). This was further supported by a high correlation of cellular composition between snRNA-seq and snATAC-seq and the presence of the same eight cell types in the majority of samples (Extended Data Fig. 2d,e). To explore regulatory information provided by the snATAC-seq, we performed transcription factor footprinting analysis using cell-type-specific pseudo-bulk ATAC-seq profiles. This revealed footprinting-based binding activity of known transcription factors such as MEF2C (ref. 7) in cardiomyocytes, CEBPD)[8] in myeloid cells, FOS–JUNB[9] in fibroblasts and SRF10 in vascular smooth muscle cells (vSMCs), which correlated with the expression of their predicted target genes in snRNA-seq data (Extended Data Fig. 2f). Together, our integrative analysis of transcriptomic and chromatin accessibility data defined a robust catalogue of cell types in the adult human heart across multiple modalities and samples.

Molecular mapping of cell types in space

Using these data, we first identified overrepresented biological processes for each major histomorphological region (control, remote zone, border zone, ischaemic zone and fibrotic zone) using spatially variable genes (Supplementary Table 7). We identified cardiac muscle contraction in remote zones and controls, with adaptive immune system in the border and ischaemic zones and with matrisome processes in the fibrotic zones (Extended Data Fig. 2g). Overall, this analysis confirmed that the spatial data clearly reflect typical zones of biological processes following acute human myocardial infarction.

Since each spatial transcriptomics spot captured a group of cells, we increased its resolution by estimating the cell-type compositions of each spot. To this end, we deconvoluted each spot on the basis of the annotated snRNA-seq data from the same sample (Fig. 1h, Supplementary Figs. 2e–g and 3a,b, Supplementary Tables 8 and 9 and Methods). The estimated cell-type compositions from spatial transcriptomics of each patient generally agreed with their respective observed compositions in the snRNA-seq and snATAC-seq data (Extended Data Fig. 2h). We then estimated signalling pathway activities with PROGENy (Methods) for each spot from the spatial gene expression data. The comparison of spatially localized pathway activities with the estimated cellular abundance per spot enabled us to link the information on spatial cell composition to cellular function for each slide. For example, in areas with an abundance of fibroblasts, we detected increased TGFβ signalling activity, and in ischaemic regions, increased myeloid cell abundance occurred in areas of higher NFκB signalling activity (Fig. 1h,i).

Mapping the information obtained from the snATAC-seq data to space resulted in spatially resolved footprinting-based transcription factor binding activity, as exemplified by the previously described transcription factors associated with cardiomyocytes (for example, MEF2C; ref. 7), myeloid cells (for example, CEBPD[8] and ATF1[11]), fibroblasts (for example, FOS–JUNB[9]) and vSMCs (for example, SRF10) (Fig. 1j). To test the association of genetic variants with cell types, we performed enrichment analysis based on cell-type-specific pseudo-bulk ATAC-seq profiles and cardiomyopathy-related single nucleotide polymorphisms (SNPs) obtained from genome-wide association studies[12] (GWAS). We focussed on SNPs relevant to left ventricular function, since we hypothesized that these might provide the most biologically relevant information for the cellular composition of myocardial tissue. This analysis revealed that SNPs associated with stroke volume and left ventricular end-diastolic volume were enriched in endothelial cells (Extended Data Fig. 2i), consistent with the role of the endothelial cells in cardiac relaxation and dilation[13]. SNPs associated with left ventricular end-systolic volume and left ventricular ejection fraction were enriched in cardiomyocytes, supporting the relationship between contraction and these left ventricular measures. We also visualized the spatial distribution of GWAS signals by mapping SNPs associated with left ventricular ejection fraction to each spot from spatial transcriptomics (Extended Data Fig. 2j). In summary, our integrated spatial atlas enabled us to map cell-type abundance, signalling pathway activities, transcription factor binding activity and GWAS signals across the complete spectrum of cardiac tissue zonations, providing an in-depth view at tissue remodelling processes following myocardial infarction in humans.

Spatial organization of myocardial tissue

To explore the spatial organization of the myocardial tissue, we leveraged the spatial transcriptomics data. Unsupervised clustering of spots from all samples on the basis

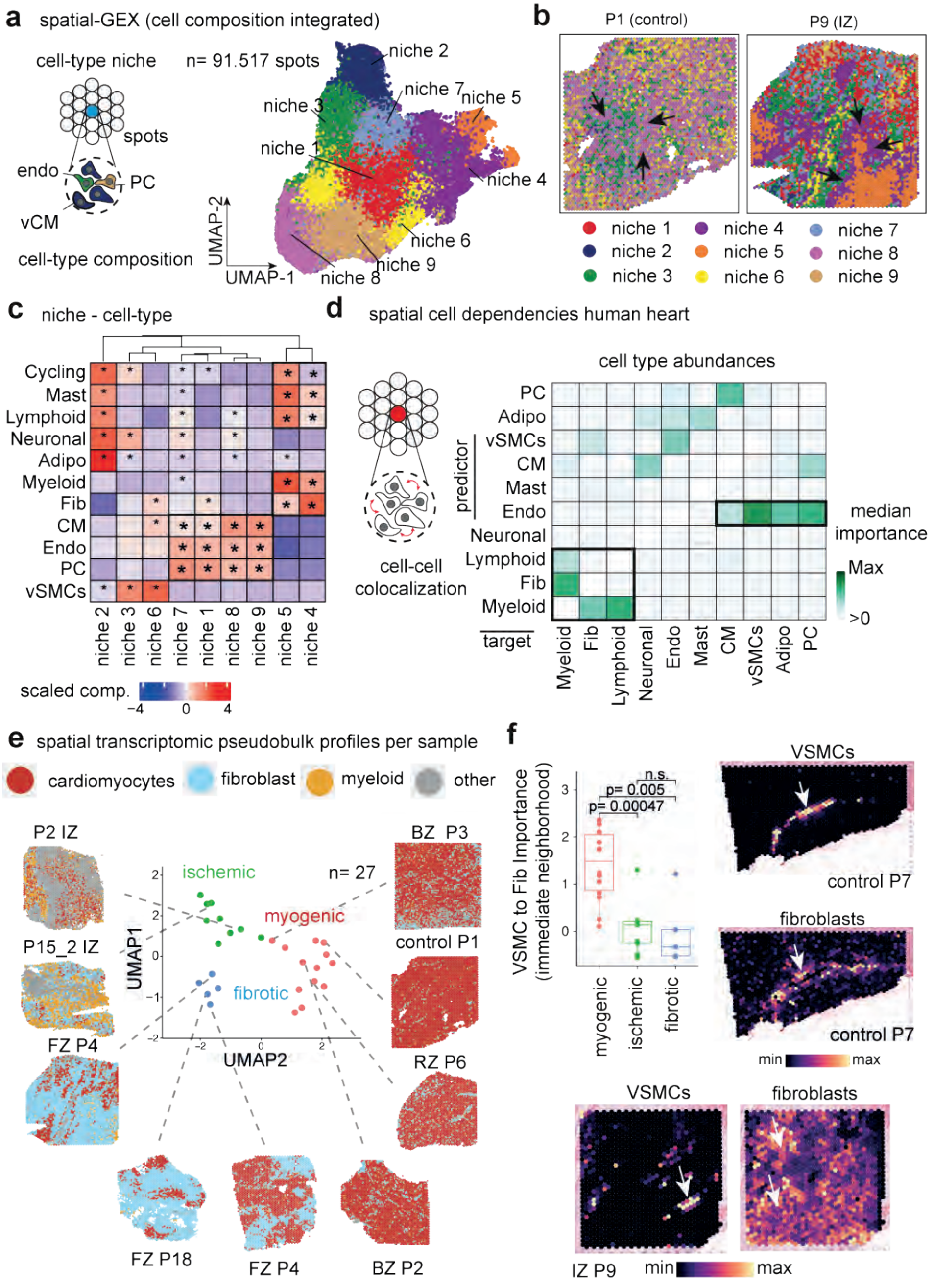

a
spatial-GEX (cell composition integrated)
cell-type niche
n= 91.517 spots
spots
endo
PC
vCM
cell-type composition
UMAP-2
UMAP-1
niche 2
niche 3
niche 7
niche 5
niche 1
niche 4
niche 6
niche 9
niche 8
b
P1 (control)
P9 (IZ)
niche 1
niche 2
niche 3
niche 4
niche 5
niche 6
niche 7
niche 8
niche 9
c
niche - cell-type
Cycling
Mast
Lymphoid
Neuronal
Adipo
Myeloid
Fib
CM
Endo
PC
vSMCs
niche 2
niche 3
niche 6
niche 7
niche 1
niche 8
niche 9
niche 5
niche 4
scaled comp.
−4
0
4
d
spatial cell dependencies human heart
cell type abundances
cell-cell colocalization
predictor
PC
Adipo
vSMCs
CM
Mast
Endo
Neuronal
Lymphoid
Fib
Myeloid
target
Myeloid
Fib
Lymphoid
Neuronal
Endo
Mast
CM
vSMCs
Adipo
PC
median importance
Max
>0
e
spatial transcriptomic pseudobulk profiles per sample
cardiomyocytes
fibroblast
myeloid
other
P2 IZ
P15_2 IZ
FZ P4
BZ P3
control P1
RZ P6
FZ P18
FZ P4
BZ P2
ischemic
myogenic
fibrotic
n= 27
UMAP1
UMAP2
f
VSMC to Fib Importance (immediate neighborhood)
n.s.
p= 0.005
p= 0.00047
myogenic
ischemic
fibrotic
VSMCs
control P7
fibroblasts
control P7
min
max
VSMCs
fibroblasts
IZ P9
min
max

Fig. 2

a, Schematic of cell-type niche definition and UMAP of spatial transcriptomics spots based on cell-type compositions. **b**, Mapping of cell-type niches in a control and an ischaemic zone sample. Arrows show niche 8 (left) and niche 4 (right). **c**, Scaled median cell-type compositions (comp.) within each niche. Asterisks indicate increased composition of a cell type in a niche compared with other niches (one-sided Wilcoxon rank sum test, adjusted (adj.) $P < 0.05$). Bold asterisks and outlines show the tissue modules discussed in the main text. **d**, Median importance of cell-type abundance in the prediction of abundances of other cell types within a spot. **e**, UMAP of all patient samples from spatial transcriptomics and visualization of abundance of the major cell types in myogenic (control, remote zone and border zone), ischaemic and fibrotic groups. **f**, Top left, importance of vSMC abundance in the immediate neighbourhood for prediction of fibroblast (Fib) abundance in myogenic, ischaemic and fibrotic groups (adj. P-value using a two-sided Wilcoxon rank-sum test is shown). In all box plots in this Article, the centre line corresponds to the median, the bottom and top hinges delineate the first and third quartiles, respectively, the top whisker extends from the hinge to the largest value no further than 1.5× the inter-quartile range (IQR) from the hinge and the bottom whisker extends from the hinge to the smallest value at most 1.5× IQR from the hinge; data beyond the end of the whiskers are outlying points and are plotted individually. Myogenic group: $n = 14$, ischaemic group: $n = 9$, fibrotic group: $n = 5$. Deconvoluted vSMCs and fibroblast abundance in a myogenic sample (top right) and in an ischaemic sample (bottom). For details on visualization, statistics and reproducibility, see Methods. NS, not significant. Adipo, adipocytes; CM, cardiomyocytes; PC, pericytes; Endo, endothelial cells.

of their cell-type compositions identified nine clusters, which we defined as major cell-type niches (Fig. 2a and Extended Data Fig. 3a–d). We hypothesized that these niches represent potential structural building blocks that are shared between different slides and could facilitate comparisons between subjects. Visualization of these niches in space revealed that some niches aligned closely with the underlying sample condition; for example, cell-type niche 8 was equally distributed across a control slide, whereas cell-type niche 5 localized to distinct regions on the ischaemic slide (Fig. 2b). We then tested the overrepresentation of the annotated cell types derived from snRNA-seq in the cell-type niches. We observed 4 myogenic cell-type niches (1, 7, 8 and 9), which were enriched with cardiomyocytes, endothelial cells, and pericytes (Fig. 2c); an inflammatory cell-type niche (niche 5); and a fibrotic cell-type niche (niche 4) containing fibroblasts, myeloid and lymphoid cells. The fibrotic cell-type niche (4) contained a higher proportion of fibroblasts, whereas the inflammatory cell-type niche (5) contained more myeloid and lymphoid cells (Fig. 2c). Finally, we observed niches associated with rare cell types of the myocardium, such as vSMCs (cell-type niches 3 and 6), adipocytes, lymphoid and cycling cells (cell-type niche 2) (Fig. 2c and Extended Data Fig. 3d). Our integrated results provide a comprehensive description of cellular colocalization events, enabling downstream molecular comparisons within this atlas across all tissue zonations. We next tested whether the abundances of major cell types within spots could be predicted by their spatial context described by the cell-type compositions of their neighbourhood. We evaluated three different neighbourhood area sizes using MISTy: (1) the importance of cell-type abundances within a spot (colocalization) (Fig. 2d), (2) in the local neighbourhood (radius of 1 spot), and (3) in an extended neighbourhood that expanded to a radius of 15 spots. We observed that endothelial cells were the most predictive of the abundance of vSMCs, pericytes, adipocytes and cardiomyocytes within all spots, probably reflecting dependencies between cell types of the vasculature (Fig. 2d). Lymphoid and myeloid cells showed strong dependencies with each other in line with zones of immune cell infiltration and inflammation—similarly captured by cell-type niche 5 (Fig. 2d). Notably, we observed strong dependencies between myeloid cells and fibroblasts, which were strongly co-enriched in niche 4 (Fig. 2d and Extended Data Fig. 3e), in line with a known key role of macrophages in fibroblast activation14 and fibroblasts in macrophage attraction[15]. Between immediate and extended neighbouring spots (Extended Data Fig. 3f–h), we observed stronger dependencies between cells associated with the cardiac vasculature (vSMCs, endothelial cells, pericytes and fibroblasts) indicating that the myocardial vascular network dominates cardiac tissue structural organization.

To link tissue organization to function, we analysed spatial dependencies between signalling pathways and cell types. Modelled importance of colocalized pathways captured relationships between PI3K and p53 signalling (Extended Data Fig. 4a–e), which showed a mutually exclusive spatial distribution (Extended Data Fig. 4c). Both pathways were related to the abundance of cardiomyocytes (Extended Data Fig. 4a). PI3K signalling in cardiomyocytes controls the hypertrophic response to preserve cardiac functions[16], whereas p53 is known to act as a master regulator in cardiac homeostasis[17]. Spatial segregation of these cardiomyocyte-related pathways points towards functional cardiomyocyte heterogeneity. We observed colocalized and extended neighbourhood relationships of known key pathways in fibrosis including TGFβ and NFκB predicted by fibroblasts, and JAK–STAT and NFκB predicted by immune cells (Extended Data Fig. 4a–e). Overall, cardiomyocytes were the best predictor cell types of the activities of the estimated pathways. Hypoxia and WNT pathways showed a colocalization to cardiomyocytes in ischaemic specimens (Extended Data Fig. 4b–e), highly consistent with the cardiomyocyte differentiation events occurring after myocardial infarction[18]. Our results compiled principles of tissue organization of the human heart that relate to coordinated cellular processes and provide a basis for comparative analysis.

Structural variation of cardiac tissue

To identify general tissue differences during remodelling after myocardial infarction, we compared the samples of distinct histomorphological regions, time points and individuals at the molecular and compositional level. We defined three major sample groups: myogenic-enriched (including control, border zone and remote zone), fibrotic-enriched (including all fibrotic zone samples, except one) and ischaemic-enriched (including all ischaemic zone samples) samples. Hierarchical clustering of their pseudo-bulk spatial transcriptomics supported this grouping and was displayed as a UMAP embedding (Fig. 2e and Extended Data Fig. 4f). Co-clustering of control, border zone and remote zone samples can be explained by the large abundance of functional myocardial tissue within these specimens (Fig. 2e). Since the pseudo-bulk profile of each spatial transcriptomic dataset combines information of multiple cell types, we next tested how differences in cellular composition determined by all modalities (that is, snRNA-seq, snATAC-seq and spatial transcriptomics) are associated with these three groups. Ischaemic-enriched samples showed a larger proportion of myeloid, lymphoid and cycling cells, with the lowest proportions of cardiomyocytes, representing cellular compositional changes expected after myocardial infarction. By contrast, fibroblasts and vSMCs were enriched in fibrotic-enriched samples (Extended Data Fig. 4g). These results indicate that the spatial transcriptomic data align with major histomorphological sample annotation and capture compositional hallmarks following myocardial infarction across our datasets.

We then analysed whether the cell-type compositional changes between sample groups were also reflected as changes in the spatial dependencies between the major cell types in spatial transcriptomics. To this end, we contrasted the importance, previously computed using MISTy, of each major cell type in predicting the others in the three different neighbourhood area sizes (colocalization, immediate and extended neighbourhood) between the three different sample groups (Extended Data Fig. 4h). We observed an increased spatial dependency in the immediate neighbourhood between lymphoid and myeloid cells in ischaemic samples compared with myogenic-enriched samples, reflecting the expected role that immune cell interactions have in cardiac repair following myocardial infarction (Extended Data Fig. 4i). Moreover, an increased colocalization of cardiomyocytes and pericytes in fibrotic-enriched samples revealed an exclusion of pericytes from scar tissue areas (Extended Data Fig. 4j). Similarly, the distribution of fibroblasts was better predicted by the presence of vSMCs in the immediate neighbourhood only in myogenic-enriched samples, where fibroblasts surrounded the vasculature, in contrast to ischaemic and fibrotic tissue specimens, where more extensive tissue scarring processes were captured (Fig. 2f).

We next compared compositions of cell-type niches between groups and observed differences in six out of nine cell-type niches (Extended Data Fig. 4k). Cell-type niches 8 and 9 (Extended Data Fig. 4k–l), mostly representing cardiac muscle structures, were more present in myogenic- and fibrotic-enriched samples compared with ischaemic-enriched samples, whereas cell-type niche 7, enriched in cardiomyocytes and pericytes (Extended Data Fig. 4k), was reduced in fibrotic-enriched samples. Niche 4, mainly associated with fibrotic structures (more fibroblasts than myeloid cells and thus termed fibrotic niche), was observed in higher proportions in fibrotic-enriched samples, whereas niche 5 (more myeloid cells than fibroblasts and thus termed inflammatory niche) was mainly present in ischaemic-enriched samples (Extended Data Fig. 4k). In summary, the major cell-type niches enabled us to categorize and compare interindividual spatial differences. Overall, this demonstrates the importance of cardiac vasculature in defining

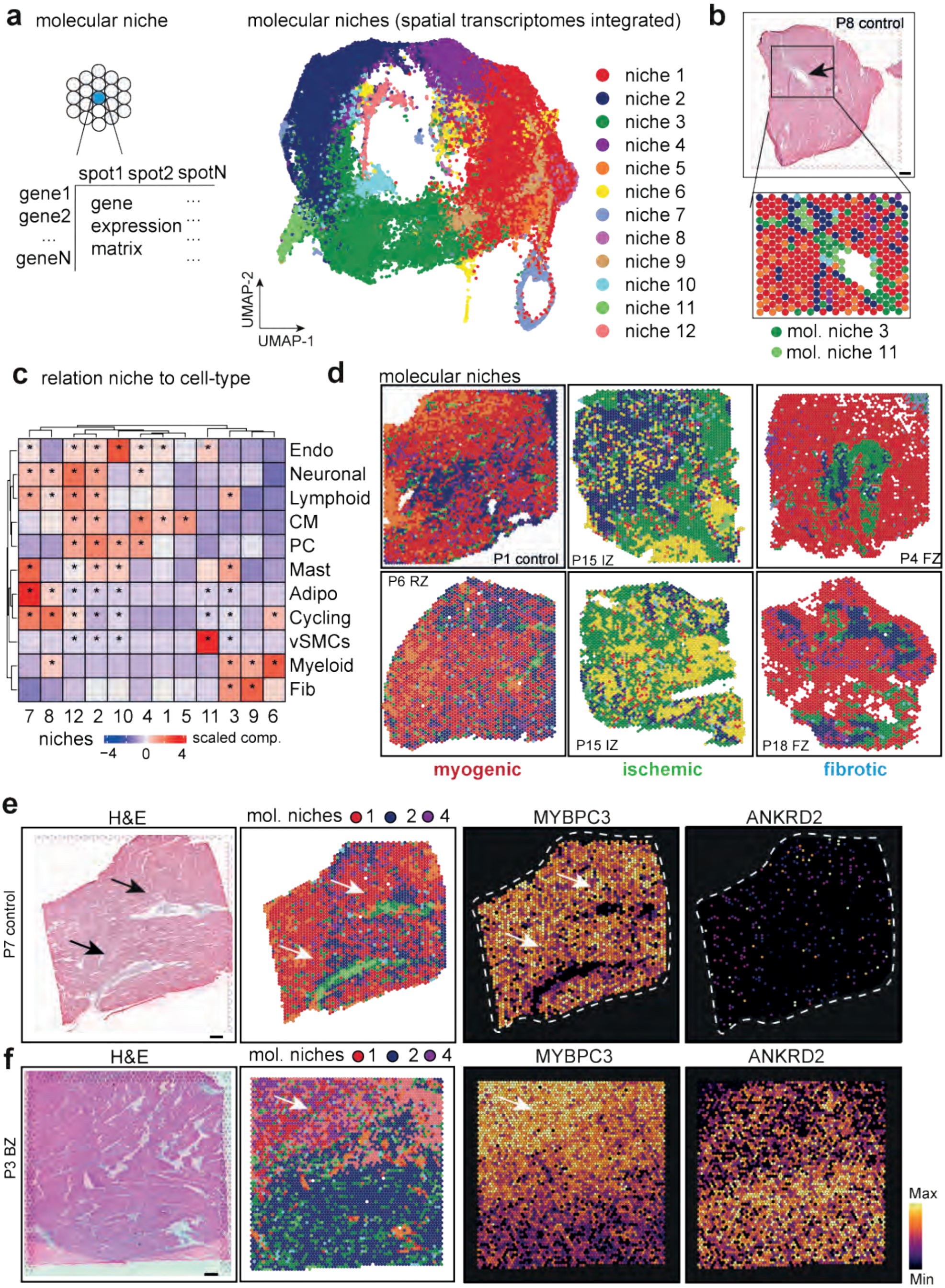

a
molecular niche
spot1 spot2 spotN
gene1
gene2
...
geneN
gene
expression
matrix
molecular niches (spatial transcriptomes integrated)
UMAP-2
UMAP-1
niche 1
niche 2
niche 3
niche 4
niche 5
niche 6
niche 7
niche 8
niche 9
niche 10
niche 11
niche 12
b
P8 control
mol. niche 3
mol. niche 11
c
relation niche to cell-type
Endo
Neuronal
Lymphoid
CM
PC
Mast
Adipo
Cycling
vSMCs
Myeloid
Fib
7 8 12 2 10 4 1 5 11 3 9 6
niches
scaled comp.
−4 0 4
d
molecular niches
P1 control
P15 IZ
P4 FZ
P6 RZ
P15 IZ
P18 FZ
myogenic
ischemic
fibrotic
e
H&E
mol. niches 1 2 4
MYBPC3
ANKRD2
P7 control
f
H&E
mol. niches 1 2 4
MYBPC3
ANKRD2
P3 BZ
Max
Min

Fig. 3

a, Schematic of molecular niche definition and UMAP of spatial transcriptomics spots based on gene expression. **b**, Spatial mapping of molecular niches. Arrows highlight molecular niche 11 (enriched in *MYH11*$^+$ vSMCs) surrounded by molecular niche 3 (enriched in *PDGFRA*$^+$ fibroblasts). **c**, Scaled median cell-type compositions within each molecular niche. Asterisks indicate increased composition of a cell type in a niche compared with other niches (one-sided Wilcoxon rank-sum test, adj. $P < 0.05$). **d**, Distribution of molecular niches in three different patient groups. Note the differential abundance of molecular niches 1 (red) and 6 (yellow). **e**,**f**, Haematoxylin and eosin (H&E) staining and visualization of molecular niches 1, 2 and 4 and gene expression (*MYBPC3* and *ANKRD2*) of a control (**e**) and a border zone (**f**). Scale bars, 10 mm. For details on visualization, statistics and reproducibility, see Methods.

the overall myocardial architecture and the unique spatial dependencies of fibroblasts and myeloid cells, which facilitates gaining molecular insights of disease-specific spatial tissue remodelling.

Molecular variation following infarction

To study the molecular differences between similar tissue structures in an unbiased manner across samples, we generated a set of molecular niches by clustering of spots on the basis of their gene expression (Fig. 3a,b and Extended Data Fig. 5a–d). We identified molecular niches associated with inflammatory and fibrotic processes (molecular niches 3, 6 and 9), vSMCs (molecular niche 11) and myogenic-enriched regions (molecular niches 1, 2, 4, 5 and 12) (Fig. 3c). The molecular niches enriched in cardiomyocytes were depleted in ischaemic-enriched samples, whereas the fibrotic- and inflammatory-enriched molecular niches were depleted in myogenic-enriched samples (Fig. 3d and Extended Data Fig. 5e,f). The vSMC-enriched molecular niche 11 had a more distinct cell-type marker gene expression of vSMCs (MYH11) compared with the cell-type defined niche 6 (Fig. 3b versus Extended Data Fig. 3d).

Of note, we observed molecular niches that enabled us to differentiate border zone, remote zone and control samples (Extended Data Fig. 5g), which were indistinguishable using the major cell-type niches (Extended Data Fig. 4m). Molecular niche 3, enriched in fibroblasts and immune cells, was more present in remote zones and border zones compared with control samples. Moreover, we observed differences in the proportions of the molecular niches 1, 2 and 4 among border zone, remote zone and controls (Extended Data Fig. 5g). These three molecular niches were enriched mainly in cardiomyocytes (Fig. 3c), but with a distinct molecular profile: among the top 5 upregulated genes of niche 2 was XIRP1, which encodes an intercalated-disc ion-channel-interacting protein and RRAD, which encodes a GTPase known to regulate L-type Ca2+ channels and contractile functions of the heart19; molecular niche 4 was enriched for SLC8A1 (also known as NCX1), which encodes the Na+/Ca2+ exchanger that is the major regulator of the Ca2+ efflux in cardiomyocytes and is critical to maintain Ca2+ homeostasis during excitation–

contraction coupling20, and MPC1, which encodes mitochondrial pyruvate carrier, a known mitochondrial metabolic regulator of heart function21 (Extended Data Fig. 5h). Overall, molecular niche 1 was enriched in control and remote zone samples and niche 2 was enriched in the damaged tissue areas in border zone samples (Fig. 3e,f and Extended Data Fig. 5g). We observed slight changes in enrichment of molecular niches 2 and 4, and a depletion of niche 1 in border zones compared with controls (Extended Data Fig. 5g,i,j), suggesting that differences in cardiomyocyte phenotypes might also be present between these groups. In summary, the comparison of molecular niches pointed towards subtle changes between the remote myocardium and controls, and expected differences between border zone and both controls and remote zone that were not detectable in the cell-type niche comparison. Overall, this suggested the existence of functional differences between cardiomyocyte states in our data.

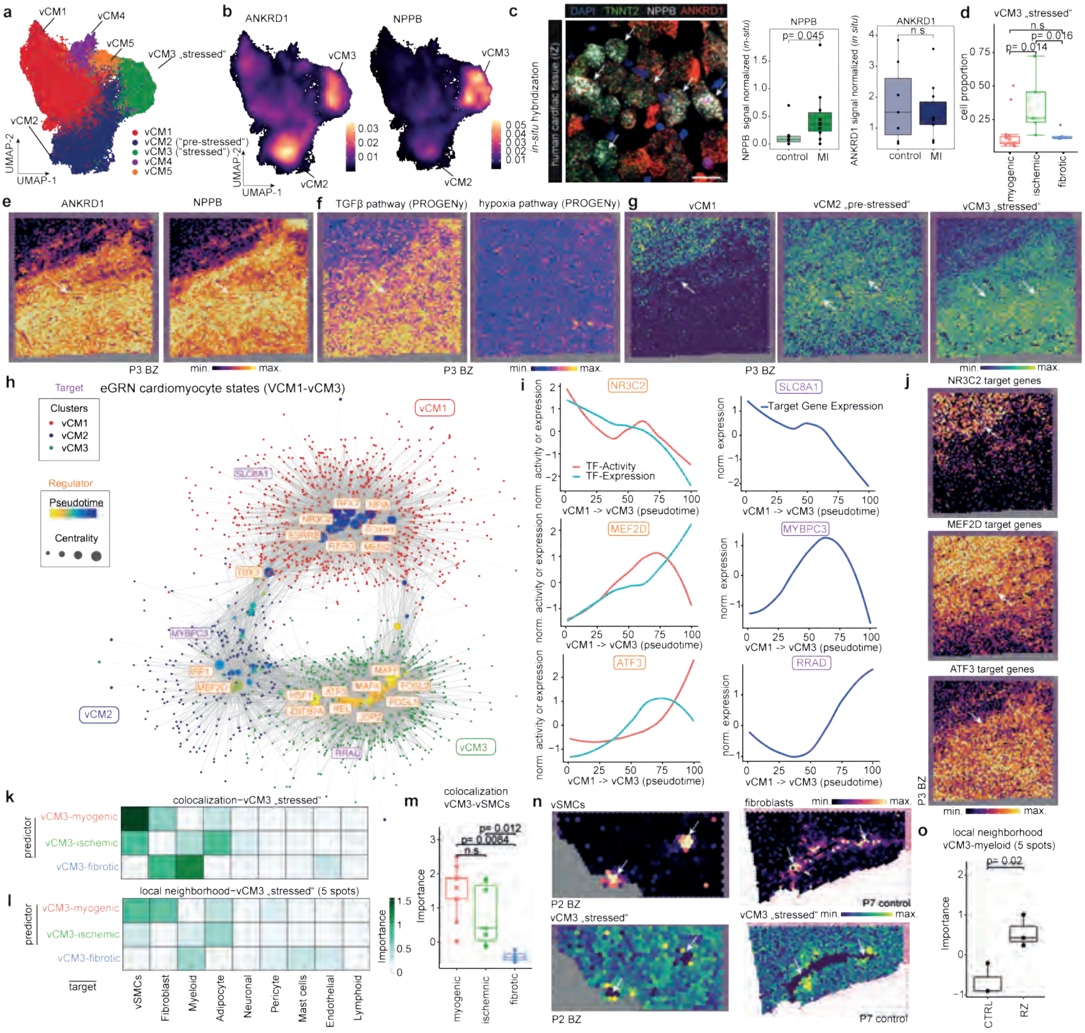


Fig. 4

a, Sub-clustering of cardiomyocytes. **b**, Gene expression of *ANKRD1* and *NPPB*. **c**, Left, smFISH staining of vCM3 marker genes. Scale bar, 50 µm. Right, quantification of *NPPB* and *ANKRD1* signal relative to the *TNNT2* signal. Two-sided Wilcoxon rank-sum test (control donors: *n* = 7, patients with myocardial infarction (MI): *n* = 10). **d**, Proportion of stressed vCM3 cells. Wilcoxon rank-sum test (unpaired, two-sided; myogenic: *n* = 13 myogenic, ischaemic: *n* = 7, fibrotic: *n* = 4). **e–g**, Expression of *ANKRD1* and *NPPB* (**e**), TGFβ and hypoxia signalling activities (**f**) and expression of vCM-state marker genes (**g**) in a border zone sample. **h**, eGRN analysis including vCM1, vCM2 and vCM3. Each node represents a transcription factor (regulator) or a gene (target). **i**, Transcription factor activity and expression over pseudotime. Norm., normalized. **j**, Expression of transcription factor target genes in the border zone sample, as in **e. k,l**, Mean importance of the abundance of major cell types within a spot (**k**) and the local neighbourhood (within a 5-spot radius) (**l**) in the prediction of vCM3 in spatial transcriptomics. **m**, Importance of vSMC abundance predict vCM3 in myogenic, ischaemic and fibrotic groups within a spot (adj. *P*-value, two-sided Wilcoxon rank-sum test; myogenic: *n* = 9, ischaemic: *n* = 7, fibrotic: *n* = 4). **n**, Deconvoluted abundance of vSMCs or fibroblasts and vCM3 state scores in a border zone (left) and a control human heart (right). **o**, Importance of myeloid cell abundance in the local neighbourhood for predicting vCM3 in control and remote zone samples (two-sided *t*-test; controls: *n* = 3, remote zones: *n* = 3). For details on visualization, statistics and reproducibility, see Methods.

Disease-specific cardiomyocyte states

To further investigate distinct cardiomyocyte states, we aimed to understand the molecular heterogeneity of cardiomyocytes after myocardial infarction. We co-embedded the snRNA-seq and snATAC-seq data from cardiomyocytes into a common low-dimensional space and clustered the cells (Extended Data Fig. 6a). This uncovered five cell states of ventricular cardiomyocytes (vCM1–5), spanning multiple samples and modalities (Fig. 4a and Supplementary Table 10). Differential gene expression analysis revealed a significant upregulation of ANKRD1 in both vCM2 and vCM3, whereas NPPB showed a distinct upregulation and increased chromatin accessibility in vCM3 (Fig. 4b and Extended Data Fig. 6b,c). We validated this upregulation by single-molecule fluorescence in situ hybridization (smFISH) in an independent patient cohort (Fig. 4c and Extended Data Fig. 6d). Both NPPB and ANKRD1 have been reported to be upregulated in the border zone after myocardial infarction in mice22. vCM2 additionally showed enhanced expression of MYH7 (Extended Data Fig. 6b), a cardiomyocyte-associated stress gene that encodes the β-myosin heavy chain23. Thus, we annotated the vCM2-state as 'pre-stressed'. In addition, we observed a higher correlation between ion-channel-related genes and vCM1 marker genes compared with 'stressed' vCM3 marker genes in spatial transcriptomics, which further highlights the functional differences between these two cardiomyocyte states (Extended Data Fig. 6e). Accordingly, we annotated the vCM3 state as stressed. Moreover, when comparing the differential expression of individual genes belonging to these ion-channel-related gene sets in snRNA-seq data, we observed mostly upregulations in vCM1 compared with vCM3 (Extended Data Fig. 6f,g). Cellular composition comparison between sample groups revealed that vCM1 was associated with myogenic-enriched samples and vCM3 was significantly associated with ischaemic-enriched samples. This was validated in an independent cohort using in situ hybridization, suggesting that these cardiomyocyte states represent distinct cellular stress states within the acute myocardial infarction phase (that is, vCM1, 'non-stressed'; vCM2, 'pre-stressed'; and vCM3, 'stressed') (Fig. 4c,d and Extended Data Fig. 6h,i).

Next, we checked vCM marker genes in spatial transcriptomics in border zone samples, since spatial remodelling of this area is inextricably linked to the recovery of cardiac function. Interestingly, despite homogenous H&E staining and unique molecular identifier (UMI) distribution across spots (Supplementary Fig. 2g), we observed extensively heterogeneous spatial gene expression patterns of ANKRD1 and NPPB (Fig. 4e). Pathway analysis of the spatial gene expression data indicated an increased TGFβ signalling activity within the injured area (lower right), but a homogeneous distribution of hypoxia pathway activity (Fig. 4f). Mapping of cell states to space in a border zone sample revealed that vCM1 were solely located in the top left uninjured corner, vCM2 were located in the middle–top area, serving as a transition zone from injured towards remote myocardium, and vCM3 were primarily located below the transition zone within the injured area (Fig. 4g). Of note, such a spatially distributed pattern was also observed in another border zone sample, indicative of a similar remodelling process (Extended Data Fig. 6j).

Variability of cardiomyocyte states

To infer an enhancer-based gene-regulatory network (eGRN), we leveraged our multi-omics data to further investigate molecular mechanisms differentiating the relevant cardiomyocyte states (that is, vCM1–vCM3) (Methods and Supplementary Table 11). To this end, we paired the cells between snATAC-seq and snRNA-seq data and studied gene-regulatory changes along the cellular continuum from vCM1 to vCM3 (Extended Data Fig. 7a). Next, we estimated an enhancer-mediated transcription factor–target network by considering transcription factor activity (from snATAC-seq), expression of transcription factor and target genes (from snRNA-seq), and motif-supported peak-to-gene links (Extended Data Fig. 7b–d). Clustering of these transcription factors to the target network revealed three major modules, with each corresponding to a distinct cardiomyocyte state (Extended Data Fig. 7e).

We next used network analysis to visualize and detect major transcription factors (Fig. 4h). We identified the mineralocorticoid receptor (NR3C2), a major target of therapy for common heart failure, as a major regulator of the vCM1 state (Fig. 4h). Decreased NR3C2 expression has been associated with the development of severe heart failure and cardiac fibrosis24, and we observed decreased transcription factor binding activity and gene expression along the pseudotime of vCM1 to vCM3 differentiation (Fig 4i). Target genes of NR3C2 include several ion channel genes (such as SLC8A1), which also showed decreased gene expression along the pseudotime axis (Fig. 4i). Notably, these target genes were also differentially expressed in cardiomyocyte-enriched molecular niches

(Fig. 3e,f and Extended Data Fig. 5h) and aligned spatially in the border zone with the vCM1 state (Fig. 4j). Notably, we also observed transcription factors (TBX3 and MEF2D) that were associated with pre-stressed stages of cardiomyocyte differentiation (Fig. 4h). Our analysis suggests that MEF2D, a cardiomyocyte factor controlling pacemaker function[25], regulates the expression of the sarcomere protein MYBPC3 (Fig. 4i). MYBPC3, in turn, has been reported to regulate cardiomyocyte proliferation postnatally[26]. Of note, we identified MYBPC3 independently in our spatial data as being enriched in molecular niche 1 (Fig. 3e and Extended Data Fig. 5h).

We also identified ANKRD1, a mediator of cardiomyocyte response to stress[27], as a target of MEF2D, suggesting a key regulatory role of MEF2D in the transition from vCM1 to vCM327 (Extended Data Fig. 7f). For vCM3 (stressed cardiomyocytes), we identified ATF3 as a regulator of the GTPase and Ca2+ regulator gene RRAD (Fig. 4h). We independently identified RRAD in molecular niche 2 (Extended Data Fig. 5h), which supports its relevance as a spatially differentially expressed gene of a distinct cardiomyocyte state, especially in border zone samples (Fig. 4i). We additionally identified the transcriptional regulator JDP2—which has a function in preventing cardiomyocyte hypertrophy and cell death[28]—as an important regulator of the vCM3 cardiomyocyte state, with TGFB2 as one of its target genes (Extended Data Fig. 7g,h). In summary, our cardiomyocyte states and major transcription factor regulators identified from the integrated snRNA-seq and snATAC-seq data reflect expression patterns associated with molecular niches supporting spatial changes of cardiomyocyte states during remodelling.

We next estimated the cell dependencies of the stressed cardiomyocyte state vCM3 with other cell types within each spatial spot and its local neighbourhood (radius of five spots) between sample groups (Fig. 4k–o). We observed that there was an increased importance of vSMCs in predicting vCM3 within a spot in myogenic and ischaemic samples (Fig. 4k), whereas fibroblasts and myeloid cells had a larger role in fibrotic samples (Fig. 4k). The local neighbourhood modelling of vCM3 revealed that the abundance of fibroblasts better explained vCM3 in myogenic-enriched samples compared with fibrotic samples (Fig. 4l and Extended Data Fig. 7i). To gain further insight, we visualized the dependencies of vSMCs and fibroblasts on vCM3 in myogenic-enriched samples and observed that their colocalization occurred in the perivascular niches (Fig. 4n). Overall, this demonstrates that the stressed cardiomyocyte state vCM3 occurs in the perivascular niche of larger blood vessels, highlighting the interaction of mesenchymal cells[29] of the perivascular niche with stressed cardiomyocytes in this tissue area. Furthermore, we noticed that when comparing remote zone with control samples, stressed vCM3s are best predicted by myeloid cells (Fig. 4o). This underlines the importance of immune–cardiomyocyte interactions that could additionally explain the increased arrhythmia susceptibility in the remote regions of the post-infarct heart, since it has been shown that cardiac macrophages influence normal and aberrant cardiac conduction30. Our results showed that the stressed-cardiomyocyte vCM3 can be found in distinct spatial cell-type neighbourhoods enriched by different compositions of vSMCs, fibroblasts, adipocytes or myeloid cells.

Cardiac endothelial cell heterogeneity

Co-embedding of snRNA- and snATAC-seq data identified five subtypes of endothelial cells from all major vascular beds, namely capillary endothelial cells, arterial endothelial cells, venous endothelial cells, lymphatic and endocardial endothelial cells (Fig. 5a, Extended Data Fig. 8a–d, and Supplementary Table 12). Subtype-based pseudo-bulk ATAC-seq signals also revealed distinct chromatin accessibility of these marker genes (Extended Data Fig. 8c). Our analysis suggested POSTN as a characteristic marker for endocardial endothelial cells, which we validated using smFISH (Extended Data Fig. 8e). Analysis of cell proportions among the myogenic-enriched, ischaemic-enriched and fibrotic-enriched samples revealed a reduction of capillary endothelial cells in the ischaemic samples associated with a concordant increase in venous endothelial cells (Fig. 5b and Extended Data Fig. 8f,g). Furthermore, we observed that lymphatic endothelial cells were overall less abundant than the other populations, as expected, but were significantly increased in the ischaemic zone, suggesting an increased abundance of lymphatics modulating the immune response following cardiac injury[31] (Fig. 5b).

We modelled the association of the different endothelial cell subtypes with the abundances of the other major cell types in spatial transcriptomics. We observed that the markers of arterial endothelial cells were best predicted by vSMCs within a spot and in the local neighbourhood (radius of five spots) reflecting the anatomy of arteries and arterioles in the heart (Fig. 5c,d and Extended Data Fig. 8h). Moreover, the expression of markers of capillary endothelial cells were best predicted by the presence of pericytes in the tissue, in line with the known presence and role of pericytes in direct contact with capillary endothelium32 (Extended Data Fig. 8i). The other endothelial subtypes were mainly predicted by the presence of fibroblasts within a spot and in the local neighbourhood (Extended Data Fig. 8h). Additionally, we observed that the abundance of myeloid cells correlated with the expression of markers of lymphatic endothelial cells (Extended Data Fig. 8h). Focusing on molecular niche 10, which contained the highest cell proportion of endothelial cells and additionally pericytes and mast cells (Extended Data Fig. 8j), we observed a significant enrichment of capillary endothelial cells (Extended Data Fig. 8k). Pathway analysis revealed a significantly higher hypoxia and TGFβ signalling activity in ischaemic and in fibrotic samples, underlining the importance of these processes in chronic fibrotic cardiac remodelling processes (Extended Data Fig. 8l). Pathways important for endothelial signalling in homeostasis such as PI3K and TRAIL showed a reduction in the fibrotic and ischaemic groups, respectively, highlighting

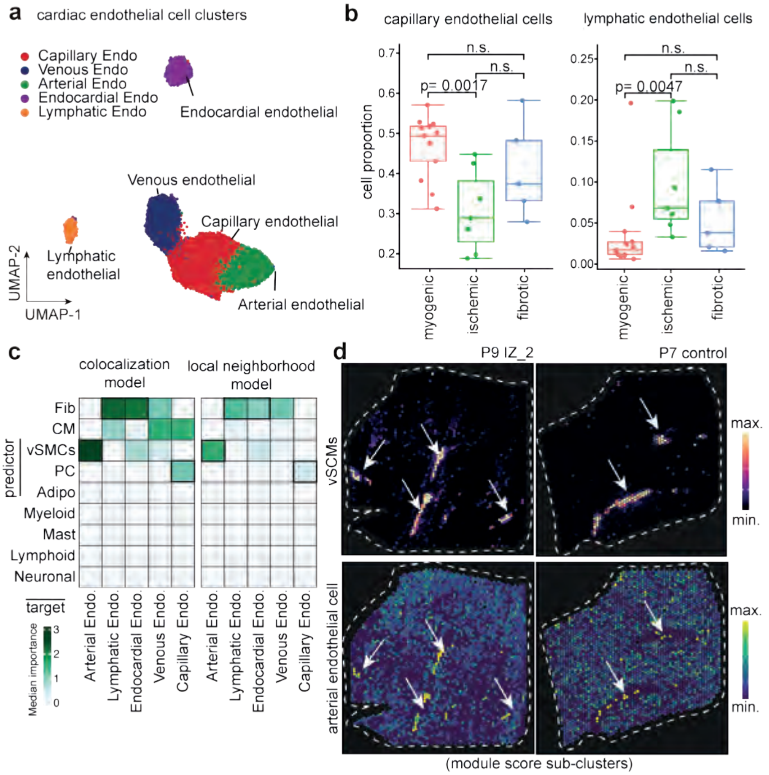


Fig. 5

a, Sub-clusters of human endothelial (endo) cells using the integrated snRNA-seq and snATAC-seq data. **b**, Comparison of capillary endothelial cells and lymphatic endothelial cell proportion between donor and patient groups. Wilcoxon rank-sum test (unpaired, two-sided; myogenic: *n* = 13, ischaemic: *n* = 7, fibrotic: *n* = 4). **c**, Median importance of the abundance of major cell types within a spot (left) and the local neighbourhood (effective radius of 5 spots) (right) in the prediction of endothelial cell-state scores in spatial transcriptomics. **d**, Spatial distribution of the abundance of vSMCs and the state score of arterial endothelial cells in an ischaemic (left) and control (right) sample. Arrows point at colocalization events. For details on visualization, statistics and reproducibility, see Methods.

further the differential endothelial cell signalling changes. Gene set enrichment analysis further revealed an altered metabolism (for example, fatty acid metabolism and oxidative phosphorylation) of this endothelial cell niche in diseased samples which was further associated with an increased inflammatory response via the TNF and NFκB pathways and increased apoptosis signalling[33] (Extended Data Fig. 8m). In summary, we resolved all major endothelial cells states, localized them in space and described their spatial dependencies. Further, we identified a spatial niche enriched in capillary endothelial cells with complex metabolic and signalling changes.

Cardiac myofibroblast differentiation

To dissect molecular and cellular mechanisms of fibrogenesis in the human heart, we clustered all fibroblasts using the integrated snRNA-seq and snATAC-seq data and identified four sub-clusters (Fib1–4) (Fig. 6a, Extended Data Fig. 9a and Supplementary Table 13). Fib1 was marked by SCARA5, which we recently reported as a marker for myofibroblast progenitors in the human kidney[34]. Fib2 was marked by POSTN, COL1A1 and FN1, which, together with the fact that this population expresses most extracellular matrix (ECM)-related genes, suggests that Fib2 indeed comprises terminally differentiated myofibroblasts (Fig 6b and Extended Data Fig. 9a–c). Notably, Fib2 also exhibited an upregulation of RUNX1, which we recently reported as being involved in kidney myofibroblast differentiation35. Overexpression of RUNX1 in human heart PDGFRβ-expressing cells led to increased myofibroblast differentiation and matrix expression (Extended Data Fig. 9d). We validated the presence of high SCARA5 expression in fibroblasts by co-staining with the pan-fibroblast and myofibroblast marker COL15A1 as well as POSTN and COL1A1 in human heart tissues, and demonstrated that POSTN is significantly enriched in COL1A1+ cells compared with SCARA5+ cells (Extended Data Fig. 9e). Visualization of these markers in our spatial transcriptomics dataset suggested that Fib1 and Fib2 were enriched in mutually exclusive regions of the heart following injury (Fig. 6c and Extended Data Fig. 9f). Additionally, we observed that Fib1 comprised the highest proportion in myogenic-enriched samples, whereas Fib2 (myofibroblasts) were significantly enriched and Fib3 slightly reduced in ischaemic samples (Fig. 6d and Extended Data Fig. 9 g,h).

To precisely understand differentiation trajectories of fibroblasts and transfer this knowledge to the human data, we performed inducible lineage tracing in mice using the pan-mesenchymal Cre driver Pdgfrb-CreER (crossed to a R26-tdTomato reporter) combined with scRNA-seq at different time points following myocardial infarction (Extended Data Fig. 9i–l). We integrated and annotated the cells by label transfer (Fib1–4) from human to mouse (Extended Data Fig. 9m,n). We observed an overall increase of the Fib2 population and collagens and ECM genes over time, whereas the Fib1 proportion was decreased, pointing towards a differentiation trajectory from SCARA5+ fibroblasts (Fib1) to myofibroblasts (Fib2) in mice (Extended Data Fig. 9o,p). Based on these observations, we inferred a pseudotime trajectory from Fib1 (SCARA5+) to Fib2 (myofibroblast) in the human samples, which was further supported by an increased enrichment of the ECM score (Fig. 6e,f) and of ECM biological gene ontology processes consistent with fibroblast-to-myofibroblast differentiation (Extended Data Fig. 9q).

To understand the regulatory mechanisms of these stromal cell differentiation processes we inferred a fibroblast eGRN (Fig. 6g, Extended Data Fig. 10a,b and Supplementary Table 14). Clustering resolved two eGRN modules that each corresponded to

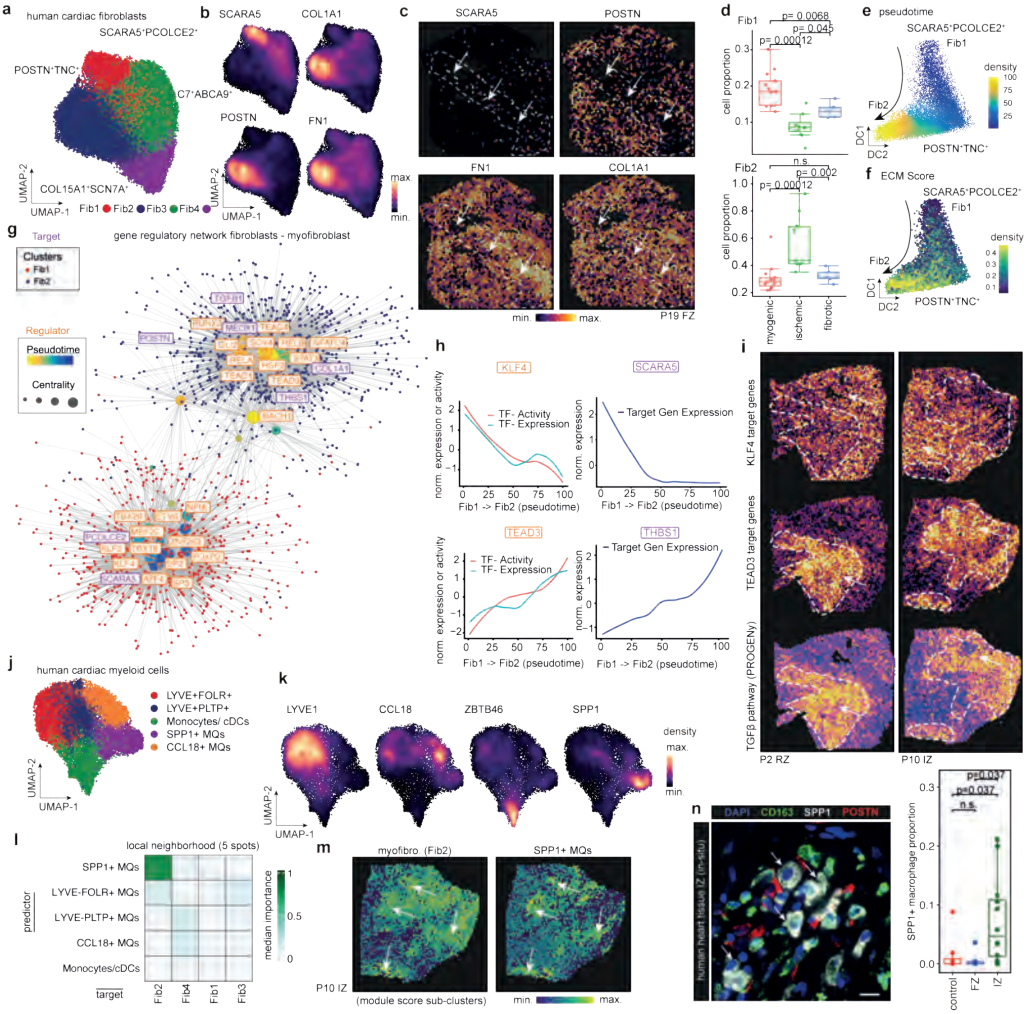


Fig. 6

a, UMAP of human cardiac fibroblasts (integrated snRNA-seq and snATAC-seq data). **b**, Expression of *SCARA5*, *COL1A1*, *POSTN* and *FN1*. **c**, Visualization of the markers in spatial transcriptomics data. **d**, Comparison of Fib1 and Fib2 compositions. Wilcoxon rank-sum test (unpaired, two-sided; myogenic: *n* = 13; ischaemic: *n* = 8, fibrotic: *n* = 5). **e**, Diffusion map of Fib1 and Fib2 populations. Colours refer to pseudotime points. **f**, Same as **e**, with colours referring to ECM score. **g**, eGRN analysis, including Fib1 and Fib2. Each node represents a transcription factor (regulator) or gene (target). Targets are coloured by clustering results and regulators are coloured by pseudotime with maximum transcription factor activity. The size of regulator nodes represents centrality. **h**, Transcription factor activity and expression over pseudotime and their corresponding target gene over pseudotime. **i**, Visualization of *KLF4* and *TEAD3* target genes and TGFβ pathway activity in a remote zone (left) and ischaemic zone (right) sample. **j**, UMAP of sub-clusters of human cardiac myeloid cells using the integrated snRNA-seq and snATAC-seq data. cDC, classical dendritic cell; MQ, macrophage. **k**, Gene expression of *LYVE1*, *CCL18*, *ZBTB46* and *SPP1*. **l**, Median importance of myeloid cell states in the local neighbourhood in the prediction of fibroblast cell states. **m**, Cell-state scores of myofibroblasts (Fib2) and *SPP1*⁺ MQs in a remote zone sample. Arrows point to regions where there is an observed colocalization. **n**, In situ staining of *CD163*, *POSTN* and *SPP1* on human cardiac myocardial infarction tissue. Arrows indicate *CD163+SPP1+* macrophages near myofibroblasts. Scale: 10 µm. Quantification of *SPP1*+ macrophages relative to *CD163*+ macrophages from the in situ hybridization images (adj. *P*-value from a two-sided Wilcoxon rank-sum test, *n* = 8 control group, *n* = 6 fibrotic group, *n* = 12 ischaemic group). For details on visualization, statistics and reproducibility, see Methods.

a distinct fibroblast state (Extended Data Fig. 10c) and identified potential regulators of myofibroblast differentiation (Fig. 6g). Among the transcription factors regulating the Fib1 module was KLF4, which regulates diverse cellular functions including cellular growth arrest, and is also one of the original reprogramming factors of induced pluripotent stem cells. Our network analysis highlighted the role of KLF4 in regulating SCARA5 and PCOLCE2 expression in Fib1, and it also targets MBLN1, an important regulator of cardiac wound healing36 and fibroblast-to-myofibroblast transition[37]. Concordantly, we observed reduced KLF4 binding activity and reduced SCARA5 expression in our pseudotime analysis (Fig. 6h), highlighting the role of KLF4 as a putative inhibitor of fibroblast activation. Among the transcription factors identified in the Fib2 module were TEAD3 (an effector of the Hippo pathway), GLI2 (in the hedgehog pathway) and RUNX2, which have been previously identified as regulators of myofibroblast differentiation[38] (Fig. 6h and Extended Data Fig. 10d,e). Our network analysis revealed that both TEAD3 and GLI2 regulate bona fide myofibroblast target genes including COL1A1, TGFB1 and POSTN. Additionally, our network analysis identified the key anti-angiogenic regulator THBS139 as a direct target of TEAD3 and the recently identified cardiac fibrosis regulator MEOX1 in human cardiac myofibroblasts[40]. We next visualized the expression of the KLF4 and TEAD3 target genes in spatial transcriptomics slides and observed gradients and mutually exclusive spatial expression in defined cardiac regions of fibrotic responses, highlighting their differential spatial activity in the human heart (Fig. 6i and Extended Data Fig. 10d).

Fibro-myeloid spatial relations

Myeloid-derived cells have been reported to have key roles in cardiac remodelling following myocardial infarction[41]. To understand their heterogeneity, we sub-clustered them using the multi-omic data and identified five sub-clusters across all myocardial infarction samples (Fig. 6j,k, Extended Data Fig. 11a–d and Supplementary Table 15). We observed that two clusters showed expression of resident myeloid cell markers42 (LYVE- and FOLR-expressing myeloid clusters), as well as a CCL18- and SPP1-expressing macrophage cluster and a monocyte and classical dendritic cell cluster (Fig. 6j and Extended Data Fig. 11b–d). We used an independent snRNA-seq dataset of three acute human myocardial infarction samples as reference for validation and found high concordance in terms of myeloid cell populations based on marker gene expression (Extended Data Fig. 11e). Cell proportion analysis revealed an increased abundance of a macrophage population defined by SPP1 expression in the ischaemic sample group, whereas CCL18+ macrophages were increased in fibrotic samples (Extended Data Fig. 11f). SPP1+ macrophages have been described in pulmonary fibrosis and COVID-1943,44, and recent work suggests a role of these cells in cardiac tissue remodelling in zebrafish[45]. We observed an upregulation of CD36 in the SPP1+ myeloid population; CD36 encodes a macrophage receptor known to be important for macrophage phagocytosis, binding to apoptotic and dead neutrophils and having a unique role in cardiac remodelling following myocardial infarction[46] (Extended Data Fig. 11b). Indeed, smFISH staining of SPP1+ macrophages suggests increased phagocytic activity, since multiple intracellular vacuoles could be observed (Extended Data Fig. 11g,h). Quantification of multiplex in situ hybridization of SPP1, TREM2 and CCR2 in human myocardial infarction tissue specimens revealed that approximately half of all TREM2-expressing myeloid cells also express SPP1, whereas CCR2+ myeloid cells where less frequent (Extended Data Fig. 11i). Cell-dependency analyses of myeloid cell states revealed a close interaction for two identified LYVE+ resident macrophage populations, whereas the disease-enriched SPP1+ macrophages predicted the presence of CCL18+ macrophages (Extended Data Fig. 11j,k).

Following acute myocardial infarction, an inflammatory response is triggered, resulting in tissue remodelling that can lead to heart failure47. It has been demonstrated that SPP1 itself can activate fibroblasts in vitro[48], highlighting the fibro-myeloid signalling interaction as a crucial driver of the cardiac remodelling process. To further gain insights about the spatial dependencies of the myeloid and fibroblasts states, we modelled their marker expression using the spatial transcriptomics data. We observed that the presence of SPP1+ macrophages better predicted all fibroblasts states compared to other myeloid cell states, with a higher importance for myofibroblasts within a spot and in the local neighbourhood (Fig. 6l and Extended Data Fig. 12a). Myofibroblast marker expression aligned with a gradient of expression of the markers of SPP1+ macrophages (Fig. 6m). This pattern was also recovered by our cell-type niche definition, in which the inflammatory niche 5 was surrounded by the fibrotic-rich niche 4 (Extended Data Fig. 12b), which we could confirm by a higher expression of SPP1+ macrophages and myofibroblast marker genes in niche 5 compared with niche 4 (Extended Data Fig. 12c). As our data pointed towards a clear spatial association of myeloid cells and fibroblasts, and spatially associated cells are presumably more likely to communicate with each other, we next used receptor–ligand interaction analysis to study their cellular crosstalk. We observed an overall complex myeloid–fibroblast interaction (Extended Data Fig. 12d), and detected distinct changes in crosstalk between SPP1+ macrophages and Fib2. This included increased PDGF-C, PDGF-D and THBS1 signalling in ischaemic versus myogenic samples and increased ADAM17 and TGFB1 in fibrotic versus myogenic samples (Extended Data Fig. 12e). Of note, we observed enhanced TGFβ1 signalling in ischaemic versus myogenic samples towards Fib3 (Extended Data Fig. 12f). To validate the spatial interaction of SPP1+ macrophages and Fib2, we performed RNA in situ hybridization on human cardiac tissues following myocardial infarction and could confirm the spatial interaction and enrichment of SPP1+ macrophages in an independent tissue cohort (n = 26 patients) using an orthogonal method (Fig. 6n and Extended Data Fig. 12g).

In summary, we have decoded cellular fibroblast and myeloid heterogeneity and spatial modelling of the fibro-myeloid cell states, revealing a unique interaction of SPP1+ macrophages with myofibroblasts across the different stages of human cardiac tissue remodelling.

Discussion

In multicellular organs, such as the human heart, normal cellular function and tissue homeostasis depend on the interaction between neighbouring individual cell types. Single-cell technologies can profile the molecular heterogeneity of the different cell types and changes that occur during disease. However, without spatial context it is unclear how these different cell types interact in space to coordinate tissue functions. Here we provide a comprehensive map of the human heart at early and late stages after myocardial infarction compared to control hearts (non-transplanted donor hearts) by integrating spatial transcriptomics with single-nucleus gene expression and chromatin accessibility data.

Our computational analyses enabled an increased resolution of spatial transcriptomics by estimating cell-type compositions for each location and by estimating pathway activities, mapping transcription factor binding activities, and projecting GWAS SNPs. These different layers of biological information enabled us to link the organization in human heart tissue specimens of different histomorphological regions, different time points after myocardial infarction and different individuals to cellular functions. Here we characterized inflammatory and fibrotic remodelling events that differentiated functional myocardium from ischaemic and chronically remodelled tissue. We explored the effects that these remodelling events had on cardiac architecture, specifically on the vasculature and the dependencies between fibroblasts and myeloid cells. Furthermore, we identified spatial enrichment of different functional states of myogenic regions in control, remote and border zones that were not captured by looking at cell-type compositions or histology only.

Analysis of the integrated snRNA-seq and snATAC-seq data identified different cell states and subtypes for cardiomyocytes, endothelial cells, fibroblasts and myeloid cells. We observed distinct cardiomyocyte cell states associated with spatial distribution, pathway activity and disease condition. Leveraging our multi-omic data, we inferred an eGRN and identified potential regulators of cardiomyocytes and fibroblasts, which were also reflected in spatial transcriptomics data. Our data revealed a distinct niche of the border zone surrounding the injured myocardium, with a sharp border between injured and uninjured cell types and were marked by a gradient of ANKRD1 and NPPB expression. Late-stage remodelling after myocardial infarction was driven by fibrosis, with fibroblast-to-myofibroblast differentiation in distinct tissue areas. Our data provide novel insights into myofibroblast differentiation in human hearts after myocardial infarction, with distinct gene expression and gene-regulatory programmes driving this process. In addition, we decoded the fibroblast myeloid cellular heterogeneity after human myocardial infarction and identified a distinct cellular dependency between myofibroblasts and activated phagocytic macrophages (SPP1+CD36+). The combination of spatial technologies with single-cell data represented an opportunity to study how cardiac cell states are influenced by their tissue microenvironment. The identified interactions between cell types largely reflect the spatial organization of the tissue and, although many other factors are involved, these interactions provide hypotheses for further analysis. Of note, we observed high levels of cell death in the ischaemic samples, as expected, and thus also higher levels of ambient RNA, which could introduce a bias in the analyses. Furthermore, we cannot exclude an overestimation of cardiomyocytes in our cell-type proportion analysis, since about 25% of adult human cardiomyocytes are binucleated[49], although multiple nuclei in a cell are reported to be transcriptionally homogenous[50].

We envision that our publicly available atlas will serve as a reference for future studies integrating single-cell genomics and epigenomics with spatial gene expression data of the human heart. Furthermore, we believe that our data will facilitate the understanding of spatial gene expression and gene-regulatory networks within the human myocardium and will be a resource for future studies that aim to understand the function of distinct cardiac cell types in cardiac homeostasis and disease.

Methods

Ethics

The local ethics committee of the Ruhr University Bochum in Bad Oeynhausen, the RWTH Aachen University, Utrecht University and WUSTL approved all human tissue protocols (no. 220-640, EK151/09, 12/387 and no. 201104172 respectively). Human myocardial tissue was collected from non-transplanted donor hearts, patients after myocardial infarction undergoing heart transplantation, implantation of a total artificial heart or left ventricular assist device (LVAD) implantation. The study met all criteria of the code of conduct for responsible use of human tissue that is used in the Netherlands. The collection of the human heart tissue was approved by the scientific advisory board of the biobank of the University Medical Center Utrecht, The Netherlands (protocol no. 12/387). All patients provided informed consent and the study was performed in accordance with the Declaration of Helsinki. Written informed consent for collection and biobanking of tissue samples was obtained prior to LVAD implantation.

Human tissue processing and screening

Heart tissues were sampled by the surgeon and immediately frozen in liquid nitrogen. Tissues were homogenized in liquid nitrogen and 7–10 mm3 pieces were embedded in OCT compound (Tissue-Tek) and frozen on dry-ice. Ten-micrometre tissue cryosections were stained with H&E and the appropriate tissue regions were selected for further processing. In total 52 human tissue samples were screened this way and evaluated by a cardiac pathologist. For RNA quality control we minced a 3 × 3 mm3 heart tissue piece in liquid nitrogen and isolated the RNA using Qiagen RNeasy Mini kit (Qiagen) using a proteinase K digestion step as suggested in RNeasy Fibrous Tissue Mini Kit (Qiagen, 74704). RNA integrity number (RIN) analysis (Agilent) was performed using Bioanalyzer RNA 6000 Nano kits (Agilent, No. 5067). RIN ranged from >2 to a maximum of 8.8.

Spatial gene expression assay

Frozen heart samples were embedded in OCT (Tissue-Tek) and cryosectioned (Thermo Cryostar). The 10-μm section was placed on the pre-chilled Optimization slides (Visium, 10X Genomics, PN-1000193) and the optimal lysis time was determined. The tissues were treated as recommended by 10X Genomics and the optimization procedure showed an optimal permeabilization time of 12 or 18 min of digestion and release of RNA from the tissue slide. Spatial gene expression slides (Visium, 10X Genomics, PN-1000187) were used for spatial transcriptomics following the Visium User Guides. Brightfield histological images were taken using a 10X objective on the Nikon Eclipse TiE and a Leica Aperio Versa 200 scanner. Stitching of the raw images was performed using the NIS-Elements software. Next generation sequencing libraries were prepared according to the Visium user guide. Libraries were loaded at 300 pM and sequenced on a NovaSeq 6000 System (Illumina) as recommended by 10X Genomics.

Single-nuclei isolation of human hearts

Single-nuclei isolation was performed as previously described51. Briefly, heart tissue was cut into small pieces (0.5 mm3) in a sterile petri dish on ice and transferred to a tissue homogenizer. Nuclei isolation buffer 0.5 ml (EZ lysis buffer, NUC101, Sigma-Aldrich) plus RNase inhibitor (Protector RNase Inhibitor, Roche) were added to the tissue, and 10-15 strokes with pestle A were applied followed by 10–15 strokes of pestle B. The nuclei were stained with DAPI and FANS sorted using a Sony SH800 to enrich the nuclei. Nuclei isolation from three acute myocardial infarction samples from the WUSTL biobank was performed as described52.

scRNA-seq

Nuclei suspensions with a concentration ranging from 400–1000 nuclei per μl were loaded into the chromium controller (10X, Genomics, PN-120223′) on a Single Cell B chip

(10X Genomics, PN-120262) and processed following the manufacturer' original protocol to generate single-cell gel beads in the emulsion. The sequencing library was generated using the Chromium Single cell 3′ reagent Kit v3 (10X, PN-1000092) and Chromium i7 Multiplex Kit (10X Genomics, PN-120262). Quality control for the constructed library was performed by Tape Station. Libraries were sequenced on NovaSeq targeting 50,000 reads per cell 2 × 150 paired-end kits using the following read length: 28 bp Read1 for cell barcode and UMI, 8-bp I7 index for sample index, and 91-bp Read2 for transcript.

sc-ATAC-seq

The remaining nuclei after processing for 3′ scRNA-seq assay were centrifuged at 500g at 4 °C for 5 min and resuspended in 10 μl of nuclei suspension buffer. After tagmentation the nuclei suspension was loaded on the Chromium Chip E (10X Genomics, PN-1000082) in the Chromium controller according to the manufacturer's protocol. The library was sequenced on an Illumina NovaSeq 6000 using the following read length: 50 bp Read1 for DNA fragments, 8 bp for i7 index for sample index, 16 bp i5 index for cell barcodes, and 50 bp Read2 for DNA fragments.

RNA in situ hybridization and image quantification

In situ hybridization was performed using formalin-fixed paraffin embedded tissue samples and the RNAscope Multiplex Detection KIT V2 (RNAscope, cat. no. 323100) and RNAscope 4-Plex Ancillary Kit (RNAscope, cat. no. 323120) following the manufacturer's protocol with minor modifications. The antigen retrieval was performed for 30 min at 99 °C in a water bath (VWR). Tissue pretreatment and washing was performed as suggested by the RNAscope staining protocol. The following probes were used for the RNAscope assay: Hs-CD163 cat. no. 417061-C1, Hs-CCR2 cat. no. 438221-C1, Hs-ANKRD1 cat. no. 524241-C1, Hs-POSTN cat. no. 575941-C1, Hs-Col15a1 cat. no. 484001-C2, Hs-Col1a1 cat. no. 401891-C2, Hs-PECAM1-O2 cat. no. 487381-C2, Hs-NPPB cat. no. 448511-C2, Hs-TREM2 cat. no. 420491-C2, Hs-SPP1 cat. no. 420101-C2, Hs-NPR3 cat. no. 431241-C3, Hs-POSTN cat. no. 409181-C3, Hs-SCARA5 cat. no. 574781-C3, Hs-TNNT2 cat. no. 518991-C3, Hs-SPP1 cat. no. 420101-C4 and Hs-NFE2L1 cat. no. 53850.

Nuclei quantification of H&E stained Visium slides

To quantify nuclei from the H&E staining, we used VistoSeg53, an automated MATLAB pipeline for image analysis. Using this pipeline, the individual TIFF files were used for nuclei segmentation using k-means colour-based segmentation in the image processing toolbox. Next, the binary images were refined with the refineVNS() function for accurate detection of the nuclei. Then a CSV and JSON file was generated that contained the metrics to reconstruct the spot grid to allow for nuclei quantification per 10X Visium detection spot. Counting of nuclei was performed with the countNuclei() function. The images were checked individually with the spotspotcheck() function. Code is available at http://research.libd.org/VistoSeg.

Animal model of myocardial infarction

Myocardial infarction was performed as previously described54. In brief, 12-week-old male and female C57Bl/6J Pdgfrb-creER;tdTomato mice were subjected to chronic left anterior descending artery ligation. The mice were anaesthetized using isoflurane (2–2.5%). The mice were injected 30 min before surgery with metamizole (200 µg g−1 body weight) subcutaneously. Then they were intubated and ventilated with oxygen using a mouse respirator (Harvard Apparatus). Before incision, we injected bupivacaine (2.5 µg g−1 body weight) subcutaneously and intercostally for local analgesia. Then a left thoracotomy was performed, and myocardial infarction was induced by ligature of the left anterior descending artery with 0/7 silk (Seraflex, IO05171Z). The ribs, muscle layer and skin incision were closed. Metamizole was administered for three days via drinking water (1.25 mg ml−1, 1% sucrose) post-surgery. All mice were housed under standardized conditions in the Animal Facility of the University Hospital Aachen (Germany). The operating procedure was in accordance with European legislation and approved by local German authorities (LANUV, reference no. 81-02.04.2017.A410.). Mice were euthanized at different time points (sham, 4 days, 7 days and 14 days). As control, hearts from sham-operated, age-matched mice were taken (2 sham female and 2 sham male mice).

Inducible fate-tracing experiments

For inducible fate tracing, male and female Pdgfrb-creER;tdTomato mice (8 weeks of age) received tamoxifen (3 mg intraperitoneally) 4 times followed by a washout period of 21 days and were then subjected to myocardial infarction surgery or sham (12 weeks of age) as described above and euthanized at 4 days, 7 days and 14 days after surgery.

Echocardiography

Left ventricular heart function was determined by echocardiography performed on a small-animal ultrasound imager (Vevo 3100 and MX550D transducer, FUJIFILM Visualsonics). Recordings of short and long cardiac axis were taken in B mode (2D real-time) using a 40 MHz transducer (MX550D). During the procedure, mice were anaesthetized with 1–2% isoflurane. Ejection fraction (EF) and global longitudinal strain (GLS) were recorded and analysed with the VevoLab Software. The Simpson method was used to assess EF. The GLS was measured in the B-mode of the long axis.

smFISH spot quantification and nuclear segmentation

Images for smFISH were exported in native Nicon format (.nd2). Images were split by channel using bfconvert55 for further processing. RNA spots were quantified using the command line version of Radial Symmetry-FISH (RS-FISH)56. The sigma parameter from RS-FISH, determining spot size, was set to 2.9 for all images. Threshold settings in RS-FISH were manually determined for each channel and were set to the following values for cardiomyocyte state analysis: channel 1 (TNNT2) = 0.0107, channel 2 (ANKRD1) = 0.005, channel 3 (NPPB) = 0.0066. To remove spot counts resulting from low signal, high background images, we removed spots with an intensity lower than the 25th percentile of the channel intensity distribution across all images and applied a minimum intensity threshold of 600. For the quantification of CD163+SPP1+ macrophages, while we were not able to perform full cell segmentation, we performed nuclear segmentation using Mesmer57 with pre-trained nuclear segmentation models to identify all detectable nuclei in each image based on DAPI staining. We subsequently assigned spots to the closest nuclei based on euclidean distance and classified cells as positive or negative for the different markers (POSTN, CD163 and SPP1). Cells with more than 2 spots for a given marker were considered positive for that marker.

Masson trichrome staining

Masson's trichrome staining was conducted using a ready-to-use kit (Trichrome Stain (Masson) Kit, HT15, Sigma-Aldrich) as described by the manufacturer.

Antibodies and immunofluorescence staining

Heart tissues were fixed in 4% formalin for 4 h at room temperature and then embedded in paraffin. For staining slides were blocked in 5% donkey serum followed by 1 h of incubation with the primary antibody, washing 3 times for 5 min in PBS, and subsequent incubation of the secondary antibodies for 45 min. Following DAPI (4′,6′-diamidino-2-phenylindole) staining (Roche, 1:10.000) the slides were mounted with ProLong Gold (Invitrogen, cat. no. P10144). The following antibodies were used: anti-ACTA2(aSMA)-Cy3 (C6198,1:250, Sigma-Aldrich), anti-SEMA3G (HPA001761, 1:100, Sigma-Aldrich), AF647 donkey anti-rabbit (1:200, Jackson Immuno Research).

Confocal imaging

Acquisition of images was performed using a Nikon A1R confocal microscope using 40× and 60× objectives (Nikon). Image processing was performed using the Nikon Software or ImageJ58.

Generation of a human PDGFRB + cardiac cell line

PDGFRB+ cells were isolated from a 69-year-old male patient, undergoing left ventricular assist device surgery. To generate a single-cell suspension, the tissue was homogenized in a gentleMACS dissociator (Miltenyi) and digested with liberase (200 µg ml−1, Roche cat. no. 5401020001) and DNase (60 U ml−1), for 30 min at 37 °C. After filtering the cell suspension (70 µm mesh), cells were stained in two steps using a specific PDGFRB antibody (R&D cat. no. MAB1263 antibody, dilution 1:100) followed by Anti-Mouse IgG1-MicroBeads solution (Miltenyi, cat. no. 130-047-102). Following MACS isolation, cells were cultured in DMEM media (Thermo Fisher cat. no. 31885) for 20 days and immortalized using SV40-LT and HTERT. Retroviral particles were produced by transient transfection of HEK293T cells using TransIT-LT (Mirus). Two types of amphotropic particles were generated by co-transfection of plasmids pBABE-puro-SV40-LT (Addgene #13970) or xlox-dNGFR-TERT (Addgene #69805) in combination with a packaging plasmid pUMVC (Addgene #8449) and a pseudotyping plasmid pMD2.G (Addgene #12259). Retroviral particles were 100x concentrated using Retro-X concentrator (Clontech) 48 h post-transfection. Cell transduction was performed by incubating the target cells with serial dilutions of the retroviral supernatants (1:1 mix of concentrated particles containing SV40-LT or rather hTERT) for 48 h. Subsequently at 72 h after transduction, the transduced PDGFRb+ cells were selected with 2 µg ml−1 puromycin for 7 days.

Lentiviral overexpression of RUNX1

The human cDNA of RUNX1 was PCR amplified using the primer sequences 5′- atgcgtatccccgtagatgcc −3′ and 5′- tcagtagggcctccacacgg −3′. Restriction sites and N-terminal 1xHA-Tag were introduced into the PCR product using the primer 5′- cactcgaggccaccatgtacccatacgatgttccagattacgctcgtatccccgtagatgcc −3′ and 5′- acggaattctcagtagggcctccacac −3′. Subsequently, the PCR product was digested with XhoI and EcoRI and cloned into pMIG (pMIG was a gift from W. Hahn) (Addgene plasmid #9044 ;http://n2t.net/addgene:9044; RRID:Addgene_9044). Retroviral particles were produced by transient transfection in combination with packaging plasmid pUMVC (pUMVC was a gift from B. Weinberg (Addgene plasmid #8449)) and pseudotyping plasmid pMD2.G (pMD2.G was a gift from D. Trono (Addgene plasmid #12259 ; http://n2t.net/addgene:12259; RRID:Addgene_12259)) using TransIT-LT (Mirus). Viral supernatants were collected at 48–72 h post-transfection, clarified by centrifugation, supplemented with 10% FCS and Polybrene (Sigma-Aldrich, final concentration of 8 µg ml−1) and filtered with a 0.45-µm PES filter membrane (Millipore; SLHP033RS). Cell transduction was performed by incubating the PDGFB+ cells with viral supernatants for 48 h. eGFP-expressing single cells were sorted with a SH800 Cell Sorter.

Quantitative PCR with reverse transcription

Cell pellets were collected and washed with PBS followed by RNA extraction using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Two-hundred nanograms total RNA was reverse transcribed with High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and quantitative PCR with reverse transcription was carried out as described29 Data were analysed using the 2Ct method. The primers used are listed in Supplementary Table 18.

Preprocessing of snRNA-seq, snATAC-seq and spatial transcriptome data

For snRNA-seq data, CellRanger software (v6.0.2) was used to perform the alignment with default options. Since the input consists of nuclei, we enabled the option '–include-introns' to include intronic reads. For snATAC-seq data, the CellRanger ATAC pipeline (v2.2.0) was used with the default settings. For spatial transcriptome data, the SpaceRanger software (v1.3.2) was used to pre-process the sequencing data. The option '–reorient-images' was enabled to allow for automatic image alignment. hg38 was used as the reference genome for human data alignment.

snRNA-seq data processing

To identify the major lineages representative of all of our specimens, we created a single-nuclei atlas analysing and integrating each snRNA-seq dataset using Seurat59 (v4.0.1).

Each dataset went through identical quality control processing. We discarded nuclei (1) in the top 1% in terms of the number of genes, (2) with less than 300 genes and less than 500 UMIs, (3) with more than 5% of mitochondrial gene expression, and (4) doublets as estimated using scDblFinder (v1.4.0)60 with default parameters. Count matrices were log-normalized for downstream analyses using a scaling factor of 10,000. We calculated a dissociation score for each cell using Seurat's module score functions with a gene set provided by O'Flanagan et. al.61 and discarded the nuclei that belonged to the top 1%. To generate an integrated atlas of all samples, log-normalized expression matrices were merged and dimensionality reduction was performed on the collection of the top 3,000 most variable genes that were shared with most of the samples using principal component analysis (PCA). To select the collection of shared variable genes between samples, first we estimated the top 3,000 most variable genes per sample and then selected the top 3,000 most-recurrent genes from them across all samples. PCA correction was performed with harmony62 (v1.0) using as covariates the patient, sample, and batch labels. A shared nearest neighbour (SNN) graph was built with the first 30 principal components using Seurat's FindNeighbors, and the cells were clustered with a Louvain algorithm with FindClusters. A high resolution (1) was selected to generate a large collection of nuclei clusters to capture representative major cell lineages, even if present in low proportions. Cluster markers were identified with Wilcoxon tests as implemented in Seurat's FindAllMarkers function. Final assignment of cells to major cell lineages was based on literature marker genes. We filtered out small clusters (median number of nuclei across filtered clusters = 269) with low gene count distributions (median counts across filtered clusters = 756) or feature recovery (median number of genes across filtered clusters = 695), with marker genes that could not be assigned to known cell types of the heart. To visualize all nuclei in a two-dimensional embedding, a UMAP was created with Seurat's RunUMAP function using the first 30 principal components of harmony's PCA correction embedding. Major cell-type markers were estimated by performing differential expression analysis of cell-type and patient-specific pseudo-bulk profiles. Pseudo-bulk profiles were calculated by summing up the counts of all cells belonging to the same cell type and patient. Profiles coming from less than 10 cells or profiles from which the maximum gene expression was of less than 1,000 counts per library were discarded. Differentially expressed genes were calculated by fitting a quasi-likelihood negative binomial generalized log-linear model as implemented in edgeR (v3.32.1)63 (false discovery rate (FDR) < 0.15). Each cell type was compared against the rest.

Comparison with independent healthy and ischaemic human heart cell atlases

We compared our generated atlas with another reference human single-nuclei RNA-seq atlas4 at the molecular and compositional level. The counts matrix was downloaded directly from https://www.heartcellatlas.org and we selected the data coming from single-nuclei and left ventricle samples. Nuclei with fewer than 200 genes, and genes expressed in less than 3 nuclei were excluded. Log normalization with a scaling factor of 10,000 was performed with scanpy's64 (v1.7.0) normalize_total function.

To evaluate our major cell-type annotation, we calculated the enrichment of the top 200 marker genes based on log fold change of each cell type defined in the reference atlas in the list of the top 200 marker genes of each of our defined cell types with hypergeometric tests. Marker genes of the reference atlas were calculated with Wilcoxon tests as implemented in scanpy's 64 (v1.7.0) rank_genes_groups (adj. P < 0.01). Each cell type was compared against the rest. To evaluate the compositional stability of our control samples, we calculated the Pearson correlation between the median proportion of each shared cell type of the reference atlas and our control, border zone, and remote zone samples. Similarly, we compared our atlas to an independent collection of human heart nuclei derived from three ischaemic specimens. First, we analysed and integrated the smaller collection of samples using identical procedures as the ones used in our provided atlas. After nuclei clustering, we assigned each cluster to a cell type using literature markers. Cell-type markers were calculated with Wilcoxon tests (adj. P < 0.01) and the top 200 genes based on log fold change were selected. Marker overlap and compositional stability comparison with ischaemic specimens from our atlas were performed as described previously.

snATAC-seq data processing

To control the data quality, the fragment files were used as input for the package ArchR (v1.0.1)65, and low-quality cells were filtered out based on transcription start site (TSS) enrichment (> 4) and the number of unique fragments (> 3,000 and < 100,000). Doublets were identified and removed by using the functions addDoubletScores and filterDoublets from ArchR with default settings. Next, peaks were identified by using the function addReproduciblePeakSet for each sample. All peaks were merged to create a union peak set of which each peak was annotated as distal, promoter, exonic and intronic. A count matrix was constructed with the function addPeakMatrix. For dimensionality reduction, the method scOpen (v1.0.0)35 was used to generate a low-dimensional matrix of the cells. The algorithm Harmony62 was applied to correct the batch effects and integrate the data and UMAP was used to generate a 2D embedding for visualization. Cells were clustered using the Leiden algorithm with a resolution of 1. To annotate the clusters, a gene activity score matrix was created using the function addGeneScoreMatrix and marker genes were detected for each cluster using the function getMarkerFeatures. The same markers from snRNA-seq data were used to annotate the clusters.

Comparison between snRNA-seq and snATAC-seq data

The Seurat66 label-transferring approach was used to compare the annotation of snRNA-seq and snATAC-seq. To do so, the snRNA-seq data were used as reference and the function FindTransferAnchors was applied to identify a set of anchors using gene expression from snRNA-seq and gene activity score from snATAC-seq. Next, the cell labels from snRNA-seq were transferred to snATAC-seq by running the function TransferData. An adjusted rand index was calculated to evaluate the agreement between annotated and predicted cell labels for snATAC-seq data.

Cell-type-specific transcription factor binding and regulon activity

To estimate transcription factor binding activity for each major cell type identified from snATAC-seq data, we first aggregated the reads within each cell type and created a pseudo-bulk profile. Next, we used MACS2 (v2.2.7)67 to perform peak calling and removed the peaks from chrY, mitochondrial and unassembled 'random' contigs. We then predicted the transcription factor binding sites and estimated transcription factor binding activity using HINT-ATAC (v0.13.2)68 based on the JASPAR2020 database69. We linked the transcription factor binding sites to the nearest genes to create a cell-type-specific transcription factor–gene interaction. The number of ATAC-seq reads in the region with 100 bp up-stream and downstream of the the transcription factor binding site were used to indicate how strong the interaction was: each transcription factor–gene interaction was weighted as the ratio between the number of ATAC-seq reads around the transcription factor binding site associated with that gene and the maximum number of reads observed in any binding site of the transcription factor. All interactions with a weight larger than 0.3 were considered in downstream analysis. This generated weighted and filtered cell-type-specific regulons. To infer a transcription factor regulon activity score, we estimated the mean expression of the target genes in each cell-type-specific regulon. Cell-type pseudo-bulk profiles were filtered to contain only genes with at least 10 counts in 5% of the samples, before the estimation of normalized weighted means using decoupleR's70 (v1.1.0) wmean function with 1,000 permutations. Regulon activities were standardized and correlated with transcription factor binding activities using Spearman correlations. The minimum correlation of 0.5 was used as threshold and the top 5 transcription factors per cell type were selected for visualization.

Cell-type-specific GWAS signal enrichment

GWAS summary statistics for 4 MRI based left ventricle function parameters12 were downloaded from the Cardiovascular Disease Knowledge Portal (https://cvd.hugeamp.org/). For each phenotype, GWAS summary statistics were clumped with Plink (v1.9)71 to identify index SNPs (clump-p1 = 0.0001, clump-kb = 250, clump-r2 = 0.5) using the European samples from 1000 Genomes as a reference population.

Next, we lifted over the coordinates of index SNPs from hg19 to hg38 using the LiftOver tools. For each major cell type, we generated an average chromatin accessibility profile by using snATAC-seq data from all cells. The cell-type-specific GWAS signal enrichment was performed using gchromVAR (v0.3.2)72 and enrichment scores were normalized to z-scores. P-values were calculated based on the z-scores and were corrected by the Benjamini–Hochberg method.

Cell-type-specific integration of snATAC-seq and snRNA-seq data and sub-clustering

For each major cell type that was recovered by both snATAC-seq and snRNA-seq, we aimed to identify sub-clusters spanning multiple samples and modalities. To do so, we devised a multi-step approach to integrate and cluster the data by controlling quality from sample-, cell-type- and modality-specific aspects.

(1) To minimize the sample-specific effects, we only considered samples with a minimum number of cells in both snATAC-seq and snRNA-seq: for cardiomyocytes and endothelial cells (n_cells_ATAC > 300 and n_cells_RNA > 400); for fibroblasts (n_cells_ATAC > 100 and n_cells_RNA > 400); and for myeloid (n_cells_ATAC > 50 and n_cells_RNA > 200). This step controls for samples with low recovery of cells in a particular modality.

(2) To further filter cell-type-specific low-quality cells from snRNA-seq and snATAC-seq data, we integrated the samples as selected in step 1 using Harmony to correct batch effects from patients and regions based on PCA space (30 dimensions) for snRNA-seq and LSI space (30 dimensions) for snATAC-seq data. We then clustered the cells using Seurat (resolution = 0.4) for each modality independently. We next excluded the clusters that were (i) enriched in a single sample; (ii) showed a lower

data quality; (iii) showed a higher doublet score compared with others. Specifically, for cardiomyocytes, we removed 3 clusters from snATAC-seq data: 2 clusters (481 cells) were enriched in a single sample and another cluster (171 cells) showed a low number of unique fragments (average = 8,102). For fibroblasts, we removed 1 cluster (49 cells) from snATAC-seq (98% of cells from a single sample) and 1 cluster (1,172 cells) from snRNA-seq (average doublet score of 0.12). This step controls for cell type and modality-specific low-quality cells.

(3) We next integrated the cells from snATAC-seq and snRNA-seq data. To this end, we used the gene activity score matrix of snATAC-seq estimated by ArchR and the gene expression data from snRNA-seq data as input for canonical component analysis by Seurat. The integrated data were projected into a PCA space (30 dimensions) and Harmony was used to correct the batch effects from samples and modalities. This step generated a co-embedded and batch-corrected dataset composed of cells from snRNA-seq and snATAC-seq samples.

(4) For each major cell type, we defined the sub-clusters based on the co-embedded data using the Seurat (resolution = 0.9 or 1). Marker genes were identified by using the function FindAllMarkers. We next filtered clusters that were mainly driven by a single sample or modality. Finally, we merged and annotated the clusters based on the markers. The final statistics of the sub-clustering results for each major cell type were provided in Supplementary Table 16.

Analysis of snRNA-seq data from mouse fibroblasts

Cellranger mkfastq and count functions (version v6.0.2) with default parameters were used for demultiplexing and aligning the reads, respectively. Reads were aligned to the mouse reference transcriptome (mm10, Version=2020-A). Prior to alignment, reads for tdTomato were added to the reference. Quantified counts from each sample were aggregated and cells with counts <1,500 and >20,000 were filtered out. Further, cells with >5% reads mapped to mitochondrial genes, as well as cells with <500 genes were removed. Scrublet73 was used to detect potential doublets and only the resulting 40,495 cells with <0.2 scrublet score were kept for further analyses. The highly_variable_genes() function with seurat_v3 flavour implemented in Scanpy (version 1.8.1) was used to obtain the top 2,000 most highly variable genes. Count data was log-normalized using sc.pp.normalize_total(target_sum=1e4) followed by sc.pp.log1p(). The data was subset to the 2,000 genes, unwanted sources of variation from n_umi and mito_fraction were regressed out using sc.pp.regress_out(), and the top 30 principal components were estimated using sc.tl.pca(). Harmony was then used to account for large differences across samples using 'sample' as the batch indicator. Network neighbourhood graph was constructed using the function sc.pp.neighbors() with 30 adjusted principal components, cosine distance and n_neighbors = 10. Leiden clustering with resolution 1.0 was used to cluster the cells into 17 clusters. Marker genes were identified using the Wilcoxon test implemented in sc.tl.rank_genes_groups() function in Scanpy. Clusters were manually annotated using the marker genes. We next cleaned up the data by removing clusters with low data quality and re-clustered the data with resolution of 0.2. To annotate the cells, we used the label transfer approach from Seurat based on the sub-clustering results from human fibroblasts.

Gene-regulatory network inference for cardiomyocytes and fibroblasts

We inferred an eGRN for cardiomyocytes and fibroblasts using a multi-step approach including modality pairing, transcription factors and genes selection, and network construction.

(1) We first paired the cells between snATAC-seq and snRNA-seq based on the previously described co-embedding space using an optimal matching approach74. This method returns a matching of a snATAC-seq cell to a unique cell in snRNA-seq. Next, we produced a diffusion map75 and created trajectories in this space using the function addTrajectory from ArchR (v1.0.1)65. For cardiomyocytes, we inferred a trajectory from clusters vCM1, vCM2 and vCM3, where vCM1 were considered as roots and vCM3 as the terminal state. For fibroblasts, we built a trajectory with SCARA5+ fibroblasts as root and myofibroblasts as terminal state.

(2) Next, we predicted a single-cell-specific transcription factor binding activity score using the R package chromVAR (v1.16)76 from the snATAC-seq data based on motif from the JASPAR2020 database69. In contrast to HINT-ATAC, chromVAR provides transcription factor activity scores at single-cell level. We next selected transcription factors that display concordant binding activity (snATAC-seq) and its gene expression (snRNA-seq) (Pearson correlation > 0.1). This analysis identified 65 transcription factors for cardiomyocytes and 44 transcription factors for fibroblasts. We considered these transcription factors to be potential regulators. We sorted the transcription factors along the trajectory as defined in step 1 and assigned a pseudotime label to each transcription factor. Next, we selected highly variable genes using the snRNA-seq data along the trajectories as described65. We kept the top 10% variable genes and considered them as potential transcription factor targets.

(3) To associate regulators with targets (that is, transcription factors with genes), we explored the correlation of peak accessibility and gene expression to identify peak-

to-gene links. Specifically, for each gene, we consider peaks that are within 125 kb on either side of the transcription start sites, while excluding the promoter regions. This analysis generated a list of enhancer-to-promoter links. We only considered significantly correlated links (FDR < 0.0001) with a positive correlation as before65. Finally, we associated a transcription factor with a target gene if this gene was linked to an enhancer and this enhancer was predicted to be found by this transcription factor.

(4) To build a quantitative transcription factor–gene-regulatory network, we estimated the correlation of the transcription factor binding activity from snATAC-seq and target gene expression from snRNA-seq data and only considered those interactions with Pearson correlation >0.4. We visualized the network based on a force-layout, which places transcription factors (or target genes) with similar interactions close together. We coloured transcription factor nodes in the networks using the assigned pseudotime labels as inferred in step 2. To characterize the importance of transcription factors, we computed two measures: node betweenness (denoted by b)77 and pagerank (denoted by p)78. A final importance score for transcription factor i was calculated as:

$$\text{Importance}\, i = \sqrt{(b_i - min(b))^2 + (p_i - min(p))^2}$$

(5) Finally, to map the inferred GRN into spatial transcriptomics data, we used the target genes for each transcription factor and calculated a module score by using the function AddModuleScore from Seurat (v4.1.0).

Characterization of spatial transcriptomics datasets

Single-slide processing

Filtered feature-barcode expression matrices from SpaceRanger (v1.3.2) were used as initial input for the spatial transcriptomics analysis using Seurat (v4.0.1). Spots with less than 300 measured genes and less than 500 UMIs were filtered out. Ribosomal and mitochondrial genes were excluded from this analysis. Individual count matrices were normalized with sctransform79, and additional log-normalized (size factor = 10,000) and scaled matrices were calculated for comparative analyses using default settings.

Cell-type compositions were calculated for each spot using cell2location80 (v0.05). Reference expression signatures of major cell types were estimated using regularized negative binomial regressions and our integrated snRNA-seq atlas. We fitted a model in six downsampled iterations of our snRNA-seq atlas (30%) and generated a final reference matrix by taking the mean estimation. Each slide was later deconvoluted using hierarchical bayesian models as implemented in run_cell2location. We provided the following hyperparameters: 8 cells per spot, 4 factors per spot, and 2 combinations per spot. Additionally, for each spot we calculated cell-type proportions using the cell-type-specific abundance estimations. Cell-type compositions of the complete slide were calculated adding the estimated number of cells of each type across all spots. To compare the stability of estimated cell compositions between our different data modalities, we calculated Spearman correlations between the estimated cell type proportions of each slide and the observed cell type proportions in its corresponding snRNA-seq and snATAC-seq dataset.

Estimation of functional information from spatial data

For each spot, we estimated signalling pathway activities with PROGENy's81,82 (v1.12.0) model matrix using the top 1,000 genes of each transcriptional footprint and the sctransform normalized data. Spatially variable genes were calculated with SPARKX83 (v1.1.1) using log-normalized data (FDR < 0.001). To obtain overrepresented biological processes from each list of spatially variable genes, we performed hypergeometric tests using the set of canonical pathways provided by MSigDB84 (FDR < 0.05).

Estimation of cell death molecular footprints from spatial data

To associate the differences in nuclei capture in snRNA-seq between the different samples to cell death processes, we leveraged the information from spatial transcriptomics to estimate the general expression of genes associated to cell death for each sample. For each unfiltered slide we estimated per spot the normalized gene expression of BioCarta's84 'death pathway' and Reactome's85 'regulated necrosis pathway' using the decoupleR (v1.1.0) wmean method and the sctransform normalized data. To have a final pathway score per slide, we calculated for each slide the mean 'pathway expression' across all spots.

Mapping transcription factor binding activity and GWAS enrichment to spatial data

To visualize the transcription factor binding activity estimated from snATAC-seq data in space, we used the estimated cell type proportion calculated from cell2location scores for mapping. Specifically, for each spot i and transcription factor j, we calculated the transcription factor binding activity as follows:

$$\mathrm{ACT}_{ij} = \sum_{k=1}^{K} \mathrm{Proportion}_{ik} \times \mathrm{ACT}'_{kj},$$

where Proportionik is the estimated proportion of cell type k, K is the number of cell types, and ACT'kj is the binding activity of transcription factor j in cell type k from snATAC-seq data. An equivalent approach was used to map GWAS scores into space.

Cell-state spatial mapping

To map the functional states of each cell type into spatial locations, we leveraged the deconvolution results of each slide and the set of differentially expressed genes of each recovered cell state. Given the continuous nature of cell states, we assumed that the collection of up and downregulated genes of a cell state represented its transcriptional fingerprint and could be summarized in a continuous score in locations where we could reliably identify the major cell type from which the state was derived. For a given major cell type of interest k, we identified spots where its inferred abundance was of at least 10%. To estimate state scores associated with cell type k, we used decoupleR's (v1.1.0) normalized weighted mean method (wmean) and the set of the upregulated genes of each state defined with snRNA-seq and snATAC-seq (log fold change > 0; Wilcoxon tests, FDR < 0.05). The log fold change of each selected gene was used as the weight in the wmean function.

Analysis of ion channel-related genes

We related the expression of ion channel-related gene sets to the different cardiomyocyte cell states and their location in spatial transcriptomics. First we selected two different gene sets containing ion channel-related genes: (1) Reactome's85 'ion channel transport' and a curated list of transmural ion channels from Grant et al.86. Gene sets are provided in Supplementary Table 17. First, we calculated gene set scores for each spatial transcriptomics spot using decoupleR's wmean function. Then we correlated these gene set scores to the spatial mapping of cardiomyocyte cell states in regions where we observed at least 10% of cardiomyocytes. Additionally, we evaluated if any of the genes belonging to these gene sets were differentially expressed between the vCM1 and stressed vCM3 population using Wilcoxon tests as implemented in scran's (v1.18.5) findMarkers function (area under the curve (AUC) < 0.4, AUC > 0.6, FDR < 0.05).

Spatial map of cell dependencies

We used MISTy's87 implementation in mistyR (v1.2.1) to estimate the importance of the abundance of each major cell type in explaining the abundance of the other major cell types. Cell-type cell2location estimations of all slides were modelled in a multi-view model using three different spatial contexts: (1) an intrinsic view that measures the relationships between the deconvolution estimations within a spot, (2) a juxta view that sums the observed deconvolution estimations of immediate neighbours (largest distance threshold = 5), and (3) a para view that weights the deconvolution estimations of more distant neighbours of each cell type (effective radius = 15 spots). The aggregated estimated standardized importances (median) of each view of all slides were interpreted as cell-type dependencies in different spatial contexts, such as colocalization or mutual exclusion. Nevertheless, the reported interactions did not imply any causal relation.

Before aggregation, we excluded the importances of all predictors of target cell types whose R2 was less than 10% for each slide.

To associate tissue structures with tissue functions, we fitted a MISTy model to explain the distribution of PROGENy's pathway activities standardized scores. The multi-view model consisted of the following predictors: (1) an intrinsic view to model pathway crosstalk within a spot, (2) a juxta view to model pathway crosstalk between neighbouring spots (largest distance threshold = 5), (3) a para view estimating pathway relations in larger tissue structures (effective radius = 15), (4) an intrinsic view and (5) a para view containing cell2location estimations (effective radius = 15). These last two views model explicitly the relations between cell-type compositions of spots and pathway activities. Cycling cells and TNF were not included in the described analyses. Before aggregation, we excluded the importances of all predictors of target pathway activities whose R2 was less than 10% for each slide.

Niche definitions from spatial transcriptomics data

To identify groups of spots in the different samples that shared similar cell-type compositions, we transformed the estimated cell-type proportions of each spatial transcriptomics spot and slide into isometric log ratios (ILR)88, and clustered spots into groups. These niches represent groups of spots that are similar in cell composition and represent potential shared structural building blocks of our different slides; we refer to these groups of spots as cell-type niches. Louvain clustering of spots was performed by first creating a shared nearest neighbour graph with k different number of neighbours (10, 20, 50) using scran's89 (v1.18.5) buildSNNGraph function. Then, we estimated the clustering resolution that maximized the mean silhouette score of each cluster. We assigned overrepresented cell types in each structure by comparing the distribution of cell-type compositions within a cell-type niche versus the rest using Wilcoxon tests (FDR < 0.05). We tested if a given cell state was more representative of a cell-type niche by performing Wilcoxon tests between each niche and the rest (FDR < 0.05). Only positive state scores were considered in this analysis.

Additionally, to complement the repertoire of niches identified with cell-type compositions, we integrated and clustered the Visium spots of all slides using their log-normalized gene expression. We called these clusters molecular niches. Integration and clustering of spots was performed with the same methodology as the one used to create the snRNA-Seq atlas. A low resolution was used (0.2) to have a similar number of molecular niches as cell-type niches. Cell-type and cell-state enrichment was performed as mentioned before.

Differential expression analysis of molecular niches enriched with cardiomyocytes

Differential expression analysis between molecular niches enriched in cardiomyocytes (niche 0, niche 1, niche 3) was performed using the log-transformed expression of all spots belonging to a given niche. Wilcoxon tests were performed with scran's89 (v1.18.5) findMarkers function. Genes with a summary AUC >0.55 and FDR <0.05 were considered upregulated genes.

Differential molecular profiles of the molecular niche 10 enriched with capillary endothelial cells

Differential expression analysis between ischaemic, fibrotic and myogenic-enriched spatial transcriptomic spots was performed with Wilcoxon tests as implemented in scran's89 (v1.18.5) findMarkers function. To obtain overrepresented biological processes from upregulated genes, we performed hypergeometric tests using the set of hallmark pathways provided by MSigDB84. Normalized PROGENy's pathway activities for each spot were calculated using decoupleR's wsum method with 100 permutations on log-transformed data. Mean normalized pathway scores were calculated per slide and comparisons between groups were performed with Wilcoxon tests. Reported P-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

General differences in tissue organization

We annotated the different spatial transcriptomic slides into three groups based on histological differences with the help of pathologists: myogenic-enriched, fibrotic-enriched and ischaemic-enriched. A general comparison of the sampled patient specimens was performed at the compositional and molecular level.

Hierarchical clustering, with euclidean distances and Ward's algorithm, was used to cluster the pseudo-bulk profiles of the spatial transcriptomics datasets (replicates where merged, n = 27). Genes with less than 100 counts in 85% of the sample size were excluded for this analysis. Log normalization (scale factor = 10,000) was performed. To visualize the general molecular differences between our samples, log-normalized pseudo-bulk profiles of the spatial transcriptomics datasets were projected in an UMAP embedding.

To identify compositional differences between our sample groups, we compared cell-type and niche compositions. To identify cell-type composition changes associated to the sample groups, mean cell-type compositions across single-cell and spatial datasets were compared with Kruskal–Wallis tests (FDR < 0.1). Pairwise comparisons of sample groups were performed with the Wilcoxon test. Additionally, to test which cell-type and molecular niches had different distributions between our group samples, we performed Kruskal–Wallis tests over the compositions of cell-type or molecular niches (FDR < 0.1). Additional pairwise comparisons were performed with Wilcoxon tests (P-values adjusted with Benjamini–Hochberg procedure). For this, we only consider slides where no single niche represents more than 80% of the spots. Also, we only consider niches representing more than 1% of the composition of at least 5 slides.

To identify differences between the structurally similar tissues captured in the myogenic-enriched group, we separated the samples into remote, border, and control zones and repeated the niche composition comparison described previously.

To identify patterns of tissue organization associated with a sample group, we tested if differential cell dependencies were captured by the MISTy models used to predict cell-type abundance (see 'Spatial map of cell dependencies'). First, we filtered the standardized importance matrices of each sample's MISTy model fitted to predict the abundance of major cell-types to contain only the values of target cell types predicted with an R2 greater than 0.05. Then, for each slide we created a spatial dependency vector where each element contains the importance of each possible pair of target and predicted cell types. Finally, we tested which cell interactions had higher importances in one of the sample groups compared to the rest using Wilcoxon tests (FDR < 0.25). To prioritize interactions, we only performed pairwise comparisons between sample groups for cell-type dependencies from which the maximum median importance across all groups was greater than 0.

Estimation of the effects of the spatial context on gene expression

We used mistyR (v1.2.1) to find the associations between the tissue organization and the spatial distribution of stressed cardiomyocytes and the different endothelial, myeloid and fibroblast cell states. We hypothesized that the distribution of specific cell states in the spatial transcriptomics slides could be modelled by the cell-type composition or cell-state presence of individual spots and their neighbourhood.

For a given collection of cell states of interest, we first defined regions of interest in every single slide as the collection of spots where the inferred abundance of the cell type from which the cell state was derived was at least 10%. These regions limit the target spots used in the MISTy model, however the whole slide is used to spatially contextualise the predictors. We used as predictors the abundances of cell types estimated with cell2location or cell states scores. To account only for the effects of the activation of a cell state, the state scores of predictor cell states were masked to 0 whenever their score was lower than 0. In all models we included two classes of spatially contextualized predictive views: an intrinsic (intra) and a local neighbourhood view (para, effective radius = 5).

Specifically we fitted the following models to answer four questions:

(1) What are the main cell types whose abundance within a spot or in the local neighbourhood predict the stressed vCM3?

vCM3 ~ intra(cell-type abundance) + para(cell-type abundance)

(2) What are the main cell types whose abundance within a spot or in the local neighbourhood predict the endothelial subtypes? How do the different subtypes relate to each other?

ECsubtypes ~ intra(ECsubtypes) + para(ECsubtypes) + intra(cell-type abundance) + para(cell-type abundance)

(3) What are the myeloid cell states within a spot or in the local neighbourhood that better predict fibroblasts cell states? How do fibroblasts cell states relate to each other?

FibroblastStates~intra(FibroblastStates)+para(FibroblastStates)+intra(MyeloidStates) + para(MyeloidStates)

(4) What are the main cell types whose abundance within a spot or in the local neighbourhood predict the myeloid cell states? How do the different states relate to each other?

MyeloidStates ~ intra(MyeloidStates) + para(MyeloidStates) + intra(cell-type abundance) + para(cell-type abundance)

Specific view importances were compared between patient groups as described previously with an R2 filter of 0.1.

Cell–cell communication analysis

To estimate ligand–receptor interactions between the sub-populations of fibroblasts and myeloid cells, we extracted gene expression matrix from the integrated snRNA-seq and snATAC-seq data for each sample group (that is, myogenic, ischaemic and fibrotic) and combined the matrices from all sub-populations. We next used LIANA (v0.0.9)90, a framework that compiles the results of state-of-the-art cell communication inference methods, to infer ligand–receptor interactions. We focused on the CellPhoneDB91 ligand–receptor method with Omnipath's ligand–receptor database92 implemented in LIANA90. This was done by combining snRNA-seq samples of myogenic, ischaemic and fibrotic groups and subsetting only the fibroblasts and myeloid cells sub-states. Next, we used CrossTalker (v1.3.1)93 to find changes in cell–cell communication by contrasting ligand–receptor interactions predicted in myogenic vs. ischaemic samples and myogenic vs. fibrotic samples. The interactions considered by CrossTalkeR were obtained by filtering the output of LIANA90 ($P > 0.01$).

Visualization, statistics, and reproducibility

In data represented as box plots (Figs. 2f, 4c,d,m,o, 5b and 6d,n) the middle line corresponds to the median, the lower and upper hinges describe the first and third quartiles, the upper whisker extends from the hinge to the largest value no further than 1.5 × inter-quartile range (IQR) from the hinge and the lower whisker extends from the hinge to the smallest value at most 1.5 × IQR of the hinge, and data beyond the end of the whiskers are outlying points that are plotted individually. In Figs. 4b and 5b,k, Colours refer to gene-weighted kernel density as estimated by using R package Nebulosa94. All reported P-values based on multi-comparison tests were corrected using the Benjamini–Hochberg method95. The depicted immunofluorescence micrographs are representative (Figs. 4c and 6n). The number of samples for each group was chosen on the basis of the expected levels of variation and consistency. The depicted RNAscope, immunofluorescence micrographs are representative and were performed at least twice, and all repeats were successful. Fig. 1a contains a panel from BioRender.com.

Affiliations

[1] Institute of Experimental Medicine and Systems Biology, Medical Faculty, RWTH Aachen University, Aachen, Germany.

[2] Division of Nephrology and Clinical Immunology, Medical Faculty, RWTH Aachen University, Aachen, Germany.

[3] Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University and Heidelberg University Hospital, Bioquant, Heidelberg, Germany.

[4] Department of Urology and Pediatric Urology, RWTH Aachen University, Aachen, Germany. [5]Department of Urology and Kidney Transplantation, Martin-Luther-University, Halle, Germany.

[6] Interdisciplinary Center for Clinical Research, RWTH Aachen University, Aachen, Germany. [7]Institute of Computational Genomics, RWTH Aachen University, Aachen, Germany.

[8] Department of Developmental Biology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.

[9] Department of Cell Biology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.

[10] Department of Internal Medicine and Department of Nephrology and Transplantation, Erasmus Medical Center Transplant Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands.

[11] Department of Cardiology, RWTH Aachen University, Aachen, Germany.

[12] Department of Nephrology, University Hospital Homburg, Homburg, Germany.

[13] Institute of Pathology and Electron Microscopy Facility, RWTH Aachen University, Aachen, Germany.

[14] Department of Hematology, Erasmus Medical Center, Rotterdam, the Netherlands. [15]Department of Pathology, RIMLS, Radboudumc, Nijmegen, the Netherlands.

[16] Institute of Human Genetics, RWTH Aachen University, Aachen, Germany.

[17] Department of Urology, St Antonius Hospital, Eschweiler, Germany.

[18] III Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

[19] Institute of Cell and Tumor Biology, RWTH Aachen University, Aachen, Germany.

[20] Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, Germany.

[21] Molecular Medicine Partnership Unit (MMPU), European Molecular Biology Laboratory and Heidelberg University, Heidelberg, Germany.

[22] Department of Medicine, Division of Nephrology, Kidney Research Institute and Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA.

[23] Department of Bioengineering (Adjunct), and Department of Laboratory Medicine & Pathology (Adjunct), University of Washington, Seattle, WA, USA. [24]These authors contributed equally: Yaoxian Xu, Christoph Kuppe.

7

Adult human kidney organoids originate from CD24+ cells and represent an advanced model for adult polycystic kidney disease

Yaoxian Xu[1,24], **Christoph Kuppe**[1,2,24], Javier Perales-Patón[1,3], Sikander Hayat[1], Jennifer Kranz[1,4,5], Ali T. Abdallah[6], James Nagai[7], Zhijian Li[7], Fabian Peisker[1], Turgay Saritas[1,2], Maurice Halder[1], Sylvia Menzel[1], Konrad Hoeft[1,2], Annegien Kenter[8,9,10], Hyojin Kim[1], Claudia R. C. van Roeyen[1], Michael Lehrke[11], Julia Moellmann[11], Thimoteus Speer[12], Eva M. Buhl[13], Remco Hoogenboezem[14], Peter Boor[2,13], Jitske Jansen[1,15], Cordula Knopp[16], Ingo Kurth[16], Bart Smeets[15], Eric Bindels[14], Marlies E. J. Reinders[10], Carla Baan[10], Joost Gribnau[8,9], Ewout J. Hoorn[10], Joachim Steffens[17], Tobias B. Huber[18], Ivan Costa[7], Jürgen Floege[2], Rebekka K. Schneider[8,19], Julio Saez-Rodriguez[3,20,21], Benjamin S. Freedman[22,23] and Rafael Kramann[1,2,10]

Abstract

Adult kidney organoids have been described as strictly tubular epithelia and termed tubuloids. While the cellular origin of tubuloids has remained elusive, here we report that they originate from a distinct CD24$^+$ epithelial subpopulation. Long-term-cultured CD24$^+$cell-derived tubuloids represent a functional human kidney tubule. We show that kidney tubuloids can be used to model the most common inherited kidney disease, namely autosomal dominant polycystic kidney disease (ADPKD), reconstituting the phenotypic hallmark of this disease with cyst formation. Single-cell RNA sequencing of CRISPR–Cas9 gene-edited *PKD1*- and *PKD2*-knockout tubuloids and human ADPKD and control tissue shows similarities in upregulation of disease-driving genes. Furthermore, in a proof of concept, we demonstrate that tolvaptan, the only approved drug for ADPKD, has a significant effect on cyst size in tubuloids but no effect on a pluripotent stem cell-derived model. Thus, tubuloids are derived from a tubular epithelial subpopulation and represent an advanced system for ADPKD disease modeling.

Main

Various groups have reported kidney organoids derived from human pluripotent stem cells (hPSCs), and improved differentiation protocols have led to mini-kidney structures in a dish that contain many cellular constituents of the adult kidney[1,2,3,4]. However, hPSC-derived organoids resemble early stages of human kidney development and contain nonkidney cell types[5]. They might therefore not be the ideal system to model adult human disease or to study potential regenerative therapies. Recently, organoids of the human adult kidney have been reported as cells outgrowing from tubular fragments or from urine. They were termed tubuloids to reflect their strictly tubulo-epithelial origin and differentiation[6]. Such structures might be superior to model features of epithelial kidney disease, as they are derived from adult human kidneys. However, the exact cellular origin of these tubuloids remains elusive, and it is unclear whether they can be used to induce the phenotype for an inherited, highly prevalent human kidney disease by gene editing.

Here we report that a distinct CD24+ kidney epithelial cell population gives rise to tubuloids and that these cells possess distinct metabolic and gene regulatory programs. CD24+ cells are scattered throughout the nephron. Their proximal tubule (PT) and loop of Henle (LOH) fraction shows the strongest in vitro expansion and long-term growth of largely functional tubular structures. We also demonstrate that CD24+ cell-derived tubuloids can be used to model autosomal dominant polycystic kidney disease (ADPKD) using multiplex CRISPR–Cas9 gene editing, which leads to rapid cyst formation. Using single-cell RNA sequencing (scRNA-seq) of tissue from patients with ADPKD and healthy donors as well as gene-edited tubuloids compared to controls, we demonstrate similarities in upregulation of reported disease-driving genes. Furthermore, we demonstrate that tolvaptan treatment reduces cyst size in tubuloids, while it does not have any effect on cyst size in gene-edited induced pluripotent stem cell (iPSC)-derived organoids. Therefore, tubuloids represent an advanced model of human ADPKD and are useful for drug studies to identify new treatment candidates.

Results

CD24+ cells are metabolically distinct

CD24+ cells that coexpress CD133 have been described as a potential progenitor population in human kidney[7,8,9,10]. As reported[11], we detected CD24-expressing cells as a small scattered subset mainly among PT epithelium in human kidneys (Fig. 1a). We established primary cultures of isolated human CD13+ PT cells and primary CD24+ cells (Fig. 1b and Supplementary Fig. 1). It is widely accepted that adult progenitors reside in a niche that is defined by a low partial oxygen pressure and physiologic hypoxia[12]. We therefore asked whether CD24+ cells have a different energy metabolism as compared to regular PT epithelium (CD13+). Interestingly, we detected decreased basal and maximal oxygen consumption rates (OCRs) in CD24+ cells as compared to CD13+ PT cells and changes in several other metabolic parameters (Fig. 1c,d and Extended Data Fig. 1a–h). These data suggest specialized metabolism in CD24+ cells different from that in PT cells.

Distinct gene regulatory program in CD24+ cells

We analyzed the chromatin accessibility of CD24+ cells compared to CD13+ PT cells using an assay for transposase-accessible chromatin with sequencing (ATAC-seq). CD24+ and CD13+ cells were sorted by fluorescence-activated cell sorting (FACS; Extended Data Fig. 1i) from adult human kidney nephrectomy specimens and immediately subjected to ATAC-seq. When compared to CD13+ PT cells, CD24+ cells exhibited a distinct gene regulatory program (Fig. 1e and Extended Data Fig. 1j) with increased accessibility of genes associated with progenitor cells and dedifferentiation as well as inflammation. Gene ontology (GO) term analysis indicated cell cycle activity, cellular responses to stress and dedifferentiation (loss of normal tubule transport activity) of CD24+ cells (Extended Data Fig. 1k).

scRNA-seq of CD24+ and CD13+ cells

To understand the differences between CD13+ and CD24+ cells at the single-cell level, we performed scRNA-seq of freshly sorted CD13+ and CD24+ cells from human kidneys. Unsupervised clustering and cell type assignment revealed that CD13 strictly labels PT (segment S1–S3) and the thin limb of the LOH (Fig. 1f,h and Supplementary Fig. 2a, left), while CD24 sorting captures cells from various parts of the nephron (Fig. 1g,i and Supplementary Fig. 2a, right). Two large CD24+ populations (41.1%) showed expression of markers from the PT S3 segment and the thin limb of the LOH (Fig. 1g–i). We termed these populations scattered tubule epithelial cells (STC1, STC2). We further identified type A intercalated cells from the collecting duct (IC-A: SLC4A1+), parietal epithelial cells (PECs: VCAM1+, CLDN1+, KLK6+), distal convoluted tubule cells (DCTs: SLC12A3+) and two thick ascending limb populations of the LOH (TAL_1 and TAL_2: SLC12A1+). These data indicated that only about half of the CD24+ cells were derived from PT and the thin LOH limb and thus showed overlap with the source of CD13+ cells.

To determine whether metabolic differences of the CD24+ cells as compared to the CD13+ cells can be caused by the CD24+ cells that are located in more distal parts of the nephron (TAL, DCT, IC-A), we next sorted CD24+CD13+ double-positive cells as well as CD24−CD13+ cells and repeated the metabolic analysis. These data indicated that CD24+ cells as a subpopulation of the CD13+ cells are metabolically distinct from CD13+ cells with decreased maximal OCR and spare respiratory capacity (Supplementary Fig. 2b,c).

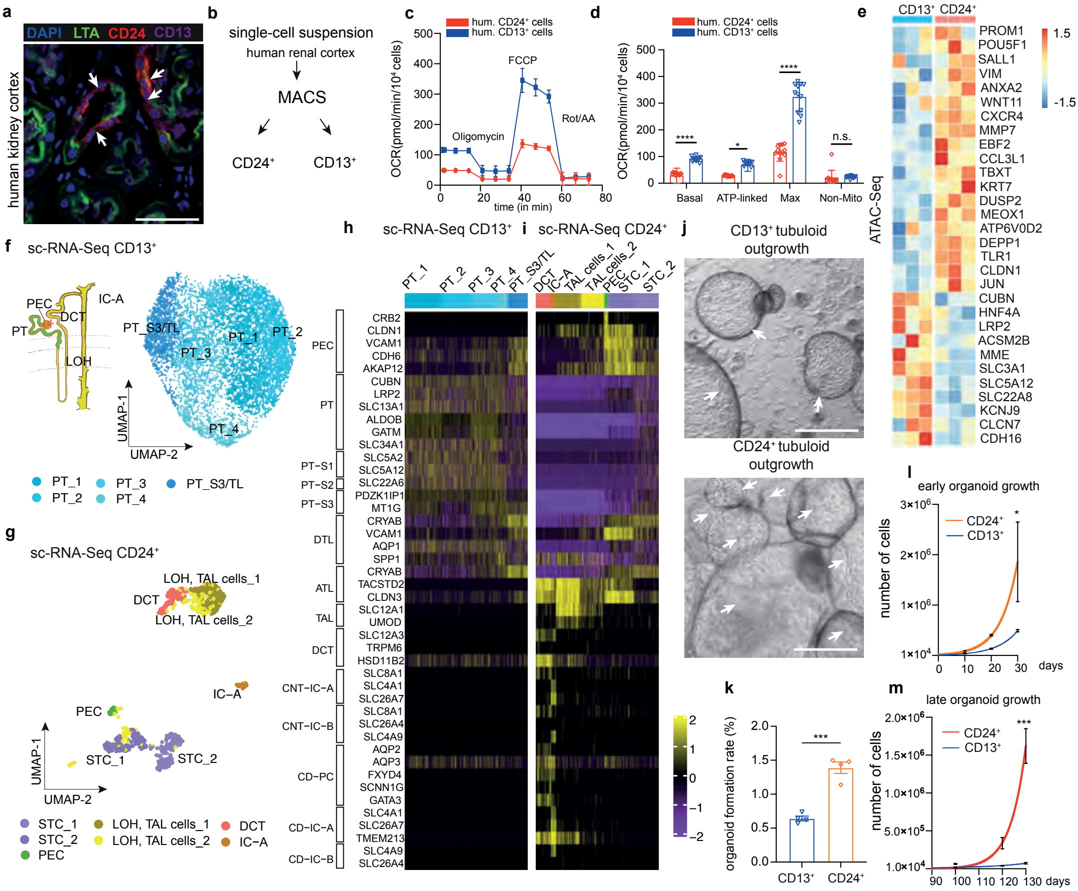


Fig.1

a, Human kidney tissue stained for LTL, CD13, CD24 (arrows) and DAPI (nuclei). Arrows indicate CD24 costaining with CD13/LTL. Scale bar, 50 µm. **b**, Scheme of cell isolation. **c**,**d**, OCR in CD24$^+$ and CD13$^+$cells. Basal, unstimulated OCR; ATP linked, oligomycin OCR; max, FCCP (carbonyl cyanide-4 (trifluoromethoxy) phenylhydrazone) OCR; non-mitochondrial (non-mito), rotenone/antimycin A (ROT/AA) OCR. Statistical analysis in $n=12$ (mean ± s.d.) (**c**) and $n=12$ (mean ± s.e.m.) (**d**); two-way ANOVA with Tukey's post hoc test, *$P=0.0226$ for ATP-linked OCR; ****$P<0.0001$ for basal and maximum OCR; nonsignificant, $P>0.9999$ for non-mitochondrial OCR (**d**). Hum., human (**c**,**d**). **e**, Heatmap displaying ATAC-seq peak count data in the TSS of selected genes of CD13$^+$ or CD24$^+$ cells. **f**, Scheme of the human nephron with PT, LOH, DCT and IC-A of the collecting duct and UMAP embedding for sorted CD13$^+$ cells from the human kidney. $n=7{,}121$ cells from the PT (PT_1, PT_2, PT_3, PT_4) and S3 segment/thin limb of the LOH (S3/TL). **g**, UMAP embedding for sorted CD24$^+$ cells from the human kidney. $n=868$ cells from the DCT, TAL cells from the LOH (TAL cells 1 and 2), PEC, STC_1 and STC_2, and collecting duct IC-A. **h**,**i**, Scaled gene expression of the reported KPMP marker genes of the identified clusters in scRNA-seq of CD13$^+$ cells (**h**) and CD24$^+$ cells (**i**). **j**, Representative difference interference contrast (DIC) microscopy images of 30-day tubuloids. Arrows mark cysts and cyst borders. Scale bars, 200 µm. **k**, Comparison of organoid formation rate. $n=4$ (mean ± s.e.m.); unpaired two-tailed *t*-test, ***$P=0.0002$. **l**,**m**, Early (**l**) and late (**m**) organoid growth curves. Two-way ANOVA with Bonferroni's post hoc test, $n=2$ from separate experiments; the graph shows the mean of the two experiments.

CD24+ cells as the source of adult kidney tubuloids

Recently reported tubuloids from adult human kidneys have been established as outgrowing structures from digested tissue6. However, the cellular origin of these tubuloids remains unknown. We sought to establish adult kidney tubuloids from isolated CD24+ cells and from regular PT (CD13+) cells. The Wnt target gene LGR5 is one of the most widely used adult stem cell markers13,14. While we did not detect expression of LGR5 in CD24+ cells, we observed higher accessibility in the transcriptional start site (TSS) of the LGR4 gene and increased expression of LGR4 mRNA in CD24+ cells (Supplementary Fig. 2d,e). We thus used Wnt3a and R-spondin 1 (RSPO1) treatment in our organoid protocol.

The organoid formation rate and population doubling of CD24+ cells were increased compared to CD13+ cells (Fig. 1j–m). We observed four phases of CD24+ cell growth and expansion into tubulo-epithelial organoids referred to as tubuloids (Fig. 2a,b, Extended Data Fig. 2a–f, Supplementary Videos 1–4 and Supplementary Notes). The tubuloid culture showed robust growth for >6 months (187 days; 23 passages). Compared to the one-phase protocol in ref. 6, we observed an increased organoid formation rate using the four-phase protocol and also more tubules within each tubuloid (Fig. 2c–e and Extended Data Fig. 2g,h). We further tried a protocol that solely uses epidermal growth factor (EGF) and fibroblast growth factor 2 (FGF2)15 and demonstrated a reduced organoid formation rate in the absence of Wnt3a and RSPO1 as compared to the four-phase protocol, confirming the importance of these factors for tubuloid formation (Extended Data Fig. 2i–k).

In summary, these data indicate that CD24+ cells are the origin of tubuloids and possess an increased in vitro growth and maintenance capacity compared to CD13+ PT cells. Because CD24+ cells can be found in several parts of the nephron, we next asked whether indeed their PT S3 and LOH fraction (CD24+CD13+) shows an enhanced organoid formation rate as compared to other CD24+ cells (CD24+CD13−), CD13+ cells without CD24 expression (CD24−CD13+) or any other kidney cell type (CD24−CD13−). Notably, we observed organoid formation only from CD24+ cells (Extended Data Fig. 3a–l), with CD24+ cells from the PT and thin limb of the LOH (CD24+CD13+) showing the best organoid formation rate as compared to CD24+ cells from any other location (CD24+CD13−; Extended Data Fig. 3m).

Tubuloids mainly represent the proximal nephron

To understand the composition and differentiation of tubuloids, we next performed scRNA-seq from early (day 21) and late (day 97) tubuloids obtained using the four-phase protocol (Fig. 2a). Most tubuloid populations had high expression of CD24 and CD13 (ANPEP), in line with their origin (Fig. 2f–i). The majority of cells (62.7%) in the early tubuloid were similar to the STC population in the human kidney (Fig. 1i) and characterized by expression of PT S3 markers such as PDZK1IP1 as well as markers of the

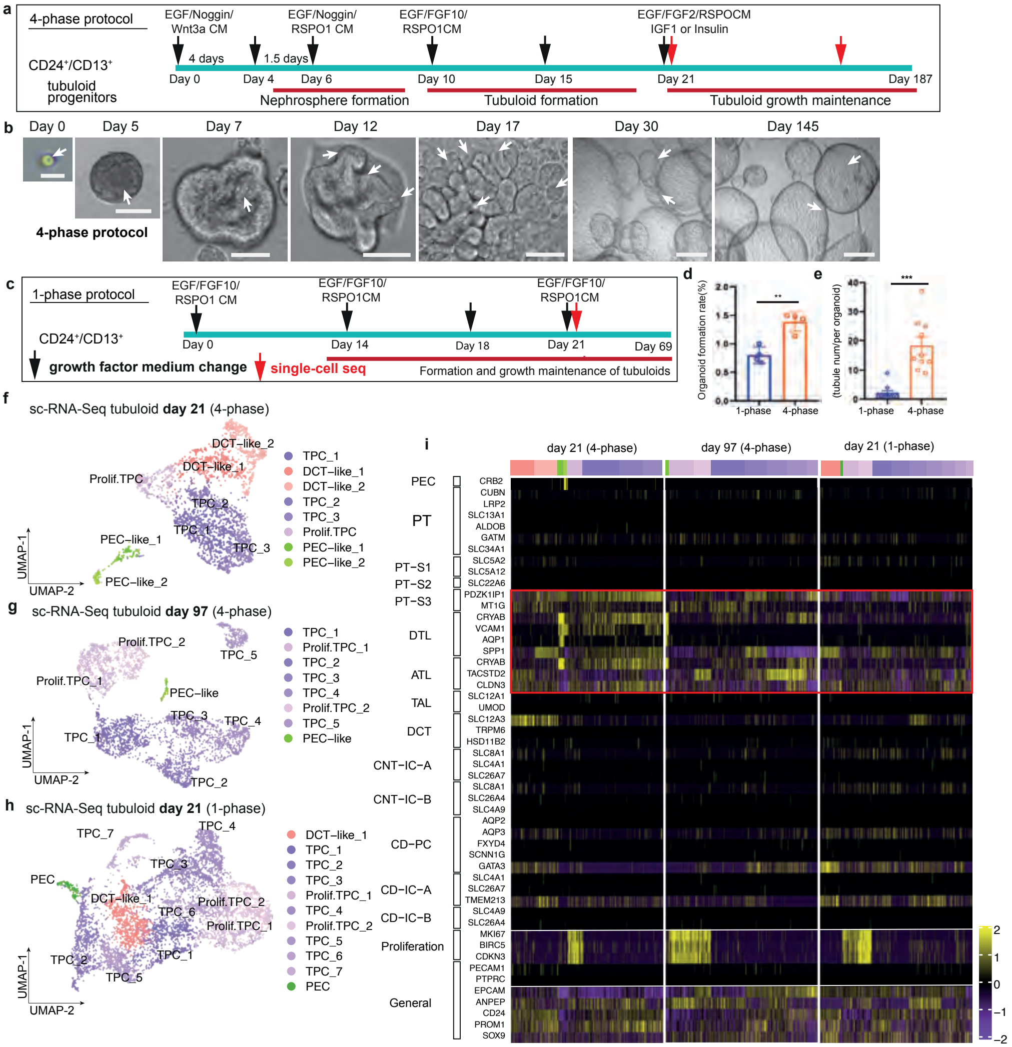


Fig. 2

a, Timeline of tubuloid generation (four-phase protocol). Black arrows indicate the change of conditioned medium (CM) and growth factors, and red arrows indicate time points for scRNA-seq. **b**, Representative images of CD24$^+$ cell-derived tubuloids using the four-phase protocol. Scale bars, 200 µm (days 30, 145), 100 µm (day 0) and 50 µm (days 5, 7, 12 and 17). **c**, Schematic timeline of tubuloid generation using the one-phase protocol. **d**,**e**, Quantification of tubuloid formation rate on day 22 (**d**) and number of tubuli within a tubuloid on day 25 (**e**) of CD24$^+$ cells comparing the four-phase and one-phase protocol. Statistical analysis in $n = 4$ (mean ± s.d.); unpaired two-tailed t-test with Welch's correction, $^{**}P = 0.0020$ (**d**) and $n = 10$ (mean ± s.d.); unpaired two-tailed t-test with Welch's correction, $^{***}P = 0.0002$; mean = 18.60 tubuli per four-phase tubuloid, 95% confidence interval = 9.464–22.34 tubuli, indicating that 95% of four-phase tubuloids contained around 9–22 tubuli (**e**). **f**, UMAP embedding of cells from an early tubuloid (day 21, four-phase protocol). $n = 2{,}291$ cells in eight clusters: TPC_1–TPC_3, proliferating TPC, DCT_1 and DCT_2, and PT/PEC_1 and PT/PEC_2. **g**, UMAP embedding of cells from a late tubuloid (day 97, four-phase protocol). $n = 3{,}693$ single cells in eight clusters: TPC_1–TPC_5, proliferating TPC (prolif. TPC_1 and TPC_2), and cells that showed markers of PEC and PT (PT/PEC). **h**, UMAP embedding of cells from an early tubuloid (day 21, one-phase protocol). $n = 5{,}631$ cells in 11 clusters identified: TPC (TPC_1–TPC_7), proliferating TPCs, (prolif. TPC_1 and TPC_2), DCT-like-1, and PT/PEC. **i**, Heatmap of marker gene expression for different nephron parts using KPMP marker genes as well as selected general marker genes and proliferation markers of the early-stage (day 21) and late-stage (day 97) tubuloids (four phase) and an early (day 21) tubuloid (one-phase protocol). For details on statistics and reproducibility, see Methods.

thin limb of the LOH (descending thin limb and ascending thin limb (ATL); Fig. 2i and Supplementary Fig. 3a). The annotation was based on marker genes from the Kidney Precision Medicine Project (KPMP)16. We termed these populations tubuloid progenitor cells (TPC_1–TPC_3 and proliferating TPC) as these are the cells that form the largest parts of the tubuloids. We further detected smaller populations of DCT cells (DCT-like 1 and DCT-like 2) characterized by SLC12A3 expression and two small populations that showed expression of PT and PEC marker genes (PT/PEC_1 and PT/PEC_2; Fig. 2f,i, Supplementary Fig. 3a and Supplementary Notes).

Notably, we observed strong expression of genes that are associated with proliferation (MKI67, BIRC5, CDKN3) only among the TPC populations (proliferating TPC; Fig. 2i). This suggested that primarily cells derived from the PT (S3) and thin limb of the LOH sustain long-term growth in the tubuloids. In line with this, we primarily detected TPC and proliferating TPC (TPC_1–TPC_5, proliferating TPC_1 and TPC_2) in the late-stage tubuloid (day 97, 97.9% of cells; Fig. 2g,i and Supplementary Fig. 3b). We further detected one small PT/PEC population in the late-stage tubuloid (Fig. 2g,i, Supplementary Fig. 3b and Supplementary Notes). We also performed scRNA-seq of a tubuloid generated from isolated CD24+ cells that were subjected to the published one-phase protocol6. Again, the majority of the cells (85.6%) showed a transcriptional marker profile consistent with the PT S3 segment and the thin limb of the LOH and were thus annotated as TPC (TPC_1–TPC_7), with only TPC populations showing expression of proliferation marker genes (proliferating TPC_1 and TPC_2) (Fig. 2h,i and Supplementary Fig. 3c). We further detected one population with minor expression of SLC12A3 that we termed DCT-like 1 as well as one small PT/PEC population (Fig. 2h,i, Supplementary Fig. 3c and Supplementary Notes). Of note, due to dedifferentiation, the cell type annotation within these organoids/ tubuloids is difficult, and similarly, in the data reported by ref. 6, no classical marker genes are used for the cell type annotation. However, using the reported KPMP guideline marker genes for the human nephron, the annotation indicates that the majority of tubuloids are derived from CD24+ cells within the PT S3 and LOH.

Tubuloids consist of functional polarized tubule epithelium

Immunostaining of tubuloids on day 21 confirmed that CD24+ cell-derived tubuloids retain epithelial differentiation and express E-cadherin (CDH1) and β-catenin (Fig. 3a and Supplementary Fig. 4a,b,f). Staining of the sodium/potassium ATPase (Na+/K+-ATPase) showed a distinct basolateral pattern, while F-actin was expressed at the apical side of the cells (Fig. 3b), indicating cell polarity within the tubuloids. Staining for acetylated tubulin indicated primary cilium formation at the luminal surface (Fig. 3c). We observed a Villin1 (VIL1)-positive brush border toward the lumen of the tubuloids (Fig. 3d). By electron microscopy, epithelial cells in the tubuloids showed typical features of human

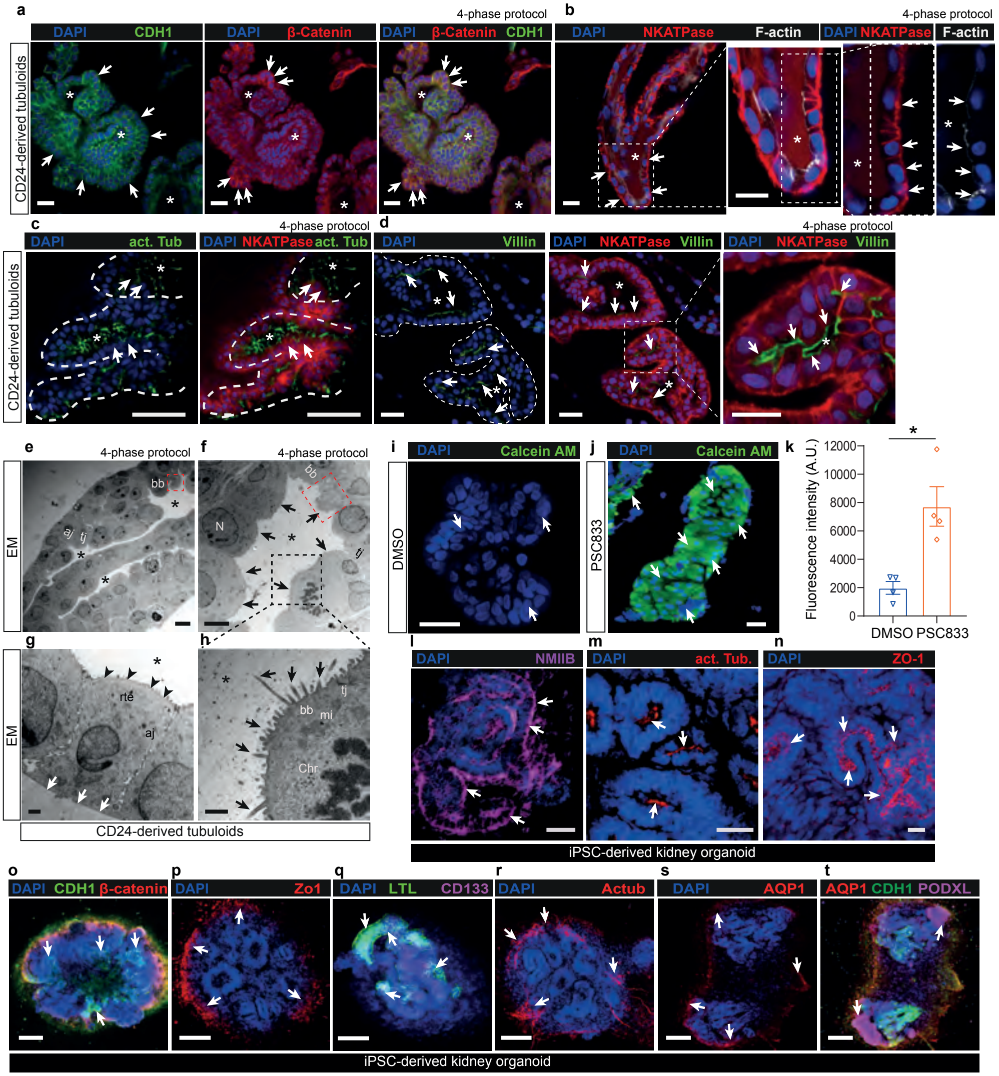
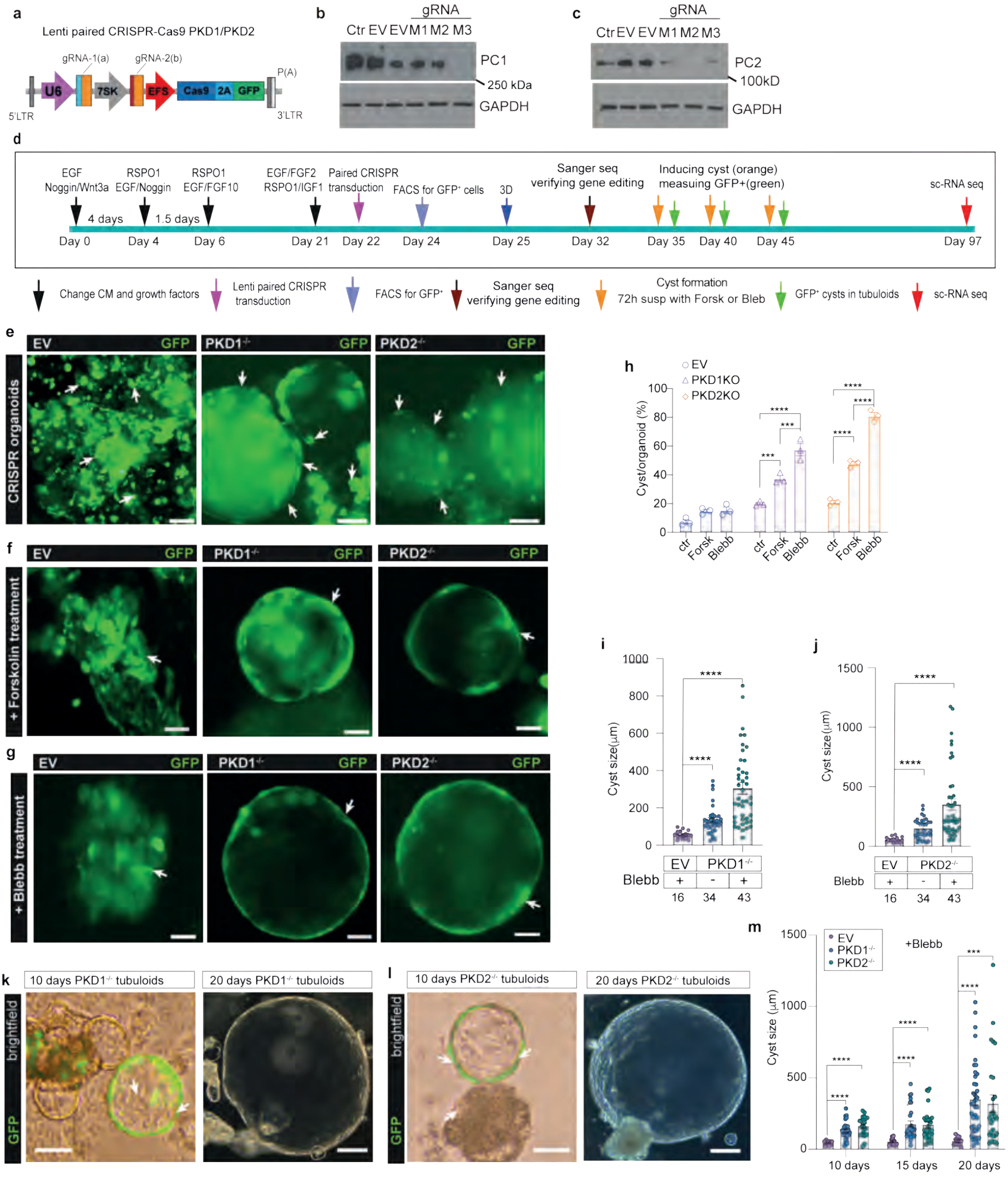


Fig. 3

a, Representative images of tubuloids (four phase) derived from CD24$^+$ cells stained for CDH1 (E-cadherin), β-catenin and DAPI (nuclei). Scale bars, 50 µm. **b**, Representative images of tubuloids (four phase) stained for basolateral Na$^+$/K$^+$-ATPase, apical filamentous actin (F-actin) and DAPI. Scale bars, 50 µm. **c**, Representative images of tubuloids (four phase) stained for apical acetylated tubulin (Ac-tub), Na$^+$/K$^+$-ATPase and DAPI. Scale bars, 50 µm. **d**, Representative images of tubuloids (four phase) stained for Villin1 (brush border), Na$^+$/K$^+$-ATPase and DAPI. Scale bars, 50 µm. **e**–**h**, Representative transmission electron microscopy (TEM) images of tubuloids (four phase, day 21). Arrows mark cell borders and microvilli. N, nucleus; TJ, tight junction; AJ, adherens junction; BB, brush border; Mi, mitochondria; Chr, chromosome; RTE, renal tubule epithelium; red dashed squares indicate brush border; black dashed squares mark cilia. Black arrows mark the apical side and white arrows mark the basolateral side of epithelial cells. Scale bars, 100 µm (**e**), 5 µm (**f**) and 1 µm (**g**,**h**). **i**–**k**, Representative images (**i**,**j**) and quantification (**k**) of intracellular calcein-AM accumulation in tubuloids in the presence of the P-gp transporter inhibitor PSC833 or vehicle (0.2% DMSO). Unpaired two-tailed *t*-test with Welch's correction, *n* = 4 (mean ± s.e.m.), **P* = 0.0207. Scale bars, 50 µm (**i**,**j**). AU, arbitrary units. **l**–**t**, Representative images of human iPSC-derived kidney organoids stained for NMIIB (**l**), acetylated tubulin (**m**,**r**), Zo1 (**n**,**p**), CDH1 and β-catenin (**o**), CD133 and LTL (**q**), AQP1 (**s**) and AQP1 with PODXL (**t**). DAPI was used to counterstain nuclei in all images. Scale bars, 50 µm (**o**–**t**), 25 µm. Asterisks in all images indicate renal tubule lumen (**a**–**h**). Arrows indicate positive staining in immunofluorescence images (**a**–**d**,**i**,**j**,**l**–**t**). For details on statistics and reproducibility, see Methods.



adult PT epithelium, including tight and adherens junctions, lumen formation and polarization (Fig. 3e–h). CD24+ cell-derived tubuloids showed Zonula occludens protein 1 (Zo1) expression (Supplementary Fig. 4c). Nonmuscle myosin IIB (NMIIB), a motor protein interacting with the actin cytoskeleton, was also strongly expressed in tubuloids

Fig. 4

a, Scheme of lentiviral paired CRISPR–Cas9. **b**,**c**, Representative western blots of gene-edited tubuloids. M1, two U6-driven paired gRNAs; M2 and M3, U6- and 7SK-driven paired gRNAs; Ctr, no transduction; GAPDH was used as a loading control. Uncropped western blots in Supplementary Fig. 18a–d. **d**, Timeline of gene editing in tubuloids. Blebb, blebbistatin; Forsk, forskolin. **e**, Representative images of EV, *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$ tubuloids at 10 days after transduction. Scale bars, 100 µm (**e**(middle and right)) and 50 µm (**e** (left)). **f**–**h**, Representative images (**f**,**g**) and quantification of cyst formation rate (**h**) using two-way ANOVA with Bonferroni's post hoc test in $n = 3$ (mean ± s.e.m.), nonsignificant for EV forskolin or blebbistatin versus Ctr; *PKD1*$^{-/-}$ forskolin versus Ctr, ***$P = 0.0007$; *PKD1*$^{-/-}$ blebbistatin versus forskolin, ***$P = 0.0001$; *PKD1*$^{-/-}$ blebbistatin versus Ctr and *PKD2*$^{-/-}$forskolin or blebbistatin versus Ctr and blebbistatin versus forskolin, ****$P < 0.0001$ (**h**). Scale bars, 100 µm (**f** (left), **g** (middle)) and 50 µm (**f** (middle and right), **g** (left and right)). Ctr tubuloids were treated with DMSO. **i**,**j**, Quantification of cysts with Welch and Brown–Forsythe and Welch ANOVA test post hoc Tamhane's T2 in $n = 43$ *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ tubuloids + blebbistatin and $n = 34$ *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ tubuloids alone versus $n = 16$ EV tubuloids + blebbistatin (mean ± s.d.), ****$P < 0.0001$ in PKD1$^{-/-}$ tubuloids for blebbistatin versus tubuloids alone and blebbistatin or tubuloids alone versus EV (**i**) and in *PKD2*$^{-/-}$ tubuloids for blebbistatin or tubuloids alone versus EV (**j**); ***$P = 0.0004$ in *PKD2*$^{-/-}$ tubuloids for blebbistatin versus tubuloids alone (**j**). **k**,**l**, Representative images of *PKD1*$^{-/-}$ (**k**) and *PKD2*$^{-/-}$ (**l**) tubuloids at days 10 and 20 treated with blebbistatin. Scale bars, 200 µm (**k**,**l** (right)) and 100 µm, (**k**,**l** (left)). **m**, Statistical analysis of cysts using two-way ANOVA with Bonferroni's post hoc test at day 10 in $n = 26$ *PKD1*$^{-/-}$ or $n = 21$ *PKD2*$^{-/-}$ versus $n = 9$ EV tubuloids (mean ± s.d.), day 15 in $n = 29$ *PKD1*$^{-/-}$ or $n = 26$ *PKD2*$^{-/-}$ versus $n = 11$ EV tubuloids (mean ± s.d.) and day 20 in $n = 38$ *PKD1*$^{-/-}$ or $n = 27$ *PKD2*$^{-/-}$ versus $n = 11$ EV tubuloids (mean ± s.d.), ***$P = 0.0004$ at day 20 for *PKD2*$^{-/-}$ versus EV tubuloids; ****$P < 0.0001$ at day 10 for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ versus EV tubuloids, day 15 for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ versus EV tubuloids and day 20 for *PKD1*$^{-/-}$ versus EV tubuloids. For details on statistics and reproducibility, see Methods.

(Supplementary Fig. 4d). The epithelium of the tubuloids also showed expression of CD133 and aquaporin 1 (AQP1), while the podocyte marker podocalyxin (PODXL) was not expressed (Supplementary Fig. 4e,f), as expected. The percentage of tubuloids that were positively stained for the markers described above was quantified, and human kidney tissue was used as validation of staining (Supplementary Fig. 4g,h).

P-glycoprotein (P-gp) is an ATP-dependent efflux transporter located at the apical membrane of the renal PT and can be blocked using the compound PSC833. To evaluate whether CD24+ cell-derived tubuloids contain P-gp transporter function, we exposed them to PSC833 and then incubated them with the P-gp substrate calcein-AM. Fluorescent calcein prominently accumulated in the tubuloids when the P-gp efflux transport function of the tubuloids was inhibited by PSC833 (Fig. 3i–k), confirming the activity of P-gp.

We next compared the adult kidney tubuloids with iPSC-derived kidney organoids (Fig. 3l–t). As expected, iPSC-derived organoids contained segments of tubule-like structures with expression of E-cadherin and β-catenin and PT-like segments that stained positive for Lotus tetragonolobus lectin (LTL) and AQP1 (Fig. 3l–t), while various other parts did not stain for these markers. Zo1, NMIIB and acetylated tubulin were also detectable in tubule-like segments of the iPSC organoids (Fig. 3l–t). In sharp contrast to the tubuloids, a population of epithelial cells in the iPSC-derived organoids expressed the podocyte marker PODXL (Fig. 3t). Taken together, we demonstrate a strict tubule epithelial origin and differentiation of tubuloids in contrast to iPSC-derived kidney organoids, which differentiate into various different parts of the kidney as well as other nonkidney off-target cell types5.

Modeling adult polycystic kidney disease in tubuloids

One key advantage of organoids is human disease modeling in the dish. While this has been successfully accomplished with hPSC-derived kidney organoids for certain disease states, tubuloids may constitute a more accurate model for the adult human tubule and do not contain early developmental stage cell types or off-target cell types from differentiation protocols.

ADPKD is the most common hereditary kidney disease[17] and accounts for about 10% of all patients with end-stage renal disease. Mutations in two large multi-exon genes, PKD1 and PKD2, cause the disease. Tolvaptan has recently been approved for treatment, but potent and truly curative therapies with no or limited adverse effects are still missing. We therefore aimed to model ADPKD in tubuloids to develop an in vitro platform that resembles cyst formation. As CRISPR–Cas9-induced insertions and deletions (indels) are often still in frame, which reduces knockout efficiency[18], we established a lentiviral multiplex CRISPR cloning vector (Supplementary Figs. 5a and 6) for effective transduction of organoids and then used it to establish a paired CRISPR–Cas9 construct for targeting two locations of either the PKD1 or PKD2 gene locus (Fig. 4a and Supplementary Figs. 5b,c, 7 and 8). We compared different promoter assembly strategies for the paired lentiviral PKD1 and PKD2 gene editing (Supplementary Figs. 5d–i and 9–12). Western blot analysis indicated that a U6- and 7SK-driven guide RNA (gRNA) expression system led to sufficient knockout of PC1 protein (M3 clone; Fig. 4b) or PC2 protein (M2 and M3 clones; Fig. 4c), whereas using a U6 promoters for both gRNA pairs did not result in sufficient protein loss (M1 clone; Fig. 4b,c).

We therefore used U6- and 7SK-driven gRNA expression for gene editing of PKD1 and PKD2 in tubuloids and observed considerable GFP expression at 48 h after transduction (Supplementary Fig. 13a). As a control (empty vector (EV)), we used the same virus without gRNA. We verified transduction efficacy by flow cytometry (Fig. 4d,e and Supplementary Fig. 13b–i). Efficient knockout was validated in sorted GFP+ cells from tubuloids on day 10 after transduction (Fig. 4d,e) by PCR (Extended Data Fig. 4a–d) and Sanger sequencing, indicating a deletion of 265 bp in the PKD1 gene (Extended Data Fig. 4e and Supplementary Fig. 7) and 165 bp in the PKD2 gene (Extended Data Fig. 4f and Supplementary Fig. 8).

Cyst formation in PKD1 −/− and PKD2 −/− tubuloids

We noticed cyst formation in the paired PKD1 or PKD2 CRISPR–Cas9 transduced tubuloids on day 10 after transduction (Fig. 4e and Supplementary Video 5). However, cyst formation rates remained very low at 9–20%. It has been reported that elevated renal cyclic AMP (cAMP) levels promote cyst growth and that the cAMP agonist forskolin can induce rapid and dose-dependent cyst formation in organoids derived from PKD−/− hPSCs19. Furthermore, the myosin II ATPase inhibitor blebbistatin as well as removal of adherent cues reportedly increases cyst growth in hPSC-derived kidney organoids[19,20]. Interestingly, we observed a significant increase in the cyst formation rate by both substances using a suspension culture (Fig. 4f–h). Blebbistatin treatment yielded the highest cyst formation rate (Fig. 4h) and also increased the cyst size (Fig. 4i,j). Next, we also quantified the percentage of GFP+ cysts, which was markedly higher in gene-edited PKD−/− tubuloids (40–50%, on day 10 after transduction as compared to control; Extended Data Fig. 4g and Supplementary Notes). We next asked whether blebbistatin treatment of tubuloids in three-dimensional (3D) culture can affect subsequent cyst growth in suspension culture. Interestingly, the cysts that originated from PKD−/− tubuloids exhibited an increase in size proportional to the timespan of prior blebbistatin exposure (Fig. 4k–m).

Cysts in PKD-knockout tubuloids resemble human ADPKD tissue

Confocal analysis confirmed that cyst formation occurred in GFP+ cells of PKD−/− tubuloids but not in control tubuloids (Supplementary Fig. 14a–j). The GFP signal intensity was increased within the cysts (Supplementary Fig. 14c). Cysts of PKD−/− tubuloids in the absence or presence of blebbistatin were formed by single- or double-lined epithelial cells, which costained for Zo1 and NMIIB (Supplementary Fig. 14b–h) as well as E-cadherin (CDH1) and acetylated tubulin (Supplementary Fig. 14i,j). The cyst lining epithelial cells were oriented with the primary cilia toward the cyst lumen (Supplementary Fig. 14i). When comparing the cyst morphology of the gene-edited tubuloid with cysts in human ADPKD kidney tissue, we observed striking similarities (Supplementary Fig. 15a–d).

Single-nucleus RNA sequencing of human ADPKD tissue

We next performed single-nucleus RNA-seq (snRNA-seq) of human ADPKD tissue ($n = 3$ patients with PKD1 or PKD2 mutations) and two kidney donor biopsies as healthy controls (Fig. 5a and Extended Data Fig. 5a–c). We observed an enrichment of specific immune cell clusters in ADPKD (Extended Data Fig. 5d,e). Using differential gene expression analysis, we observed increased expression of genes with a proposed role in processes associated with ADPKD, including upregulation of AKAP12 in the PT and LOH, MET and LRRK2 in the DCT and VCL in the DCT and collecting duct (Fig. 5b, Extended Data Fig. 5f and Supplementary Notes).

Using PROGENY, we observed increased hypoxia, mitogen-activated protein kinase (MAPK), epidermal growth factor receptor (EGFR), nuclear factor (NF)-kB and tumor necrosis factor α (TNFα) signaling activity, particularly in the LOH and collecting duct clusters of the human ADPKD tissue (Extended Data Fig. 5g). We detected increased

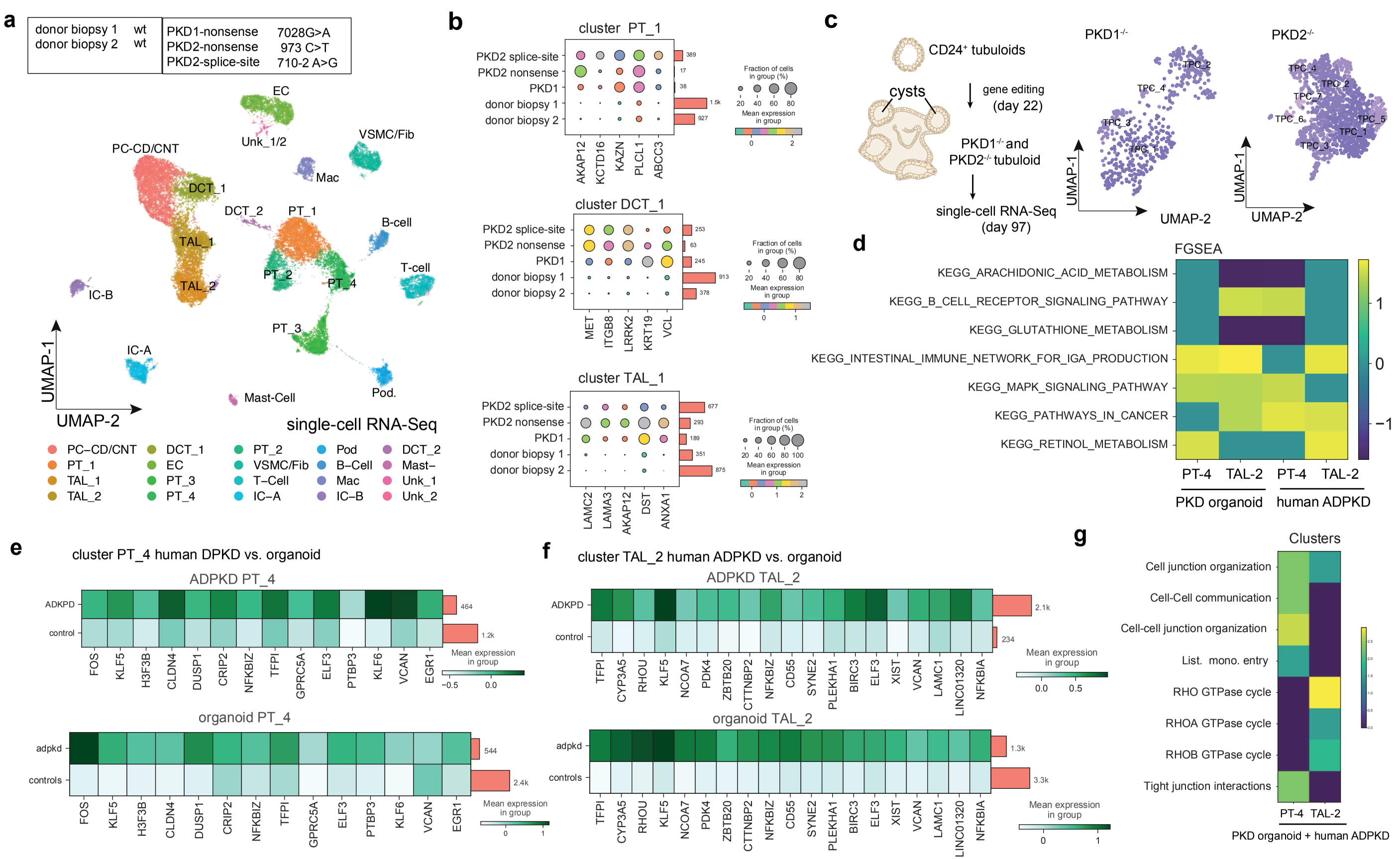

a
donor biopsy 1 wt
donor biopsy 2 wt
PKD1-nonsense 7028G>A
PKD2-nonsense 973 C>T
PKD2-splice-site 710-2 A>G
EC
Unk_1/2
VSMC/Fib
Mac
PC-CD/CNT
DCT_1
DCT_2
PT_1
B-cell
TAL_1
TAL_2
PT_2
PT_4
T-cell
IC-B
PT_3
IC-A
Mast-Cell
Pod.
UMAP-1
UMAP-2
single-cell RNA-Seq
PC–CD/CNT
PT_1
TAL_1
TAL_2
DCT_1
EC
PT_3
PT_4
PT_2
VSMC/Fib
T–Cell
IC–A
Pod
B–Cell
Mac
IC–B
DCT_2
Mast–
Unk_1
Unk_2
b
cluster PT_1
PKD2 splice-site
PKD2 nonsense
PKD1
donor biopsy 1
donor biopsy 2
AKAP12
KCTD16
KAZN
PLCL1
ABCC3
389
17
38
1.5k
927
Fraction of cells in group (%)
20 40 60 80
Mean expression in group
0 2
cluster DCT_1
MET
ITGB8
LRRK2
KRT19
VCL
253
63
245
913
378
0 1
cluster TAL_1
LAMC2
LAMA3
AKAP12
DST
ANXA1
677
293
189
351
875
20 40 60 80 100
0 1 2
c
CD24+ tubuloids
cysts
gene editing (day 22)
PKD1-/- and PKD2-/- tubuloid
single-cell RNA-Seq (day 97)
PKD1-/-
PKD2-/-
TPC_1
TPC_2
TPC_3
TPC_4
TPC_5
TPC_6
TPC_7
UMAP-1
UMAP-2
d
FGSEA
KEGG_ARACHIDONIC_ACID_METABOLISM
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY
KEGG_GLUTATHIONE_METABOLISM
KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION
KEGG_MAPK_SIGNALING_PATHWAY
KEGG_PATHWAYS_IN_CANCER
KEGG_RETINOL_METABOLISM
1
0
−1
PT-4
TAL-2
PT-4
TAL-2
PKD organoid
human ADPKD
e
cluster PT_4 human DPKD vs. organoid
ADPKD PT_4
ADKPD
control
FOS
KLF5
H3F3B
CLDN4
DUSP1
CRIP2
NFKBIZ
TFPI
GPRC5A
ELF3
PTBP3
KLF6
VCAN
EGR1
464
1.2k
Mean expression in group
−0.5 0.0
organoid PT_4
adpkd
controls
544
2.4k
0 1
f
cluster TAL_2 human ADPKD vs. organoid
ADPKD TAL_2
ADKPD
control
TFPI
CYP3A5
RHOU
KLF5
NCOA7
PDK4
ZBTB20
CTTNBP2
NFKBIZ
CD55
SYNE2
PLEKHA1
BIRC3
ELF3
XIST
VCAN
LAMC1
LINC01320
NFKBIA
2.1k
234
0.0 0.5
organoid TAL_2
adpkd
controls
1.3k
3.3k
0 1
g
Clusters
Cell junction organization
Cell-Cell communication
Cell-cell junction organization
List. mono. entry
RHO GTPase cycle
RHOA GTPase cycle
RHOB GTPase cycle
Tight junction interactions
PT-4
TAL-2
PKD organoid + human ADPKD

Fig. 5

a, Overview of the human kidney tissue and *PKD1* and *PKD2* genotypes used for snRNA-seq and UMAP embedding of *n* = 26,509 single cells from the five human kidney samples. Labels refer to identified cell types. EC, endothelial cells; Mac, macrophages; Fib, fibroblasts; PT, proximal tubular cells; Pod, podocytes; vSMC, vascular smooth muscle cells; IC-A/B, intercalated cells A/B; DCT, distal convoluted tubular cells; PC-CD/CNT, principal cells of connecting tubule/collecting duct; TAL, thick ascending limb tubular cells; unk, unknown. WT, wild type. **b**, Top five most upregulated genes in human ADPKD versus donor biopsies. **c**, Top, scheme of gene editing and scRNA-seq in tubuloids. Bottom, UMAP embedding of *n* = 496 single cells from *PKD1* gene-edited tubuloids (four phase, *PKD1*$^{-/-}$) (left) and *n* = 1,483 cells from *PKD2* gene-edited tubuloids (*PKD2*$^{-/-}$) (right). **d**, Common dysregulated pathways obtained from gene set enrichment with KEGG pathway analysis using all differentially expressed genes in human ADPKD versus donor biopsies in PT_4 and TAL_2 cells with the cells of the tubuloids that mapped to PT_4 and TAL_2 using Symphony. NES, normalized enrichment score. **e**,**f**, Select top commonly upregulated genes in human ADPKD versus donor biopsies and the cells that mapped to PT_4 (**e**) and TAL_2 (**f**) in gene-edited tubuloids (*PKD*$^{-/-}$ tubuloids) versus control tubuloids. **g**, Gene set over-representation analysis with Reactome pathways for the common upregulated genes in human tissue (ADPKD versus donor biopsies) and tubuloids (*PKD*$^{-/-}$ versus control). List. mono. entry, *Listeria* monocytogenes entry.

expression of mTOR, MYC and cAMP target genes in several epithelial cell populations from the ADPKD tissue (Extended Data Fig. 5h), suggesting that these epithelial populations represent cyst-lining epithelial cells. Gene set enrichment and GO term analyses pointed toward proinflammatory pathways in various immune cell populations and enrichment of terms associated with cytoskeleton, focal adhesion and tight junctions as well as MAPK signaling in epithelial cell clusters (Extended Data Fig. 6).

Over-representation analysis of the ADPKD data compared to the healthy kidney data pointed toward phosphoinositide 3-kinase (PI3K)–Akt and MET signaling in PT, PI3K–Akt, BRAF, MAPK and RHO GTPase signaling in DCT and MET and MAPK signaling among various other pathways in TAL and collecting duct epithelium (Supplementary Fig. 16). Many of these pathways have been reported as key players in ADPKD[21,22,23,24].

CrossTalkeR[25] analyses indicated increased signaling between epithelial cell types and immune cells as well as fibroblasts in ADPKD (Extended Data Fig. 7a-f). We further performed subclustering analysis of the epithelial populations (Extended Data Fig. 8 and Supplementary Notes). Overall, these data provide an unbiased snRNA-seq atlas of human ADPKD as compared to healthy human kidneys.

Transcriptomic comparison between PKD-knockout tubuloids and human tissue

To compare the gene-edited ADPKD tubuloid model to the human disease, we next performed scRNA-seq from tubuloids after transduction with the *PKD1*- or *PKD2*-targeted paired CRISPR–Cas9 editing construct (Fig. 5c and Supplementary Fig. 17a-c). We mapped the tubuloid cell clusters to the human kidney tissue data using Symphony[26] and observed that most tubuloid clusters mapped to PT_4 and TAL_2 (Extended Data Fig. 9a-e). This analysis also confirmed our annotation and indicated that tubuloids are mainly representing the proximal part of the nephron. Staining of human ADPKD kidney tissue

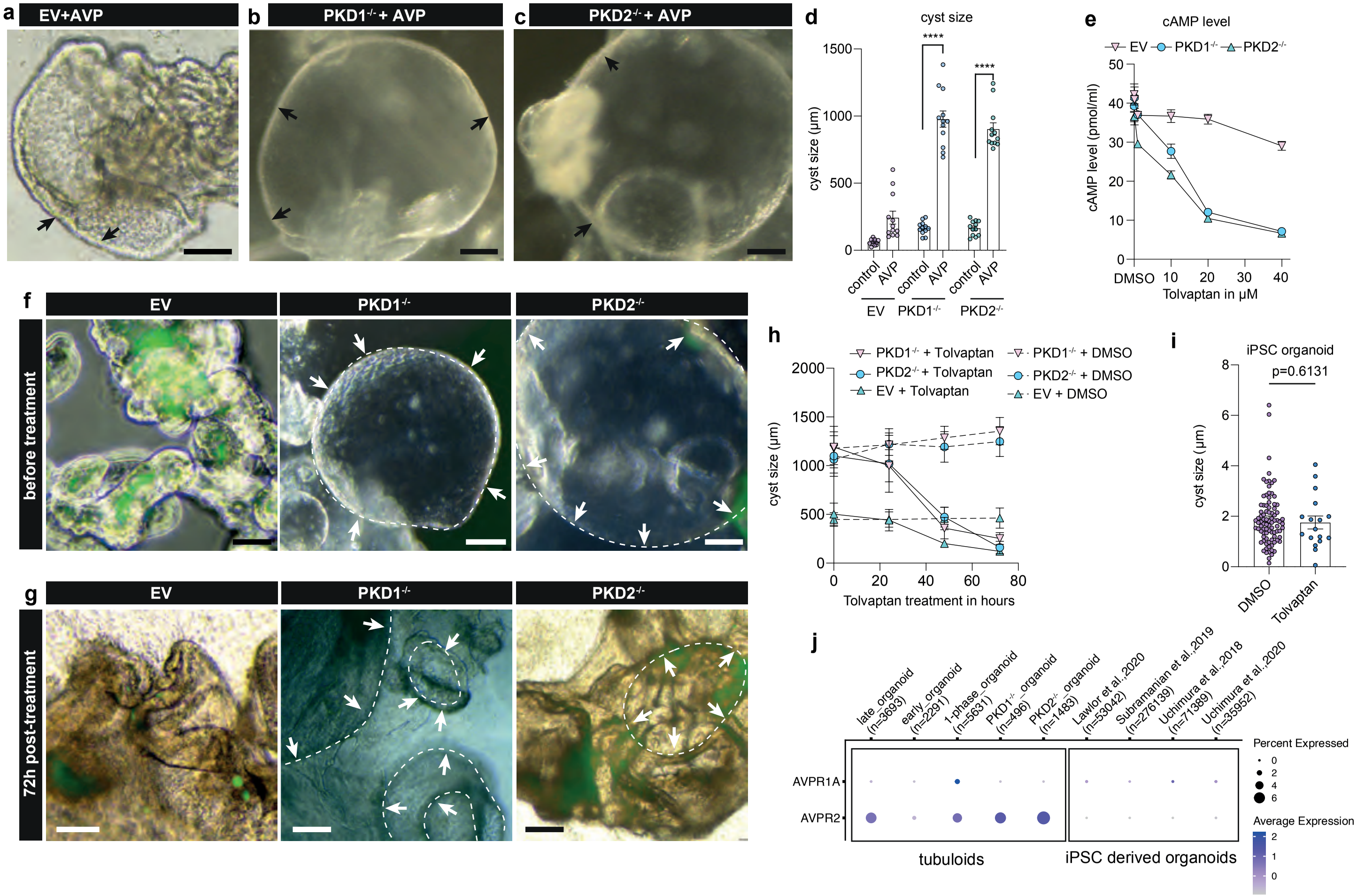

a
EV+AVP
b
PKD1-/- + AVP
c
PKD2-/- + AVP
d
cyst size
cyst size (µm)
1500
1000
500
0
****
****
control
AVP
control
AVP
control
AVP
EV
PKD1-/-
PKD2-/-
e
cAMP level
EV
PKD1-/-
PKD2-/-
cAMP level (pmol/ml)
50
40
30
20
10
0
DMSO
10
20
30
40
Tolvaptan in µM
f
before treatment
EV
PKD1-/-
PKD2-/-
g
72h post-treatment
EV
PKD1-/-
PKD2-/-
h
PKD1-/- + Tolvaptan
PKD2-/- + Tolvaptan
EV + Tolvaptan
PKD1-/- + DMSO
PKD2-/- + DMSO
EV + DMSO
cyst size (µm)
2000
1500
1000
500
0
0
20
40
60
80
Tolvaptan treatment in hours
i
iPSC organoid
p=0.6131
cyst size (µm)
8
6
4
2
0
DMSO
Tolvaptan
j
late_organoid (n=3693)
early_organoid (n=2291)
1-phase_organoid (n=5631)
PKD1-/-_organoid (n=496)
PKD2-/-_organoid (n=1483)
Lawlor et al.,2020 (n=53042)
Subramanian et al.,2019 (n=276139)
Uchimura et al.,2018 (n=71389)
Uchimura et al.,2020 (n=35952)
AVPR1A
AVPR2
tubuloids
iPSC derived organoids
Percent Expressed
0
2
4
6
Average Expression
2
1
0

Fig. 6

a–**d**, Representative images (**a**–**c**) and quantification of cyst size (**d**) in EV, *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$tubuloids treated with AVP versus control. Two-way ANOVA with Bonferroni's post hoc test in $n = 12$ (mean ± s.e.m.), $*P = 0.0165$ for EV + AVP versus control; $****P < 0.0001$ for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ + AVP versus control (**d**). Scale bars, 200 µm (**b**,**c**) and 100 µm (**a**). **e**, Quantification of cAMP in EV, *PKD1*$^{-/-}$and *PKD2*$^{-/-}$ tubuloids treated with tolvaptan (0.1–40 µM) or DMSO (0.0) as control. Two-way ANOVA with Tukey's post hoc test in $n = 3$ (mean ± s.d.), $P = 0.9779$ in EV, $P = 0.5354$ in *PKD1*$^{-/-}$ and $P > 0.9999$ in *PKD2*$^{-/-}$ versus control with 0.1 µM tolvaptan; $P = 0.4963$ for EV, $P = 0.5975$ for *PKD1*$^{-/-}$ and $P = 0.0956$ for *PKD2*$^{-/-}$ with 1 µM tolvaptan; nonsignificant for 10–40 µM tolvaptan in EV; $P = 0.1868$ for *PKD1*$^{-/-}$and $****P < 0.0001$ for *PKD2*$^{-/-}$ with 10 µM tolvaptan; $*P = 0.0136$ for *PKD1*$^{-/-}$ and $****P < 0.0001$ for *PKD2*$^{-/-}$ with 20 µM tolvaptan; $***P = 0.0007$ for *PKD1*$^{-/-}$ and $**P = 0.0015$ for *PKD2*$^{-/-}$ with 40 µM tolvaptan. **f**–**h**, Representative images (**f**,**g**) and quantification (**h**) of EV, *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$ tubuloids subjected to tolvaptan or DMSO with different durations. Two-way ANOVA with Tukey's post hoc test in $n = 3$ (mean ± s.e.m.); for tolvaptan treatment: 0.0 h: $****P < 0.0001$ for *PKD1*$^{-/-}$ versus EV and $***P = 0.0001$ for *PKD2*$^{-/-}$ versus EV; 24 h: $***P = 0.0003$ for *PKD1*$^{-/-}$ versus EV and $***P = 0.0002$ for *PKD2*$^{-/-}$versus EV; 48–72 h: nonsignificant in cyst size for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ versus EV ($P = 0.4014$ for 48 h and $P = 0.5268$ for 72 h in *PKD1*$^{-/-}$ versus EV; $P = 0.0818$ for 48 h and $P = 0.9403$ for 72 h in *PKD2*$^{-/-}$ versus EV); 24–72 h for DMSO: $****P < 0.0001$ for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ versus EV; comparing 72 h after treatment with before treatment: nonsignificant for EV and $****P < 0.0001$ for *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ (**h**). Scale bars, 200 µm (**f**,**g**). **i**, Quantification of cyst size in iPSC-derived ADPKD organoids (*PKD2* gene editing) treated with tolvaptan or DMSO. $n = 88$ versus $n = 17$ (mean ± s.e.m.), unpaired two-tailed *t*-test. **j**, Expression of *AVPR1A* and *AVPR2* in tubuloids and published scRNA-seq datasets from iPSC-derived organoids.5[41,42,43] For details on statistics and reproducibility, see Methods.

indicated that cysts can also be derived from proximal parts of the nephron (Extended Data Fig. 9f-h), in line with the literature[27,28,29].

We then compared all genes that were differentially expressed in PT_4 and TAL_2 of the human tissue to the cells that were mapped to PT_4 and TAL_2 from the tubuloids (>94% of the tubuloid cells; Extended Data Fig. 10a). Common enriched pathways included MAPK signaling and retinol metabolism, among others (Fig. 5d and Extended Data Fig. 10b). MAPK signaling has been reported to be active in cyst-lining cells, and it has been suggested that cAMP could contribute directly to extracellular signal-regulated kinase (ERK) activation via protein kinase A (PKA), Rap-1 and B-Raf to promote cyst growth[23,30]. Retinoic acid has been demonstrated to induce transcription of *PKD1* (ref. 31), and transgenic mice overexpressing a functional human *PKD1* gene develop renal cysts[32,33].

We next focused on the top common markedly and differentially expressed genes in PT_4 and TAL_2 between human tissue (ADPKD versus healthy donor biopsies) and the tubuloid-derived cells that mapped to human tissue PT_4 and TAL_2 using Symphony. Among the common markedly upregulated genes in human disease and gene-edited tubuloids as compared to controls, we identified *SYNE2*, *PLEKHA1*, *BIRC3*, *RHOU* and *EGR1* (Fig. 5e,f and Supplementary Notes). Common downregulated genes included *ANPEP*, *S LC20A1* and *CDH6*, among others, pointing toward epithelial dedifferentiation (Extended Data Fig. 10c).

To compare the common Reactome terms enriched in PT_4 and TAL_2, we performed an over-representation analysis of the markedly common upregulated genes in gene-edited tubuloids and human disease. This analysis indicated enrichment of terms such as tight junction interaction and cell junction organization in PT_4 and RHO GTPase cycle in TAL_2 (Fig. 5g). Tight junction composition has been reported to be altered in ADPKD[34], and impaired formation of desmosomal junctions has also been reported in ADPKD[35]. Furthermore, work from several groups has demonstrated an important role of the Rho family of GTPases in cystogenesis[24,36].

In summary, our data indicate that *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$ tubuloids resemble human ADPKD cyst formation and cyst morphology and that some of the reported molecular mechanisms of ADPKD are also altered in the *PKD*$^{-/-}$ tubuloids. Of note, ADPKD in humans develops over many years in a heterogeneous environment with immune cells and inflammation as well as mesenchymal cells, perfusion and altered filtration, which we are obviously lacking within the gene-edited tubuloids.

Tolvaptan reduces cyst size in ADPKD tubuloids

Increased intracellular cAMP levels have a central role in ADPKD, and vasopressin promotes intracellular cAMP generation via its AVPR2 receptor. Tolvaptan has been demonstrated to lower cAMP levels as an AVPR2 antagonist, resulting in reduced cyst growth and disease progression[37]. Incubation with arginine vasopressin (AVP) markedly increased the growth of *PKD*$^{-/-}$ cysts (Fig. 6a-d), while we detected a dose-dependent effect of tolvaptan on cAMP levels in *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$ tubuloids (Fig. 6e). Based on these results and cytotoxicity data (Extended Data Fig. 10d), we next used a dosage of 15 µM tolvaptan to study the potential effect on cyst size. We observed a time-dependent effect of tolvaptan treatment on cyst size in both *PKD1*$^{-/-}$ and *PKD2*$^{-/-}$ tubuloids (Fig. 6f-h). Notably, we did not observe an effect of tolvaptan treatment on iPSC-derived ADPKD cysts (Fig. 6i), in line with published experiments[38]. The reason for this might be that AVPR2, the primary target of tolvaptan, is expressed in tubuloids, while it is not expressed in iPSC-derived kidney organoids (Fig. 6j and Extended Data Fig. 10e,f). Interestingly, we detected an increased expression level of AVPR2 in *PKD1*$^{-/-}$ or *PKD2*$^{-/-}$ tubuloids as compared to control tubuloids (Extended Data Fig. 10g), in line with studies by Torres et al.[39]. and others[40] showing upregulation of AVPR2 in ADPKD. Single-molecule fluorescence in situ hybridization and immunostaining suggested widespread AVPR2 expression in cyst-lining cells in human ADPKD tissue (Extended Data Fig. 10h-j).

Discussion

Tubuloids have been recently reported as a tool to study human kidney epithelial homeostasis and disease6. However, their exact cellular source in the human kidney remained unclear. In this study, we report that kidney tubuloids originate from CD24+ cells. Notably, we could demonstrate that other renal cells are not able to generate tubuloids, while CD24+ cells from the PT and LOH have the overall highest organoid formation capacity and outcompete CD24+ cells from other nephron parts in long-term tubuloid cultures. Human CD24+ cells maintain low rates of oxygen metabolism in vitro and display a distinct gene regulatory program with increased accessibility of various genes that have previously been associated with a progenitor-like phenotype. While early-stage tubuloids contained proximal and distal tubule cells with a potential minor contribution of PECs, only cells from the S3 part of the PT and the downstream thin limb of the LOH ultimately expanded and formed the vast majority of late-stage tubuloids with features of a functional and polarized tubule. We further demonstrate the use of a four-phase tubuloid protocol that results in a higher organoid formation rate and more tubules within a given organoid as compared to the published one-phase protocol. Tubuloids generated with the four-phase protocol presented here contain different parts of the tubule with PT, LOH and DCT at early time points, while we only identified PT and thin limb of the LOH at late time points. The original tubuloid paper10 reports similar findings and also the presence of collecting duct cells, which we did not observe. This difference might be due to the use of different marker genes for cell type annotation and is likely also due to the use of different protocols for generation of tubuloids from purified CD24+ cells in our study compared to culture of tubular fragments by Schutgens et al.[10].

hPSC-derived kidney organoids have recently emerged as a tool for disease modeling. However, these organoids still contain various cell types that lack kidney-specific differentiation (off-target). Transcriptionally, they also appear to represent an earlier developmental stage than the adult human kidney, as recently shown by scRNA-seq[5,20]. Therefore, it is questionable whether disease modeling in hPSC-derived kidney organoids generated by currently available protocols sufficiently mimics certain features of the adult human kidney situation and thus can serve for disease modeling, target identification and validation approaches. Kidney tubuloids derived from CD24+ cells may be more useful to model features of diseases such as ADPKD that originate from the tubule epithelium and thus can become a valuable translational tool to study disease mechanisms and identify new therapeutics. Indeed, comparing snRNA-seq data with human ADPKD and control tissue and scRNA-seq data of gene-edited PKD1$^{-/-}$ or PKD2$^{-/-}$ tubuloids suggests specific similarities within important pathways associated with cystogenesis. Furthermore, we could demonstrate that tolvaptan treatment reduces cyst size in PKD$^{-/-}$ tubuloids, while it does not affect cyst size in iPSC-derived PKD$^{-/-}$ kidney organoids.

Taken together, our data indicate that CD24+ cells are the source of kidney tubuloids and that these adult kidney tubuloids represent an advanced model of adult human polycystic kidney disease that will hopefully be useful for the development of interventional strategies.

Methods

Ethics statements and patient tissue collection

The study complies with all relevant ethical regulations and was approved by the ethical board of the RWTH Aachen University (EK016/17) and the Erasmus Medical Center, Rotterdam (no. 196.927/2000/235, MEC20130-188). For full details on patients, see Supplementary Notes.

Plasmid construction

The LentibbCas9v2eGFP vector was assembled using LentiCRISPRv2GFP originally developed by David Feldser's laboratory (82416, Addgene) as a backbone. We used paired gRNAs for targeting exon 36 and 37 of the PKD1 gene or exon 1 of the PKD2 gene. Lentiviral paired gRNA CRISPR–Cas9 engineering assembly was performed similarly to recent publications44 with some modifications: first, gRNA-1 or gRNA-a was cloned into pX330 expression vectors, provided by Feng Zhang's laboratory (42230, Addgene); gRNA-2 or gRNA-b was introduced into the ph7SK-gRNA expression vector, developed by Charles Gersbach's laboratory (53189, Addgene); and fragments containing gRNA-1 and gRNA-2 expression cassettes (or fragments including gRNA-a or gRNA-b expression cassette) were then simultaneously transferred into upstream EFS promoter of the LentibbCas9v2eGFP vector by Golden Gate assembly. For full details, see Supplementary Notes.

Preparation of Wnt3a and RSPO1 conditioned medium

Preparation of Wnt3a and RSPO1 conditioned medium was performed similarly to previous studies45,46. For full details, see Supplementary Notes.

Isolation of cells from human kidneys

Human nephrectomy tissue specimens were used to establish CD13+ and CD24+ primary cell culture and tubuloids. The tissue was dissected and minced. The fragments were digested with 1 mg ml−1 of collagenase (C-4-22, Millipore) in DMEM/F12 medium with DNase I (D5025, Sigma) plus 1:50 Liberase (540102001, Roche) for 45 min at 37 °C with shaking at 160 r.p.m. using a thermal shaker. The digested fragments were passed through 70-µm and 40-µm cell strainers (352350 and 431750, Corning). MACS isolation of CD24+ cells was performed using the CD24 microBeads kit (130-095-951); CD13+ cells were first labeled by biotin-conjugated mouse anti-human CD13 antibody (130-119-572) and then anti-biotin microBeads (130-090-485, Miltenyi Biotec). CD24+ cells or CD13+ cells were purified by positive selection with LS columns (130-042-401, Miltenyi Biotec). For FACS, the freshly isolated kidney single-cell suspension was first incubated with 5 µl of Human TruStain FcX (422302, Biolegend) in 100 µl of cell suspension at room temperature for

10 min and then incubated with Pe/Cy7 or BV421 anti-human CD24 antibody (311120 and 311122, Biolegend) and PE anti-human CD13 antibody (301704, Biolegend). Following this, the cells were washed twice with cell staining buffer (420201, Biolegend), resuspended in 500 µl of cell staining buffer and sorted using an SH800 Cell Sorter (Sony, Biotechnology). For full details, see Supplementary Notes.

Primary kidney tubular cell culture

Isolated CD24+ cells or CD13+ cells were seeded to T25 flasks in advanced DMEM/F12 medium supplemented with 20 ng ml–1 EGF (AF-100-15, Peprotech) and 500 ng ml–1 insulin (8923023, Sanofi) plus 1% B27 minus vitamin A, 1% penicillin/streptomycin, 1% l-glutamine and 20 mM HEPES (Thermo Fisher Scientific). The primary cell cultures were split after 7–8 d. The P2 cells were used for various experimental analyses.

Cell viability, metabolism assays and ATAC-seq

To optimize ATAC-seq library preparation with decreased mitochondrial DNA content and low cell numbers, we developed a two-step lysis method. CD24+ or CD13+ cells (1,000–8,000) were FACS sorted from human kidneys and centrifuged at 500g for 5 min. Pellets were then resuspended in 50 µl ice-cold hypotonic buffer, incubated for 3 min on ice and centrifuged at 500g for 9 min. Pellets were lysed in 50 µl lysis buffer plus 0.01% digitonin, centrifuged at 500g for 9 min and resuspended in 50 µl of a transposase reaction mix, including 25 µl 2XTD buffer, 0.5 µl tagment DNA enzyme 1 and 24.5 µl nuclease-free water. The transposition reaction was incubated at 37 °C for 30 min. Following this, the transposed DNA was purified and eluted in 15 µl nuclease-free water. Transposed DNA was amplified by two rounds of PCR using NEBNext 2× Master mix with custom Nextera PCR primers. The quality of the library was checked by Agilent D1000 ScreenTape on the 2200 TapeStation system. The ATAC-seq libraries of CD24+ or CD13+ primary kidney cells were loaded on an Illumina NextSeq 500 for 75-bp paired-end sequencing. For full details, see Supplementary Notes.

Adult kidney tubuloid culture

Freshly purified cells were plated into single wells of 12-well plates with 50% Wnt3a conditioned medium in advanced DMEM/F12 medium supplemented with 50 ng ml–1 EGF, 5 ng ml–1 Noggin (120-10C, Peprotech), 10 µM Y27632 (S1049, Selleckchem) and 2% B27 minus vitamin A. On day 4, single cells were prepared using Accutase (A6964, Sigma) and resuspended at 5 × 104 cells in 50 µl of 10% RSPO1 conditioned medium plus 50 ng ml–1 EGF, 5 ng ml–1 Noggin, 10 µM Y27632 and 2% B27 minus vitamin A with 150 µl of Matrigel (356231, Corning) on ice. The cell–Matrigel mixture was transferred into tissue culture plates (40 µl per well in 24-well plates or 25 µl per well in a µ-Slide 8-well chamber (80826, ibidi)). After 36–48 h, the RSPO1/EGF/Noggin conditioned medium was replaced by organoid differentiation medium composed of 10% RSPO1 conditioned medium supplemented with 100 ng ml−1 FGF10 (100-26, Peprotech) and 50 ng ml−1 EGF and cultured for 15 days, with a medium exchange every 2–3 d. For long-term culture, the tubuloids were grown in maintenance medium consisting of advanced DMEM/F12 supplemented with 50 ng ml−1 EGF, 50 ng ml−1 FGF2, 100 ng ml−1 IGF1, 500 ng ml−1 insulin and 2 % B27 minus vitamin A. For full details, see Supplementary Notes.

iPSC-derived kidney organoid differentiation and staining

hPSC stocks were maintained in mTeSR1 medium with daily medium changes and weekly passaging using Accutase or ReLeSR (STEMCELL Technologies, Vancouver). For differentiation into organoids, iPSCs (WTC-11 cell line; Coriell, GM25256) bearing knockout mutations in PKD2 were plated at 2,000 cells per well in 24-well plates or 200 cells per well in 384-well plates, precoated with 300 µl of DMEM-F12 containing 0.2 mg ml−1 Matrigel and sandwiched the following day with 0.2 mg ml−1 Matrigel in 500 µl of mTeSR1 (STEMCELL Technologies, Vancouver) to produce scattered, isolated spheroid colonies. Coating and plating of 384-well plates was performed using a Matrix Wellmate liquid handling robot. Forty-eight hours after sandwiching, hPSC-derived spheroids were treated with 12 µM CHIR99021 (Tocris Bioscience) for 36 h in 1,000 µl of advanced RPMI + 1× Glutamax + Pen-strep (all from Thermo Fisher Scientific) and then changed to RB (Advanced RPMI + 1× Glutamax + 1× B27 Supplement, all from Thermo Fisher Scientific). For full details, see Supplementary Notes.

Real-time RT–qPCR

Total RNA was purified from primary kidney cells or tubuloids using the RNeasy mini kit (74104, Qiagen). Five hundred nanograms of RNA was used as a template to synthesize cDNA with the SuperScript III First-Strand Synthesis System (18080051, Thermo Fisher Scientific). RT–qPCR was performed in quadruplicate with cDNA (1:10 dilution), 300 nM primers and iTAQ SYBR Green Supermix (172-524, Bio-Rad) using the CFX Connect Real-time PCR Detection System (Bio-Rad). GAPDH was used as the housekeeping gene. For full details, see Supplementary Notes.

Transmission electron microscopy

The tubuloids were embedded in pure epon and polymerization of epon was performed at 90 °C for 2 h. Ultrathin sections were cut using an ultramicrotome (Reichert Ultracut S, Leica), and contrast was enhanced by staining with 0.5% uranyl acetate and 1% lead citrate (both EMS). The sections were visualized using an acceleration voltage of 60 kV with a Zeiss Leo 906 transmission electron microscope (Carl Zeiss). For full details, see Supplementary Notes.

P-gp transport assay

Tubuloids were disrupted using a P200 pipette and cultured overnight with 5 µM P-gp inhibitor PSC833 (4042, Tocris) in organoid maintenance medium or 0.2% DMSO in organoid maintenance medium as control. After washing in Hank's buffer (14025050, Gibco), tubuloids were treated with 1 µM calcein-AM (C1430, Invitrogen) in Hank's buffer supplemented with 5 µM P-gp inhibitor or 0.2% DMSO for 1 h at 37 °C. After washing, tubuloids were fixed in 4% paraformaldehyde (PFA), and they were counterstained with 1 µg ml–1 DAPI for 15 min. P-gp transport image analyses were performed on an LSM 710 (Zeiss). For full details, see Supplementary Notes.

PKD gene editing in tubuloids

Lentiviral particles were produced by transfecting 293FT cells (R70007, Thermo Fisher Scientific) grown in 60-mm dishes with 1 µg lentiviral paired CRISPR–Cas9, 750 ng psPAX2 (12260, Addgene) and 250 ng pMD2.G (12259, Addgene) using TransIT-LT1 (MIR 2300, Mirusbio). The filtered virus-containing supernatants were used to infect tubuloids. For full details, see Supplementary Notes.

Western blotting

Cells were lysed in RIPA buffer and boiled at 95 °C for 5 min. After quantification, they were loaded onto 4–15% mini-Protean TGX gels (4568086, Bio-Rad) and transferred to PVDF membranes (162-0177, Bio-Rad). The membranes were blotted for PC1 (rabbit polyclonal antibody, 1:2,000, ABT128, Millipore) or PC-2 (mouse monoclonal antibody, 1:1,000, sc-47734, Santa Cruz Biotechnology). GAPDH was used for a loading control (mouse monoclonal antibody, 1:2,000, NB300-221, NovusBio). For full details, see Supplementary Notes.

Cytogenesis assays

PKD1–/–, PKD2–/– and EV tubuloids were plated in 24-well plate or µ-Slide 8-well chamber as above and cultured in a 3D culture system with EGF, FGF2, IGF1 and insulin in RSPO1 conditioned medium. Cysts were counted at ×4 magnification (smaller cysts were confirmed by ×10 or ×20 magnification) using a Nikon Eclipse Ts2 inverted microscope on day 21 after initiating 3D culture. Chemical cyst formation was stimulated by liquid handling robots as described previously19,20 with some modifications. In detail, tubuloids were cultured in 3D with EGF/FGF2/IGF1/insulin conditioned medium and moved to 48-well plates in suspension culture on days 10, 15 and 20 of 3D culture after transduction, respectively. Then, 5 µM forskolin or 10 µM blebbistatin was added to the culture in suspension and cultures were incubated for 72 h with a medium exchange every 2–3 d. The cysts were imaged using a Nikon Eclipse Ts2 inverted microscope, and cyst size was quantified using Fiji-ImageJ. For full details, see Supplementary Notes.

Immunostaining and imaging

For whole-mount staining of tubuloids, samples were washed in PBS and fixed with 4% PFA in a μ-Slide 8-well chamber for 20 min. After blocking in 5% normal donkey serum at room temperature for 1 h, tubuloids were incubated with primary antibodies in antibody dilute buffer (0.3% Triton X-100 and 1% BSA in PBS) at 4 °C overnight, followed by washing and incubation at 4 °C overnight with the secondary antibodies and subsequent DAPI staining (1 mg ml–1 at 1:1,000). For immunofluorescence staining of paraffin sections, the sections were deparaffinized and rehydrated and then antigen retrieval was performed using a microwave in 1× Antigen Unmasking Solution (H-3300, Vector labs) followed by staining using standard protocols. For full details, see Supplementary Notes.

RNA in situ hybridization

In situ hybridization was performed as previously described47. For full details, see Supplementary Notes.

Cell viability assay

The CellTiter-Glo 3D cell viability assay was used as previously described with some modifications48. For full details, see Supplementary Notes.

cAMP assay

PKD–/– tubuloids and EV tubuloids were cultured in organoid growth medium supplemented with 1 μM AVP. After a 3-day incubation with AVP, five different doses (0.1–40 μM) of tolvaptan dissolved in DMSO or DMSO (vehicle) was added in triplicate. To measure intracellular cAMP level, PKD–/– tubuloids and EV tubuloids treated for 72 h were collected and pelleted at 300g for 5 min and then incubated in 500 μl of Cell Recovery Solution on ice for 1 h. The cells were counted in organoid growth medium via a hemocytometer and subsequently washed twice in PBS. Each condition received 1 × 105 cells transferred into tubes of 1.5 ml, with cells washed three times in PBS, resuspended in 125 μl of 1× cell lysis buffer and frozen at –20 °C. Following two additional freeze-thaw cycles, cells were centrifuged at 600g (4 °C) for 10 min and the supernatants were carefully transferred into 96-well plates. The cAMP levels were measured at 450 nm in a plate reader (Cytation 5, Biotek) and assessed according to the manufacturer's instructions for the cAMP Parameter assay kit (KGE002B, R&D Systems).

Time-course imaging of cyst size after tolvaptan treatment

PKD1 and PKD2 gene-edited tubuloids and EV tubuloids were cultured in organoid growth medium supplemented with 1 μM AVP as described above. After a 3-day incubation of tubuloids with AVP, the tubuloids were treated with tolvaptan or DMSO (vehicle) in triplicates. For time-course imaging, the tubuloids were treated with 15 μM tolvaptan or DMSO plus 1 µM AVP in triplicates. Images were taken on an inverted Nikon ECLipse, Ts2-FL immediately before treatment and at 24-, 48- and 72 h after treatment.

scRNA-seq of tubuloids

Single-cell suspensions of tubuloids and primary human kidney cells were run on a Chromium Single Cell Chip kit with subsequent library preparation using 10x Genomics reagents (PN-120236, PN-120237, PN-120262). The library quality was determined using D1000 ScreenTape on the 2200 TapeStation system (Agilent Technologies). Sequencing was performed on an Illumina Novaseq platform using S1 and S2 flow cells. For full details, see Supplementary Notes.

Nuclei isolation

For snRNA-seq, nuclei were isolated with Nuclei EZ Lysis buffer (NUC-101, Sigma-Aldrich) supplemented with protease inhibitor (Roche) and RNase inhibitor (AM2696, Life Technologies). Samples were homogenized using a Dounce homogenizer (885302-0002, Sigma-Aldrich) in 1 ml of ice-cold Nuclei EZ Lysis buffer and incubated on ice for 2 min with an additional 1 ml of lysis buffer. The pellet was resuspended and washed with 4 ml of buffer. Following centrifugation, the pellet was resuspended in nuclei suspension buffer (1× PBS, 1% BSA, 0.1% RNase inhibitor). For full details, see Supplementary Notes.

Genotyping of PKD1 and PKD2 genes from ADPKD kidney tissue

For analysis of the PKD1 (NM_001009944.3) and PKD2 (NM_000297.4) genes, a custom-targeted next-generation sequencing panel was used. Library preparation was done with the Lotus DNA Library Prep Kit, and a probe-based capture protocol was used to enrich target regions (IDT, Custom Gene Panel). Subsequent sequencing of pooled libraries was performed on a MiSeq sequencing platform (Illumina). Annotation and bioinformatic prioritization of variants was performed using KGGSeq (v1.0; http://pmglab.top/kggseq/index.htm). For full details, see Supplementary Notes.

Analysis of ATAC-seq data

Generation of fastq files and adaptor removal were completed with the Illumina software bcl2fastq (v2.20) (https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html). Library complexity was evaluated with Preseq (version 2.0)49. Sequence quality control was performed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were aligned to the human reference genome (GRCh38/hg38) using bwa (version 0.7.17)50 with default parameters. Mitochondrial reads were filtered out using samtools51 (version 0.1.19). Peak calling was performed using common software MACS2 (version 2.2.5)52. BigWig track files for visualization of peaks in genome browsers were generated using the rtracklayer package (version 1.44.2)53 and normalized with 1 million as a scaling factor. After generating a merged peak file from all samples, peak annotation was performed using ChIPseeker (version 1.20.0)54. For downstream analysis, only peaks overlapping a 1-kb region around a TSS were selected. The featureCounts software from the Rsubread package (version 1.34.7)55 was used to count reads mapping on a region overlapping the selected peaks. Analysis of differential chromatin accessibility was performed using DESeq2 (version 1.24.0)56. For full details, see Supplementary Notes.

Analysis of single-cell and single-nucleus RNA-seq data

Demultiplexing and alignment were performed using the cellranger mkfastq utility with the default/mandatory parameters. FastQ files were aligned to the human reference genome (GRCh38 assembly), and the UMI expression was quantified using CellRanger (10x Genomics, version 3.1). The dataset was analyzed using the Seurat R package (version 3.1.0)57. The cell type assignment among the resulting cell clusters was manually curated. For this, gene specificity scores and gene conditional probabilities for each cluster were obtained using genesorteR (version 0.3.1)58. The human tissue samples were integrated using Harmony (v.1.0) with default parameters and considering every sample library and laboratory as an individual batch59. Differentially expressed genes were tested by pseudobulking expression profiles59. Each subset of pseudobulk profiles was processed separately by a filtering step of lowly expressed genes, Trimmed Mean of M-values normalization60. Then, the count data were fitted using the negative binomial generalized linear model with quasi-likelihood dispersion estimation and tested for differences using the edgeR quasi-likelihood pipeline (v.3.26.7)61. Differentially expressed genes were considered at a false discovery rate of 5% after multiple-testing correction. The enrichment of biological pathways was tested using fgsea (v.1.0.1)62, with an ad hoc collection of gene sets related to ADPKD (MYC targets, mTOR1 signaling and cAMP) from MSigDB63, on the ranking of differential expression of each cystic cell population. Footprint-based pathway activity was estimated using PROGENy64,65, applied on the ranking of differential expression. MAST (v1.10.0) was used to find differentially expressed genes in each major cluster and perform gene set enrichment analysis with KEGG, Reactome, PID and BIOCARTA, and over-representation analysis of upregulated genes in ADPKD cells as compared to control cells as described below. To compare cell types found in the human tissue to the tubuloid cells from the organoids, Symphony (v.0.1.0)26 was used to compress the integrated reference of the human tissue samples, which was created with Harmony before. Ligand–receptor analysis was performed using the CellphoneDB method implemented in LIANA66, followed by Crosstalkr to identify the most relevant ligand–receptor pairs per condition. For full details, see Supplementary Notes.

Statistical analysis and reproducibility

Data are presented as mean ± s.e.m. if not specified otherwise in the legends. A comparison of two groups was performed using an unpaired t-test. For multiple group comparison, one-way ANOVA with Bonferroni's multiple-comparison test or two-way ANOVA with Bonferroni or Tukey's multiple-comparisons test was applied. Statistical analyses were performed using GraphPad Prism 8 (GraphPad Software), and a P value of less than 0.05 was considered significant. The number of samples for each group was chosen on the basis of the expected levels of variation and consistency. The depicted RNAscope, immunofluorescence micrographs and western blot micrographs are representative. All studies were performed at least twice, and all repeats were successful.

Data availability

Processed gene expression values from the scRNA-seq are available at https://doi.org/10.6084/m9.figshare.11786238. Processed data from the ATAC-seq analysis are available for peer review under the private link at https://figshare.com/s/728705bc42446275044d in FigShare. These data have a reserved DOI (https://doi.org/10.6084/m9.figshare.11848281). All raw data are available in the controlled EGA access repository EGAS00001006551. Source data are provided with this paper.

Code availability

The code for reproducible analysis of the single-cell data is available at https://github.com/saezlab/Xu_tubuloid and https://github.com/KramannLab/kidney_human_organoids in GitHub repositories. The computer code to reproduce the analysis of ATAC-seq data is available at https://github.com/ATA82/ATAC_Seq_Xu in GitHub. The specific software and methods used for the analysis are described in the README file of the repository.

APPENDICES

Discussion

Conclusion

Valorisation

CV

Papers

Acknowledgements

Discussion

The advancement of experimental and computational technologies in recent years has enabled the ability to dissect biological systems at the single-cell level, providing unprecedented insights into the complexity and diversity of cellular populations (**Chapter 1**). Single-cell RNA sequencing represents a powerful technology to decode the complex cellular heterogeneity of tissues as the human kidney. Several studies have utilized this technology to characterize the cellular heterogeneity of human healthy kidney tissue, yet a study focusing on diseased human kidney tissue was lacking. Unlike traditional transcriptomics techniques that measure the average expression of genes across many cells, scRNA-Seq allows for the measurement of gene expression at the single-cell level. Kidney fibrosis represents a common consequence of various kidney injuries and diseases and is associated with significant loss of kidney function due to progression of the disease (Duffield et al. 2013; Meng, Nikolic-Paterson, and Lan 2016; Duffield 2014). Understanding the cellular and molecular mechanisms underlying kidney fibrosis is critical for the development of new treatments. Myofibroblasts represent the culprit cell type which promote fibrosis including the deposition of excessive amounts of extracellular matrix components (Falke et al. 2015). However, the origin of myofibroblasts in the kidney has remained unclear. Furthermore, recent advances in multi-omic technologies have made it possible to study the molecular changes that occur in tissues not only based on singlecell data but including spatial coordinates into the data analysis (Moses and Pachter 2022; Chen, Teichmann, and Meyer 2018). Particularly myocardial infarction (MI), also known as heart attack, is a leading cause of death worldwide (Jayaraj et al. 2019). We thought that a better understanding of the molecular changes that occur in the heart during an MI is critical for the development of new treatments for this very important human disease. Next to the spatial information chromatin accessibility plays a critical role in regulating gene expression (Buenrostro et al. 2015; Ma et al. 2020). Through recent developments the study of chromatin accessibility at single-cell level has become available generating data of thousands of cells in an unbiased manner. Even though several computational methods have been developed for the analysis of single-cell ATAC-Seq. including Cell Ranger ATAC, SCRAT (Ji, Zhou, and Ji 2017) and scARCHEs (Lotfollahi et al. 2022) these methods have limitations in their ability to accurately predict chromating accessibility states and linking them to distinct transcription factor activities. Two major studies (**Chapter 3+6**), “Decoding myofibroblast origins in human kidney fibrosis” and “Spatial multi-omic map of human myocardial infarction” we have utilized these above described technologies to gain a deeper understanding of human kidney and heart diseases. In addition, “Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen” (**Chapter 4**) describes a computational method for analyzing whole genome epigenomic data from single-cell ATAC sequencing experiments. The study "Decoding myofibroblast origins in human kidney fibrosis" (**Chapter 3**) used single-cell sequencing techniques to investigate the origins of myofibroblasts, a cell type that plays a key role in the development of kidney fibrosis. Here we performed single-cell transcriptomics and epigenomics sequencing assays on renal tissue samples from both healthy individuals and those with kidney fibrosis. We found that myofibroblasts in fibrotic kidneys were derived from both interstitial fibroblasts and pericytes, rather than solely from fibroblasts as previously thought. Furthermore, we identified distinct epigenetic signatures that distinguished myofibroblasts from fibroblasts and pericytes, providing new insights into the mechanisms of myofibroblast activation.

The study "Spatial multi-omic map of human myocardial infarction" (**Chapter 6**) used a combination of single-cell sequencing and spatial transcriptomics to create a highresolution map of cardiac remodeling after myocardial infarction. We analyzed multiple physiological zones of myocardium from patients with myocardial infarction and controls at different time points using single-cell gene expression, chromatin accessibility and spatial transcriptomics. By integrating data from these different technologies, we were able to evaluate cardiac cell-type compositions at increased resolution and identify disease-specific cardiac cell states.

The study "Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen" (**Chapter 4**) describes a computational method for analyzing data from single-cell sequencing experiments, particularly ATAC-seq (Assay for Transposase- Accessible Chromatin) data. ATAC-seq is a technique for measuring chromatin accessibility, which can provide insight into gene regulation and cellular diversity. The authors of the study demonstrate that scOpen can accurately predict chromatin accessibility states in various cell types and can also identify subtle changes in accessibility associated with different cell states. The method is based on a deep learning model that utilizes a convolutional neural network (CNN) to analyze single cell ATAC-seq data. We trained the model on a large dataset of single-cell ATAC-seq data from multiple cell types and used it to predict chromatin accessibility in new single-cell ATAC-seq datasets. The results of the study show that scOpen can accurately predict chromatin accessibility states in various renal cell types, including proximal tubular cells, podocytes and immune cells like macrophages. We also demonstrate that scOpen can identify subtle changes in accessibility associated with different cell states, such as the transition from a pluripotent to a differentiated state. Thus scOpen, a computational method for estimating chromatin accessibility from single-cell ATAC-seq data, can be used to identify cell types and subpopulations, and to study the dynamics of chromatin accessibility during development or disease.

In addition, single-cell sequencing and spatial sequencing have been used to study the kidney organoids, which are laboratory-grown three-dimensional structures that mimic the structure and function of the kidney. The study "Adult human kidney organoids originate from CD24+ cells and represent an advanced model for adult polycystic kidney disease" (**Chapter 7**) used single-cell sequencing and spatial sequencing to study the kidney organoids and found that they originate from CD24+ cells and can be used as an advanced model for adult polycystic kidney disease.

Understanding the underlying mechanisms of complex diseases, such as heart and kidney diseases, requires the integration of multiple layers of data including genetic, epigenetic, transcriptomic, and proteomic information. However, the sheer volume of data generated by high-throughput sequencing technologies can be overwhelming, making it difficult to extract meaningful insights. This is where multi-omics integration comes into play.

In recent years, there has been a growing interest in using multi-omics integration to generate mechanistic hypotheses. This approach combines information from multiple layers of data in order to gain a more comprehensive understanding of the underlying biology. One study, "Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses" (**Chapter 5**) highlights the importance of integrating data with prior knowledge in order to generate meaningful hypotheses. The study's main focus is on integrating data from multiple sources in order to identify genes that are associated with specific diseases. We used a combination of transcriptomic data, genetic data, and prior knowledge to identify genes that are likely to be involved in the development of heart and kidney diseases. Further we used a machine learning approach to integrate the data and identify genes that are associated with specific diseases. We found that by integrating data with prior knowledge, we were able to identify genes that were not previously known to be associated with the disease. This highlights the importance of integrating data from multiple sources in order to identify new targets for drug development. Furthermore, the study also demonstrated the value of using a causal inference method to integrate the data. This approach allows for the identification of causal relationships between genes and diseases, which can provide a deeper understanding of the underlying biology. This can help to identify new targets for drug development, which can ultimately improve the treatment of heart and kidney diseases. Overall, the study highlights the importance of multi-omics integration in the identification of new targets for drug development. By integrating data from multiple sources with prior knowledge, researchers can gain a more comprehensive understanding of the underlying biology of complex diseases. This can ultimately lead to the development of new and more effective treatments.

Overall, these studies demonstrate the power of single-cell sequencing and spatial transcriptomics in uncovering new insights into human disease. The ability to analyze individual cells and their spatial context allows for a more detailed understanding of cellular heterogeneity and the underlying mechanisms of disease. In addition, the integration of multiple omics data and the use of computational methods like scOpen are essential in the interpretation and understanding of these large and complex datasets. The knowledge gained from these studies will undoubtedly lead to new opportunities for the development of targeted therapies for kidney and heart diseases.

Conclusion

In conclusion, the studies “Spatial multi-omic map of human myocardial infraction” and “Decoding myofibroblast origins in human kidney fibrosis” have provided valuable insights into the mechanisms of human heart and kidney disease. The former study has generated an integrative high-resolution map of human cardiac remodelling after myocardial infarction using single-cell gene expression, chromatin accessibility, and spatial transcriptomic profiling, revealing disease-specific cardiac cell states and their dependencies on other cell types. The latter study has decoded the origins of myofibroblast in fibrosis by using single-cell RNA sequencing, identifying novel myofibroblast subtypes and their lineage relationships.

Additionally, the study “Adult human kidney organoids originate from CD24+ cells” which validated the use of adult human kidney organoids as an advanced model for the study of adult polycystic kidney disease. This study has highlighted the potential of organoids in the study of disease mechanisms and drug development.

Furthermore, the study “Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses” demonstrated the potential of integrating multiomics data with prior knowledge to generate mechanistic hypotheses in disease.

Overall, these studies have highlighted the power of single-cell sequencing and spatial sequencing in understanding disease mechanisms and the potential of integrating multi-omics data with prior knowledge for generating mechanistic hypotheses in disease. These findings pave the way for more advanced studies in the future and hold promise for the development of new and more effective therapies.

Valorisation

The studies "Spatial multi-omic map of human myocardial infraction" and "Decoding myofibroblast origins in human kidney fibrosis" have significant potential for drug discovery and therapeutic development. The use of cutting-edge single-cell sequencing and spatial sequencing techniques, as well as the integration of multiple omics data, allows for a more comprehensive and detailed understanding of the underlying molecular mechanisms of these human diseases for the first time.

One key aspect of the myocardial infarction study is the identification of disease specific cardiac cell states, which can be targeted for therapeutic intervention. For example, it has been recently demonstrated that engineered CAR-T cells directed against the antigen FAP (fibroblast activating protein) on fibroblasts can be used therapeutically to treat heart fibrosis (Ruel et al. Science). Additionally, the study provides an integrative molecular map of human myocardial infarction, which can serve as a valuable reference for the field and pave the way for advanced mechanistic and therapeutic studies of cardiac disease.

The kidney fibrosis study also contributes to the understanding of the underlying molecular mechanisms of disease, specifically in the identification of specific cell populations that drive fibrosis. By pinpointing these myofibroblast origins, new targets for therapeutics can be identified and developed. NKD2, a WNT-regulator, has already been identified and using in-vitro organoid models showed involvement in regulating fibrosis pathways. Furthermore, the study highlights the potential of using organoids as a model for disease, which can aid in drug discovery and development.

The "Adult human kidney organoids originate from CD24+ cells and represent an advanced model for adult polycystic kidney disease" and COSMOS paper also provide insights into the use of organoids in disease modeling, specifically in the case of polycystic kidney disease. This can provide a valuable tool for drug discovery and understanding disease progression.

Additionally, the "Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses" demonstrates the importance of integrating multiple omics data with prior knowledge to generate more accurate mechanistic hypotheses, which can aid in the identification of new drug targets. Using such models based on spatially resolved multi-omics data will be an important step forward towards reaching personalized medicine not only in the field of cancer but also other fields like nephrology and cardiology.

Overall, these studies demonstrate the power of cutting-edge technologies and multiomics data integration in advancing our understanding of human diseases, leading to new opportunities for drug discovery and therapeutic development.

References

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." Nature 523 (7561): 486–90.

Chen, Xi, Sarah A. Teichmann, and Kerstin B. Meyer. 2018. "From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture." Annual Review of Biomedical Data Science, July. https://doi.org/10.1146/annurev-biodatasci-080917-013452.

Duffield, Jeremy S. 2014. "Cellular and Molecular Mechanisms in Kidney Fibrosis." The Journal of Clinical Investigation 124 (6): 2299–2306. Duffield, Jeremy S., Mark Lupher, Victor J. Thannickal, and Thomas A. Wynn. 2013. "Host Responses in Tissue Repair and Fibrosis." Annual Review of Pathology 8 (January): 241–76.

Falke, Lucas L., Shima Gholizadeh, Roel Goldschmeding, Robbert J. Kok, and Tri Q. Nguyen. 2015. "Diverse Origins of the Myofibroblast—implications for Kidney Fibrosis." Nature Reviews. Nephrology 11 (4): 233–44.

Jayaraj, Joshua Chadwick, Karapet Davatyan, S. S. Subramanian, and Jemmi Priya. 2019. "Epidemiology of Myocardial Infarction." Myocardial Infarction 10.

Ji, Zhicheng, Weiqiang Zhou, and Hongkai Ji. 2017. "Single-Cell Regulome Data Analysis by SCRAT." Bioinformatics 33 (18): 2930–32.

Lotfollahi, Mohammad, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, et al. 2022. "Mapping Single- Cell Data to Reference Atlases by Transfer Learning." Nature Biotechnology 40 (1): 121–30.

Ma, Sai, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, et al. 2020. "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin." Cell 183 (4): 1103–16.e20.

Meng, Xiao-Ming, David J. Nikolic-Paterson, and Hui Yao Lan. 2016. "TGF-β: The Master Regulator of Fibrosis." Nature Reviews. Nephrology 12 (6): 325–38.

Moses, Lambda, and Lior Pachter. 2022. "Museum of Spatial Transcriptomics." Nature Methods, March. https://doi.org/10.1038/s41592-022-01409-2.

Curriculum Vitae

Personal Data

Title	PD Dr. med., MHBA
First name	Christoph
Name	Kuppe
Current position	1. Emmy-Noether Research Group Leader, UKA 2. Attending Physician, Department of Nephrology, UKA
Current institution/site	Department of Internal Medicine and Nephrology, Institute of Experimental Medicine and Systems Biology, University Center RWTH Aachen, Germany
Identifiers/ORCID	0000-0003-4597-9833

Qualifications and Career

Stages		**Periods and Details**
Degress		
School	2001 - 2004	High School degree with distinction "Best of class 2004", Mariann-Hiller Missionary Gymnasium, Germany
	2001 - 2002	Cody High School, Parliamentarian Exchange Program, Cody, Wyoming, USA
University	2004 – 2010	School of Medicine, RWTH Aachen University, Germany
Board Exam	2010	State board examination in Medicine, Germany
Doctorate	2011	Medical Doctoral, MD, Department of Nephrology, RWTH Aachen University, Prof. Dr. J. Floege/Prof. Dr. M. Moeller; disputation: summa cum laude, (highest distinction)
Habilitation	2021	Venia legendi Internal Medicine and Nephrology, Pathomechanisms of glomerular and tubule-interstitial kidney fibrosis, University Medical Center Aachen, RWTH Aachen, Germany; supervisor: Prof. R. Kramann

Professional Carrer – Previous academic positions and career grants

2012 – 2016	Postdoctoral Research Fellow, Department of Nephrology Aachen University, Prof. Moeller and visiting scientist Department of Nephropathology, Prof. Gröne, DKFZ, Heidelberg
2018 – 2021	Postdoctoral Research Fellow, Prof. R. Kramann, RWTH Aachen, supported by DGIM Clinician Scientist fellowship

2021	DFG Emmy Noether Program – group leader, functional genomics on kidney disease, RWTH Aachen University, Institute of Experimental Medicine and System Biology
2022	ERC Starting grant, DECODE-DKD, diabetic kidney disease
2023	EKFS Clinician Scientist Professorship, Cellular Dynamics and Translational Systems Biology, RWTH Aachen University, Institute of Experimental Medicine and System Biology

Professional Career – Clinical education

2011 – 2017	Residency, nephrology fellowship, Department of Medicine II, Nephrology, Prof. Floege, UKA
2017	Board certification for Internal Medicine and Nephrology
2017 – 2019	Master of Health and Business Administration, University of Erlangen-Nürnberg, Germany
2022 – today	Senior Attending, Department of Medicine II, Nephrology, UKA

Supplementary Career Information

I am a nephrologist by training with strong background in basic and translational research. For my medical thesis I was honored with the Borchersmedal of the RWTH and received the prestigious Georg-Haas prize from the German Society of Nephrology. I transitioned from mouse model/lineage tracing-based research to using a multi-omic approach to study disease progression in the kidney. I recently employed single-cell genomics and spatial multiomics methods and computational approaches to decode the cellular and gene regulatory complexity of human kidney disease and myocardial infarction. In the next years, the starting phase of my own independent research group, I will focus on using cutting edge technologies (single cell genomics, spatial multi-omics, CRISPR screens) to develop predictive and interpretable models CKD. One of the main goals is to understand how cell-fate decisions are encoded by the transcriptome and which include intrinsic and extrinsic signals have on the fate decision of cells in disease. The overall aim is that my research lab fosters translation and generates a real patient benefit.

Engagement in the Research System

2023 – present selected as **Editorial Board member KI-ISN** (Kidney International Journal and International Society of Nephrology) designed for early career professionals and academics working in the field of nephrology, dialysis and transplantation.

Reviewing activities and participation in boards and panels

Journal Reviewer: >10 journals including Nature Communications, American Journal of Nephrology, Kidney International, Scientific Reports, PNAS

Thesis Reviewer (PhD): RWTH Aachen, DE; University Clinic Hamburg Eppendorf, DE

Grant and Abstract Reviewer: German Research Foundation (DFG), Israel Science Foundation (ISF)

Memberships of scientific societies

2022 Member of the International Society of Nephrology (ISN)
2021 Fellow of the American Society of Nephrology (FASN)
2020 Member of the European Renal Association (ERA-EDTA)
2015 Member of the German Society of Internal Medicine (DGIM) and German Society of Nephrology (DGFN)

Presentations at national and international scientific meetings (selection)

2023 Dutch-German Meeting on Translation Cardiology, Würzburg, Germany
2022 Epigenomics of Common Diseases, WCS Conferences, Cambridge, UK
2022 Introduction to single cell and spatial omics, AHA Meeting, Chicago, USA
2021 Multi-Level Single-Cell Omics of Human Myocardial Infarction and Cardiac Fibrosis, AHA, virtual
2021 ESC Congress 2021, Single cell and spatial biology of human heart, virtual
2021 New insights in renal fibrosis using functional genomics, DGFN, Germany
2020 DGK, Virtual Cardiology Symposium, Spatial Multiomics in human MI

Teaching activities

2018-today	Nephro Skills Course for 5th year medical students
2020-today	Physical Examination for medical students
2018-today	Clinical skills Lab: Dialysis and CKD, ICU Class, RWTH Aachen
2017-2018	ICU Medicine, Division of Gastroenterology, RWT Aachen University,
2012-2016	Interactive Patient cases with simulations, AIXTRA Clinical Skills Lab, RWTH Aachen University
2011-2017	Nephro Skills Course and Kidney research class, Aachen University
2011-2017	Clinical Investigation in Nephrology, RWTH Aachen University, Aachen
2008-2009	Preparation Course Anatomy, RWTH Aachen Medical School

Supervision of Researchers in Early Career Phases

Xiaoting Zhang, M.D. thesis, 2018-2021, now physician scientist China
Xian Liao, M.D. thesis candidate, 2020- ongoing
Paul Kießling, PhD candidate, 2022- ongoing
Osman Goni, PhD, 2023- ongoing
Emilia Scheidereit, M.D. thesis candidate, 2022- ongoing
Feng Zihao, M.D. thesis candidate, 2023- ongoing

Academic Distinctions

2023 Theodor-Frerichs-Price of the German Society of Internal Medicine
2022 Carl-Ludwig-Award of the German Society of Nephrology (DGfN)
2022 Life-Science Bridge Award Aventis Foundation
2021 Bernd-Sterzel Award for Basic Research in Nephrology, (DGfN)
2021 Fellowship of the American Society of Nephrology (FASN)
2014 Stipend of Novartis Foundation
2012 Borchersmedal of the RWTH Aachen University
2011 Georg Haas Prize, German Kidney Disease Centers e.V.

Publications

Published articles: 50
H-Index: 21
Citations: 214
Top Ten Publications:

1. Xu Y*, **Kuppe C***, Perales-Patón J, et al. Adult human kidney organoids originate from CD24+ cells and represent an advanced model for adult polycystic kidney disease. ***Nat Genet.*** 2022 Nov;54(11):1690-1701. doi: 10.1038/s41588-022-01202-z. Epub 2022 Oct 27. PMID: 36303074

2. **Kuppe C***, Ramirez Flores RO*, Li Z*, et al. Spatial multi-omic map of human myocardial infarction. ***Nature.*** 2022 Aug;608(7924):766-777. doi: 10.1038/s41586-022-05060-x. Epub 2022 Aug 10. PMID: 35948637.

3. Abdelbary MMH*, **Kuppe C***, Michael SS, Krüger T, Floege J, Conrads G. Impact of sucroferric oxyhydroxide on the oral and intestinal microbiome in hemodialysis patients. ***Sci Rep.*** 2022 Jun 10;12(1):9614. doi: 10.1038/s41598-022-13552-z. PMID: 35689007.

4. Li Z*, **Kuppe C***, Ziegler S, Cheng M, Kabgani N, Menzel S, Zenke M, Kramann R, Costa IG. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. ***Nat Commun.*** 2021 Nov 4;12(1):6386. doi: 10.1038/s41467-021-26530-2. PMID: 34737275.

5. Dugourd A*, **Kuppe C***, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, Vieira V, Bekker-Jensen DB, Kranz J, Bindels EMJ, Costa ASH, Sousa A, Beltrao P, Rocha M, Olsen JV, Frezza C, Kramann R, Saez-Rodriguez J. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. ***Mol Syst Biol.*** 2021 Jan;17(1):e9730. doi: 10.15252/msb.20209730. PMID: 33502086.

6. **Kuppe C**, Perales-Patón J, Saez-Rodriguez J, Kramann R. Experimental and computational technologies to dissect the kidney at the single-cell level. ***Nephrol Dial Transplant.*** 2022 Mar 25;37(4):628-637. doi: 10.1093/ndt/gfaa233.

7. **Kuppe C***, Ibrahim MM*, Kranz J, Zhang X, Ziegler S, Perales-Patón J, Jansen J, Reimer KC, Smith JR, Dobie R, Wilson-Kanamori JR, Halder M, Xu Y, Kabgani N, Kaesler N, Klaus M, Gernhold L, Puelles VG, Huber TB, Boor P, Menzel S, Hoogenboezem RM, Bindels EMJ, Steffens J, Floege J, Schneider RK, Saez-Rodriguez J, Henderson NC, Kramann R. Decoding myofibroblast origins in human kidney fibrosis. ***Nature.*** 2021 Jan;589(7841):281-286. doi: 10.1038/s41586-020-2941-1. Epub 2020 Nov 11. PMID: 33176333.

8. Tajti F*, **Kuppe C***, Antoranz A, Ibrahim MM, Kim H, Ceccarelli F, Holland CH, Olauson H, Floege J, Alexopoulos LG, Kramann R, Saez-Rodriguez J. A Functional Landscape of CKD Entities From Public Transcriptomic Data. ***Kidney Int Rep***. 2019 Nov 13;5(2):211-224. doi: 10.1016/j.ekir.2019.11.005. Feb. PMID: 32043035

9. **Kuppe C**, Leuchtle K, Wagner A, Kabgani N, Saritas T, Puelles VG, Smeets B, Hakroush S, van der Vlag J, Boor P, Schiffer M, Gröne HJ, Fogo A, Floege J, Moeller MJ. Novel parietal epithelial cell subpopulations contribute to focal segmental glomerulosclerosis and glomerular tip lesions. ***Kidney Int***. 2019 Jul;96(1):80-93. doi: 10.1016/j.kint.2019.01.037. Epub 2019 Feb 27. PMID: 31029503.

10. **Kuppe C**, Rohlfs W, Grepl M, Schulte K, Veron D, Elger M, Sanden SK, Saritas T, Andrae J, Betsholtz C, Trautwein C, Hausmann R, Quaggin S, Bachmann S, Kriz W, Tufro A, Floege J, Moeller MJ. Inverse correlation between vascular endothelial growth factor back-filtration and capillary filtration pressures. ***Nephrol Dial Transplant***. 2018 Sep 1;33(9):1514-1525. doi: 10.1093/ndt/gfy057. PMID: 29635428.

Acknowledgements

I am deeply indebted and thankful to countless individuals who have journeyed with me, provided guidance, lent support, kindled motivation, brought joy, and instilled inspiration during my challenging exploration of science and laboratory life over the past few years.

Dear Professor Floege, since the inception of my residency in your division back in 2011, your exceptional support and mentorship have been a constant and invaluable part of my journey. You've enabled me to seamlessly blend clinical practice in internal medicine and nephrology while concurrently launching a career in basic and translational research. Your teachings on how to craft a paper and present scientific work to a wider audience, have been monumental in my development. Your swift and insightful suggestions regarding grant and paper writing have been and continue to be incredibly beneficial. Your assistance has significantly honed my paper writing and grant application skills. Every time I prepare for a talk, I recall your presentation on "how to ruin your presentation," which helps me maintain a coherent narrative and avoid overloading each slide with information. Your guidance in our meetings, discussions, and throughout my clinical and scientific career has been priceless. I am profoundly thankful for all of this and deeply appreciate your continued mentorship.

Dear Professor Kramann, Dear Rafael,
I vividly remember the apprehension I felt prior to our initial meeting discussing single-cell RNA sequencing of human tissues in 2018. I was completely convinced that becoming a part of your lab and receiving mentorship from you would be the most beneficial step in my scientific journey. Your guidance during my years in your lab has been nothing short of extraordinary. I deeply appreciate the openness of your lab, allowing me to engage in groundbreaking experiments. I am profoundly grateful for all our scientific discussions and your guidance in my scientific career. The strength and confidence you instilled in me for science, for designing experiments, and for testing hypotheses, have made me more secure in my ambitions and in my path to reach them. Your lab has struck the perfect balance, both scientifically and interpersonally. It's a place where everyone collaborates without friction, where constant laughter mixes with the occasional drama, and wonderful extracurricular activities abound. Your patience and dedication, evident in the countless nights you spent training me initially on Fluorescence-Activated Cell Sorting (FACS), have been a testament to your commitment as a mentor. I want to express my deepest gratitude for giving me the opportunity to work in your lab and for guiding me to where I stand today. I am sincerely thankful for your ongoing mentorship and support.

Dear Prof Schurgers, dear Leon, thanks for being incredibly helpful and supporting my start in Maastricht and for supporting my PhD. I am joyfully looking back at our first meetings in Maastricht shortly after starting my own lab in Aachen and your support and guidance. I am looking forward to various fascinating collaborative projects together with you.

To my beloved family, specifically my parents, my sister Annika and brother Thomas, I extend my heartfelt thanks for your love and steadfast support through the years of my education and academic pursuits. I am certain that without your immense backing, I would not have reached where I am today. Your consistent readiness to offer help, open your doors, and lend a listening ear has been a source of immense comfort to me. For all of this, I am profoundly grateful!

To Jennifer, my dearly beloved wife – my achievements and where I stand today would not have been possible without you! They are deeply intertwined with your presence in my life and constant support. Our connection, our love, excessed what can be expressed in words.

To Emma and Frieda our wonderful children. You are both the perfect antidote to the stresses of academic journeys. A smile or hug from you both has been a most powerful motivator. Mummy and Daddy love you more than words can express.