Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. Hyojin Kim

born in: Seoul, South Korea

Oral examination: 08-11-2021

Transcriptomics data analysis from renal fibrosis

Referees: Prof. Dr. Benedikt Brors

Prof. Dr. Julio Saez-Rodriguez

Abstract

Around 10 % of the world population suffers from chronic kidney disease. While the initial stimulus of kidney injury may vary, fibrosis represents the common end-stage of nearly all kidney diseases. However, the pathogenesis of renal fibrosis remains not well understood due to the complexity of the kidney tissue, heterogeneity of kidney cells as well as high heterogeneity across patients. The kidney is one of the most complex organs and consists of a multitude of different cell types such as podocytes, proximal tubular cells, distal tubular cells, endothelial cells, pericytes, fibroblasts and myofibroblasts. Myofibroblasts have been previously identified as the central conductors in fibrosis. After kidney injury, myofibroblasts expand and produce excess extracellular matrix, which in return can lead to pathologic tissue remodeling and loss of kidney function.

In order to better understand the pathogenesis of renal fibrosis on a cellular level, I analyzed single-cell RNA sequencing (RNA-seq) data of renal perivascular cells, which were isolated by fluorescence activated cell sorting (FACS) using the cell markers Gli1, Ng2, Myh11, Pdgfrb and Cd31. FACS, as well as single cell library generation was performed in collaboration by Christoph Kuppe, MD, PhD of the RWTH Uniklinik Aachen. The cell markers Gli1, Ng2, Myh11, Pdgfrb are common markers for identification of fibroblasts, pericytes, endothelial cells and epithelial cells, while Cd31 is used as a marker of endothelial cells within the perivascular niche of the kidney. Based on the sorted cells, I performed cell-type specific functional studies including analysis of pathway and transcription factor activity as well as ligand receptor interaction analysis. When interpreting outputs, I integrated renal fibrosis-related ligands, receptors, pathways and transcription factors all together. Based on the integrated data, I was able to identify 6 key biological motives in fibrosis, which were supported by prior studies. Additionally, I identified several driver genes of myofibroblast differentiation in renal fibrosis. Literature studies confirmed that 40 of these genes were previously identified as driver genes of myofibroblast differentiation or fibrosis-related genes.

In a second step, I conducted bulk-level microarray data analysis of chronic kidney disease samples to identify potential candidates for drug repositioning. By reversely matching the disease signatures to datasets of drug-treated cell lines, I identified 20 small molecules as drug-repositioning candidates for 9 different kidney diseases. One of the drugs, “Nilotinib”, was already approved by the FDA. Nilotinib is known to ameliorate renal fibrosis in rats by inhibiting *Pdgfr signaling.* Consistent with this, the single-cell study also identified that the *Pdgfa-Pdgfrb* interaction with subsequent JAK-STAT downstream signaling is a key pathway leading to renal

fibrosis.

In summary, I analyzed renal fibrosis-causing biological pathways, transcription factors, ligand receptor interaction and cell differentiation on a single cell level. To better understand the pathogenesis of fibrosis, I interpreted results by combining biological pathways, transcription factors and ligand receptor interaction analysis, and collapsed these into well-known pathological motives, which are consistent with prior studies on kidney fibrosis. Additionally, I identified several novel candidate genes that may play a central role in pericyte (or fibroblast) to myofibroblast differentiation. Some of these genes will be validated experimentally. At a bulk data level, I performed drug repositioning analysis for 9 different chronic kidney diseases and identified the FDA-approved drug “Nilotinib” as a candidate for drug repositioning for kidney fibrosis. This work opens up new possibilities to understand the pathogenesis of renal fibrosis on a single cell level and enables drug-repositioning for renal fibrosis on single cell level.

Zusammenfassung

Ungefähr 10 % der Weltbevölkerung leidet an einer chronischen Niereninsuffizienz. Obwohl der initiale Stimulus einer akuten Niereninsuffizienz variiert, stellt die Nierenfibrose das gemeinsame Endstadium fast aller Nierenerkrankungen dar. Nichts destotrotz bleibt die Pathogenese der Nierenfibrose aufgrund der Komplexität des Nierengewebes, der Heterogenität der Nierenzelltypen und der großen Heterogenität von Patienten nicht gut verstanden. Die Niere stellt eines der komplexesten Organe dar und besteht aus einer Vielzahl unterschiedlicher Zelltypen wie Podozyten, proximalen Tubuluszellen, distalen Tubuluszellen, Endothelzellen, Perizyten, Fibroblasten und Myofibroblasten. Myofibroblasten wurden in vorherigen Arbeiten bereits als Schlüssel-Zellpopulation in der Pathogenese der Fibrose identifiziert. Nach einer Nierenschädigung expandieren Myofibroblasten und produzieren massiv extrazelluläre Matrix, welche wiederum zu einem pathologischen Gewebeumbau und Verlust der Nierenfunktion führt.

Um die Pathogenese der Nierenfibrose auf Zell-Ebene besser zu verstehen, analysierte ich Einzel-Zell RNA Sequenzierungs-Datensätze von perivaskulären Zellen aus murinen Nieren, welche mittels fluoreszenzaktivierte Zellsortierung (FACS) unter Verwendung der perivaskulären Zellmarker Gli1, Ng2, Myh11 und Cd31 isoliert wurden. Sowohl FACS als auch die Erstellung der Einzelzell-Bibliotheken wurden in Zusammenarbeit von Dr. Christoph Kuppe der RWTH Uniklinik Aachen durchgeführt. Die Zellmarker Gli1, Ng2, Myh11, Pdgfrb sind verbreitete Marker zur Identifizierung von Fibroblasten, Perizyten, Endothelzellen und Epithelzellen, während Cd31 als Marker für Endothelzellen in der perivaskulären Nische der Niere verwendet wird. Mit den isolierten und Einzel-Zell sequenzierten Zellen führte ich auf Einzel-Zell Ebene funktionelle Analysen aus, wie unter anderem die Analyse der Signalweg- und Transkriptionsfaktor Aktivität und Liganden-Rezeptor-Interaktionsanalysen. Zur Auswertung der Ergebnisse integrierte ich Signalwegs-Aktivität, Transkriptionsfaktor-Aktivität und die Ergebnisse der Liganden-Rezeptor Interaktionsanalyse. Anhand der integrierten Daten konnte ich 6 biologische Schlüsselprozesse identifizieren, welche im Konsens mit früheren Studien sind. Darüber hinaus identifizierte ich mehrere Treibergene einer Myofibroblasten-Differenzierung in der Pathogenese der Nierenfibrose. Literaturstudien bestätigten, dass 40 Gene davon bereits als Treibergene einer Myofibroblasten-Differenzierung oder Fibrose-verwandte Gene identifiziert wurden.

In einem zweiten Schritt führte ich eine Microarray-Datenanalyse von Nierenproben mit

chronischer Niereninsuffizienz durch, um Medikamente für eine potentielle Repositionierung zur Therapie der Nierenfibrose zu identifizieren. Durch den umgekehrten Abgleich der Krankheitssignaturen aus den Microarray Datensätzen mit Sequenzierungsdaten von mit Medikamenten behandelten Zelllinien identifiziere ich 20 Medikamente als potentielle Kandidaten für 9 verschiedene Nierenerkrankungen. Eines der Medikamente, "Nilotinib", wurde bereits von der FDA zugelassen. Frühere Studien konnten zeigen, dass Nilotinib in Ratten eine Nierenfibrose reduziert indem es den Pdgfr Signalweg inhibiert. In Übereinstimmung damit konnte ich in der Einzelzellstudie nachweisen, dass die Interaktion von Pdgfa und Pdgfrb sowie der nachgeschaltete JAK-STAT-Signalweg ein Schlüsselprozess ist, welcher eine Nierenfibrose induziert.

Zusammenfassend habe ich in meiner Promotion Nierenfibrose-induzierende Signalwege, Transkriptionsfaktoren, Liganden-Rezeptor-Interaktion und Zell-Differenzierung auf Einzel-Zell Ebene analysiert. Um die Pathogenese der Fibrose besser zu verstehen, integrierte ich biologische Signalwege, Transkriptionsfaktoren und Liganden-Rezeptor-Interaktionsanalysen und fasste diese zu bekannten pathologischen Prozessen zusammen, welche in Übereinstimmung mit bestehenden Studien in der Nierenfibrose sind. Darüber hinaus identifizierte ich mehrere Gene, welche möglicherweise eine zentrale Rolle in der Differenzierung von Perizyten (oder Fibroblasten) zu Myofibroblasten spielen. Verschiedene dieser Gene werden zusätzlich experimentell validiert. Mittels Bulk-RNA Datensätzen führte ich schließlich Medikamenten-Repositionierungsanalysen für 9 verschiedene Nierenerkrankungen durch und identifizierte das von der FDA zugelassene Medikament "Nilotinib" als einen Kandidaten für eine Repositionierung für Nierenfibrose. Diese Arbeit eröffnet neue Möglichkeiten, die Pathogenese der Nierenfibrose auf der Einzel-Zell-Ebene zu verstehen und ermöglicht die Medikamenten-Repositionierung für Nierenfibrose auf einem Einzel-Zell-Level.

Contents

List of Figures

List of Tables

1. Introduction

Transcriptomics indicate entire transcript sets, i.e., all ribonucleic acid (RNA) molecules expressed in a given cell or tissue [1]. RNA molecules, such as messenger RNAs (mRNAs), microRNAs, etc., transfer genetic information from DNA to proteins within the cytosol of cells or regulate gene expression, making them key to understanding cellular processes, development, and disease [2].

Microarrays allow for the rapid detection of RNA by probing the cDNA (transcripts) with known, short sequences on microchips. This technology has been used to aid drug development by monitoring changes in gene expression in response to drug treatments [3]. Unfortunately, microarray technology is dependent on probe design, which can vary between hybridization approaches [4].

In contrast to microarray, RNA sequencing (RNA-seq) directly reads the cDNA sequence, promoting the discovery of novel transcripts and splice junctions [2],[4]. Although the process was initially low-throughput and expensive, high-throughput DNA sequencing methods developed around 2008 have allowed the generation of numerous RNA-seq data sets. Researchers have applied these abundant RNA-seq data sets for estimating cell-type composition using various deconvolution tools, such as CIBERSORT [5]. Using well-studied transcriptomics profiles for each cell type, RNA-seq data can be deconvoluted or regenerated into a cell types-by-gene matrix. Unfortunately, data set heterogeneity limits these applications.

To address this heterogeneity, single-cell level data collection was initiated in 2009 [6]. Single-cell sequencing technologies isolate a single cell from samples, amplify the entire genome, and identify the cDNA, which enables diverse groups of heterogeneous cell types to be distinguished and the discovery of new cell populations. This method expands biologically functional studies to the cell-type level and can be used to infer cellular communication and differentiation between various cell types. Single-cell level studies are also more effective than bulk-level transcriptomic data when interpreting complex diseases and can be used to translate between animal and human investigations. For example, Crinier et al. used single-cell RNA-seq (scRNA-seq) to identify blood and spleen natural killer (NK) cell subsets in mice and humans, with 2 major subtypes found to be similar between the 2 organisms [7].

1.1 Preprocessing single-cell RNA-seq data from mouse renal fibrosis

Single cell RNA-seq technique

Since its introduction in 2009, a range of new single-cell sequencing techniques has been developed. Over the past 10 years, the throughput (number of cells) of single-cell techniques has increased from 10 cells to 1,000,000 cells, which was made possible by advancements in single-cell isolation techniques [8]. Understanding the diverse cell isolation techniques and associated sequencing methods used in scRNA-seq studies is essential because the throughput and read depth affect the computational analysis and biological interpretations.

Single cell isolation

Conventional scRNA-seq relies on micropipette, fluorescence-activated cell sorting (FACS), or laser capture microdissection to isolate single cells. These methods are suitable for detecting rare cells using fluorescent reporters or cell surface markers. This plate-based approach minimizes the capture of multiple cells at the expense of throughput.

More recent isolation techniques, such as droplet encapsulation, have enhanced throughput [9]. Droplet encapsulation dilutes cell suspensions to a specific concentration, allowing for the probabilistic encasing of single cells into droplets (water in oil) [9]. Since 2015, this approach has been used with 3 major sequencing platforms: inDrop, Drop-seq, and 10X (Chromium)[8]. Compared to conventional, plate-based approaches that isolate approximately 100–1,000 cells, these droplet-based techniques can detect 1,000–10,000 cells in a single experiment. However, high cell concentrations are required to maximize the number of cells detected and isolated in droplets because of low capture efficiency [9].

Isolation techniques are known to affect the performance of sequencing platforms [10]. FACS-based sequencing platforms, such as Smart-seq2 and massively parallel RNA single-cell sequencing (MARS-seq), can process approximately 100–1,000 cells, while the droplet-based Drop-seq, inDrop, and 10x increase performance to approximately 1,000–10,000 cells.

Sensitivity and accuracy

In 2017, Svensson et al. compared the technical sensitivity and accuracy of approximately 30 scRNA-seq techniques [11]. This study used the identical spike-in RNA standards at known

concentrations across the experiments to compute technical sensitivity and accuracy as defined below [11].

(1) "Sensitivity: the number of input spike-in molecules at the point at which the probability of detection reaches 50 %" [11]
(2) "Accuracy: the Pearson product-moment correlation (R) between estimated expression levels and actual input RNA-molecule concentration (ground truth)" [11]

High sensitivity permits the detection of weakly expressed genes. High accuracy detects expression variation corresponding to the true biological differences in mRNA abundance across cells [11]. Svensson et al. found that traditional bulk RNA sequencing was more accurate than scRNA-seq protocols. While some scRNA-seq protocols displayed high, stable accuracy, others (genome and transcriptome sequencing (G&T-seq), cell expression by linear amplification and sequencing (CEL-seq), and MARS-seq) were unstably and variably accurate across experiments [11]. All single-cell protocols had higher sensitivity than bulk sequencing, particularly SMARTer, CEL-seq2, STRT-seq, and inDrop, indicating their capacity to detect single-digit, input spike-in molecules [11]. The group also found out that sequencing depth affects sensitivity more than accuracy. From this analysis, the microwell-based methods, single-cell universal poly(A)-independent RNA-seq (SUPeR-seq), which uses a total-RNA protocol, and CEL-seq2, which amplifies cDNA by in vitro transcription rather than PCR, performed best [11].

Comparing Drop-seq, inDrop, and 10X

Microwell-based scRNA-seq methods have advantages such as low cost and high throughput [12]. Microwell-based scRNA-seq methods are low cost and high throughput [12]. However, the lack of commercially available protocols has prevented microwell-based scRNA-seq techniques from being widely adopted [12]. In contrast, droplet microfluidics has rapidly developed to process dozens of thousands of droplets/second and produce millions of droplets, thus increasing throughput and reducing costs [12]. There are currently 3 main droplet-based systems for scRNA-seq: inDrop [13],[14],[15],[16], Drop-seq [17], and 10X Genomics Chromium (10X) [18]. All 3 methods use similar approaches for generating the droplets, differentiating individual cells, and employing unique molecular identifiers (UMI) for bias correction but differ in bead manufacture, barcode design, and cDNA amplification [12]. Zhang et al. compared these 3 systems using the same cell samples and data analyses [12]. They found 10X to have higher sensitivity, detecting approximately twice as many UMIs as inDrop and Drop-seq [12].

The computational analysis of single cell RNA-seq data

scRNA-seq data are characterized by their large size, heterogeneity, and excessive zeros. As such, scRNA-seq data must be treated differently than conventional bulk data analyses used for microarray or RNA-seq data.

The first step of analyzing large, sparse data sets is reducing the dimension of the features and samples. Data features (e.g., genes) can be reduced to those explaining the variance in the data set (highly variable genes) using a nonparametric approach. Butler et al. used the Seurat: R toolkit to select these features through variance-stabilizing transformation (vst) [19]. They applied LOESS nonparametric regression to the mean-variance relationship of log-scaled data to obtain a LOESS fitted model, which was used to select the 2,000 most variable genes [19]. Data were then scaled so that the mean expression across cells was 0 with a variance of 1, resulting in an almost normal distribution [19].

Principal component analysis (PCA) can then be used to reduce the dimensionality of the scaled data. Reduced single-cell data can be clustered. A high computing system is generally required to avoid slow processing associated with higher numbers of cells. Thus, efficient clustering or classification methods should be employed in a CPU system. Hierarchical and partitional clustering are the major clustering techniques, with k-means and hierarchical clustering most frequently used across fields. K-means clustering splits the data into K different clusters and measures the new center of each cluster. Each data can be assigned to a cluster which has a centroid close to the data. After all data are assigned newly, the centroids are remeasured. This clustering requires repeated steps to find the centroid that explains the data clusters. Hierarchical clustering is also unsuitable for heterogeneous data sets. This clustering clusters the two closest data together, and then integrates the two most similar clusters. This process is also iterative and continues until one big cluster is made up. Iterative steps don't fit the clustering for high-dimensional datasets because of hgh computing time. Therefore, high-dimensional data sets often require other clustering approaches, such as density-based and graph-based clustering.

Density-based clustering is based on the assumption that a cluster is made of contiguous regions of high density or separated by areas of low density [20]. This assumption requires all clusters to have the same amount of density. While this method has been used for high-dimensional data sets, it is inappropriate for use on high-dimensional, unknown biological data sets.

Graph-based clustering is fast when handling big data sets. It uses adjacent values in a cell-cell similarity matrix to identify the 5–30 most similar neighboring cells for each cell. Communities of several single cells are then compiled to maximize the modularity score, e.g., similarity values

(edge score) within the community. Graph-based clustering begins with a single, randomly chosen node (a cell), which is joined to another node if the connection results in the highest modularity change. Communities aggregate with one another until the whole community includes all cells [21]. This step is called optimization, and the number of communities is called a resolution. Graph-based clustering fits high-dimensional and unknown biological data sets but is complicated by the choice of optimal cluster number and cannot prove a biological story. Some researchers try to generate many communities (clusters) and label them by cell type, which is useful because identical cell types can then be merged, or considered as independent sub-cell types, or filtered out if they are not biologically meaningful.

In late 2019, Korsunsky et al. developed Harmony, a new scaling method with fuzzy clustering [22]. Harmony has been used to correct normalized data sets or PCA embedding matrices. It is similar to k-means clustering but measures the probability of cluster membership ranging from 0 to 1 rather than binary [23]. Fuzzy clustering assumes that single cells can belong to more than 1 cluster. This approach has performed well when integrating multiple scRNA-seq data sets.

Annotation is the most important step in scRNA-seq data analysis as it affects subsequent analyses, including biological interpretations. Unfortunately, scRNA-seq data from the same tissue or animal model can differ due to technical and biological factors, such as the sequencing machines, cell cycle stage, and period in the life span. Ibrahim et al. addressed this by using the number of cells expressing marker genes in a given cluster rather than using the expression value itself[24]. This binary approach alleviates technical effects. The authors also developed a new concept, specificity score, which captures how exclusively and highly a gene is expressed in a given cluster using a Bayesian approach[24]. Using these 2 concepts, biologists can map each cluster to each cell type based on their knowledge of the biology.

Biological interpretation on single cell RNA-seq data

Functional-level studies

There are several ways to interpret scRNA-seq data functionally, including gene set enrichment analysis (GSEA), Enrichr, AUCell, AddModuleScore, and PROGENy [25], [26], [27], [28], [29], [30]. Since its development in 2005, GSEA has frequently been used for bulk-level data analysis [25]. For a given gene set, GSEA uses a running-sum statistic that increases for each gene in the ranked data that is a member of the gene set and decreases if it is not in the gene set to calculate an enrichment score (ES). The method calculates the maximum deviation from 0 by a weighted Kolmogorov–Smirnov-like statistic [25]. Enrichr is similar to GSEA and ranks biological pathways or terms based on the shared number of genes with a public database, such as GO

Biological Process (2018), using a proportion test modified from the Fisher exact test [26], [27]. In 2017, AUCell, a GSEA tool for scRNA-seq, was developed, which applied the area under the curve (AUC) concept to the ES [28]. AUCell calculates the AUC score of the recovery curve for the top 5 % of genes ranked in each cell [28]. For this purpose, AUC corresponds to the proportion of genes in the gene set that are highly expressed in each single cell, with cells expressing many genes in a specific pathway having higher AUCs. This helps identify active gene sets (pathways) across all data sets or user-defined cell populations. In 2016, Tirosh et al. developed AddModuleScore, which calculates the average expression levels of a given gene set at a single-cell level and subtracts the aggregated expression of a randomly-selected feature set [29]. PROGENy measures the activity score of 14 biological pathways, including TGFb, WNT, PI3K, Trail, p53, and NFkB, in a given cell population using a specialized weight matrix corresponding to the correlation matrix between genes and the pathways [30]. High correlation values indicate that a gene was positively regulated by a particular pathway and vice versa. PROGENy scores are averaged across cells in a given cell population for each pathway while retaining a data set's biological context [30].

Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) has been used widely to infer the relative activity of transcription factors in each sample[31]. The activity of each transcription factor is calculated through its targets, which are identified from well-known databases, such as the cell-context-specific interaction network (interactome) or DoRothEA[32]. Holland et al. demonstrated that VIPER, in conjunction with the DoRothEA database, performed well for the functional interpretation of scRNA-seq data [30].

Cell differentiation

There have been several tools to perform pseudotime analysis. Pseudo-time analysis is a computational approach to explore cell dynamics, such as the cell differentiation by ordering single cells along developmental trajectories. Several tools have been developed to perform pseudo-time analysis.

Diffusion maps assume that single cells follow diffusion-like dynamics during differentiation [33]. Like other scRNA-seq tools, it uses a cell-to-cell distance matrix but preserves the non-linear structure of the data through density normalization [33]. Diffusion maps have been applied to single-cell data for dimension-reduction and ordering of cells along the differentiation path [33].

Slingshot identifies the lineage structure using a cluster-based minimum spanning tree (MST) [34]. MST maps connections between adjacent clusters to infer lineages. Slingshot can be

performed in an unsupervised or semi-supervised manner by specifying initial point clusters [34]. If the initial time point is specified, Slingshot maximizes the number of clusters shared between lineages.

Two major tools, Velocyto[35] and scVelo[36], use an RNA velocity approach to infer the trajectory of gene expression. RNA velocity relies on the concept that a gene's pre-mRNA to mRNA ratio indicates whether its expression will increase or decrease. Specifically, if mRNA abundance is higher than pre-mRNA, mRNA will decrease to maintain homeostasis at the mRNA level. While Velocyto assumes that all genes have a common splicing rate, scVelo considers that this can be violated if the data set consists of several heterogeneous cell types with diverse kinetics. scVelo measures the full transcriptional dynamics of splicing kinetics with a likelihood-based dynamic model. By extending deterministic models (i.e., Velocyto) to probabilistic events, scVelo considers that RNA velocity occurs in a wide variety of systems normally found in development [36]. scVelo also estimates the rates of transcription, splicing, and degradation at a gene level [36].

Another pseudo-time analysis, PAGA, was reported by Wolf et al.[37]. PAGA generates a network whose nodes correspond to clusters (cell populations) and whose edge weights represent the connectivity between clusters [37]. This population-based network connects 2 clusters if the number of inter-edges exceeds the number of randomly generated inter-edges, such that the connection weight represents the confidence supporting an actual connection [37]. Wolf et al. modified the original diffusion pseudo-time approach and applied it to PAGA to identify lineages at a cell population network level [37].

Intercellular communication

Cell-to-cell communication is typically categorized into 4 signaling systems: paracrine signaling, autocrine signaling, endocrine signaling, and signaling by direct contact. Paracrine signaling occurs between 2 adjacent cells. Autocrine signaling refers to a cell targeting itself by hormones or chemical messengers. Endocrine signaling occurs between 2 distant cells.

Many researchers have developed new tools to infer cell-to-cell interactions using single-cell data with prior knowledge of receptor-ligand interactions. Several databases, including FANTOM5 [39], ICELLNET [40], and OmniPath [41], report ligand-receptor pairs. FANTOM5 is a visualization and query tool for ligand-receptor interactions in humans[39]. As part of the FANTOM5 project, Ramilowski et al. presented a map of cell-to-cell communication between 144 human cell types, based on the expression profiles of 642 ligands and their 589 receptors [39]. Using known ligand-receptor pairs and public protein-protein interaction (PPI) information, they generated 2,422 interaction pairs [39]. In 2020, Noël et al. introduced ICELLNET, which

manually curated ligand-receptor pairs focused on immune-related pathways, such as growth factors, cytokines, chemokines, immune checkpoints, hormones, Notch signaling, and antigen binding [40]. While this database includes only a small number of interactions (~400), it is meaningful for inferring cell-to-cell communication from an immunological point of view [40]. OmniPath takes other public databases providing numerous, diverse PPIs from more than 115 databases. It provides useful tools, including an R package, a Cytoscape plug-in, and a Python module [41]. OmniPath also provides various metadata, including the number of papers supporting the ligand-receptor pairs and their related biological functions.

Several algorithms, such as CellPhoneDB and CellChat, have been developed to identify potential interactions. CellPhoneDB infers interactions based on the percentage of cells (1–99 %) expressing ligand-receptor pairs between cell populations [42]. This tool provides a measurement of the *p*-value for each pair used to infer cell-to-cell communication [42]. CellChat, a more advanced version of CellPhoneDB, was developed in 2021 [43]. CellChat's database includes ligand-receptor information from the KEGG Pathway database and antagonists interfering with specific interactions [44]. It categorizes interactions into 3 types: paracrine / autocrine, extracellular matrix (ECM)-receptors, and cell-to-cell contacts. Each ligand-receptor interaction is annotated with the functionally-related signaling pathway from KEGG [43]. CellChat was reported to outperform CellPhoneDB in identifying stronger interactions [43].

Renal fibrosis

Kidneys are complex organs with complicated structures, including glomeruli, tubulo-interstitium, and vasculature. Each section of the kidney has its own function and associated cell types, making it challenging to understand kidney diseases. Approximately 12 % of the world's population suffers from chronic kidney disease (CKD) [45], which has a significantly high mortality rate [46].

Renal fibrosis, inflammation and myofibroblasts

Renal fibrosis is the known common end-point of CKD. Renal fibrosis is the functional decline of the kidney due to excessive epithelial injury and inflammation. At the cellular level, epithelial cells and their vascular capillaries are lost, and activated myofibroblasts, matrix, and inflammatory cells are accumulated [47]. Activated macrophages are crucial mediators during acute inflammation. They generate large quantities of profibrotic factors and modify the microenvironment. As inflammation continues, a combination of infiltrating leukocytes and activated intrinsic renal cells lead to the production of profibrotic cytokines and growth factors [49]. This leads to the recruitment and activation of myofibroblasts and subsequent accumulation

of ECM, which is the hallmark of renal fibrosis [48],[49],[50]. Renal ECM involves a network of collagens, elastin, glycoproteins, and proteoglycans, which are potential specific, non-invasive renal fibrosis biomarkers [50].

Renal fibrosis, a perivascular niche and pericytes

The perivascular niche corresponds to the microenvironments around a vessel [51]. It is heterogeneous, being composed of numerous, diverse cell types, including pericytes, endothelial cells, and immune cells [51],[52].

Pericytes are multi-functional cells embedded in the middle of the basement membrane of capillaries wrapped around endothelial cells [53]. In the kidney, pericytes are associated with glomeruli and cortical and medullary peritubular capillaries. Pericytes serve multiple functions, including scaffolding cells for development, maintaining vasculature, and contributing to intercellular signaling along the vessel or between vessels [54].

Due to their interaction with diverse cells, pericytes are considered important to the pathogenesis of kidney disease [52]. Yang et al. studied the relationship between pericytes and renal fibrosis using putative endothelial progenitor cells (pEPCs), which are known to alleviate fibrosis [55]. They reported that pEPCs attenuated renal fibrosis by decreasing the migration of pericytes and their differentiation into myofibroblasts [55].

Biological pathways in renal fibrosis

TNF-α, TGF-β, connective tissue growth factor (CTFG), IL-6 (JAK/STAT), Wnt/β-catenin signaling, p53, mitogen-activated protein kinase (MAPK), and NF-κB have been associated with renal fibrosis. TNF-α and TGF-β regulate NF-κB, p53, and CTFG, which in turn induce JAK/STAT, NF-κB, MAPK, and Wnt signaling[56]. All of these pathways are connected by c-Jun N-terminal kinases (JNKs), which include JNK1 (*Mapk8* in mice), JNK2 (*Mapk9* in mice), and JNK3 (*Mapk10* in mice). JNK is a downstream protein of the noncanonical TGF-β, canonical PDGF, and noncanonical Wnt signaling pathways and augments skin fibrosis via crosstalk between these signaling pathways. JNK promotes secretion of TGF-β and crosstalk with STAT3 (JAK/STAT) to activate pro-fibrosis [57]. In this process, STAT3 and TGF-β are strongly associated with MAPK signaling. JNK, p38, EGFR, RAS/ERK, and PI3K signaling are also involved. During renal fibrosis, vascular endothelial growth factor (VEGF) and hypoxia contribute to the disease. However, these pathways are localized to endothelial cells, pericytes, and podocytes in the perivascular space of the kidney.

Transforming growth factor-β (TGF-β) signaling

Transforming growth factor-β (TGF-β) has been known as a central mediator of diverse cellular processes including growth, differentiation, wound repair, apoptosis and the pathogenesis of fibrosis [58]. In renal fibrosis, TGF-β has been considered as a potent profibrotic key in excessive accumulation of extracellular matrix proteins leading to renal fibrosis [58]. There are 3 major isoforms of TGF-β, TGF-β1, TGF-β2 and TGF-β3, all of which are expressed in the kidney and have been believed to induce ECM protein production in renal fibrosis [58]. However, recent studies demonstrated that TGF-β2 and TGF-β3 are likely to be involved in antifibrotic effects [58].

Regarding TGF-β signaling, the TGF-β1 ligand binds to TβRII, assembles a heteromeric complex of TβRII, phosphorylates the kinase domain of TβRI and leads to the activation of the receptor-activated or regulatory Smads, Smad2 and Smad3 [58]. Smad4 forms the complex with both Smad2 and Smad3 and then moves into the nucleus in order to regulate the expression of target genes. All Smad proteins don't cooperate with TGF-β signaling like Smad2 and Smad3. For example, Smad7 negatively regulates TGF-β signaling by recruiting E3 ubiquitin ligases [58]. The TGF-β signaling activates not only (1) Smad2/3 but also the (2) Ras-Raf-MEK-ERK pathway (called MAPKK-ERK), (3) NF-κB pathway and (4) TGF-β-activated kinase 1 (TAK1) related pathway leading to the activation of MKK4-JNK and MKK3-p38 pathways. MKK4-JNK activates transcription factors activator protein-1 (AP-1) and MKK3-p38 pathways stimulate transcription factor 2 (ATF-2), respectively [58]. Except for Ras-Raf-MEK-ERK pathways, all of these four downstream are known to promote renal fibrosis.

As well, TGF-β1 activates p53 (*Trp53* in mice) phosphorylation which in turn, interacts with activated SMADs and leads to the subsequent binding of p53/SMAD3 to target promoters [59]. p53 phosphorylation is one of key causative factors because p53 upregulates ALK5 (*Tgfbr1* in mice), SMAD3 (*Smad3* in mice), TGF-β1 (*Tgfb1* in mice), TGF-β3 (*Tgfb3* in mice), CTGF (*Ccn2* in mice), CCN2, α-smooth muscle actin (α-SMA, *Acta2* in mice) and plasminogen activator inhibitor-1 (PAI-1, *Serpine1* in mice) which contribute to a complex feed-forward loop to keep a profibrotic renal microenvironment [59]. One regulator of p53 function in the context of renal injury is the serine/threonine kinase tumor suppressor ataxia telangiectasia mutated (ATM, *Atm* in mice) [59].

Recent studies demonstrated that anti-TGF-β1 by deleting TβRII has not shown promising results for treating renal fibrosis [58]. This is caused by the effect of the inactive form of TGF-β1 (called latent form of TGF-β1). The latent TGF-β1 transgenic mice increased the

expression of Smad7 which inhibits the NF-κB while deactivating the TGF-β, which can be interpreted as tissue homeostasis [58]. As well, Tgf-β1-null mice displayed severe inflammatory responses with massive penetration of both lymphocytes and macrophages in many organs, which indicates TGF-β1 has anti-inflammatory effects. [58]. TGF-β1 promotes autophagy through the TAK1-MKK3-p38 signaling pathway, which induces the intracellular degradation of collagen and protects cells against cell apoptosis by this induction of autophagy with the activation of TAK1 and AKT [58]. This mechanism is important because the autophagy deficiency has been characterized in progressive renal fibrosis, especially, mice deficient in autophagic protein Beclin 1 and Podocyte-specific deletion of the *Atg5* gene resulted in renal injury [58].

TGF-β signaling has diverse feedback regulation through downstream signaling. TGF-β1 induces Smad7 which in turn, negatively regulates TGF-β1 [58]. Smad2 deletion leads to Smad3 phosphorylation, which allows Smad3 to bind to a collagen promoter (COL1A1, Col1a1 in mice) and finally, auto-induce TGF-β1 [58]. However, over-expressed Smad2 mitigates TGF-β1-induced Smad3 phosphorylation and type 1 collagen accumulation [58]. Klotho protein (*Kl, Klb* in mice), one of endogenous modulators of TGF-β signaling upregulates autophagy reaction of TGF-β1. Proteoglycans are the key components of the extracellular matrix, but proteoglycan "decorin" is the antagonist of TGF-β1 by binding to active TGF-β1 [58].

Wnt/β-catenin signaling

Wnt signal transduction has been known to regulate injury repair, pathogenesis of diverse human disease and embryogenesis [60]. A Wnt ligand binds to a seven-pass transmembrane Fz receptor with lipoprotein receptor related protein 6 (LRP6, *Lrp6* in mice) or LRP5 (*Lrp5* in mice). The Wnt-Fz-LRP6 complex phosphorylates LRP6 via recruiting the scaffolding protein Dishevelled (Dvl), activates Axin complex, and in turn, stabilizes β-catenin which travels into the nucleus to regulate Wnt target genes by forming complexes with TCF/LEF [60], [61]. Secreted frizzled-related protein 4 (SFRP4, Sfrp4 in mice), an endogenous extracellular Wnt antagonist, inhibits the activation of β-catenin and in the end, attenuates renal fibrosis in UUO mice [60]. It was believed that the stimulated Wnt signaling caused the accumulation of β-catenin (*Ctnnb1* in mice) which lead to the upregulation of target genes including *c-Myc (Myc* in mice), *Twist (Twist1* in mice), *TCF1(Tcf7, Tcf1, Hnf1a* in mice), and fibronectin *(Itgb1, Fn1, Itga5,* etc in mice) in the study on renal epithelial cells [62]. As well as these genes, Snail1 *(Snai1, Snai2, Sani3* in mice), plasminogen activator inhibitor-1 (*PAI-1, Serpine1* in mice), matrix metalloproteinase 7 (*Mmp7* in mice), and multiple components of the renin–angiotensin system (*RAS*), such as angiotensinogen (*Agt* in mice), renin (*Ren1, Ren2* in mice), angiotensin

converting enzyme (*Ace, Ace2* in mice), and *angiotensin receptor type 1* (*Agtr1a* in mice) have been highlighted as keys of fibrosis-related genes of Wnt signaling [60].

Like Transforming growth factor-β (TGF-β) signaling, Klotho protein is also considered as an important antagonist protein in the Wnt signaling of renal fibrotic models. This protein has been known to be upregulated in the renal tubular epithelium of the healthy kidney, but downregulated in CKD [60]. The Klotho dysregulates the Wnt/β-catenin signaling, reduces the deposition of extracellular matrix and diminishees the cytokine production [60]. However, TGF-β1 can block Klotho and then promote β-catenin to induce myofibroblasts activation leading to renal fibrosis [60].

Janus kinase/signal transducer and activator of transcription (JAK/STAT) signaling

The JAK/STAT pathway includes JAK1-JAK3 and receptor tyrosine kinase 2 and seven STAT proteins, STAT1,2,3,4, STAT5a, STAT5b and STAT6) where STAT3 has been known as a key factor of this pathway [63]. JAK-STAT signaling mediates cytokines, epidermal growth factor (EGF) and platelet-derived growth factor (PDGF) [63]. IL-6 cytokines start the JAK-STAT signaling by binding to the receptor, inducing dimerization of glycoprotein 130 receptors which in turns, phosphorylates STAT3 at Tyrosine 705th site [63]. This phosphorylated STAT protein translocates to the nucleus and regulates it's target gene expression [63]. The binding sites of STAT3 are related to fibroblasts activation and profibrotic pathways such as lipocalin 2 (*Lcn2*), tissue inhibitor of metalloproteinase 1 (*Timp1*), and PDGF-B (*Pdgfb*) [63]. As well as these genes, it has been known that STAT3 targets the promoter of kidney injury molecule-1 (*KIM-1* in humans, *Havcr1* in mice). STAT3 also binds to diverse transcription factors including Nanog, c-Myc, and Twist [63].

The JAK/STAT pathway is inhibited by the protein inhibitor of activated STAT3 (*PIAS3* in humans, *Pias3* in mice) by binding the activated STAT3 dimers [63]. Krüppel-like factor 4 (*KLF4* in humans, *Klf4* in mice), a zinc finger transcription factor, negatively regulates JAK/STAT pathway in podocytes in the similar way of PIAS3 protein [63]. As well, seven protein tyrosine phosphatases (PTPs in humans, *Ptprm, Ptpn1, Ptprt, Ptprk, Ptpru* in mice) also inhibit JAK/STAT pathway by dephosphorylation of STAT3 [63]. The suppressor of cytokine signaling (SOCS, *Socs1* and *Socs3* in mice) induced by STAT activation is also involved in the negative feedback loop of STAT activation [63].

vascular endothelial growth factor (VEGF) signaling

Endothelial cells secrete paracrine signals including VEGF and Ang-1 to retain their homeostasis. Both VEGF and Ang-2 induce that pericytes surrounding endothelial cells detach from the vessel walls, making empty space for endothelial cells to generate new branches. In endothelial cells, VEGF upregulates the Notch ligand Delta-like protein 4 (DLL4) which leads to sprouting vessels, in turn, activates Notch signaling in neighboring cells to promote the elongation of these cells. At the same time, this Notch signaling reduces the expression of VEGFR2 in order to enable only newly generated tip cells to react to Notch signals [64].

As well as podocytes, VEGF is also expressed in the thick ascending limb (TAL), the proximal and distal tubules with a lesser extent. VEGF receptors (VEGFR1 and VEGFR2) exist on endothelial cells in both the peritubular capillaries and glomerular capillary loops. Like the relationship between podocytes and endothelial cells within the glomerulus, VEGF contributes to maintaining peritubular capillary health in tubules by upregulating hypoxia-inducible factor (HIF)-mediated transcriptional under the lack of oxygen conditions [64].

Here, HIF is a master regulator in hypoxia, upregulating the transcription of more than 100 genes including Epo, glucose transporters, VEGF, and glycolytic enzymes. This transcriptional factor is a heterodimer of HIF-α and HIF-β. HIF-β is always expressed regardless of oxygen, but HIF-α expression is strongly regulated by the amount of oxygen. One of two active isoforms of HIF-a, HIF-1α is increased in response to hypoxia, and in turn, augments the HIF-1 response in tubular epithelial cells. The other active isoforms of HIF-a, HIF-2α are expressed in endothelial cells and interstitial fibroblasts [64].

The effect of HIF protein should be interpreted by the area of the kidney where renal fibrosis occurs. The specific deletion of HIF-1α in proximal tubular cells alleviates kidney fibrosis and macrophage infiltration in a mouse unilateral ureteral obstruction (UUO) model which indicates that HIF-1α in proximal tubular cells promotes kidney fibrosis. However, other studies demonstrated that excessive amounts of HIF-1α and HIF-2α in mutant mice had lower blood urea nitrogen with attenuated glomerular and tubulointerstitial damage but similar tubulointerstitial fibrosis [64].

Tumor necrosis factor alpha (TNF-a)

In macrophages, renal tubular cells, and mesangial cells, TNF-α is expressed to regulate damage, inflammation and cell death signaling mediated by c-Jun and NF-κB [56]. TNF-α is

overexpressed for healing especially by TNFR2 [56]. An animal study, it has been known that TNF-α plays a major role in renal inflammation and fibrosis by promoting M1-like macrophages, releasing monocyte chemoattractant protein-1 (MCP-1/CCL2), interleukin-1β, and TGF-β1 in glomerular disease and in the end, augmenting the acute kidney injury to CKD transition. [56].

Activation of mitogen-activated protein kinase (MAPK) and p38

There are 3 major MAPK pathways, (1) TGFβ/p38, (2) mitogen-activated protein kinase (MAPK), and (3) P13k/AKT/mTOR signaling pathways. The TGFβ/p38 mitogen-activated kinase can be activated by the TGFβ pathway, which can subsequently upregulate TP53 [65]. TGFβ/p38 is independent of the canonical SMAD signaling.. The canonical MAPK kinase pathway is initiated by the extracellular growth factors (GFs), inflammatory cytokines and stress which activate receptor tyrosine kinases (RTKs) on the cell membrane [65] [66]. For example, epidermal growth factor receptor EGFR activates the MAPK pathway. *RAS, RAF* and *MEK* are the order of downstream pathways which in turn stimulates the ERK1*/2* transcription factor activator [65]. As an example, there are p38 mitogen-activated protein kinase (MAPK) pathways and c-Jun N-terminal kinase (JNK) MAPK [65]. RTKs and RAS also activate The P13K/AKT/mTOR cascade which is involved in cell growth [65]. Here, The p38 mitogen-activated protein kinase (MAPK) pathway has been known to be involved in the proinflammatory and profibrotic mediators productions related to inflammation, apoptosis, and fibrosis [67]. Studies with mouse UUO models and IgA nephropathy patients showed that COL1 and phosphorylated p38 protein expression in the kidneys were increased, whereas these proteins were downregulated in mice with p38 MAPK inhibitors [67].

Single-cell studies of renal fibrosis

In 2018, Park et al. analyzed approximately 44,000 cells from the whole kidney of healthy mice [68]. They distinguished 21 major cell types and identified a novel transitional cell type existing between intercalated cells (ICs) and principal cell types (PCs)[68]. Cell trajectory analysis revealed that the Notch signaling pathway mediated these transitions. They also noted that these transitions were associated with metabolic acidosis, which is commonly found in CKD[68].

In 2020, Rudman-Melnick et al. published a comprehensive atlas of single-cell transcriptional changes during acute kidney injury (AKI) in a mouse unilateral ischemia/reperfusion (UIR) model [69]. They traced transcriptional changes and reported potential, novel markers of AKI [69]. Surprisingly, they found that AKI induces “mixed-identity cells,” which express markers of diverse renal cell types associated with kidney development [69].

Transgenic reporter mice have been used to investigate renal fibrosis. A key system involves the use of Cre recombinases isolated from bacteriophage P1, which catalyzes the ligation and cleavage of DNA at its specific nucleotide recognition sites (loxP) [69]. Mouse lines that link the expression of Cre to a gene of interest can be combined with a reporter (e.g., tdTomato) that is induced upon Cre activation, allowing for controlled expression of the reporter genes. Gene modulation in these systems is controlled by tamoxifen (t2), which acts as a ligand to activate t2-dependent Cre recombinases *in vivo*. Using this method, a number of time- and tissue-specific mouse mutants have been developed [70].

Kuppe et al. profiled the transcriptomes of proximal and non-proximal tubule cells from normal and fibrotic kidneys at a single-cell level to determine where ECM secretion initiated during CKD [71]. They used a diffusion map to analyze the scRNA-seq data, which indicated that myofibroblasts were generated from pericytes and fibroblasts [71]. This study used *Pdgfrb-CreER-tdTomato* transgenic reporter mice, which co-express *Pdgfrb* with tdTomato in a Cre-dependent manner, allowing for the identification of cells expressing *Pdgfrb*. *Pdgfrb* was found to be expressed in mesenchymal populations and epithelial, endothelial, and immune cells [71]. Similarly, *Gli1-CreER-tdTomato* mice were used to detect perivascular niches in the kidney [72].

To investigate renal fibrosis–related biological pathways, intercellular communication, and cell differentiation at a single-cell level, my collaborator, Dr. Christoph Kuppe at RWTH Uniklinik Aachen, and I initiated an scRNA-seq project in normal and renal fibrosis mice in 2019 (Figure 1.1.1.B). From early 2019 to late 2020, Dr. Kuppe generated *Pdgfrb-CreER-tdTomato*, *Gli1-CreER-tdTomato*, *Ng2-CreER-tdTomato*, and *Myh11-CreER-tdTomato* normal and renal fibrosis mouse lines. These transgenic reporter mice express tdTomato only in the perivascular niche because it is co-expressed with *Gli1*, *Ng2*, *Myh11*, and *Pdgfrb*, which are known to be expressed in the kidney perivascular niche. These 4 genes are normally expressed in fibroblasts, myofibroblasts, pericytes, parietal epithelial cells, and, rarely, endothelial cells. Dr. Kuppe also produced a data set from endothelial cells of renal fibrosis mice by sorting $Cd31^{+}$ cells. *Cd31* (*Pecam1* in mice) is a well-known marker for endothelial cells. Data from normal mouse endothelial cells were obtained from a public data set [73]. All data sets were analyzed using Seurat, PROGENy, DoRothEA, Harmony, CellChat, scVelo, etc (Figure 1.1.1.A). This thesis has the output from the analysis.

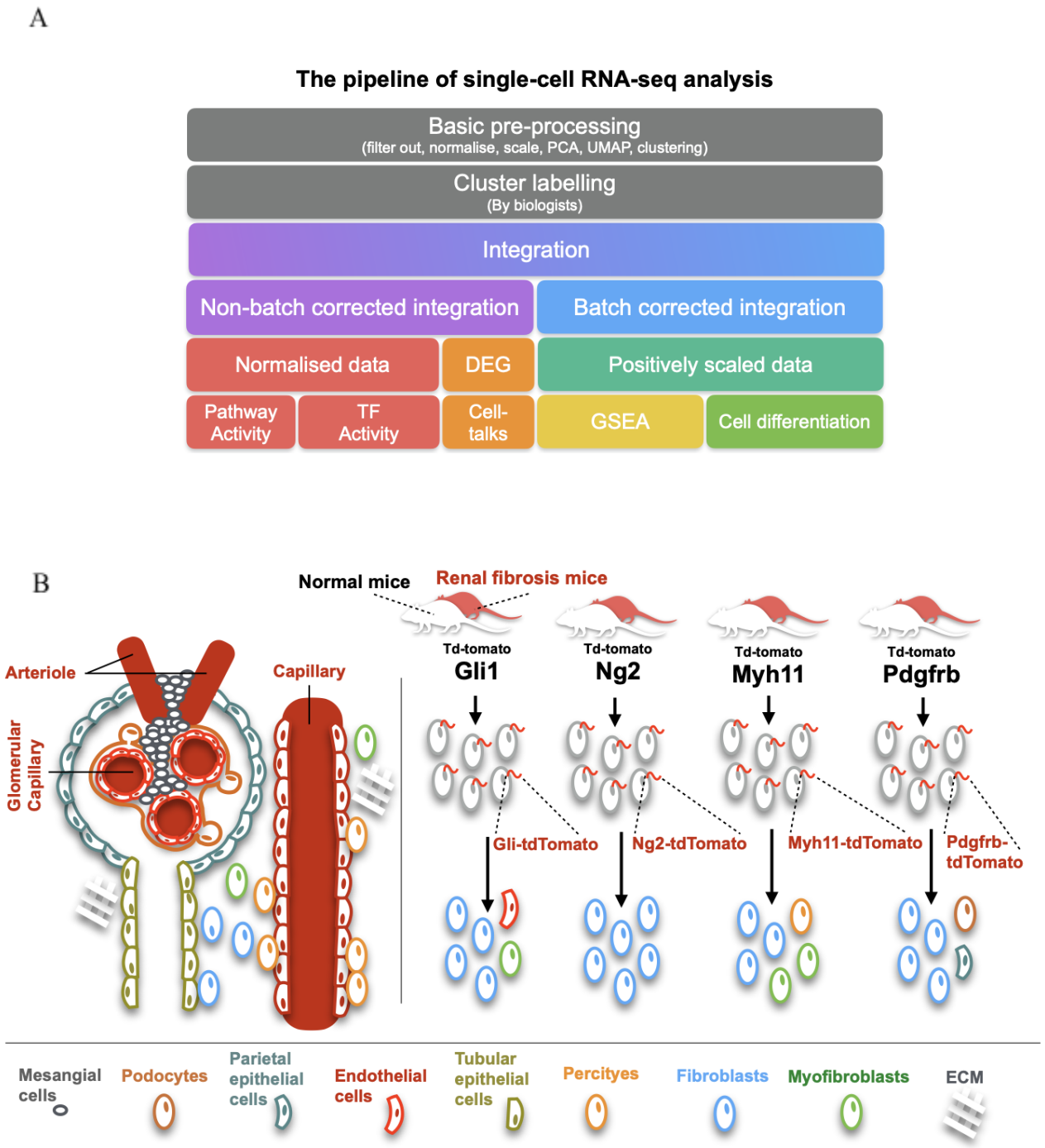


Figure 1.1.1. The experimental design and computational pipeline of this single cell RNA-seq project from mice renal fibrosis. (A) Single-cell RNA-seq data analysis consists of basic preprocessing which includes filtering out cells & genes, normalizing count-matrix datasets, reducing data by PCA & UMAP, and then clustering. For the preprocessed data set, Dr. Kuppe labelled each cluster to the cell type for 9 different data sets (the 10th data set from public data with its annotation). For the labelled data sets, I integrated datasets in 2 different ways while removing batch effects and remaining batches (non-batch corrected integration), separately. For the non-batch corrected integrated data set, pathway activities, transcription factor (TF) activities were measured, and intercellular

communication were performed. For the batch-corrected normalized data set, it had minus values as minimum. So, I scaled the data, and then took only positively scaled data. Gene set enrichment test was performed on this data, and the outputs represented the main features of cell types fully. With this biological point, cell differentiation was also conducted on the same batch-corrected positively scaled data. (B) This figure illustrates perivascular niche of kidney in the left panel, and the experimental designs of *Pdgfrb-creER-tdTomato* mice, *Gli1-creER-tdTomato* mice, *Ng2-creER-tdTomato* mice, *Myh11-creER-tdTomato* mice per each condition (white mice is normal mice, red mice is renal fibrosis mice) in the right panel. The Gli1, Ng2, Myh11 and Pdgfrb genes with tdTomato are co-expressed in fibroblasts (blue colored cells), myofibroblasts (green colored cells), percityes (orange colored cells), parietal epithelial cells (Olive colored cells) and rarely, endothelial cells (red colored cells). Cells expressing tdTomato are detected by the expression of tdTomato, and cell labeling is performed by Christoph Kuppe. The Cd31+ endothelial cells from sham and UUO mice were omitted from this figure.

1.2 Drug repositioning using microarray data

Drug repositioning involves expanding the use of approved drugs for other diseases. Conventional drug discovery consists of target identification, screening of active compounds, preclinical studies, and phase I–IV clinical trials, which can take approximately 13 years and cost in excess of $1 billion [74],[75]. Drug repositioning is faster and less costly [76]. For example, Dr. Gilla Kaplan demonstrated thalidomide's potential for treating inflammation and tumorigenesis in 1991 [77]. The pharmaceutical company, Celgene, initiated the repositioning of thalidomide for the treatment of multiple myeloma in 1998, which was approved by the FDA in 2006 [78]. As another example, Cypress Bioscience found that milnacipran, which they were developing as an antidepressant, had dual activity for relieving fibromyalgia [79].

Computational approaches in drug repositioning

There are various approaches for drug repositioning. Researchers seek to identify drugs that share protein targets (on-target), modes of action, or signal transduction pathways experimentally [76]. Gene expression patterns have also been used to understand the effects of drugs and infer drug repositioning by reversely matching to disease data.

In 2006, the microarray-based Connectivity Map (CMap) project produced 564 gene expression profiles from 5 cancer cell lines, 453 of which described their response to distinct drug treatments [80]. Using CMap, Ramsey et al. identified a new drug candidate, entinostat, for treating acute myeloid leukemia [81]. CMap has advanced numerous other instances of drug repositioning [82]. The NIH launched a new program in 2014, the Library of Integrated Cellular Signatures (LINCS), which developed a cost-effective transcriptomics assay based on Luminex bead technology and generated 1,400,000 gene expression profiles representing approximately 20,000 perturbagens across approximately 15 cell lines (LINCS-L1000) [83].

These large data sets necessitated the development of novel approaches to analyze disease or drug-treated data and visualize output, such as L1000CDS2 [84]. L1000CDS2 was developed for identifying drug candidates by comparing signatures from the disease state to LINCS-L1000. The L1000CDS2 tool queries gene expression profiles from LINCS-L1000 and converts the expression value to a characteristic direction for each gene, prioritizing the direction, rather than the magnitude, of gene expression changes. This direction is a vector 90° to the hyperplane, separating the normal and drug-treated samples (or disease) in N-dimensional space (where N corresponds to the number of genes). Based on the direction, the normal and drug-treated sample (or disease) signatures can be collapsed into 1-D values, which can be compared. This approach,

called characteristic direction [85], is similar to the concept of differentially expressed genes. The L1000CDS2 tool provides the top 50 ranked small molecules that show high cosine distances based on these profiles [84].

Transcriptomics data and clinical studies from Chronic kidney disease

The 2018 CKD medication reports (https://www.ncbi.nlm.nih.gov/books/NBK492989/) indicate several types of drugs for the treatment of CKD, including angiotensin-converting enzyme inhibitors or sartans for decreasing blood pressure, statins for reducing cholesterol, and aspirin for preventing clotting [86],[87]. However, it remains unclear whether these 3 types of medication are suitable for treating CKD. Drug repositioning for CKD, particularly computationally based drug repositioning using transcriptomics data sets, could accelerate the efficient identification of new, potential medications for CKD.

The development and advancement of microarray technology have led to the generation of several CKD transcriptomics data sets. While these data sets are smaller than for other diseases (e.g., cancers), there are several biologically meaningful public data sets in the Gene Expression Omnibus (GEO). For example, Nephroseq (https://www.nephroseq.org/) provides 26 public gene expression data from renal disease in both humans and mice, especially human data of which has detailed clinical information [88]. Tajti et al. analyzed human CKD data from 5 different public data sets and 9 different CKD subtypes, using non-tumor parts of kidney cancer nephrectomy tissues as controls [89]. In 2017, I performed drug repositioning using the L1000CDS2 tool to match expression data from these 9 CKD subtypes to the drug-treated data sets reversely in LINCS-L1000 [84],[89].

2. Methods

2.1 Single cell RNA-Seq data preprocessing from mice renal fibrosis

UMI mapping to reference genome

CellRanger (version v3.2.0) was used to map reads of fastq files generated by 10x technology to mouse reference genome (GRCm38.p6). Several softwares for read mapping and quality controls were built in CellRanger, such as STAR [90], EmptyDrops [91], etc. STAR has been used for mapping reads in bulk RNA-Seq, and EmptyDrops was developed for single cell RNA-Seq data, especially, to infer empty droplets from the data set with. CellRanger provided the quality reports for the data, such as how many cells were detected, the mean reads for each cell, the total number of reads, the percentage of reads mapped confidently to exonic regions and intronic regions, etc. I focused on two information; the reads mapped confidently to exonic regions and sequencing saturation. If the percentage of reads mapped to exonic regions was lower than 50%, and the sequencing saturation could not reach the point around 0.5, I decided that the data set was not suitable for further analysis. In other words, if many reads from the data set were mapped to the intron region in the reference genome, I considered the data set to have a lack of mRNA volume. As well, if the sequence saturation was severely low, I inferred that this data didn't have enough genes suitable for further analysis.

For the data set passed through criteria described above, I used barcodes.tsv, genes.tsv and matrix.mtx as outputs generated by CellRanger. Regarding the output, "barcode.tsv" has the information of cell barcodes corresponding to each single cell and "genes.tsv" includes all of the genes from the reference genome. The last output, "matrix.mtx" file has the information of how many times each Unique Molecular Identifier (UMI) was mapped to each gene in given single cells. Instead of using reads, using Unique Molecular Identifier (UMI) is the uniqueness of 10x technology. This Unique Molecular Identifier (UMI) was developed to solve the technical issues caused by Polymerase Chain Reaction (PCR). Polymerase Chain Reaction (PCR) has been used to amplify reads because the amount of reads from samples were very small so without PCR, it is difficult to map reads to reference genome. However, by chance, specific reads could be amplified more times than other reads. In order to solve this issue, a short length of nucleotide, UMI, is added to each read of barcodes in single cells before PCR. Here, barcodes are used for naming each single cell differently. In other words, all reads from the identical single cell have the same barcode. With this technique, I could get the output from equally amplified reads per

each single cell. Compared to other single cell projects, this project focused on the specific cells targeted by Gli1, Ng2, Myh11, Pdgfrb. In the case of the Gli1, Ng2, Myh11 and Pdgfrb, these genes were combined with the tdTomato genes when generating the transgenic reporter mice. Supplementary table 4.1.1 described tdTomato sequence.

Filtering cells and genes and normalization

After I obtained three outputs (matrix.mtx, barcode.tsv and genes.tsv) from CellRanger, I analysed them with Seurat. First step was to filter out cells based on two criterias. If some cells expressed less than 200 genes and some genes were detected in less than 3 cells, I decided that such cells and genes were not suitable for further analysis. Additionally, I defined dead cells if the reads mapped to mitochondrial genes had higher than 80 % of total read counts. This is because apoptotic cells induce the expression of mitochondrial genes and transport them into the cytoplasm in mammalian cells. Compared to RNA-sequencing from single nuclei, single cell data is based on cytoplasm so I had to check the fraction of mitochondrial genes. In general, 5 ~10 % has been used for the suitable percentage of mitochondrial genes in a single cell. However, some specific cell types of the kidney, such as proximal tubules, need more active transporting systems because of the reabsorbing filtrates passed through glomerulus. So, I used 80 % as the maximum percentage cutoff. With the filtered data, I divided UMI counts with the total number of counts for each cell, multiplied the normalized counts by a scale factor (default value is 10,000) and then log-transformed them.

Getting highly variable genes, transforming data linearly and running PCA

Normalized matrices were too big and heterogeneous to perform further analysis so that I needed to select features which explained high variance across cells. In order to select features, I used the function variance-stabilizing transformation (vst) in Seurat which uses LOESS with the mean-variance relationship of log-scaled data, and obtains a LOESS fitted model [19]. LOESS regression has been one of the nonparametric regression methods used for non-linear relation between explanatory variables and response variables [19]. With LOESS regression, Seurat found the fitted model and selected the top 2,000 highly variable genes. Seurat scaled the data by setting mean as 0 and variance as 1 for each gene [19]. Here, I gained 2,000 features as highly variable genes in a normalized, scaled data matrix. As a next step, I reduced the dimensions of cells with principal component analysis (PCA). PCA has been used to reduce big data into small one while keeping the variance of data. When I chose how many dimensions to take after reducing dimensions, I used an elbow plot which had the information about the standard deviations for each dimension. For nine different data sets, I took around 40 ~ 50 principal

components until the steep slope of the elbow plot became flat.

Graph-based clustering

After reducing dimensions, I used a function, “FindClusters” in Seurat to perform clustering [21]. This is based on graph-based clustering, especially K-nearest neighbors. There were many approaches on clustering, hierarchical clustering, density-based clustering and consensus clustering. The reason I chose graph-based clustering in Seurat was first, the data set was high dimensional sparse and heterogeneous so I didn’t use density-based clustering which assumes all density of clusters should be the same. With the same reason, I decided that measuring euclidean distance between cells was not a promising approach so I didn’t conduct hierarchical clustering. In addition, I analysed big data sets so computationally it was not suitable to use consensus clustering which requires high operations and long running time. Graph-based clustering has been known to be useful to analyse big, heterogeneous and sparse data because it uses the concept, a network. Inside this graph-based clustering, it has two processes, first is to find neighbors, second is to find a community connecting neighbors. In detail, the first starts with a similarity matrix driven by the data set reduced by PCA. Based on the similarity matrix, Seurat connects top 20 similar cells for each cell, re-weight edges of cells based on the shared neighbors (SNN), and then measures a modularity score for the connected cells [21]. When measuring the modularity score, Seurat randomly chooses one single cell as a starting point and then connects cells in a way to maximize a modularity score by the Louvain algorithm [21]. When all of the cells are connected, the modularity score becomes zero. The modularity score would be max only with an optimal partition. Sometimes, the optimal case couldn’t be the only one. In order words, I could find several different cases on ideal clusterings which maximize modularity score written below :

$$\text{Modularity score} = \sum_{n=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{K_c}{2L}\right)^2 \right]$$

where “n_c” is the number of communities (clusters) and “Lc” is the total number of links in cluster “c” and “Kc” is the total node degree in cluster “c” and “L” is the total number of links. Seurat provides several clustering outputs based on a parameter named “resolution” and also recommends using resolution diversely. Some researchers use Silhouette width to choose the most ideal clustering output. However, the data set for this project consisted of similar cell types, so I decided to take as many clusters as I could and then merge some clusters if they have similar

gene expression patterns for the marker genes. In order to move on to the next step, annotation, I made UMAP (Uniform Manifold Approximation and Projection) plots and Heatmap plots for the clustering output coloring known marker genes of the kidney [92]. UMAP has been widely used for 2-dimension reduction by building up a high-dimensional graph representation (topological representation) of the data, and then optimizing a low-dimensional graph as similar as possible.

Cluster Annotation and differentially expressed genes

After getting clusters from Seurat, I used genesorteR [20], a R package which provides the summarized information for each cluster. Instead of mapping clusters to cell types, it measured the fraction of cells which express each gene in a given cluster. If the normalized UMI count of a gene in a given cluster is higher than the median of non-zero UMI counts across the single cells, it counts the cell expressing the gene [20]. As well, it provides genesorteR's unique value, specificity score [20]. This score represents how exclusively and highly a given gene is expressed for each cluster by a Bayesian approach using three different prior probabilities as below:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Where the P(E|H) is the probability of cells expressing the gene "E" in a given cell type, "H". The P(H) is the fraction of the cell type "H" across all cells, P(E) is the probability of cells expressing the gene "E" in all cells. Instead of the specificity score, I mainly used one of the prior probability, "CondGeneProb", P(E|H). The reason I used this probability was that this score doesn't conduct comparisons between clusters. So, I could get independently summarized information on the matrix with genes in given clusters. With the matrix, the fraction of cells expressing genes ("CondGeneProb"), Christoph Kuppe, MD, PhD in RWTH Uniklinik Aachen mapped each cluster to the cell type with known gene markers. In the case of the differentially expressed gene for each cell type, "FindAllMarkers" in Seurat was used [19]. This tool applied a Wilcoxon Rank Sum test, one of nonparametric tests which don't require normally distributed data as an input.

Integration

For integration, two approaches were used, Harmony and Seurat [22], [93]. Seurat provided an integration approach using canonical correlation analysis with mutual nearest neighbors [93]. Canonical Correlation Analysis (CCA) has been used to connect two sets of multiple variables to

make high correlation between the variables. In order words, this approach minimizes the cosine distances between two linear projections from paired variables in a lower dimensional space. Like PCA, I chose the number of canonical components which explains the variance of the data set. After reducing dimensions, Seurat performed L2-normalization to the canonical correlation vectors. It represents that the direction of vectors is more important than the magnitude of the vector in a data set because the data set was reduced by CCA. As a next step, Seurat used Mutual Nearest Neighbors (MNN) [93]. This concept is based on graph-based clustering, such as Shared Nearest Neighbors. Seurat connected a single cell with five other cells which shared neighbors with the single cell. This is called "identifying anchors". Based on this approach, I integrated our data set and then re-run PCA in order to reduce the dimension of the data set. And then, I made a UMAP for visualization and then evaluated if the integration makes sense biologically or not.

As well as Seurat's standardized integration method, recently, harmony has been used widely for integration [22]. It measured the possibility of cells belonging to multiple clusters, in which the possibility ranges from 0 to 1. Harmony maximizes the diversity between clusters while minimizing the effect of batches by computing new centroids from centers of batches in a given cluster [22]. Harmony corrected normalized data sets or PCA embedding matrices. Here, I selected the first approach to use Harmony on the normalized matrix. And then, I got highly variable genes based on the corrected normalized matrix, scaled and then performed data reduction by PCA and UMAP.

After integration was complete, I computed cell-type correlation for the integrated data set. As an input, positively scaled gene expression matrices were used. genesorteR was applied to collapse the positively scaled gene expression matrix from single cell level to cell type level. genesoreR measured how many cells expressed the genes in a given cell type level. Pearson correlation was used to measure the similarity at a cell type level. The formula of measuring similarity is written below:

$$\text{cor(a,b)} = \frac{(a-\bar{a})\cdot(b-\bar{b})}{\|(a-\bar{a})\|_2 \left\|(b-\bar{b})\right\|_2}$$

where $\overline{a, b}$ are the mean of the elements of the cell type, a and b and $\cdot$ is the dot product of a and b.

Biological interpretation

Functional-level studies

The Functional level studies I performed consist of 2 main parts, measuring pathway activity

scores and transcription factor activity scores.

Pathway activity score

Computing the pathway activity score consists of 3 different types, first is to use the given weight matrix by PROGENy for each pathway, second is to get module score at a single cell level with a known gene set of pathway and the last one is to use gene set enrichment approach with public gene set data sets at a cell type level.

I measured PROGENy score for 14 biological pathways in a given cell population (Estrogen, Androgen, TGFb, WNT, PI3K, Trail, p53, NFkB, TNFa, EGFR, MAPK, JAK-STAT, Hypoxia and VEGF) [30]. I used the gene-by-pathway weight matrix for 14 pathways where the weight value indicates the correlation between genes and 14 pathways. Positively high weight means that this gene was upregulated by this pathway positively and vice versa. Briefly, this correlated weight matrix was generated from experiment data sets blocking a known pathway, such as Estrogen, TGFb, WNT, etc. With a matched normal data set to the perturbed experiments, PROGENy measured the z-score of each gene in the perturbed data set [30]. PROGENy conducted multiple linear regression between the z-score as a dependent variable and the meta information of all experiments as an independent variable and got a correlation coefficient for each gene in a given pathway [30]. Among lots of correlation coefficients, PROGENy provided a significant correlation coefficient for each pathway as a weighted matrix. I applied PROGENy to the normalized data set, and averaged PROGENy scores in a given cell population for each pathway. For visualization, I made a heatmap [30].

At a single cell level, I measured extracellular matrix (ECM) scores with a module score. I gained specific gene sets of ten different ECM related biological categories, such as collagenes, ECM glycoproteins, ECM regulators, proteoglycans, etc from MsigDB [25]. With these gene sets, I used “AddModuleScore” in Seurat [29]. This function averaged the expressions of genes set at a single cell level and then subtracted this value by the aggregated expression of 100 genes selected randomly [29]. Regarding the random selection, first, all features (genes) were binned based on the averaged values and second, the control features (genes) were selected for each bin.

For gene set enrichment tests with public gene sets, Enrichr was used [26]. Enrichr requires inputs with the type of “gene list” or “gene list with ranking” instead of gene expression values like what GSEA needs. Gene expressions at very low levels are the main feature of single cell RNA-seq so that Enrichr is more suitable for single cell level studies than GSEA. When selecting genes, top 100 highly expressed genes in a positively scaled matrix were used instead of differentially expressed genes. This is because the data sets were composed of similar cell types,

so only top 5 or 10 differentially expressed genes have meaningful differences to separate cell types. However, for the enrichment studies, I need at least 50 or 100 genes to get statistical power. For this, the top 100 highly expressed and positively scaled genes were used. AUCell was used to rank genes based on "CondGeneProb" for each cell type [28]. The Enrichr script in R was used for this study. EnrichR automatically queried the public gene set of pathways from "GO Biological Process 2018" [27].

Transcription factor activity score

When I inferred the activity of transcription factors, the Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) was used to measure the relative activity of each transcription factor [31]. The activity of each transcription factor was calculated by using its known targets, in which the targets and transcription factors were provided by the well known databases, DoRothEA [32].

Intercellular communication

For inferring intercellular interactions, quality-based, CellChat and quantity-based, ICELLNET with genesorteR were used.

First, for running CellChat, a library-size normalized matrix was used. The CellChat infers intercellular interactions by four delicate steps. First is to find differentially expressed genes for each cell population compared to others, using the Wilcoxon rank sum test (p value < 0.05) [43]. Second is to use a statistically robust mean instead of measuring averages when calculating averaged gene expression [43]. As a third, the CellChat measures the intercellular probability with a random walk based network propagation approach [43]. In other words, it maps gene from the expression matrix on a validated protein-protein interactions from STRINGdb and then it measures the probability, the main concept of which is a hill equation used in measuring the coefficient when ligands interact with receptors in a chemical reaction [94], [95]. As a last step, CellChat identifies the significant ligand-receptor interactions in the given two cell populations by permuting the group annotation of cells.

ICELLNET database and genesorteR were used to perform separate intercellular communications in order to expand the output from CellChat. ICELLNET provides manually curated pairs of ligands and receptors, especially, immune-related pathways, such as growth factors, cytokines, chemokines, hormone, checkpoints, notch signaling and antigen binding [40]. I used the source of ligand and receptors of the ICELLNET because there has been a strong correlation between immune response and renal fibrosis. I measured the percentage of cells

expressing the ligand and receptors to infer the intercellular communications in given two populations. When measuring the percentage of cells, the "CondGeneProb" of genesorteR was used [24].

Cell differentiation

In a bulk RNA-Seq field, the number of pre-mRNA (unspliced mRNA) molecules has been used to infer if the mRNA will be upregulated or repressed. The assumption is that if mRNA abundance of a gene is higher than that of pre-mRNA, the mRNA will decrease in order to keep homeostasis of the mRNA level. Computationally, the amount of both pre-mRNA and mRNA are measured when reads are mapped to the reference genome. The amount of pre-mRNA of a gene is the number of reads mapped to both the intron and exon, called this an "unsplied" read. If the read is mapped to the exon area in the genome, it is called a "spliced" read corresponding to mRNA.

This approach recently has been used for single cell RNA-Seq. data to infer cell dynamics, such as cell differentiation. This pseudotime analysis or trajectory inference has been applied to predict which cells or cell types are the highly differentiated or the origin of differentiations. This approach uses diverse methods including diffusion map and minimum spanning tree (mst). PAGA infers pseudotime based on similarity between transcriptome profiles at a cell population level and then combines this with a diffusion map [37]. Slingshot applies a Minimum Spanning Tree (MST) to cell populations [34].

Unlike the methods like PAGA and Slingshot, RNA velocity has also been used at a single cell level. There are two main tools which adopted this approach, Veloctyo [35] and scVelo [36]. The main assumption of Velocyto is that all genes have a common splicing rate [35], but scVelo disagreed with this assumption and then developed an advanced approach with a likelihood-based dynamical model [36]. scVelo assumes that for each single cell, RNA velocity could have a wide variety of dynamics like developments [36].

In order to do cell differentiation study for this project, three approaches were used, PAGA[38], scVelo[36] and Velocyto[35]. Normalized matrices of two integrated data are required as input data for PAGA and scVelo (including Velocyto). Additionally, a batch-corrected positively scaled matrix was also used (including Velocyto) because the batch-corrected normalized matrix had negative values as a minimum. First, PAGA measured the population similarity based on the modularity which is computed by the number of neighboring cells. Second, PAGA measured

pseudotime with the modified diffusion map, and then map this pseudotime to the collapsed map where each node is cell type and each edge is the modularity score [38]. In the case of the scVelo, first, matrices of integrated data sets were used as well as bam files, one of the outputs of CellRanger. Based on these two inputs, scVelo inferred pseudotime by 3 different models, deterministic, stochastic and dynamical models. Here, the deterministic model corresponds to the Velocyto model. These three models were used to infer pseudotime, and then the most suitable model was selected based on the pseudotime matched to biological knowledge. When the pseudotime model was selected, PAGA collapsed the velocity values at a cell type level, and made the directions of time flows which move from lower velocity to higher one. In detail, the direction was made by the transition probability between two single cells. This transition between two single cells was computed by comparing the gene expression changes (gene by cell) with the velocity (gene by cell) by cosine correlation [36]. As well, scVelo measured the differentially inferred velocity genes for each cell type in a given model, in order words, the genes in a given cell type which is transcriptionally regulated differently compared to all other cell types [36]. For this analysis, scVelo used a differential expression test by Welch t-test on the inferred velocity expression [36].

2.2 Drug repositioning on chronic kidney disease microarray data

Characteristic Direction approach was applied to the normalized microarray data set on 9 different human chronic kidney diseases [85],[89]. The assumption of the Characteristic Direction is that in vector spaces, the direction of vectors made by genes is more important than the magnitude like gene expression values. The Characteristic Direction approach identified the hyperplane which separates disease and control by a linear discriminant analysis. As a next step, the software found a linear line which had a 90 degree angle to the hyperplane, and calculated the cosine (cosine similarity) between the linear line and a vector of each gene [85].

I used this output as an input to L1000CDS2 [84]. This software automatically called the profiles of drug-treated cell line data in which gene expressions were already converted to cosine by the Characteristic Direction [85]. I ran the L1000CDS2 software and got the top 50 small molecules in a descending order from the highest score of 1-cosine values between two profiles of drug-treated and disease. If the score was the highest, the two profiles were the most reversely matched to each other. As an output, meta information matched to the top 50 ranked molecules were given, such as the score from 1-cosine, DrugBank link, Cell-line, Does, Time, p value, etc. I used only the name of small molecules without considering other annotations for further analysis. In order to take significant drug candidates, I calculated the adjusted p-values with Benjamini-Hochberg correction and filtered out those which had higher than 0.05 adjusted

p-values. After I found common drug candidates from at least 3 different CKD diseases, literature curations were conducted in order to know which molecules have scientific evidence on the therapeutic potential of the drug candidates for the CKD [89].

3. Results

This section describes the analysis of single-cell RNA-seq (scRNA-seq) in mouse renal fibrosis models and drug repositioning using human chronic kidney disease (CKD) microarray data sets.

3.1 Single cell RNA-Seq data preprocessing from mice renal fibrosis

Data preparation and read alignments

scRNA-seq was performed using 10 different reporter mice (Gli1, Myh11, Ng2, Pdgfrb, and Cd31), each in normal (sham) and renal fibrosis (unilateral ureteral obstruction model, UUO) mouse models. Dr. Kuppe (RWTH Uniklinik Aachen) generated 9 different data sets, with public data used for the Cd31 normal mouse data [73].

CellRanger was used to align reads from the fastq files to the reference genome and generate a gene expression count matrix. The mouse data sets were assessed using CellRanger quality reports, which include the number of detected cells and mean reads per cell (total number of sequenced reads divided by estimated number of cells). The number of sequenced cells ranged from 2,000–10,000 with 20,000–80,000 mean reads/cell across all data sets. The quality reports also include information regarding the fraction of reads mapped confidently to exonic regions, which was crucial to reviewing sequencing quality. Data sets with < 50 % of the reads mapped confidently to exonic regions were removed.

Filtering, Clustering and Annotation

After aligning with CellRanger, Seurat was used to filter cells and genes and to cluster and annotate reads. Cells with a high percentage of reads mapped to mitochondrial genes (> 80 %) and those expected to be doublets based on an unexpectedly high number of expressed genes or reads were filtered out of the data. Mitochondrial genes typically comprise < 20 % of RNA in most organs, but kidneys are known to require higher mitochondrial function, necessitating the higher (80 %) cut-off. Cells expressing fewer than 200 genes or genes detected in < 3 cells were also removed.

After filtering, Seurat was used for unsupervised clustering [93].Various numbers of clusters were generated depending on resolution. Only specific clusters expressing the tdTomato marker gene were used in subsequent analyses. As the Cd31 mice, sham (normal) and UUO (renal

fibrosis), did not express tdTomato, all clusters expressing the Cd31 gene were used in downstream analysis.

Each cluster was then assigned to a cell type (Figure 3.1.1). The Gli1 sham mice data contained pericytes, vascular smooth muscle cells (VSMCs), endothelial cells, Schwann cells, and 12 different fibroblasts, while the Gli1 UUO mice data were comprised of 2 different pericytes, VSMCs, Schwann cells, 3 different fibroblasts, 11 different myofibroblasts, neurons, and unclear cell types. In Myh11 sham mice data, adipocytes, 4 different fibroblasts, pericytes, and neurons were identified, while 10 different fibroblasts, myofibroblasts, pericytes, and unclear cell types were detected in the Myh11 UUO mice data. Pdgfrb sham mice data were annotated as neurons, pericytes, and 8 different fibroblasts, while the Pdgfrb UUO mice data were composed of 3 different myofibroblasts, 4 different fibroblasts, epithelial cells, pericytes, and podocytes. The Cd31 sham mice data contained arteriole, ascending vasa recta, capillary, descending vasa recta (DVR), postcapillary venule, and vein-related cell types (Table 3.1.1). The Cd31 UUO mice data were annotated as glomerular (glm), lymphatic (lymph), fenestrated (fnst), medullary (mdl), and DVR endothelial cells by Dr. Kuppe.

compartment I	compartment II	compartment III	sub types	subtypes based on phenotype
cortex		vein		
cortex		capillary		angiogenic, interferon, and postcapillary vein
cortex		artery		
cortex		arteriole	efferent	
cortex	glomeruli	arteriole	afferent	juxtaglomerular apparatus
cortex	glomeruli	arteriole	efferent	juxtaglomerular apparatus
cortex	glomeruli	capillary		
medulla	vasa recta	AVR		papilla, interferon
medulla	vasa recta	DVR		papilla, interferon
medulla		arteriole		
medulla		capillary		angiogenic, interferon and postcapillary vein

Table 3.1.1. The cell type labeling table of the Cd31 sham mice data set. This table shows the categorization of cell type annotation of Cd31 sham mice data sets. Arteriole is the small branch of an artery heading to capillaries. Vasa recta is the straight arterioles which enter the medulla as the descending straight arterioles (DVR), and leave the medulla to ascend to the cortex with the straight venules (AVR). Capillaries connect arteries (from heart) to vein (to heart).

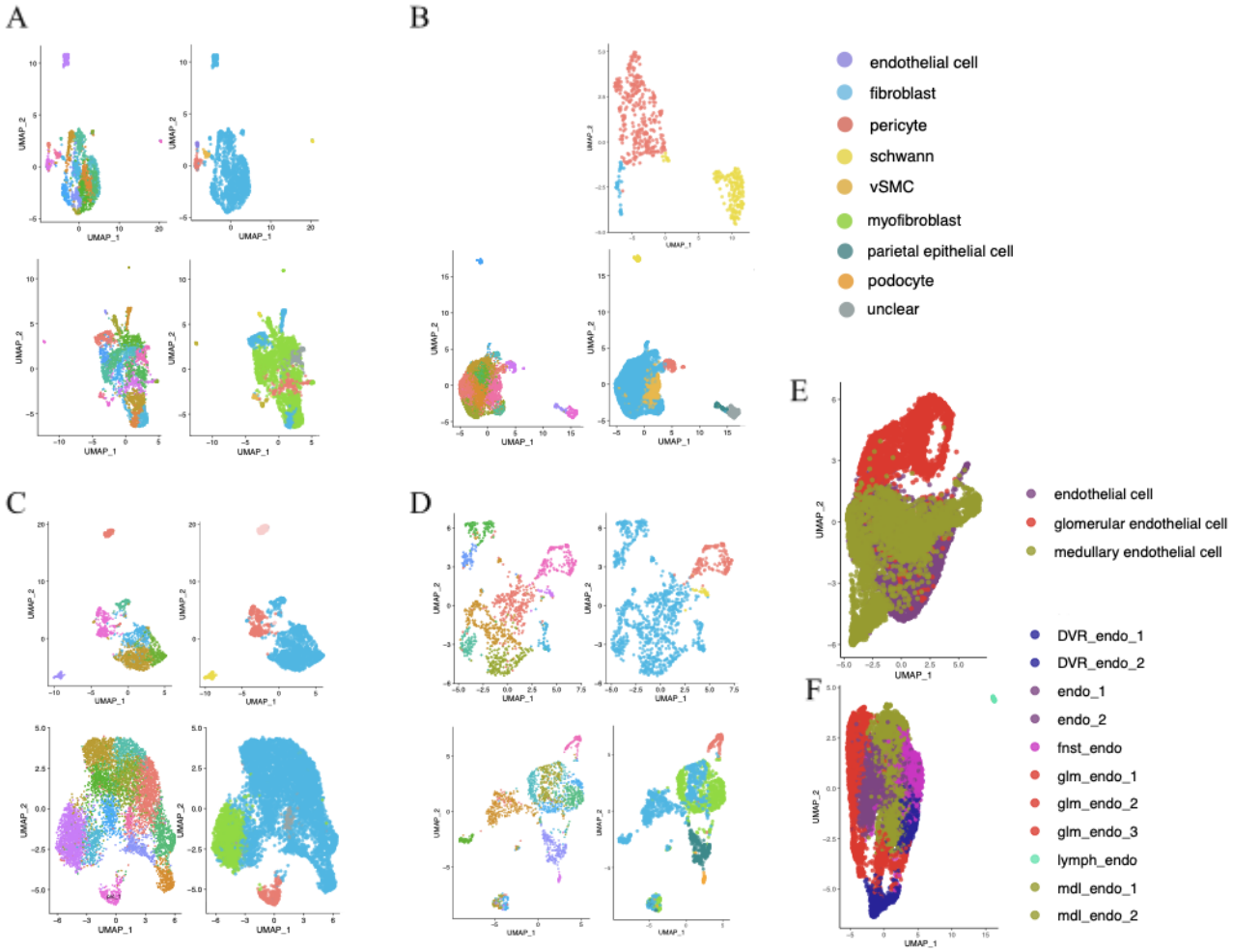


Figure 3.1.1. The clustering and annotation outputs of 10 different data sets. All figures from Figure A (Gli), B (Ng2), C (Myh11) and D (Pdgfrb) are composed of 4 sub-figures. Inside Figure A-D, the upper panel (upper left and right figures) is from normal mice (SHAM), below one (below left and below right) from renal fibrosis mice (UUO), left panel (upper left and below left) from clustering outputs, right panel (upper right and below right) from annotation colored by cell-types. Each color corresponds to different cell types. Fibroblasts are blue, myofibroblasts are green, neurons & schwann are yellow, pericytes are pink, podocytes are orange, parietal epithelial cells are olive, vascular smooth cells are dark orange, unclear are gray. Figure E is the annotation output of Cd31 normal mice, F is from Cd31 renal fibrosis mice. Purple means endothelial cells, olive for medullary endothelial cells, red for glomerular endothelial cells. In figure F, descending vasa recta endothelial cells (DVR), endothelial cells (endo), glomerular endothelial cells (glm) and medullary endothelial cells (mdl) have 2 different sub cell types marked 1 and 2.

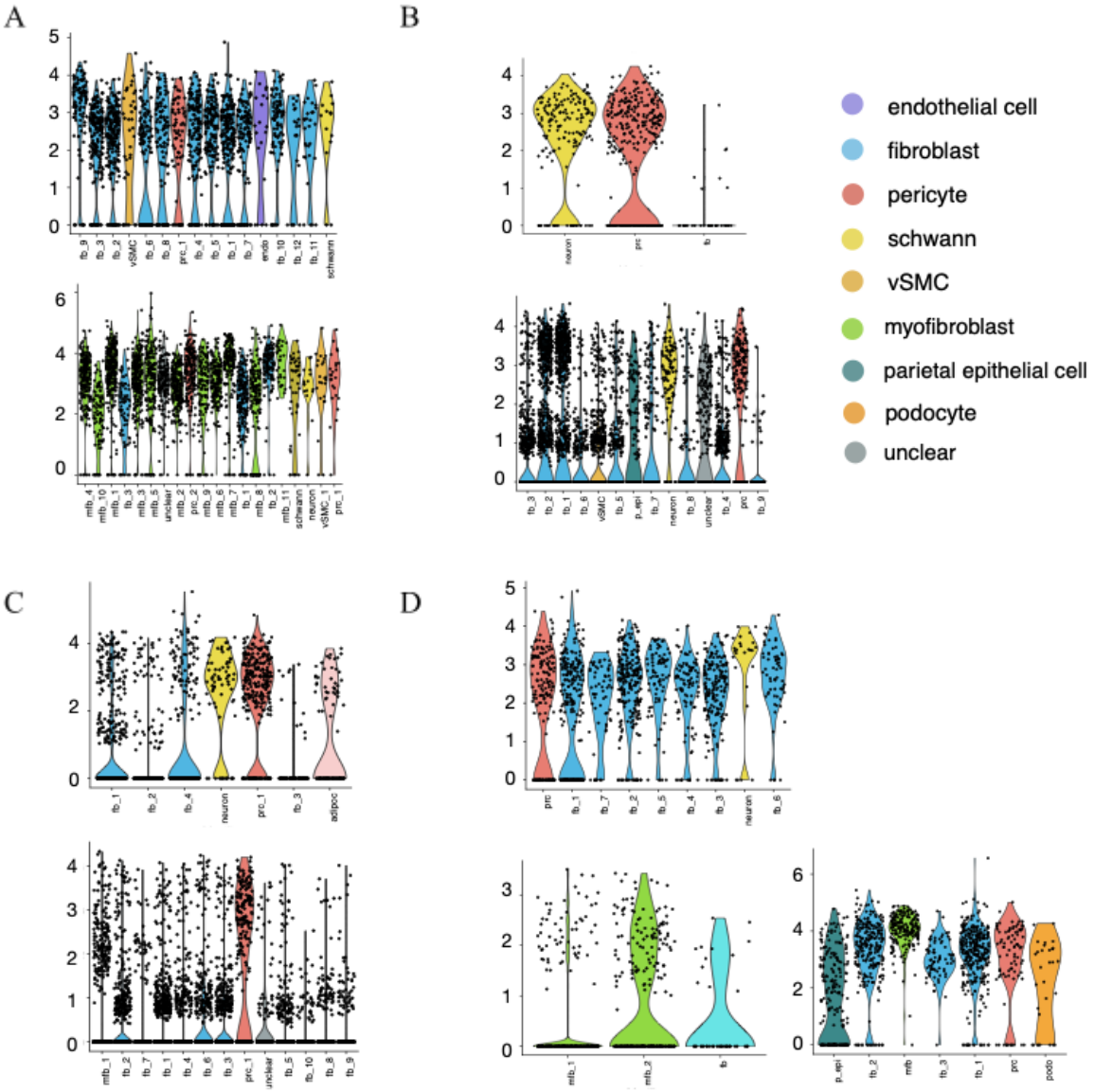


Figure 3.1.2. The expression of td-Tomato across 9 different data sets. All figures from Figure A (Gli), B (Ng2), C (Myh11), D (Pdgfrb) have two kinds of dot plots with the tdTomato expression, upper figure is from normal mice (SHAM), below one from renal fibrosis mice (UUO). In the case of the Figure D (Pdgfrb), the panel below (renal fibrosis mice) had two kinds of dots plots from 2 different data sets. Each color corresponds to different cell types. Fibroblasts are blue, myofibroblasts are green, neurons & schwann are yellow, pericytes are pink, podocytes are orange, parietal epithelial cells are olive, vascular smooth cells are dark orange, unclear are gray. These plots indicate that all annotated clusters express enough tdTomato.

The data and cell types used in this work are summarized in Figures 3.1.3 and 3.1.4. The Cd31 sham data is the largest data set, with a total of 10,186 cells. Fibroblasts are the most frequent cell type, with a total of 14,328 cells, followed by myofibroblasts with 3,915 cells. The number of genes ranges from 10,075–19,104 across all data sets.

type	Cd31_SHAM	Cd31_UUO	Gli1_SHAM	Gli1_UUO	Myh11_SHAM	Myh11_UUO	Ng2_SHAM	Ng2_UUO	Pdgfb_SHAM	Pdgfb_UUO		
total_gene	10075	17168	15285	15596	16775	18624	15977	18027	17984	19104		
total_cell	10186	8635	1645	2682	1716	7355	549	5072	1166	1689		
CD31_cRECs_SHAM	4275	0	0	0	0	0	0	0	0	0		
CD31_gRECs_SHAM	2485	0	0	0	0	0	0	0	0	0		
CD31_mRECs_SHAM	3426	0	0	0	0	0	0	0	0	0		
CK154_PDGFB_UUO	0	0	0	0	0	0	0	0	0	1159		
CK37_PDGFB_UUO	0	0	0	0	0	0	0	0	0	530	counts	percentage
DVR endothelial cell	0	1012	0	0	0	0	0	0	0	0	1012	2,486791989
adipocyte	0	0	0	0	94	0	0	0	0	0	94	0,230986608
arteriole	291	0	0	0	0	0	0	0	0	0	291	0,715075562
arteriole/afferent	158	0	0	0	0	0	0	0	0	0	158	0,388254085
arteriole/afferent/ juxtaglomerular_apparatus	193	0	0	0	0	0	0	0	0	0	193	0,474259737
arteriole/efferent	463	0	0	0	0	0	0	0	0	0	463	1,137731908
arteriole/efferent/ juxtaglomerular_apparatus	176	0	0	0	0	0	0	0	0	0	176	0,432485563
artery/large	45	0	0	0	0	0	0	0	0	0	45	0,110578695
ascending_vasa_recta	419	0	0	0	0	0	0	0	0	0	419	1,029610517
ascending_vasa_recta/ in/terferon	133	0	0	0	0	0	0	0	0	0	133	0,326821477
ascending_vasa_recta/ papilla	321	0	0	0	0	0	0	0	0	0	321	0,788794692
capillary	2764	0	0	0	0	0	0	0	0	0	2764	6,791989188
capillary/angiogenic	585	0	0	0	0	0	0	0	0	0	585	1,437523037
capillary/interferon	672	0	0	0	0	0	0	0	0	0	672	1,651308515
capillary_#1	1417	0	0	0	0	0	0	0	0	0	1417	3,482000246
capillary_#2	881	0	0	0	0	0	0	0	0	0	881	2,164885121
descending_vasa_recta	197	0	0	0	0	0	0	0	0	0	197	0,484088954
descending_vasa_recta/ papilla	209	0	0	0	0	0	0	0	0	0	209	0,513576606
endothelial cell	0	1920	20	0	0	0	0	0	0	0	1940	4,767170414
fenestrated endothelial cell	0	1121	0	0	0	0	0	0	0	0	1121	2,754638162
fibroblast	0	0	1508	397	1281	5610	35	3845	975	677	14328	35,20825654
glomerular endothelial cell	0	2240	0	0	0	0	0	0	0	0	2240	5,504361715
lymphatic endothelial cell	0	41	0	0	0	0	0	0	0	0	41	0,100749478
medullary endothelial cell	0	2301	0	0	0	0	0	0	0	0	2301	5,65425728
myofibroblast	0	0	0	1830	0	1400	0	0	0	685	3915	9,62034648
neuron	0	0	0	16	69	0	170	87	26	0	368	0,904287996
parietal epithelial cell	0	0	0	0	0	0	0	120	0	209	329	0,808453127
pericyte	0	0	65	202	272	219	344	168	165	88	1523	3,742474505
podocyte	0	0	0	0	0	0	0	0	0	30	30	0,07371913
postcapillary_venule	865	0	0	0	0	0	0	0	0	0	865	2,125568252
schwann	0	0	14	44	0	0	0	0	0	0	58	0,142523652
unclear	0	0	0	169	0	126	0	170	0	0	465	1,142646517
vSMC	0	0	38	24	0	0	0	682	0	0	744	1,828234427
vein	397	0	0	0	0	0	0	0	0	0	397	0,975549822
											40695	100

Figure 3.1.3. The summary of all data sets. This figure consists of three different tables.The green-colored table explains the number of genes and the number of cells for each data set, Gli1, Ng2, Myh11, Pdgfrb and Cd31. Blue table shows the number of cells which belong to each cell type including fibroblasts, pericytes, myofibroblasts, various types of endothelial cells, etc. Orange table shows the total number of cells per cell type across all data sets.

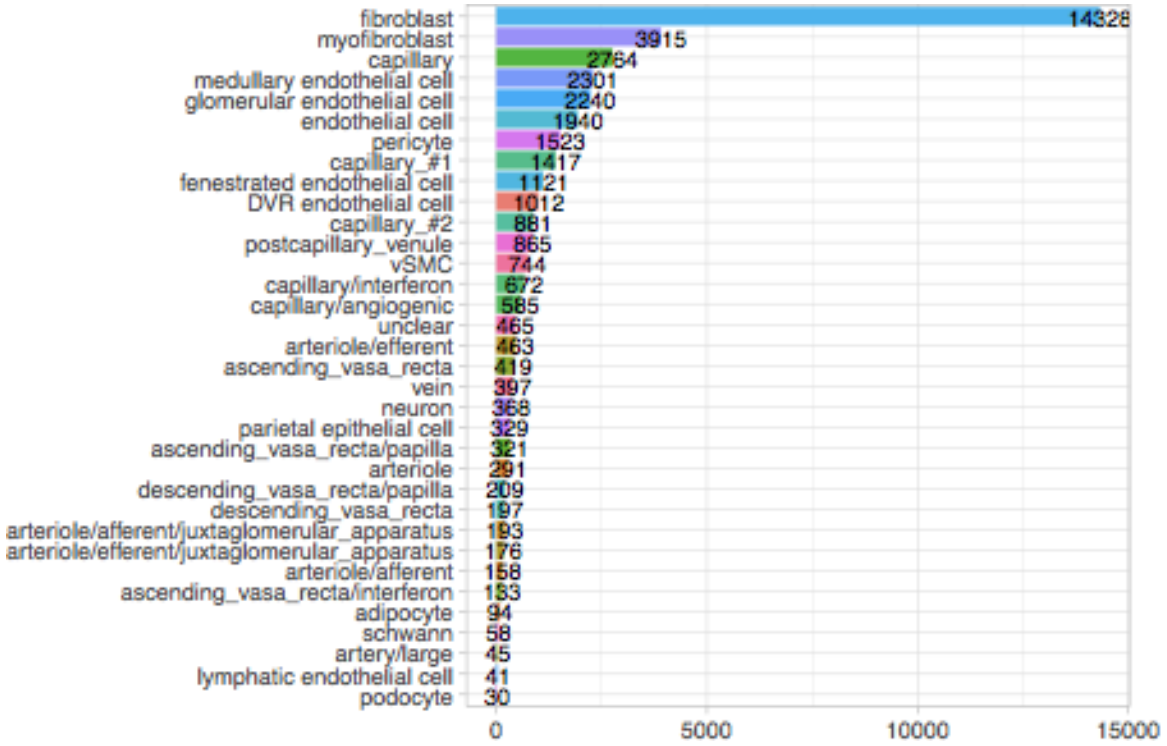


Figure 3.1.4. The summary of all data sets II. The total number of cells per cell type across all data sets ordered by the number of cells from highest to lowest.

Integration

For data integration, two different tools were used, Seurat and Harmony [93], [22]. Before the integration, cell labelings were already conducted for each 10 different reporter mice data set, so the integration approaches were evaluated by UMAP and cell-type wise correlation matrix. Figure 3.1.5.E shows the graphical overview on data integration for Figures 3.1.5-7. Figure 3.1.13 displays the overview for Figures 3.1.8-12.

In Figure 3.1.5, two UMAP and cell type-specific UMAP were displayed. Figure 3.1.5 has the integrated output of all sham mice from Gli1, Ng2, Myh11, Pdgfrb and Cd31. The left panels (A and C) are from the Seurat and the right panels (B and D) are from Harmony. The upper figures (A and B) are UMAP, and lower ones have several UMAP figures coloring only one cell type. Figure 3.1.5.A shows that Seurat integrated all data sets into one big cluster regardless of different cell types, but Figure 3.1.5.B exhibits separate clusters for each cell type such as fibroblasts, pericytes and endothelial cells. In the lower panel of the same figure (Figure C & D), A-G corresponds to adipocytes, endothelial cells, fibroblasts, glomerular endothelial cells, medullary endothelial cells, neurons, pericytes, schwann and vascular smooth muscle cells (vSMC). In Figure 3.1.5.D, I could see that (C) fibroblasts, (G) pericytes, (F) neurons and (B-E) endothelial cells have their own distinct clusters with reduced batch effect in Harmony integration output and it is better than the Seurat where all cell types gather together. The same situation is also found in the output of all integrated UUO mice data sets in Figure 3.1.6. The lower panel (Figure C&D) of the figure 3.1.6 have A-O, in order, DVR endothelial cells, endothelial cells, fenestrated endothelial cells, fibroblasts, glomerular endothelial cells, lymphatic endothelial cells, medullary endothelial cells, myofibroblasts, neurons, parietal epithelial cells, pericytes, podocytes, schwann, unclear, and vSMC. Compared to the output of Seurat in the Figure.3.1.6.A, the harmony UMAP shows clear separation between endothelial cells and non-endothelial cells. In Figure 3.1.6.D, interestingly, (H) myofibroblasts and (J) pericytes were scattered in diverse clusters because of the heterogeneity of these cell types.

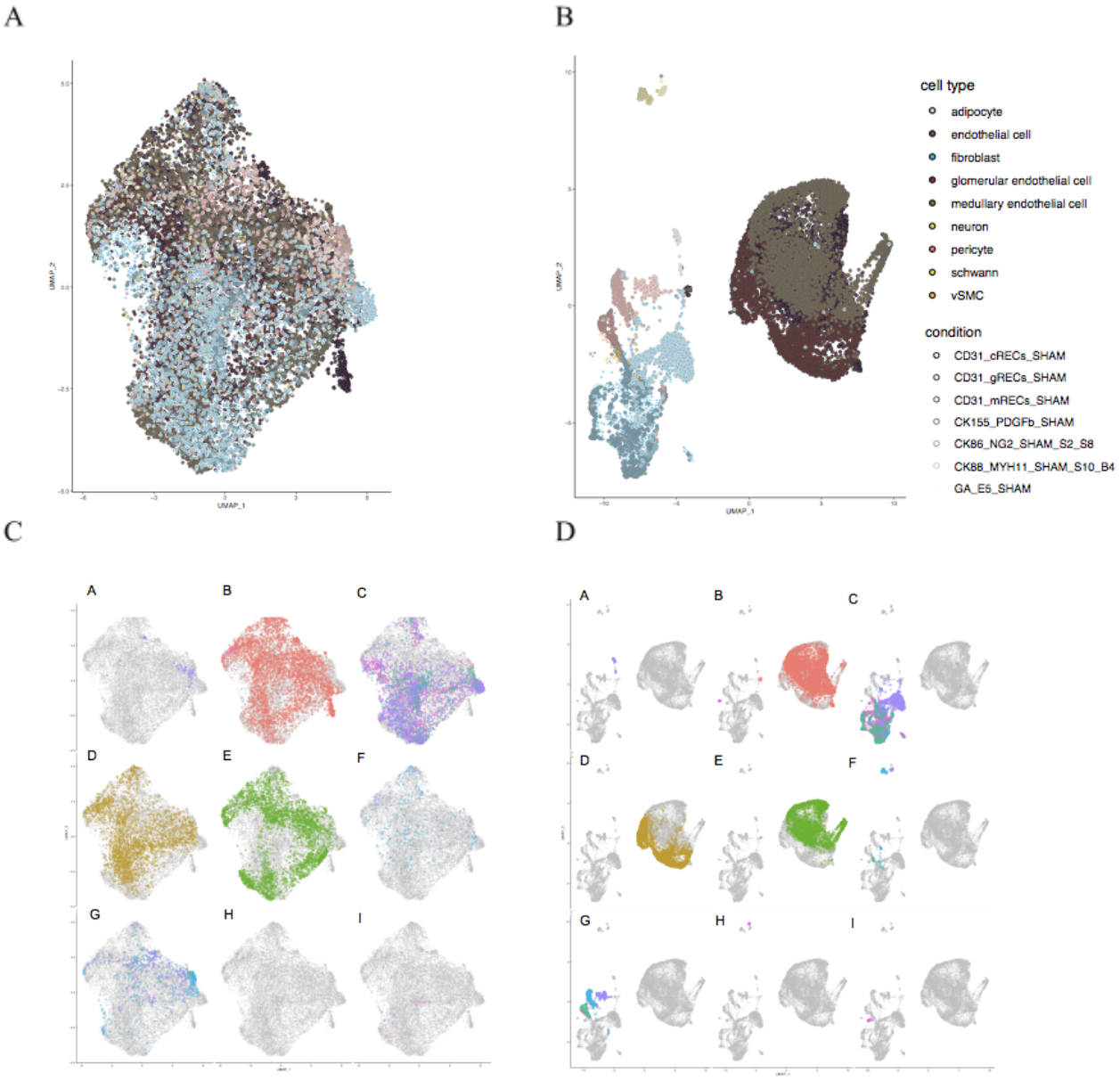

A
B
cell type
adipocyte
endothelial cell
fibroblast
glomerular endothelial cell
medullary endothelial cell
neuron
pericyte
schwann
vSMC
condition
CD31_cRECs_SHAM
CD31_gRECs_SHAM
CD31_mRECs_SHAM
CK155_PDGFb_SHAM
CK86_NG2_SHAM_S2_S8
CK88_MYH11_SHAM_S10_B4
GA_E5_SHAM
C
D
E
Td-tomato
Gli1
Td-tomato
Ng2
Td-tomato
Myh11
Td-tomato
Pdgfrb
Cd31
Normal mice
Data integration
across all
normal mice
Renal fibrosis mice
Data integration
across all
Renal fibrosis mice
Podocytes
Parietal
epithelial
cells
Endothelial
cells
Tubular
epithelial
cells
Percityes
Fibroblasts
Myofibroblasts

Figure 3.1.5. The UMAP and facets of all sham mice data sets integrated by Seurat and Harmony. The left panels (Figure A and Figure C) are from the Seurat, and the right panels (Figure B and Figure D) are from Harmony. The upper figures (Figure A and Figure B) are UMAP, and lower ones (Figure C and Figure D), facets. In facets, there are 9 different cell types, in Figure 3.1.5 C and D, (A) adipocytes, (B) endothelial cells, (C) fibroblasts, (D) glomerular endothelial cells, (E) medullary endothelial cells, (F) neuron, (G) pericytes, (H) schwann and (I) vascular smooth muscle cells (vSMC). Figure E is the graphical overview on the data integration for Figures 3.1.5-7. Data integration was conducted by three approaches, Harmony, Seurat and non-integration approach. For UMAP, Harmony, Seurat were used. For correlation matrix, Harmony, Seurat, non-integration approach were used.

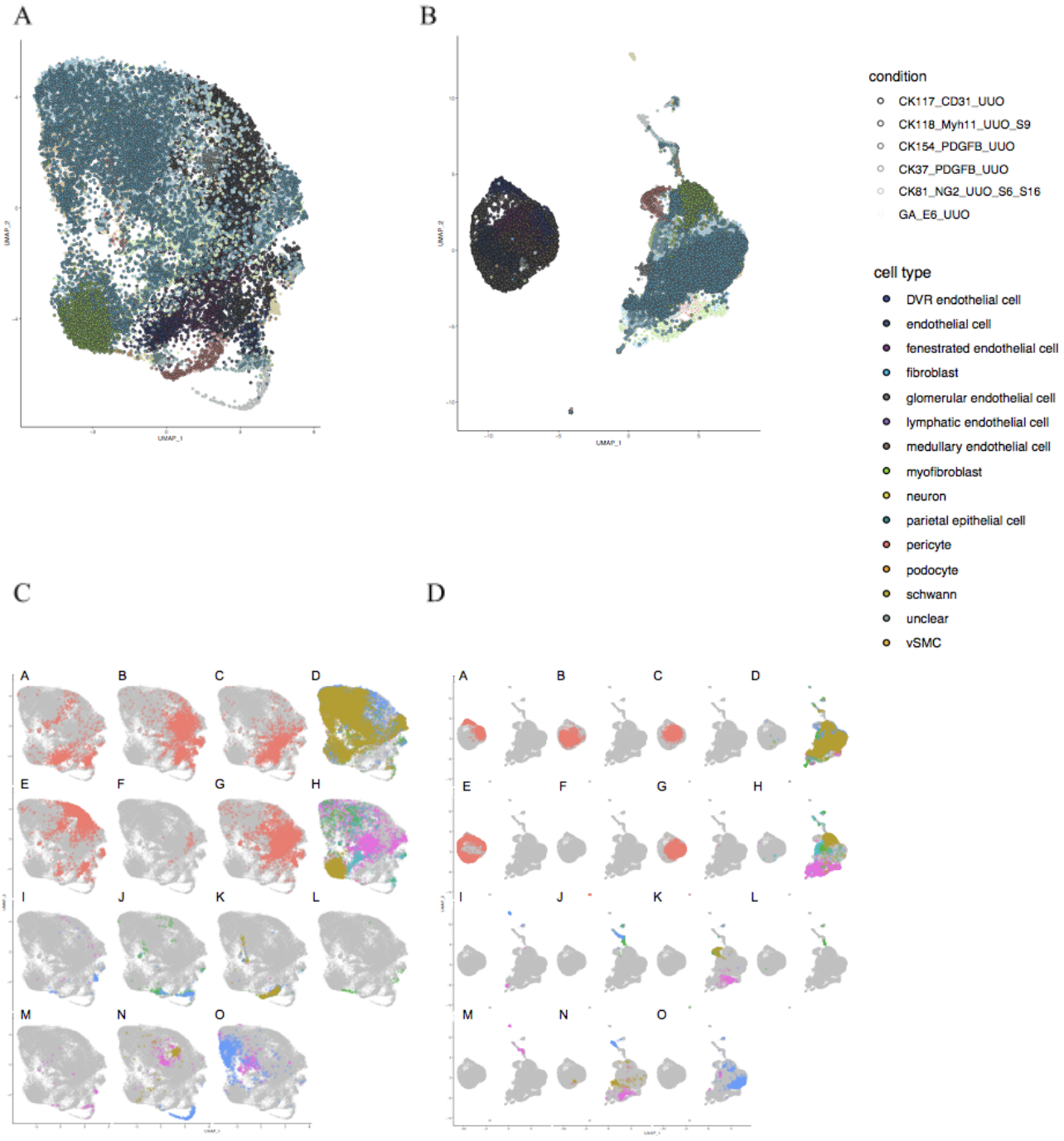


Figure 3.1.6. The UMAP and facets of all UUO mice data sets integrated by Seurat and Harmony. The left panels (Figure A and Figure C) are from the Seurat, and the right panels (Figure B and Figure D) are from Harmony.

The upper figures (Figure A and Figure B) are UMAP and lower ones (Figure C and Figure D), facets. In facets, there are 15 different cell types, (A) DVR endothelial cells, (B) endothelial cells, (C) fenestrated endothelial cells, (D) fibroblasts, (E) glomerular endothelial cells, (F) lymphatic endothelial cells, (G) medullary endothelial cells, (H) myofibroblasts, (I) neuron, (J) parietal epithelial cells, (K) pericytes, (L) podocytes, (M) schwann, (N) unclear and (O) vSMC.

In order to review if the Harmony integration has a huge artificial effect, cell type-specific correlation was measured on integrated data sets. Additionally, in order to compare the effect of Harmony, all sham and all UUO raw data sets were put together into one matrix without any integration methods. Before measuring the correlation, all data sets were normalized, scaled and then took positive values. This is because first, the scaled data set has zero as mean, so it doesn't fit measuring correlations in a cell-type specific way where all positively and negatively scaled values were averaged close to zero for the given cell type. Second, the batch-corrected normalized matrix has minus values as a minimum whereas the non-batch-corrected normalized matrix has zero value as a minimum. So, normalization matrices of these two approaches were not comparable. Third, positively scaled values indicate that the genes were upregulated over mean across all cells, in other words, which can represent the feature of the given cell types without any noise and lowly expressed values. Based on these three reasons, normalized and scaled values were not suitable for cell-type correlation measurement so I used only positively scaled values. However, even the positively scaled gene expression has a low scale between maximum and zero so instead of using the average in a given cell type for the correlation, the "CondGeneProb" of genesorteR was used. In other words, the correlation was measured in a binary way at cell-type level.

Figure 3.1.7 has 6 different cell type-specific correlation matrices measured by Pearson correlation on the "CondGeneProb" matrix computed by genesorteR [24]. The left panels (A and D) are from the Seurat, the middle (B and E) from the non-integration approach and the right (C and F) from Harmony. The upper panels (A,B and C) are from all sham mice and the lower ones (D,E and F) from all UUO mice.

In Figure 3.1.7.A (Seurat integration), only a few clusters (cluster "b","c") show strong correlations where cluster "b" and "c" include all endothelial cells from Cd31 sham. Others have weak correlation even within the same cell types in cluster "d" (18 fibroblasts from Gli1, Pdgfrb, Myh11, Ng2 and 1 neuron from Ng2). Figure 3.1.7.B (non-integration approach) have several clusters with strong correlations, in detail, cluster "a" corresponds to all endothelial cells from Cd31, cluster "b" & "c" involve fibroblasts across all data sets such as Gli1, Ng2, Myh11 and Pdgfrb and cluster "d" & "e" have neurons and pericytes from all data sets. These correlations were matched with cell-type wise biological knowledge more than the Figure 3.1.7.A. Figure

3.1.7.C (Harmony integration) has the same trend with Figure 3.1.7.B (non-integration approach), but Figure C (Harmony integration) shows a bit better cell-type correlation than Figure 3.1.7.B (non-integration approach). In Figure 3.1.7.C, cluster “b”, “c” and “d” have only fibroblasts and have more clear separation from the cluster “e” including percityes and neuron. However, in Figure B, the pericytes and neuron clusters (“d”,”e”) have over 0.5 similarity with fibroblasts clusters (“b”, “c”) compared to Figure 3.1.7.C (Harmony integration). Briefly, Harmony gathered the same (or similar) cell types more closely, and separated different cell types to each other. So, Harmony was better than Seurat, and it's a bit more preferable than a non-integration tool, which indicates that Harmony had less artificial effect compared to non-integration approach. Figure 3.1.7.D-F are the correlation matrices from all UUO mice data integrated by Seurat (D), non-integration tool (E) and Harmony (F). Like the correlation output of all sham mice data sets, those of all UUO mice data sets displayed a similar phenomenon with the correlation of all sham data integration where integration Harmony (F) is better than Seurat (D), and it’s slightly more superior than non-integration method (E).

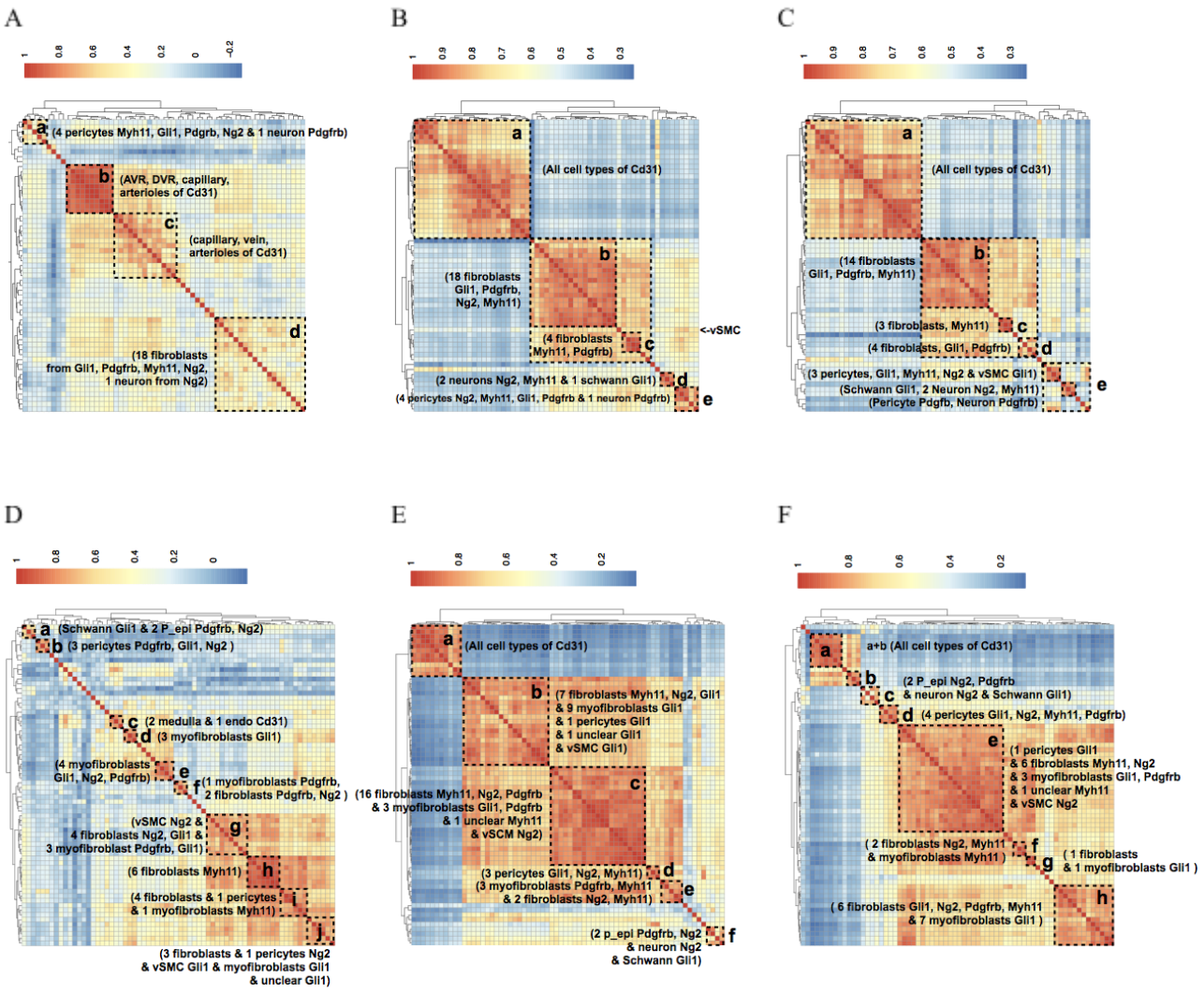


Figure 3.1.7. The cell type-specific correlation matrices on the integrated data sets. The left panels (A and D)

are from the Seurat, the middle (B and E) from the non-integration approach and the right (C and F) from Harmony. The upper panels (A,B and C) are from all sham mice and the lower ones (D,E and F) from all UUO mice. Regarding the non-integration method, the same preprocess was applied such as, normalizing, measuring highly variable genes and scaling after putting multiple data sets together into one matrix like what Seurat and Harmony did. Cell type-specific correlation was measured by Pearson correlation on "CondGeneProb" matrix computed by genesorteR.

For the smaller data set integration, Harmony was also used and tested. The graphical overview is described in Figure 3.1.13. In Figure 3.1.8-11, there are large and small UMAP figures for Gli1, Ng2, Myh11 and Pdgfrb, separately. Like when comparing the correlation matrix of integration on the all integrated data set, the same approaches were applied, but without Seurat. In Supplementary Figure 4.1.1, Figure A & C are the outputs by the non-integration tool and Figure.B & D by the Harmony. The two cell type correlation matrices look similar to each other, but Harmony made two pericytes from sham and UUO mice closer to each other than the non-integration tool in Supplementary Figure 4.1.1.B (Gli1) and D (Myh11) and 4.1.2.B (Ng2). However, Pdgfrb integration by Harmony has almost similar correlation matrices with the non-integration tool.

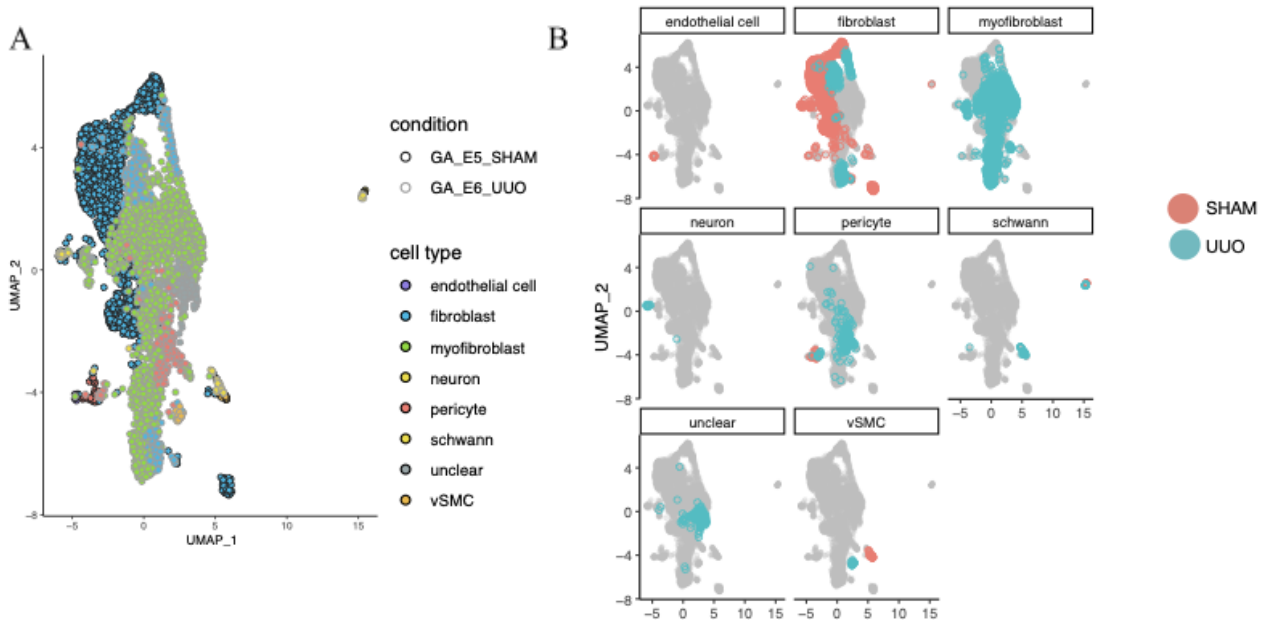


Figure 3.1.8. The UMAP and facets of Gli1 integration by Harmony. Figure A is the UMAP of Gli1 integrated by Harmony, respectively. Figure B has several UMAP figures which color only one cell type (colors are divided by the condition, UUO and SHAM).

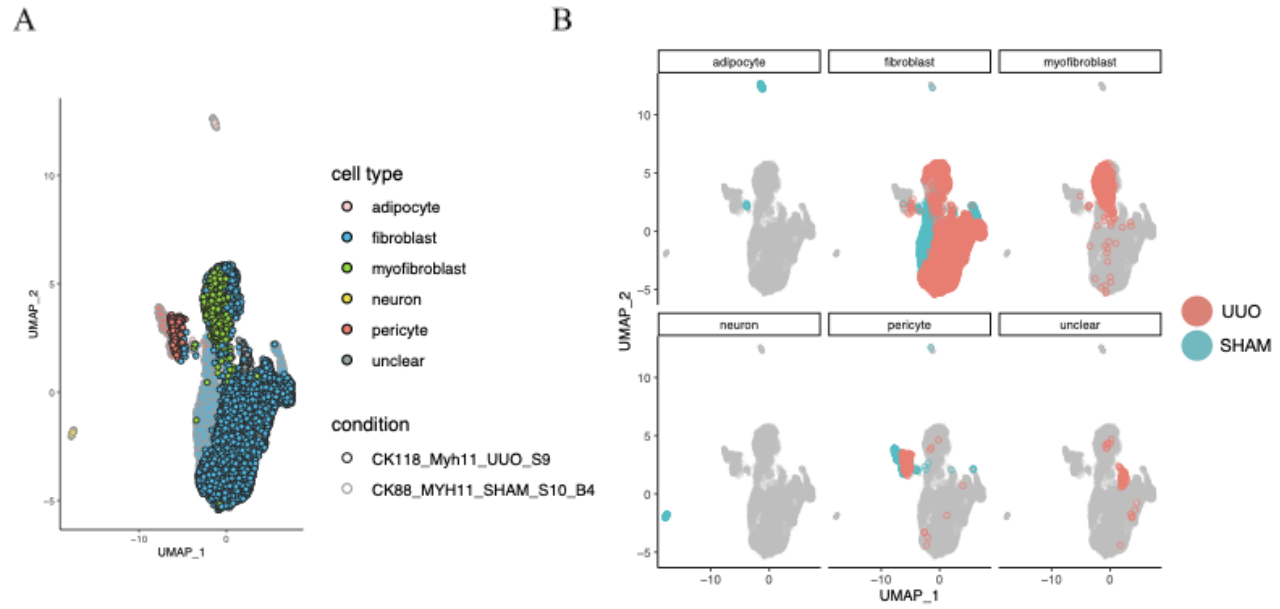


Figure 3.1.9. The UMAP and facets of Myh11 integration by Harmony. Figure A is the UMAP of Myh11 integrated by Harmony, respectively. Figure B has several UMAP figures which color only one cell type (colors are divided by the condition, UUO and SHAM).

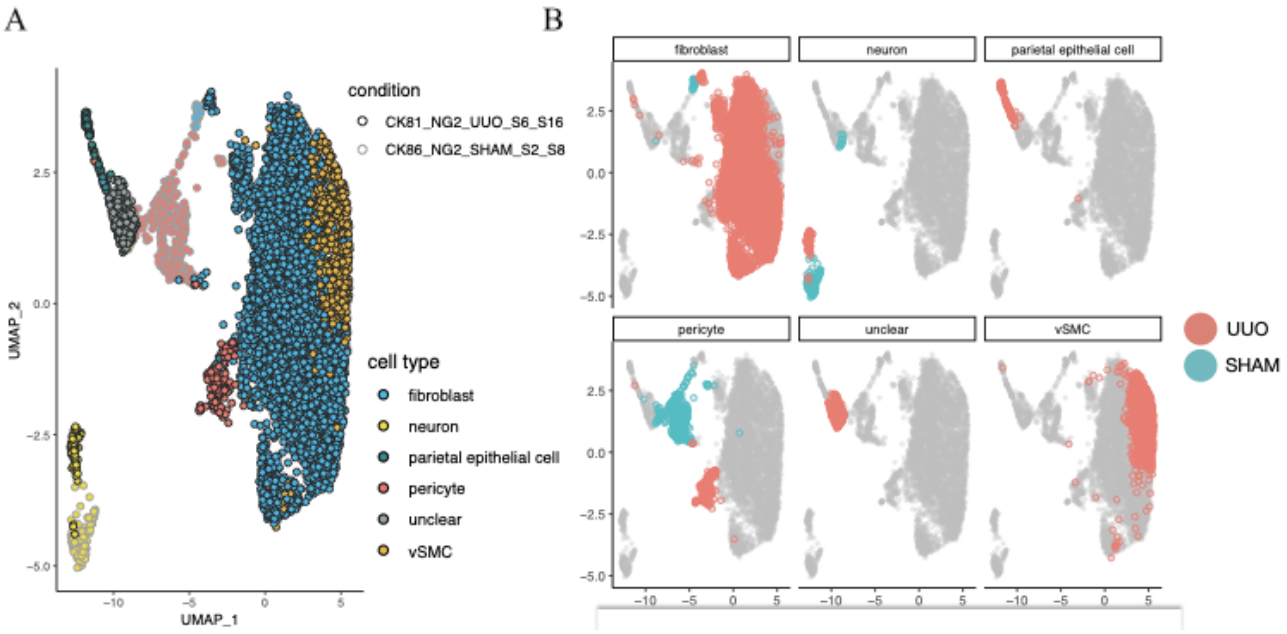


Figure 3.1.10. The UMAP and facets of Ng2 integration by Harmony. Figure A is the UMAP of Ng2 integrated by Harmony, respectively. Figure B has several UMAP figures which color only one cell type (colors are divided by the condition, UUO and SHAM).

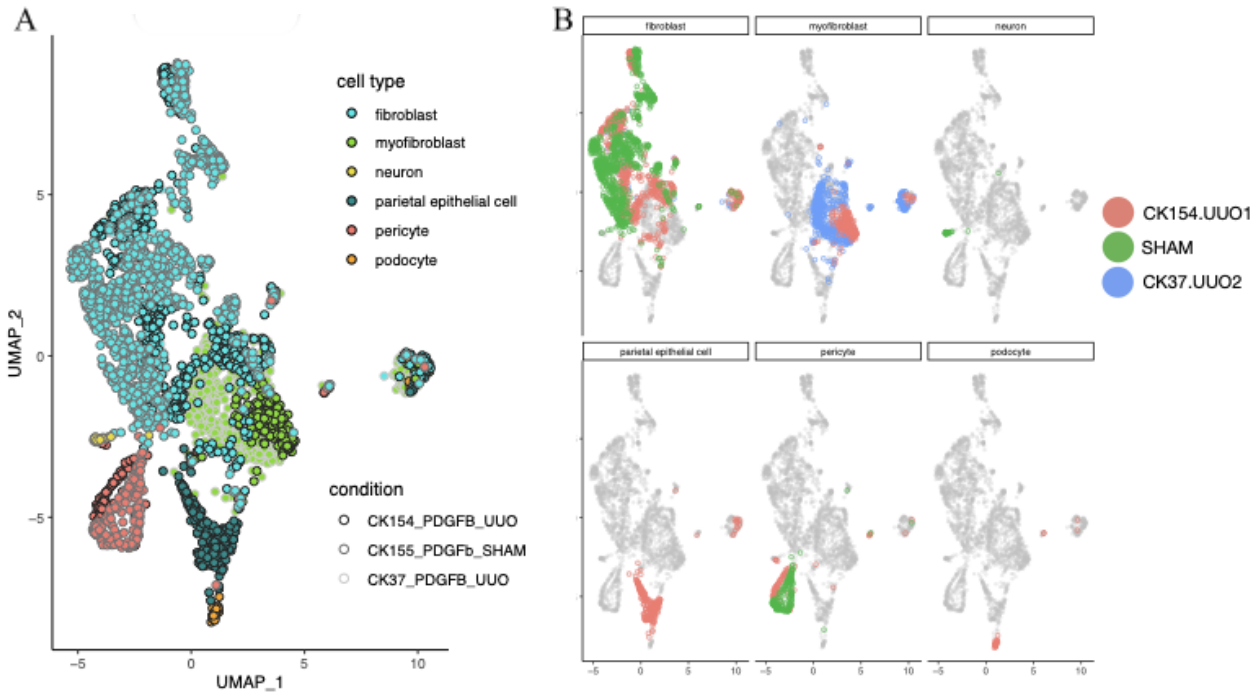


Figure 3.1.11. The UMAP and facets of Pdgfrb integration by Harmony. Figure A is the UMAP of Pdgfrb integrated by Harmony, respectively. Figure B has several UMAP figures which color only one cell type (colors are divided by the condition, UUO1, UUO2 and SHAM).

In the case of the Cd31 mice data set, Harmony separated Cd31 sham from Cd31 UUO in Figure 3.1.12 regardless of common cell types of sham and UUO mice. It represents high heterogeneity of renal endothelial cells [73]. When it comes to the integration approaches, the Harmony and non-integration approach generated similar cell-type correlation outputs to each other in Supplementary Figure 4.1.3 where the upper figure (A) was by non-integration tool and the below (B) by Harmony.

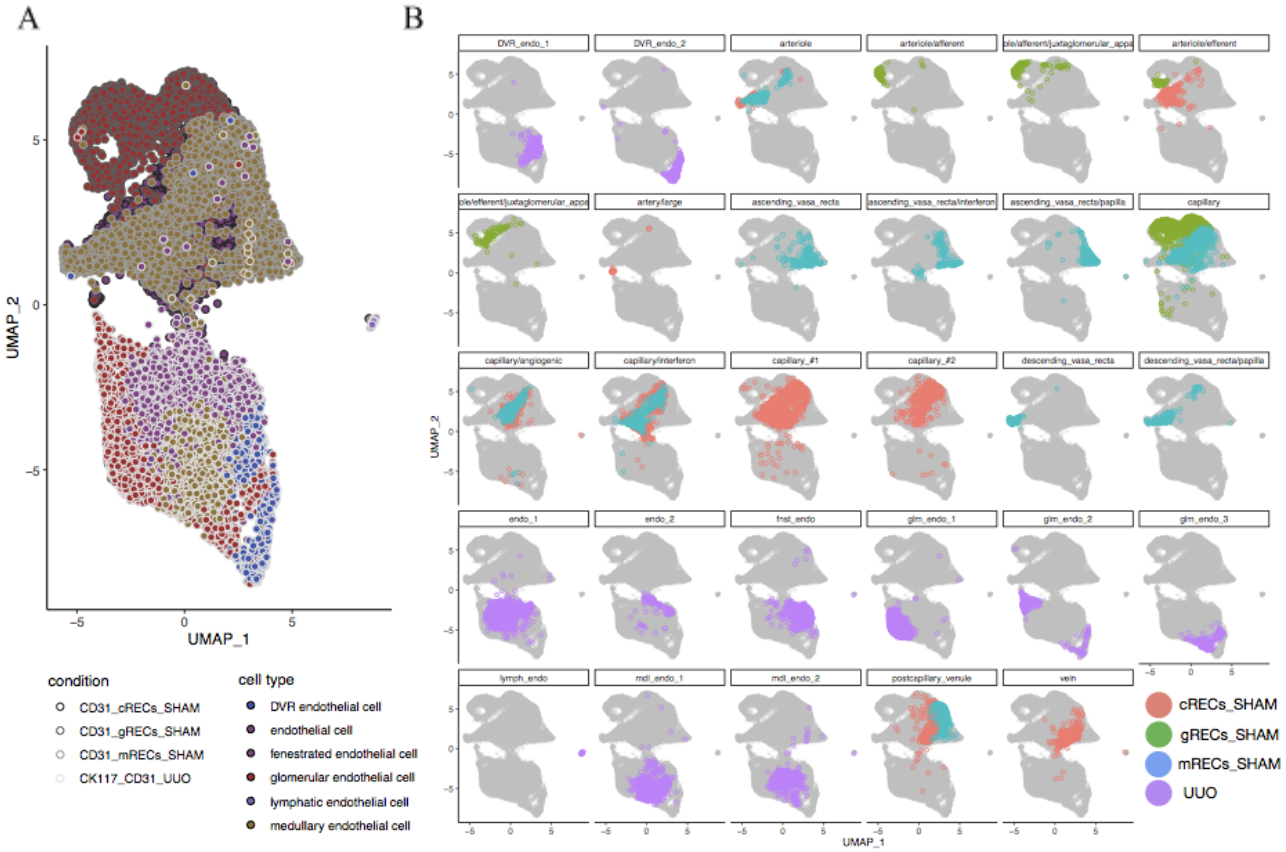


Figure 3.1.12. The UMAP and facets of Cd31 integration by Harmony. (A) The UMAP of Cd31 integrated by Harmony. (B) Several UMAP figures which color only one cell type, but color is divided by the samples, red (cortex sham), green (glomerulus sham), blue (medulla sham) and purple (UUO).

In summary, the integration study shows that first, Harmony fits the single cell data integration composed of similar cell types by gathering same (or similar) cell types closely and separating different cell types distantly. Second, it has less artificial effect when compared with the non-integration approach.

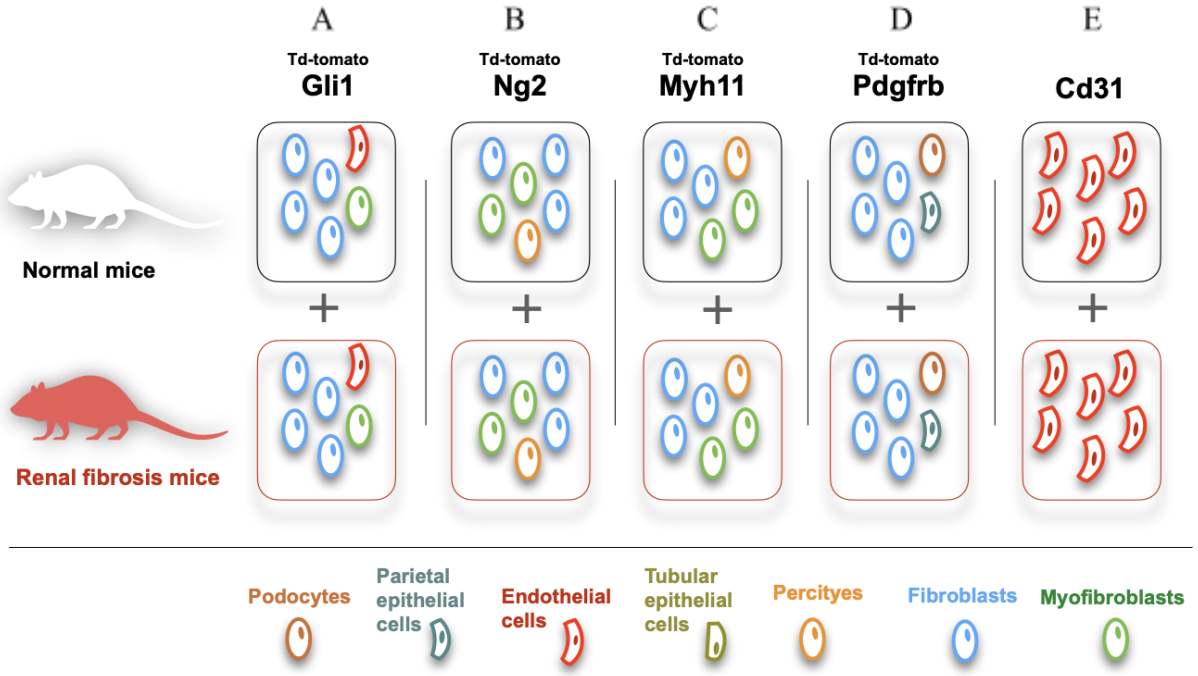


Figure 3.1.13. The graphical overview of data integration. This figure is the graphical overview on the data integration for Figure 3.1.8-12. "+" means integration. (A) Gli1 sham and UUO data integration, (B) Ng2 sham and UUO data integration, (C) Myh11 sham and UUO data integration, (D Pdgfrb sham and UUO data integration and (E) Cd31 sham and UUO data integration. Data integration was conducted by three approaches, Harmony, Seurat and non-integration approach. For UMAP, Harmony was used. For correlation matrix, Harmony, Seurat, non-integration approach were used.

Biological interpretation on single cell RNA-seq data

Functional-level studies

Pathway and transcription factor activity studies

All pathways of Progeny are linked to each other in cells, but for the efficient interpretation, these pathways could be collapsed into 3 categories, (1) c-Jun N-terminal kinase (JNK)-related (TNF-a, TGF-β, NF-κB, JAK-STAT, p53, Wnt), (2) MAPK-related (MAPK, PI3K, EGFR) and (3) endothelial cells-related (VEGF, Hypoxia). Other pathways were excluded for biological interpretation. Figure 3.1.14 displays the Progeny outputs of Gli1, Ng2 and Myh11 data sets. JNK-related pathways (TNF-a, TGF-β, NF-κB, JAK-STAT, p53, Wnt) were upregulated across cells in Gli1, Ng2, Myh11 UUO mice more than sham mice. However, PI3K (MAPK-related pathways) didn't have such a clear separation between UUO mice and sham mice data (one sample t-test, p-value=0.41, FDR=1). This implies that JNK-related pathways gave effect on the formation of renal fibrosis more than PI3K pathway.

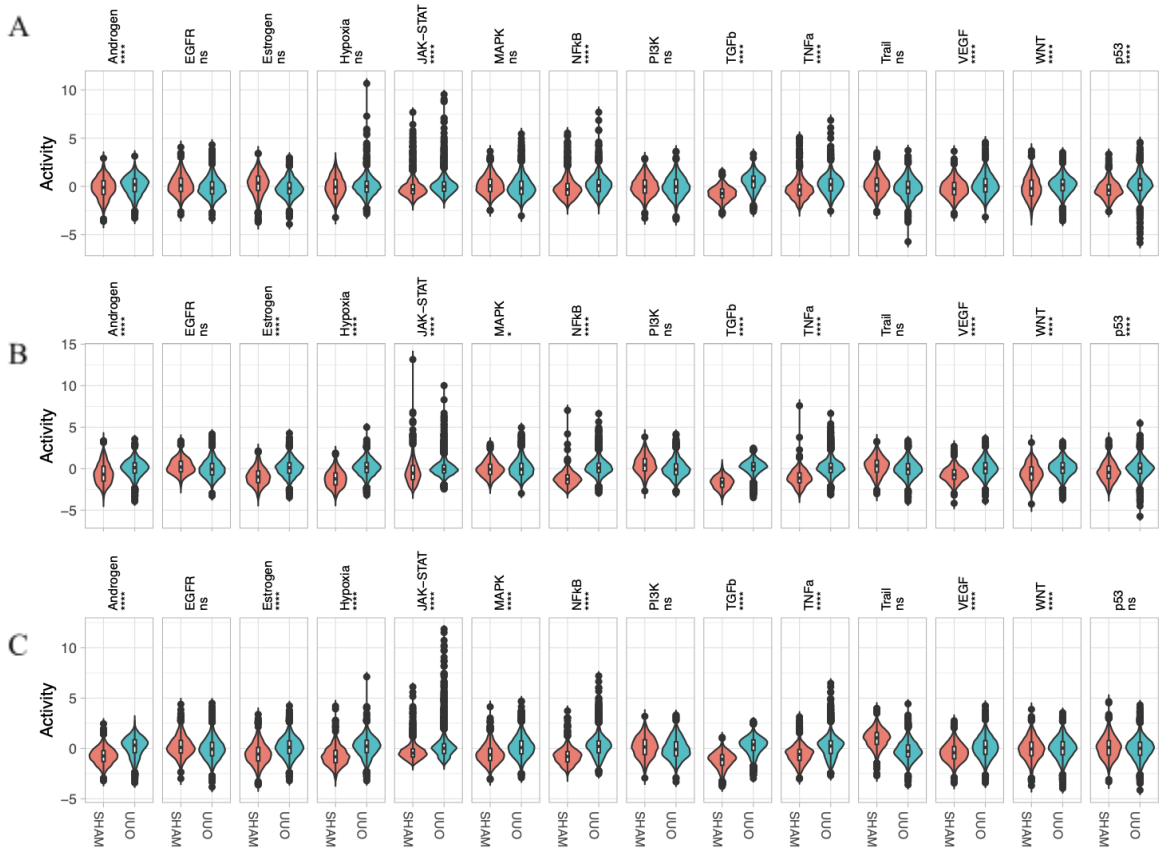


Figure 3.1.14. The PROGENy results from Gli1, Ng2 and Myh11 data sets. (A) The PROGENy output from Gli1 sham and UUO, (B) Ng2 sham and UUO and (C) Myh11 sham and UUO. The x-axis of all figures involves 14 different PROGENy pathways for each condition (SHAM and UUO), the y-axis corresponds to the PROGENy activity. The significance was measured by a one sample t-test in which the null hypothesis was that PROGENy scores in UUO were not greater than sham, and the p-value was corrected by bonferroni. ("****" if FDR < 1e-04,

"***" if 1e-04 < FDR < 0.001, "**" if 0.001< FDR < 0.01, "*" if 0.01< FDR < 0.05, "ns" if 0.05 < FDR < 1). The p-values and adjusted p-values are listed in Supplement table 4.1.2.

Figure 3.1.15 shows the highly variable transcription factors across cell types in Gli1, Ng2, Myh11, Pdgfrb data. *Atf2* (except for Myh11 UUO data), *Smad3* and *Myc* (except for Pdgfrb UUO data) were found as interesting and significant statistically (Supplements 4.1.3). These outputs are matched with the output of PROGENy, TGF-β (*Samd3, Atf2*) and MAPK (*Myc*). Those transcription factors are known as downstream-related genes of those pathways. Interestingly, structure specific recognition protein 1 (*Ssrp1*) was also found as significant in Gli1, Ng2, Myh11 UUO datasets but the literature evidence is insufficient to support the biological link between this gene and renal fibrosis.

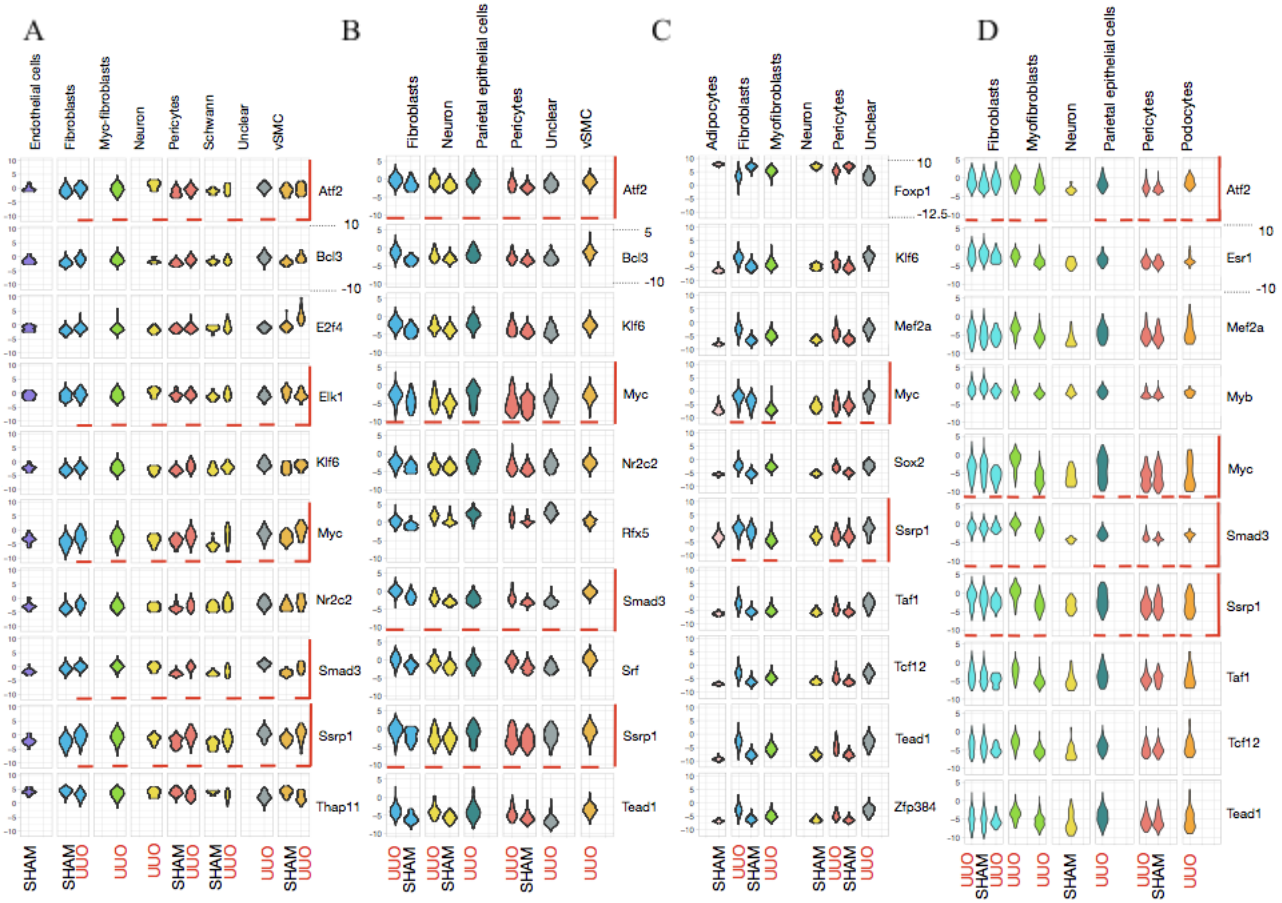


Figure 3.1.15. The DoRothEA violin plots from Gli1, Ng2, Myh11 and Pdgfrb data sets. The DoRothEA output from (A) Gli1 sham and UUO, (B) Ng2 sham and UUO, (C) Myh11 sham and UUO and (D) Pdgfrb sham and UUO. The x-axis of 4 figures involve diverse cell types divided by condition, sham and UUO. The y-axis consists of 10 highly variable transcription factors. Each color corresponds to cell types, fibroblasts (blue), myofibroblasts (green), neuron & schwann (yellow), parietal epithelial cells (dark olive), pericytes (red), podocytes (orange), vascular smooth muscle cells (dark orange), and unclear(grey). The transcription factor activity range from 10 to -10 in Gli1 sham and UUO, 5 to -10 in Ng2 sham and UUO, 10 to -12.5 in Myh11 sham and UUO, 10 to -10 in Pdgfrb sham and UUO mice data set. Biologically important transcription factors were marked by a red line. The p-values and adjusted p-values are listed in Supplements 4.1.3.

Interestingly, the Progeny and DoRothEA outputs of Pdgfrb data exhibit different patterns from Gli1, Ng2 and Myh11. Figure 3.1.16.A shows that MAPK-related pathways including MAPK and PI3K separated Pdgfrb UUO mice from sham mice rather than JNK-related pathways and EGFR. It infers that MAPK was the key factor causing renal fibrosis more than JNK-related pathways in the Pdgfrb UUO mice data set. In Figure 3.1.16.B, Cd31 data shows the VEGF and Hypoxia pathways were over-expressed and more significant in UUO mice than sham mice. NF-κB, TGF-β and TNF-a also were also highly scored pathways in Cd31 UUO mice than sham mice. It indicates JNK-related pathways with VEGF and Hypoxia, were associated with the pathogenesis of renal fibrosis in Cd31 UUO mice data, in other words, endothelial cells. Interestingly, there was a report to support this output. *In vitro* 3D collagen gel culture assays, IL-6 and TNF-a promotes inflammatory-endothelial to mesenchymal transformation (EndMT) through an Akt/NFκB-dependent pathway in both adult valve endothelium [96]. Here, I inferred that upregulated JNK-related pathways would induce renal fibrosis in a situation of up-regulated hypoxia and angiogenesis (upregulated VEGF) in endothelial cells of mice renal fibrosis.

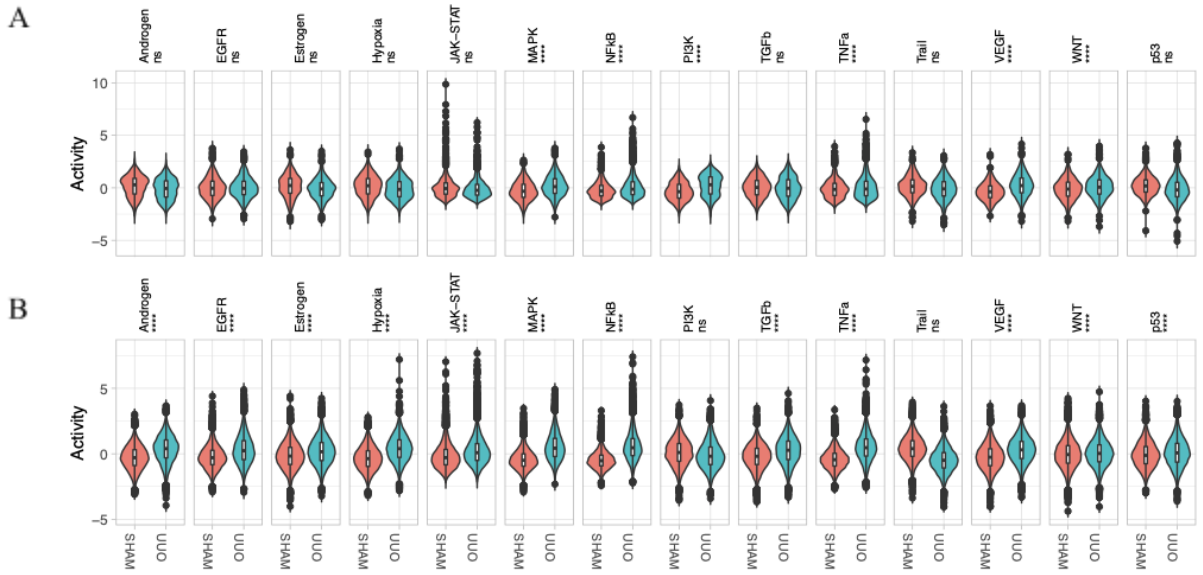


Figure 3.1.16. The PROGENy results from Pdgfrb and Cd31 data sets. There are two PROGENy outputs, (A) Pdgfrb sham and UUO, (B) Cd31 sham and UUO. The x-axis of 3 figures involves 14 different PROGENy pathways, the y-axis corresponds to the PROGENy activity. The significance was measured by a one sample t-test in which the null hypothesis was that PROGENy scores in UUO were not greater than sham, and the p-value was corrected by bonferroni. ("****" if FDR < 1e-04, "***" if 1e-04 < FDR < 0.001, "**" if 0.001< FDR < 0.01, "*" if 0.01< FDR < 0.05, "ns" if 0.05 < FDR < 1). The p-values and adjusted p-values are listed in Supplements 4.1.2.

Figure 3.1.17 has the outputs of DoRothEA of Cd31 UUO mice data set. Several up-regulated transcription factors were found in Cd31 UUO mice higher than sham mice, *Klf6, Kmt2a, Mef2a, Myc, Stat2, Taf1, Tead1* and *Tead4*. Biologically, *Klf6* has been known to upregulate the

expression of *TGF-β* and induce the process of epithelial-mesenchymal transition (EMT), one of the causes leading to renal fibrosis [97]. As well, *Klf6* enhances tubular epithelial injury by using glomerular endothelial cells [97]. Interestingly, it was matched with the DoRothEA output of Cd31 UUO mice in that *Klf6* was highly expressed across glomerular endothelial cells than other cells in Cd31 sham mice. The pathway-wise output (PROGENy in Figure 3.1.16) which shows that TGF-β signaling was upregulated across all glomerular endothelial cells in both sham and UUO mice. Here, it infers that *Klf6* transcription factors induced the activated TGF-β signaling to promote renal fibrosis in the glomerular endothelial cells. *Kmt2a* has insufficient literature studies which show the pathological relation with renal fibrosis. In the case of the *Mef2a*, this gene is one of transcription factors to activate angiogenesis directly downstream from VEGFA [98]. This prior study is matched with the PROGENy and DoRothEA output of Cd31 UUO mice data sets in that VEGF pathway activity and *Mef2a* transcription factor activity were enhanced across cells in Cd31 UUO mice, especially in glomerulus and DVR in medulla. Regarding *Myc*, this gene has been known to be increased in tubular epithelial cells when the endothelial to mesenchymal transition (EndMT) activates hypoxia. Pharmacologically, suppression of *Myc* in tubular epithelial cells alleviated fibrosis [96]. Interestingly, the hypoxia pathway had higher activity across cells in the Cd31 UUO mice data set than the Cd31 sham mice data in PROGENy output, and *Myc* transcription factor activity was upregulated in glomerulus and DVR. *Stat2, Taf1*, *Tead1* and *Tead4* have a lack of prior studies related to renal fibrosis in endothelial cells.

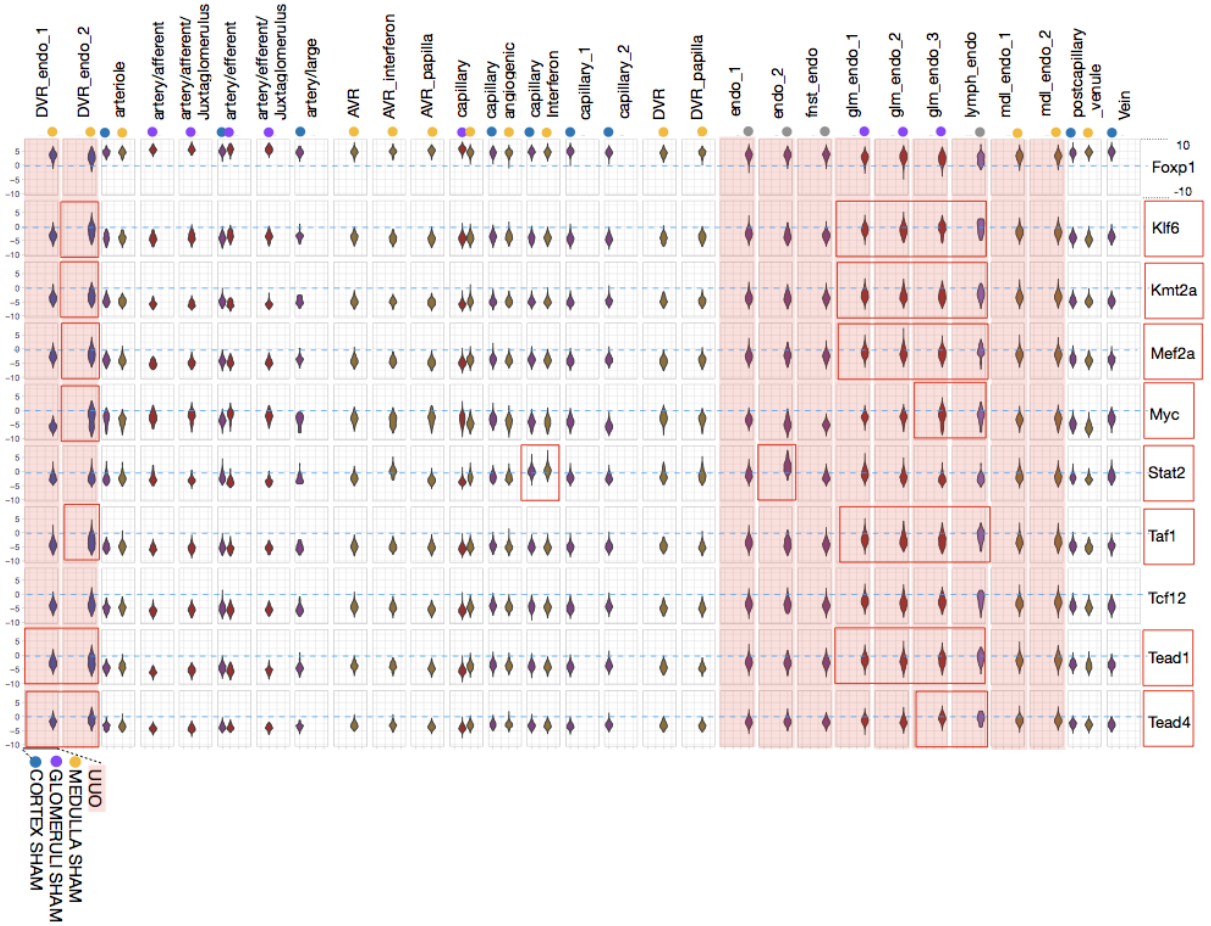


Figure 3.1.17. The DoRothEA results from Cd31 sham and UUO mice data sets. (A) The x-axis of DoRothEA outputs involve diverse cell types divided by condition, sham (Medulla sham (yellow), Glomerulus sham (purple) and Cortex sham(blue)) and UUO (red). The y-axis consists of 10 highly variable transcription factors. Red box is renal fibrosis mice (UUO). The transcription factor activity ranges from 10 to -10. All sub cell type numbers ("_1", "_2"…) are independent per each cell type and each condition. For example, sub-celltype 2 of fibroblasts ("fb_2") of SHAM is not related to the sub-cell type 2 of fibroblasts ("fb_2") of UUO.

Pathway activity studies at a single cell level and a cell type level

As a next step, the extracellular matrix (ECM) scores were computed with the function, "AddModuleScore" of Seurat across 4 different data sets, Gli1, Myh11, Ng2, Pdgfrb. Figure 3.1.18 shows the module scores of 10 different gene sets related to ECM in a given cell type. The maximum score is close to 2, and the minimum is close to zero or a negative value due to the subtraction approach in the "AddModuleScore" function. Each color corresponds to each different cell type, fibroblasts (red), myofibroblasts (green), pericytes (blue) and vascular smooth muscle cells (vSMC, purple). In Figure 3.1.18, fibroblasts or myofibroblasts had higher module scores than pericytes and vascular smooth muscle cells (VSMC) in Gli1, Ng2, Myh11 and Pdgfrb. Regarding the gene set from "NABA BASEMENT MEMBRANES", UUO mice data of Gli1, Myh11 and Ng2 UUO mice data had a higher module score than sham. In the Pdgfrb data set, myofibroblasts in UUO mice didn't have higher module scores than fibroblasts across conditions.

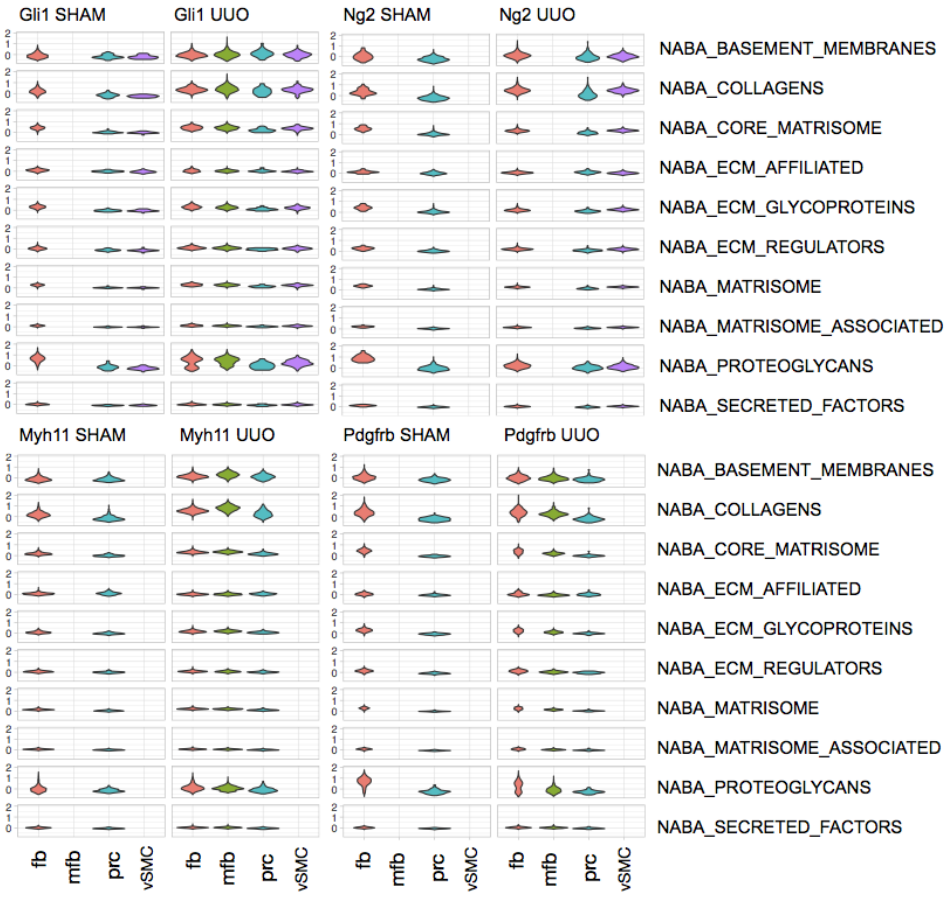


Figure 3.1.18. The ECM score of the Gli1 sham and UUO data set for the 10 different ECM-related different gene sets. These violin plots show the modularity score across all data sets for 10 different ECM-related gene sets per each cell type. Fibroblasts (red color, "fb") and myofibroblasts (green color, "mfb") had a tendency of having higher modularity than other cell types, pericytes (blue color, "prc") and vascular smooth muscle cells (purple color, "vSMC") across all data sets.

To identify significant biological pathways for each cell type, a gene set enrichment test with EnrichR was conducted. The key of this study is about how many genes are used because a few genes (< 20) can make a significant signal even if only 1 or 2 genes are shared with a gene set of known pathways, but it was the limitation of this project. This is because data sets in this project were composed of similar cell types so that it was difficult to get the enough number of differentially expressed genes (DEG). In order to find alternative genes to DEGs, I utilized the batch-corrected & positively-scaled matrix. For the datasets, I ranked the genes based on the score of the "CondGeneProb" (the percentage of cells expressing a gene per each cell type) in a given cell type and then, took top 100 genes for each cell type, in other words, it is the top 100 genes widely expressed in a given cell type. Additionally, I took only positively scaled gene expression values by collapsing all negative values as zero, so highly expressed over the average across all data sets were used too.

Figure 3.1.19 shows the enrichment results from Enrichr in Gli1, Myh11 and Ng2 data. The value (color) is the negatively log 10 scaled adjusted p.value, and the dot size corresponds to the shared number of genes with the public database "GO Biological Process 2018". Only significant pathways (adjusted.p.value < 0.05) were marked as dots. The figure 3.1.19.A shows that the positive regulation of fibroblast proliferation, positive regulation of stress fiber assembly, positive regulation of focal adhesion assembly, positive regulation of cell−matrix adhesion, neutrophil activation involved in immune response were significant across many myofibroblasts, especially, type 2, 4, 5, and these pathways were rarely detected in sham mice as significant pathways. It was interesting that these pathways have been known as the myofibroblasts' typical characteristic. As well, the negative regulation of programmed cell death was only detected in many myofibroblasts. It indicates that myofibroblasts kept remaining and then induced immune responses because the B cell differentiation and the positive regulation of lymphocyte differentiation were also identified as significant pathways in UUO (not shown in figures). In the output of integrated Myh11 data (Figure 3.1.19.B), similar pathways like Gli1 were detected as significant. Surprisingly, the negative regulation of cell death was also found in only myofibroblasts. However, regarding the positive regulation of fibroblast proliferation, it was detected as significant across many cell types in both Myh11 UUO and sham. The Ng2 integrated data set also shows a similar output like Myh11 in Figure 3.1.19.C.

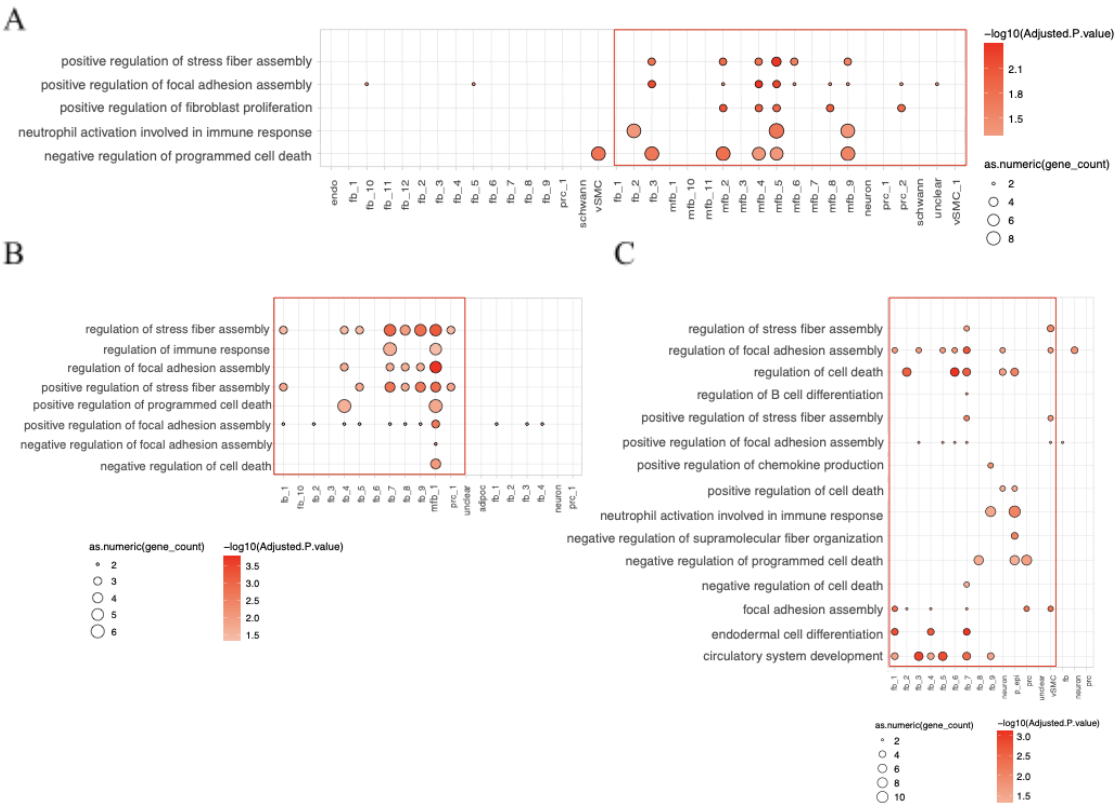


Figure 3.1.19. The GO-term enrichment score of the integrated Gli1, integrated Ng2 and integrated Myh11 data sets. The upper figure (A) is the enrichR output inferred from Gli1 data sets integrated by Harmony and the lower ones (B) is from Myh11 data sets, (C) is from Ng2 integrated by Harmony. Top100 highly and widely expressed genes in a given cell type were used.

Regarding the integrated Pdgfrb data set in Figure 3.1.20.A, the positive regulation of fibroblasts proliferation is shown as significant in only 2 cell types, fibroblasts type 1 and myofibroblasts type 2 in Pdgfrb UUO mice. The cellular response to cytokine stimulus, the endodermal cell differentiation, regulation of extrinsic apoptotic signaling pathway via death domain receptors, positive regulation of focal adhesion assembly, positive regulation of extrinsic apoptotic signaling pathway via death domain receptors, positive regulation of chemokine production and the positive regulation of cell−substrate adhesion were also identified as significant in Pdgfrb UUO mice data like other data sets. However, the negative regulation of cell death was not identified as significant pathways in both fibroblasts and myofibroblasts in Pdgfrb UUO. For the integrated Cd31 data set in Figure 3.1.20.B, several significant pathways were found, such as the positive regulation of fibroblast proliferation and cellular response to cytokine stimulus. Interestingly, the regulation of fibroblast proliferation was detected as significant across some cell types in both Cd31 sham and UUO. However, the positive regulation of fibroblast proliferation was only found in Cd31 UUO. In the case of the cytokine-related signaling, it was significant across all Cd31 UUO and sham (not shown in the figure). This is because the endothelial cells directly contact immune cells in capillaries or arteries.

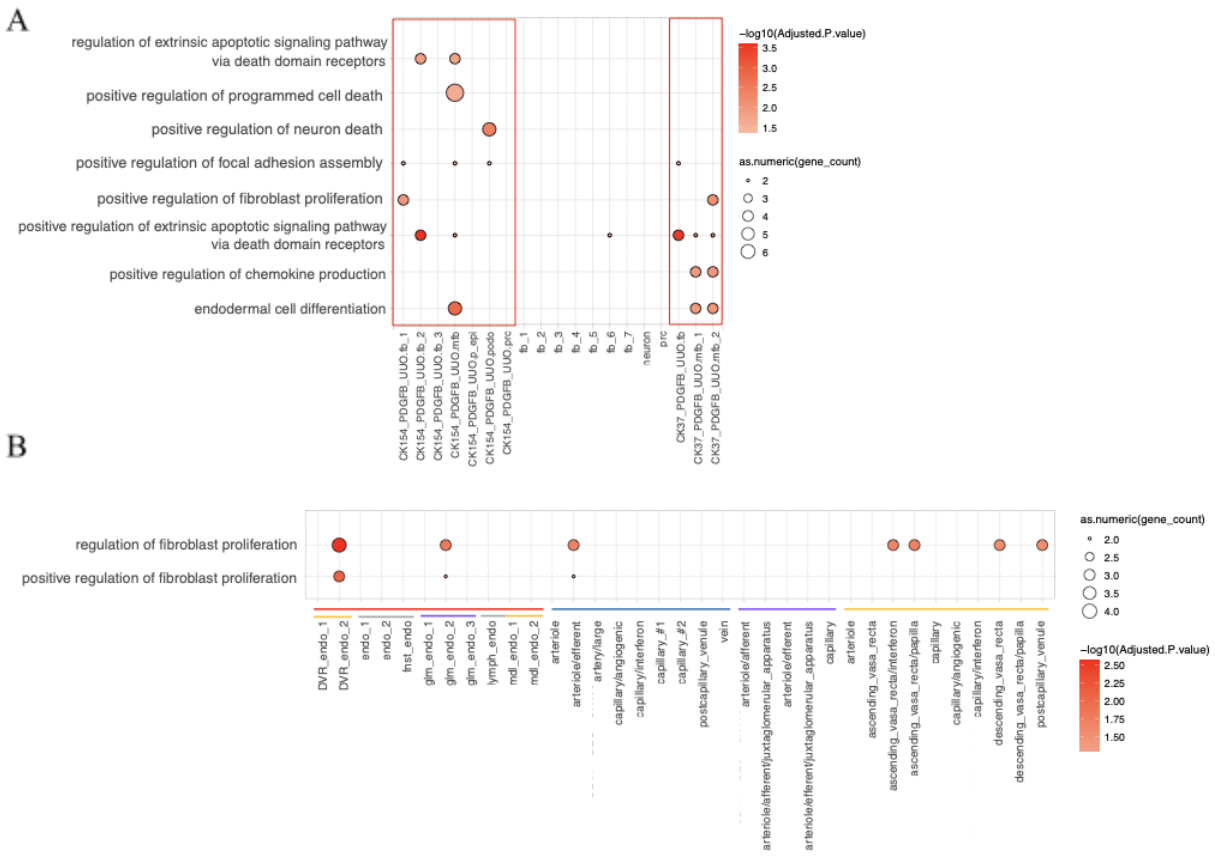


Figure 3.1.20. The GO-term enrichment score of the integrated Pdgfrb and Cd31 data sets separately. (A) The Enrichr outputs from Pdgfrb data sets integrated by Harmony. (B) EnrichR output from Cd31 data sets integrated by Harmony. Top100 highly and widely expressed genes in a given cell type were used. Red boxes (A) or lines (B) correspond to UUO mice where (below the red line) yellow means medullary endothelial cells from UUO, purple is glomerular endothelial cells from UUO and grey is global endothelial cells from UUO mice. Out of the red line, blue corresponds to cortex endothelial cells from sham, purple to glomerular endothelial cells from sham, yellow to medullary endothelial cells from sham.

Intercellular communication

In order to understand intercellular signaling across cell types, two different approaches were used, first was to infer intercellular talks based on the percentage of cells expressing the ligands and receptors (> 10% in a given two cell types) with ICELLNET database [40]. Second was to use a more delicate way to apply the hill function with differentially expressed genes in given two cell types by CellChat [43]. The outputs of the first approach are Figure 3.1.21-22 and 24. The second approach outputs are displayed in Figure 3.1.23, 25 and Supplement Figure 4.1.4-5 and 4.1.7.

In Figure 3.1.21, there are 4 outputs which show cell to cell communication in (A) Gli1, (B) Myh11, (C) Ng2, and (D) Pdgfrb sham mice data sets. Each color corresponds to biological pathways like Notch signaling (blue), growth factor (light blue), cytokine (olive), chemokine (green), checkpoint (red), hormone (pink) and non-specified signaling (grey).The x-axis consists of family, LR (ligands & receptors), sender (cell types) and receiver (cell types). The y-axis is the accumulated "CondGeneProb" by the summation of the minimum of the two "CondGeneProb" values from the given ligand and receptor.

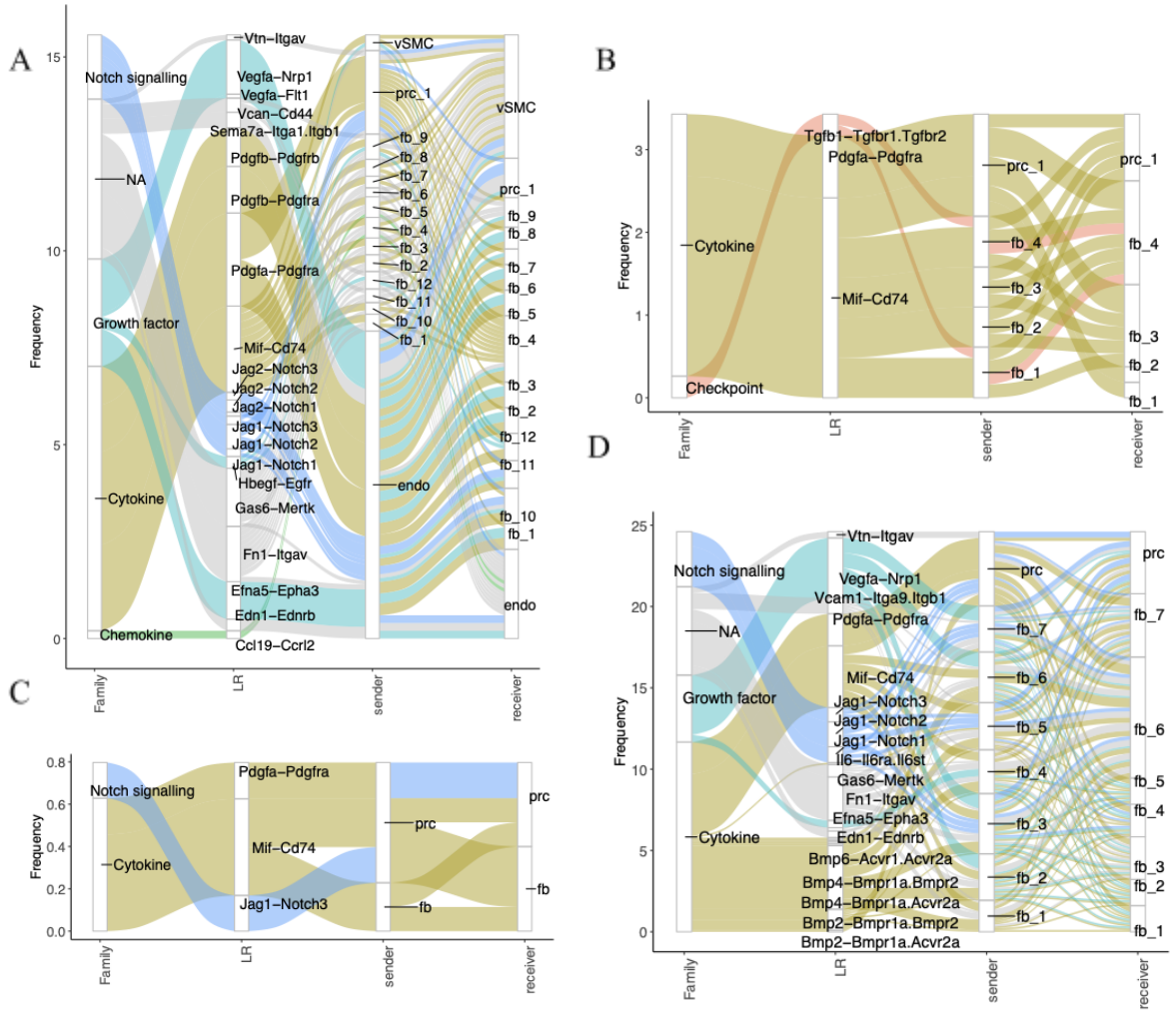


Figure 3.1.21. The intercellular communication of the Gli1, Myh11, Ng2 and Pdgfrb sham data set. (A) Gli1, (B) Myh11, (C) Ng2 and (D) Pdgfrb. Each color corresponds to biological pathways like Notch signaling (blue), growth factor (light blue), cytokine (olive), chemokine (green), checkpoint (red), hormone (pink) and non-specified signaling (grey). The x-axis consists of family, LR (ligands & receptors), sender (cell types) and receiver (cell types). The y-axis is the accumulated "CondGeneProb" of detected ligands and receptors. The thickness of the curve is the accumulated minimum percentage of cells expressing the ligand or the receptor of the interaction in a given family.

Figure 3.1.21 includes the outputs from sham mice data sets, (A) Gli1, (B) Myh11, (C) Ng2 and (D) Pdgfrb. The maximum accumulated values (Y-axis) across data sets ranged from 0.8 to 24.6, and mainly cytokine responses occupied lots of all the detected interactions of ligands and receptors. *Pdgfa-Pdgfra, Mif-Cd74* were found across all sham data sets. Datasets-wise, in Figure 3.1.21.D (Pdgfrb), all cell types had cytokine-associated immune response and the accumulated values is almost 25, the top value among all data sets.

Figure 3.1.22 shows the outputs from all UUO mice data sets, (A) Gli1, (B) Ng2, (C) Myh11 and (D) Pdgfrb. The accumulated "CondGeneProb" values had 19.6 as minimum and 66 as maximum across data sets. Chemokine-related immune responses were detected more often in UUO mice than sham mice data across all data including Gli1, Ng2, Myh11 and Pdgfrb. *Tgfb1-Tgfbr2 & Tgfbr1, Fn1-Itgbv, Mif-Cd74* and *Vegfa-Nrp1* were detected commonly across all UUO mice data sets. Regarding *Vcam1-Itgb1 & Itga9,* two receptors, *Itga9* and *Itgb1,* were identified in the Ng2, Myh11 and Pdgfrb UUO mice data set compared to sham mice data where only *Itgb1* was expressed. In addition, *Tgfb1, Tgfb2* and *Tgfb3* were up-regulated across all UUO mice than sham. In the case of the *Mif-Cd74*, macrophage migration inhibitory factor (*Mif*) is an important pathogenic factor of renal inflammation leading to the Acute Kidney Injury (AKI) via a CD74-Mif-NF-κB-dependent mechanism [100]. Interestingly, the *Mif-Cd74* interactions were also identified in all sham and UUO mice data in Figure 3.1.21 and 22, but NF-κB pathways were upregulated in the Progeny output of UUO mice than sham across all data sets in Figure 3.1.14 and 16. Here, I inferred that Mif-CD74-NF-κB-dependent mechanism contributed to leading renal fibrosis on Gli1, Ng2, Myh11, Pdgfrb UUO data. Pdgfrb UUO mice had different outputs from other Gli1, Ng2 and Myh11 UUO mice data sets. In Figure 3.1.22.D, Pdgfrb UUO mice data sets had less accumulated "CondGeneProb" than other UUO mice data. This is because the Pdgfrb gene in the td-tomato Pdgfrb^{+} UUO mice targeted different types of cells from Gli1, Ng2 and Myh11 UUO mice.

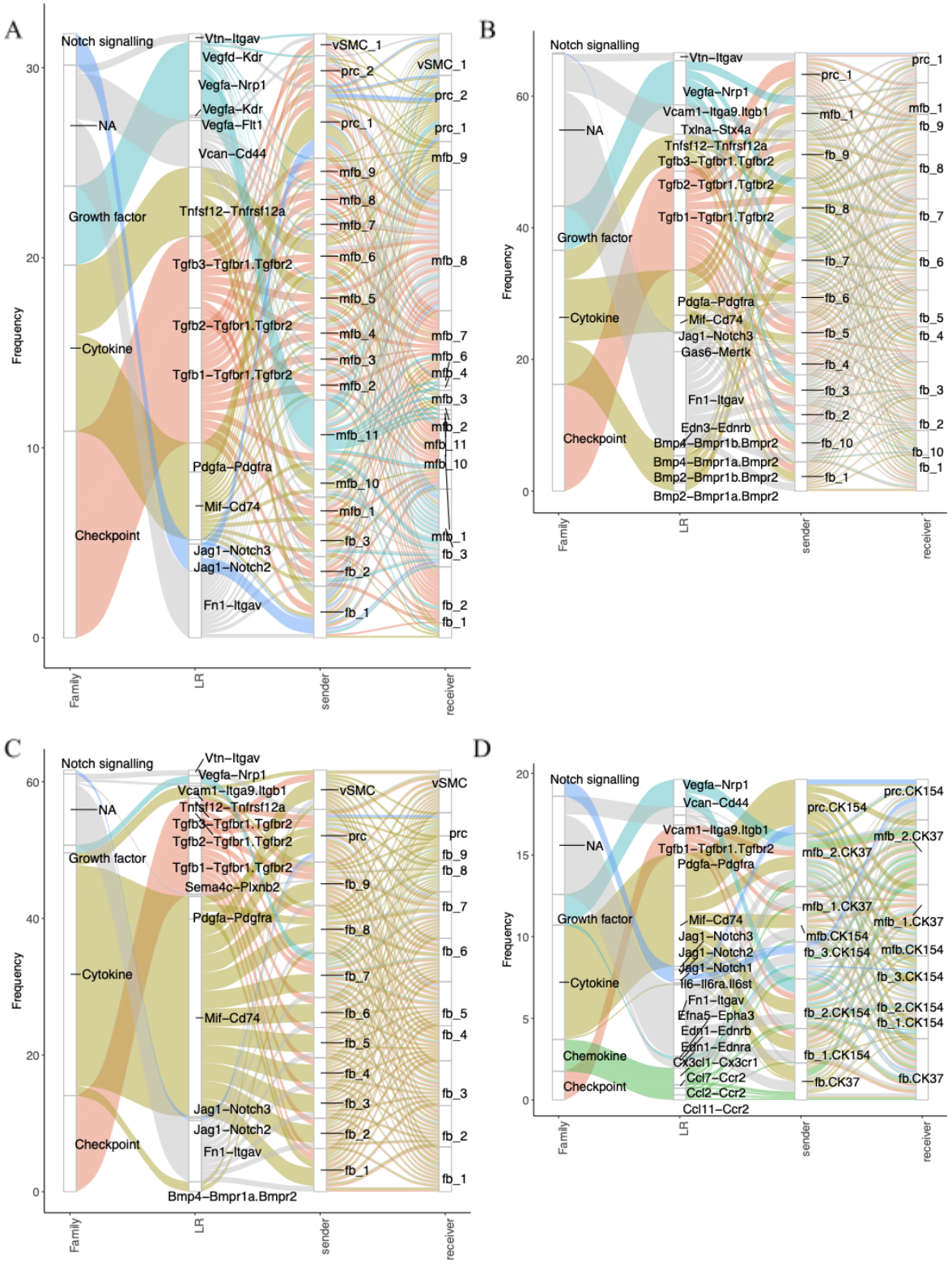


Figure 3.1.22. The intercellular communication of the Gli1, Myh11, Ng2 and Pdgfrb UUO data set. (A) Gli1, (B) Myh11, (C) Ng2 and (D) Pdgfrb. Each color corresponds to biological pathways like Notch signaling (blue), growth factor (light blue), cytokine (olive), chemokine (green), checkpoint (red), hormone (pink) and non-specified signaling (grey). The x-axis consists of family, LR (ligands & receptors), sender (cell types) and receiver (cell types). The y-axis is the accumulated "CondGeneProb" of detected ligands and receptors. The thickness of the curve is the accumulated minimum percentage of cells expressing the ligand or the receptor of the interaction in a

given family.

The number of cells in Gli1, Ng2, Myh11, Pdgfrb were diverse to each other. I checked if the number of cells affected the number of ligand-receptor interactions, in which the number of communications was divided by the number of cells. The ratio was 0.08 (144 interactions / 1645 cells) in Gli1 sham, 0.11 (297 interactions /2682 cells) in Gli1 UUO, 0.01 (6 interactions /549 cells) in Ng2 sham, 0.08 (420 interactions /5072 cells) in Ng2 UUO, 0.01 (21 interactions /1716 cells) in Myh11 sham, 0.08 (638 interactions /7355 cells) in Myh11 UUO, 0.17 (204 interactions /1166 cells) in Pdgfrb sham, 0.10 (175 interactions /1689 cells) in Pdgfrb UUO mice data set. Pdgfrb sham data has a higher ratio even though it has a lower number of cells than Pdgfrb UUO. Gli sham has the similar number of cells of Ng2 sham, but the ratio is totally different. It means that there was less effect from the cell sizes on the number of ligand-receptor interactions.

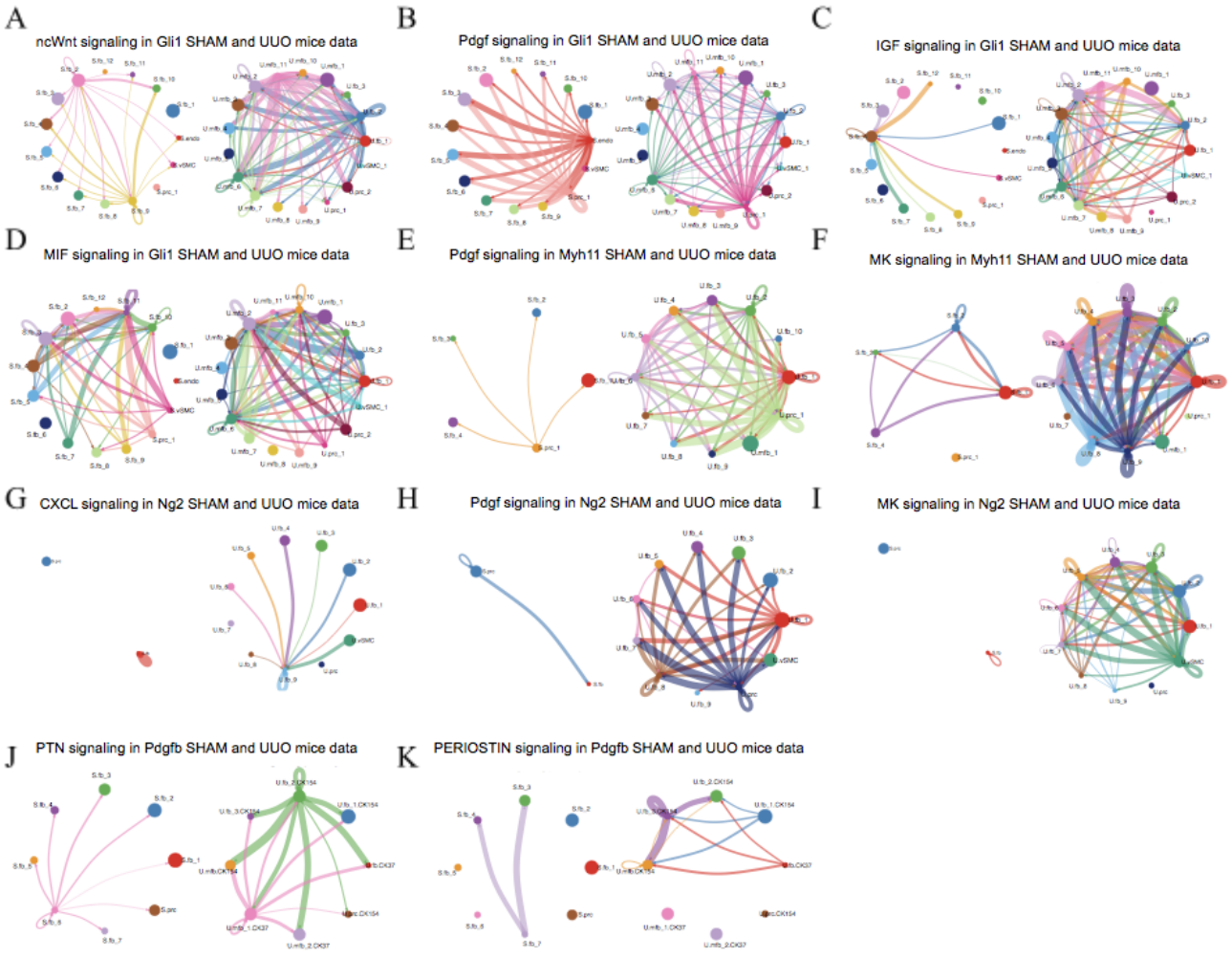


Figure 3.1.23. The CellChat intercellular communication of the Gli1, Myh11, Ng2 and Pdgfrb sham and UUO mice data sets. There are 8 signalings including non-canonical WNT (A), Pdgf (B,E,H), IGF (C), MIF (D), MK (I), CXCL (G), PTN (J) and PERIOSTIN (K) signalings. All circle plots consist of cell type as nodes with the interactions (as edges). Each edge represents the total interaction strength (or weight) between any two cell groups, which means thicker edges indicate a stronger signal. Weight refers to communication probabilities measured by the

law of mass action with the amounts of the modeled ligand-receptor. In a given two circle plot, two plots have the same maximum values of edges so two plots can be comparable to each other.

As a next step, CellChat was used [43]. CellChat database and ICELLNET databases share almost 50 percent of ligands and receptors, but CellChat infers the ligand-receptor interactions in a more delicate way while categorizing the pairs with information from the KEGG Pathway database [40]. In Figure 3.1.23, CellChat inferred that Pdgfb-related signaling was upregulated in the cell to cell communications in Gli1 UUO (Figure B) & Ng2 UUO mice (Figure H) & Myh11 UUO (Figure E) than sham mice. As well, MK-related signaling was upregulated in both Ng2 UUO (Figure I) & Myh11 UUO (Figure F), IGF signaling (Figure C) & MIF signaling (Figure D) in Gli1 UUO, and SPP1 (Not shown in the Figure 3.1.23) & PTN (Figure J) & PERIOSTIN signaling (Figure K) in Pdgfrb UUO.

In order to know if the signalings are associated with the renal fibrosis, diverse literatures were reviewed. Platelet-derived growth factors (PDGFs) are normally expressed in renal mesenchymal cells, epithelial cells and injury [101]. The PDGF receptor-expressed mesenchymal cells have autocrine and paracrine effects, which are involved in renal diseases [101]. In the MK-related signaling (midkine gene), *Mdk* (+/+) mouse models of renal ablation developed renal failure compared with *Mdk* (-/-) mice [102]. Non-canonical WNT signaling promotes the downstream c-Jun N-terminal kinase (JNK) activation which in turn, activates both non-canonical TGF-b and PDGF signaling, which leads to fibrosis in the kidney [103], [57]. However, the Insulin-like growth factor (IGF) signaling is related to the attenuation of renal fibrosis by mediating vascular homeostasis and endothelial function [104]. Mif signaling in renal fibrosis is mentioned earlier, and both PTN and PERIOSTIN signaling are written below.

In Supplementary Figure 4.1.4, there were 4 signalings as gene-wise interactions, *Pdgfa-Pdgfrb* in both Gli1 (Figure B,D) and Myh11 (Figure F,H), *Wnt5a-Fzd1* in Gli1 (Figure A,C), *Mdk-Sdc2* in Myh11 (Figure E,G). Regarding the *Pdgfa-Pdgfrb* interactions, pericytes or endothelial cells sent signals in Gli1 and Myh11 sham mice data (Figure B,F), but in the matched UUO mice data, myofibroblasts and fibroblasts were also involved as senders as well as pericytes and endothelial cells (Figure D, H) Regarding the non-canonical Wnt signaling (*Wnt5a-Fzd1*) in Gli1 and MK signaling (*Mdk-Sdc2*) in Myh11, fibroblasts or myofibroblasts acted as a sender in both sham mice and UUO mice (Figure A, C, E, G), but the singal weight was a way higher in UUO mice than sham mice.

In Supplementary Figure 4.1.5.B,D,F,H, There are 4 different signals, *Mdk-Sdc2, Pdgfa-Pdgfrb, Ptn-Sdc2* and *Postn-(Itgav+Itgb5)* found in Ng2 (Figure A,B,C,D) and Pdgfrb (Figure E,F,G,H). UUO mice had more diverse interactions than shaam across Ng2 and Pdgfrb. Biologically, there

has been a lack of studies which explain how *Mdk-Sdc2* and *Wnt5a-Fzd1* are connected to renal fibrosis directly. *Pdgfrb* has been studied as an important protein causing renal fibrosis. PDGFR‑β activation alone is sufficient to induce progressive renal fibrosis and failure [105]. Additionally, in *Pdgfd* (-/-) mice, renal interstitial fibrosis was reduced in two models of renal injury, which was associated with reduced phosphorylation of PDGFR-β and p38 (*Mapk14* in mice) as a downstream mediator [106]. The downstream of *Pdgfa-Pdgfrb* interaction is related to JAK-STAT, PI3K, Ras-MAK signaling [107]. Here, the PROGENy output of UUO mice in Gli1, Ng2 and Myh11 displays that JAK-STAT signaling was upregulated across cells in UUO rather than sham, but PI3K, MAPK were not in Figure 3.1.14. A&C. Additionally, p38 (*Mapk14*) gene was rarely expressed in the Gli1, Ng2 and Myh11 UUO data set compared to Pdgfrb UUO data in Supplementary Figure 4.1.6. So, it represents that in Gli1, Ng2, Myh11 UUO mice data set, *Pdgfa-Pdgfrb*-JAK-STAT signaling would be one of the main factors which induced renal fibrosis and this is matched to other prior study in which PROGENy showed the highest activity in the JAK/STAT pathway in *Foxd1Cre::Pdgfrb+/J* mice model where PDGFR‑β signaling was activated in renal FoxD1+ mesenchymal cells [105].

Regarding the MIF signaling, *Mif-Ackr3* interactions were identified in Gli1 UUO mice data (Not in Figures). MIF has been known to induce CD74 complexes with CXCR2, CXCR4, or CXCR7 (*Ackr3* in mice), which in turn activate chemokine expression [108]. NF-κB pathways were upregulated in Gli1 UUO than Gli sham mice (Figure 3.1.14.A). Here, it infers that the Mif-CD74-CXCR7 (*Ackr3* in mice)-NF-κB-dependent mechanism would contribute to renal fibrosis especially in Gli UUO mice. The CellChat output of Pdgfrb mice data shows different features compared to other data sets, Gli1, Ng2 and Myh11. Two signaling, PTN and PERISTON signaling were found interesting. In sham, only one fibroblast sent a signal but, in UUO, diverse myofibroblasts and fibroblasts delivered signals to other cell types (Supplementary Figure 4.1.5.E-H). Here, the interesting thing is that the ligand of PERIOSTIN signaling, Postn is known to be involved in P38 MAPK pathway leading to induce renal fibrosis [109]. This result is matched to the PROGENy output of Pdgfrb mice data where MAPK pathway was up-regulated across Pdgfrb UUO mice data rather than Pdgfrb sham mice data (Figure 3.1.16.A). p38 (*Mapk14*) gene was more expressed in the Pdgfrb UUO than the Gli1, Ng2 and Myh11 UUO data set in Supplementary Figure 4.1.6. Pleiotrophin (PTN) has been known to promote the growth, survival, and migration of various cells. However, the direct evidence between the PTN and renal fibrosis is not enough.

Cd31 sham data consisted of diverse cell types. location-wise, there were 4 main parts from glomerular (glomeruli), cortical (cortex), medullary (medulla) of kidney compartments, and vasa recta in which the capillary networks supply blood to the medulla (ascending and descending). it

was abbreviated into “g”, “c” , “m”, “AVR” and “DVR” in order. Inside of each compartment, it was divided into 4 more parts, artery, capillary, arteriole and vein. As abbreviations, “art”, “cap”, “arteriole”, and “vein” were used here, respectively. The arteriole was subdivided into two parts, afferent and efferent arteriole which brings blood to the glomerulus and carries blood away from the glomerulus, respectively. Here, arteriole/efferent and arteriole/afferent have one more sub type, juxtaglomerular apparatus. Capillary is labelled as 4 different types, capillary/angiogenic, capillary/interferon, postcapillary venule and capillary. AVR and DVR were also subdivided into papilla and interferon. This annotation was from the public paper because Cd31 sham mice data were from the public data set [73]. In the case of the Cd31 UUO data set, the labelling was simpler than the Cd31 sham mice data set. endothelial cells, medullary endothelial cells (mdl_endo), fenestrated endothelial cells (fnst_endo), glomerulus endothelial cells (glm_endo), descending vasa recta (DVR_endo) and lymphatic endothelial cells (lymph_endo).

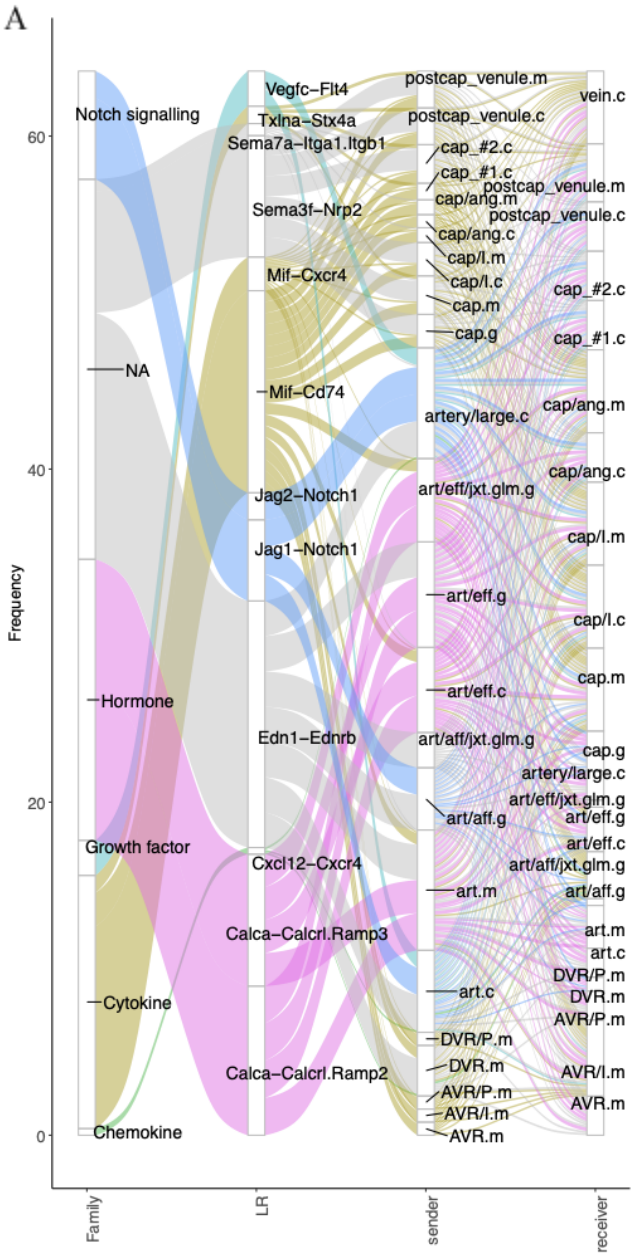


Figure 3.1.24.A is the output from Cd31 sham mice, Figure 3.1.24.B from Cd31 UUO mice data. In the sham mice data, diverse pathways were detected like hormone, cytokine, chemokine, growth factor, notch signaling-related immune signaling, etc. However, Cd31 UUO mice had non-hormone related immune responses compared to Cd31 sham mice data. Additionally, the accumulated “CondGeneProb” score of Cd31 sham (around 150) was higher than Cd31 UUO mice (around 80). When measuring the detected interactions ratio in which the number of communications were divided by the number of cells, the raio was 0.05 (577 interactions /10186 cells) in Cd31 sham, 0.05 (467 interactions / 8635 cells) in Cd31 UUO mice.

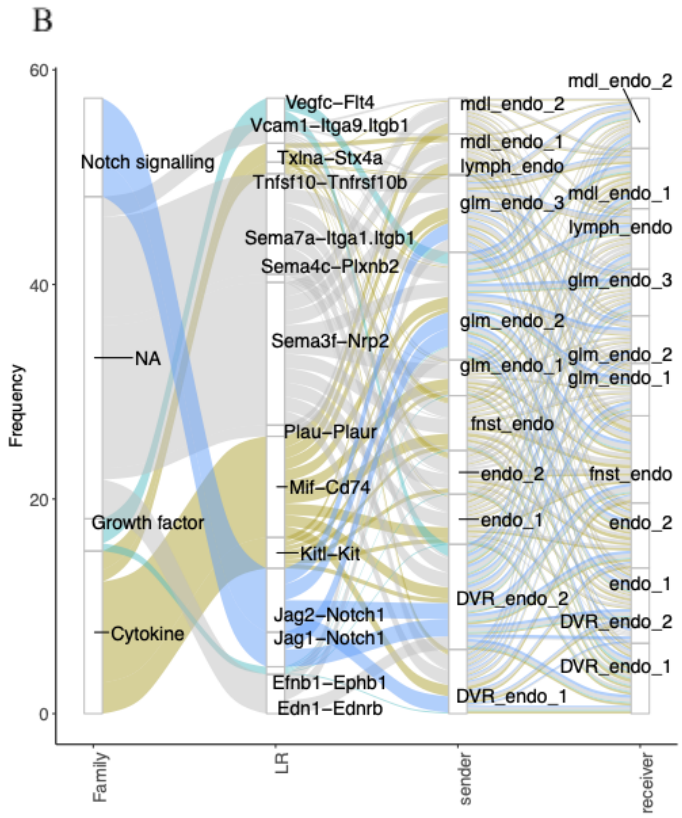


Figure 3.1.24. The intercellular communication of the CD31 sham and UUO mice data set. (A) Cd31 sham, (B) Cd31 UUO. Each color corresponds to biological pathways like Notch signaling (blue), growth factor (light blue), cytokine (olive), chemokine (green), checkpoint (red), hormone (pink) and non-specified signaling (grey). The x-axis consists of family, LR (ligands & receptors), sender (cell types) and receiver (cell types). The y-axis is the accumulated "CondGeneProb" of detected ligands and receptors. The thickness of the curve is the accumulated minimum percentage of cells expressing the ligand or the receptor of the interaction in a given family. In Figure A, Glomeruli, cortex, medullary and vasa recta (ascending and descending) are "g", "c" , "m", "AVR" and "DVR" of each. Artery, arteolie, capillary, and vein are "artery", "art", "cap", and "vein". Afferent and efferent are "aff" and "eff". Capillary is labelled as angiogenic ("ang"), interferon ("I"), postcapillary("postcap") venule and capillary. Papilla is "P" and interferon is "I". In B, there are endothelial cells (endo), medullary endothelial cells (mdl_endo), fenestrated endothelial (fnst_endo), glomerulus endothelial (glm_endo), descending vasa recta (DVR_endo) and lymphatic endothelial cells (lymph_endo)

In Figure 3.1.25, two signalings were found as interesting, APELIN and SEMA3 in Cd31 UUO mice. The APELIN (*Apln- Aplnr*) signaling was upregulated in Cd31 UUO than Cd31 sham mice data in Figure 3.1.25.A & Supplementary Figure 4.1.7.A. The SEMA3 signaling includes ligands, Sema3a (*Semaphorin 3 A*), d, f, g and it's receptors, *Nrp1+Plxna2, Nrp2+Plxna2 and Nrp2+Plxna4* ("+" means co-receptors). All of these interactions were not found in the Cd31 sham. Only *Sema3d-(Nrp1+Plxna2)* were detected in Cd31 sham mice, whereas in Cd31 UUO mice data set, *Sema3a − (Nrp1+Plxna2), Sema3d − (Nrp1+Plxna2), Sema3d − (Nrp2+Plxna2), Sema3d − (Nrp2+Plxna4), Sema3f − (Nrp2+Plxna2), Sema3f−(Nrp2+Plxna4), Sema3g−(Nrp2+Plxna2)* were detected. The Figure 3.1.25.B and Supplementary Figure 4.1.7.B include only the part of these outputs, *Sema3d* and *Sema3f* -related interactions in Cd31 sham and UUO mice data set. *Sema3d* were found in both Cd31 sham and UUO mice data, but the weight of the inferred signal is a way higher in Cd31 UUO than sham in Figure 3.1.25.B where the edge width is comparable between Cd31 sham and UUO. *Sema3f* (Supplementary Figure 4.1.7.B) and *Sema3a* (Not shown in figure) were co-expressed and interacted with

Nrp1/2+Plxna2 in Cd31 UUO but not in Cd31 sham mice data set.

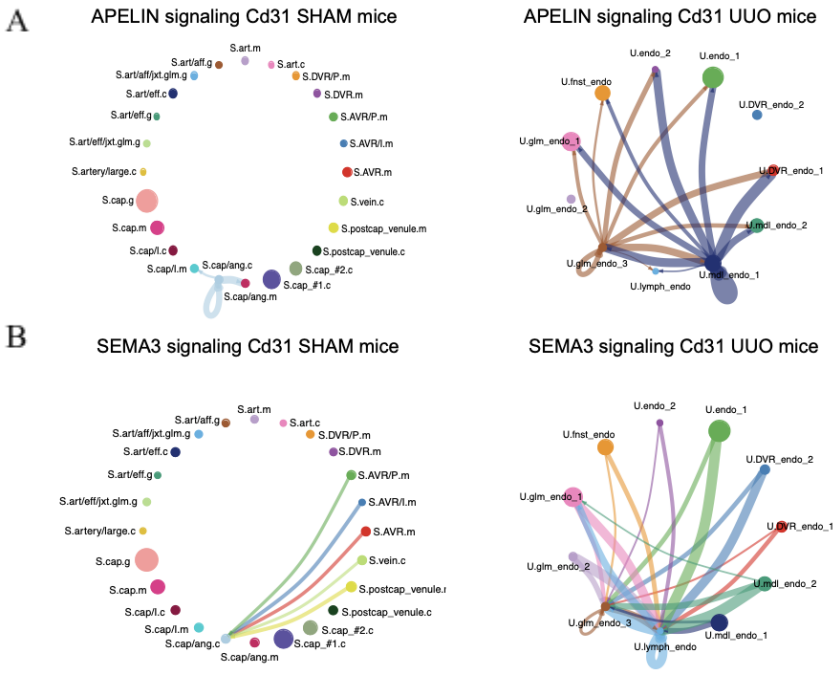


Figure 3.1.25. The CellChat intercellular communication of the CD31 sham and UUO data set. There are 2 different signalings, (A) APELIN (*Apln-Aplnr),* (B) *SEMA3 (Sema3d-(Nrp1, Plxna2), Sema3f-(Nrp2, Plxna2))*. The interaction width represents communication probabilities measured by the law of mass action with the amounts of the modeled ligand-receptor and each node corresponds to the cell type.

Biologically, The *Apelin (Apln)* has been described as a tip cell-enriched gene and activates the Apelin receptor (*Aplnr*) [110], [111]. Like VEGF and Notch signaling which regulate tightly vascular sprouting, Apelin signaling is involved in the angiogenesis independently which induces endothelial cells into a pro-angiogenic state [111]. Contrary to *Apelin*, *Sema3f* and *Sema3a* have been known to be involved in anti-angiogenesis by competing with *VEGF,* in which new blood vessels are not generated from pre-existing vessels in endothelial cells [112]. Renal failure is associated with defective angiogenesis [113]. In detail, the lack of delivery of angiogenesis and persistent hypoxia induce tissue destruction in glomerulonephritis, ischemic nephropathy, and tubulointerstitial fibrosis [114]. Additionally, it has been known that the upregulation of antiangiogenic factors would contribute to the deficiency of capillary recovery in the kidney [64]. Interestingly, the VEGF and Hypoxia signal of PRGOENy was higher across cells of Cd31 UUO than Cd31 sham (Figure 3.1.16.B). Here, I inferred that *Sema3f* and *Sema3a* would be one of the main factors leading to renal fibrosis by interacting with both Nrp1/2 and Plxna4 against high activity of VEGF in Pdgfrb mice data.

Cell differentiation

Fibroblasts are involved in wound healing by removing the fibrin clot, generating both new extracellular matrix (ECM) and collagen structures [115]. As well, fibroblasts differentiate into myofibroblasts, which reduce the margins of the wound and disappear by apoptosis after wound healing so that severe skin contraction cannot happen [116]. However, if the myofibroblasts remain abnormally, it makes excessive deposition of the extracellular matrix and finally leads to the destruction of the normal kidney and loss of renal function [117]. As well as fibroblasts, pericytes are also known to differentiate into myofibroblasts [118]. So, cell differentiation into myofibroblasts has been considered an important factor in renal fibrosis, but this differentiation has been poorly understood.

The Figure 3.1.26 has 8 different pseudo-times (Figure.C-J) and 2 umaps (Figure.A and B) of the Gli1 integrated by Harmony (Figure.A,C,E,G) and without any integration tool (Figure.B,D,F,H). Figure 3.1.26.C and D are from deterministic model, Figure E and F from stochastic model, Figure G and H are from dynamical model and the last 2 figures (Figure I and J) from an additional pseudotime, named “latent time” inferred by the same dynamical model too. In order to choose the most suitable pseudotime, the flow of colors were checked out from black to yellow if it matched with biological knowledge of differentiation to myofibroblasts. Comparing 3 different scVelo models, deterministic and stochastic models in UUO had more yellow colors (differentiated) than the dynamical model. Additionally, myofibroblasts (dotted area) of the Gli1 data integrated by Harmony (Figure 3.1.26.C) have more differentiated areas than the other cell types (out of dotted area) as well as the non-integration approaches too (Figure 3.1.26.D). The data integrated by Harmony were used for further analysis. As a next step, I interpreted the direction computed by the inferred pseudotime between every two cell types in Figure 3.1.26.K. The deterministic model (Figure 3.1.26.K) suited the biological point more than the other two models in that more number of directions moved from sham to UUO mice. Finally, based on the model, the top 100 differentially velocity-measured genes for each cell type were measured, and the output of myofibroblasts type 5 is in Figure 3.1.26.L.

In the Figure 3.1.26.L, in the myofibroblasts type 5 in Gli1 UUO mice, some interesting genes were found out, *Vegfd, Smoc2, Kcnma1, Hhip* and *Sept4*. The *Vegfd* (vascular endothelial growth factor D) has been known to induce the myofibroblasts growth, migration and the synthesis of collagens in hearts [119]. *Smoc2* (SPARC related modular calcium binding 2) has been studied to facilitate a fibroblast-to-myofibroblast transition as well as stress fiber formation, migration, proliferation and the production of extracellular matrix [120]. Silencing of either *Kcnma1* (potassium calcium-activated channel subfamily M alpha 1) showed that it mitigated the TGF-β1

activity, in turn, downregulates α-SMA which is known to induce myofibroblast differentiation [121]. *Hhip* (hedgehog interacting protein) gene activates the downstream of TGFβ1 including both Smad2 and Smad3 and induces endothelial to mesenchymal transition leading to endothelial fibrosis and apoptosis in diabetes [122]. *Sept4* (septin 4) is involved in liver fibrosis, and especially the change of *SEPT4* gene expression was observed to be associated with fibrotic changes [123].

Additionally, I inferred pseudotime in a different approach by PAGA in Figure 3.1.29.M. Unlike scVelo, PAGA measured pseudotime while using both transcriptome profiles and the prior knowledge on the origin cell type. PAGA assumed that the most differentiated cell was the fibroblasts type 9 in sham mice, and inferred that a lot of cell types in Gli1 integrated data had the same time point. However, it was not useful to do more additional analysis and didn't fit biological knowledge.

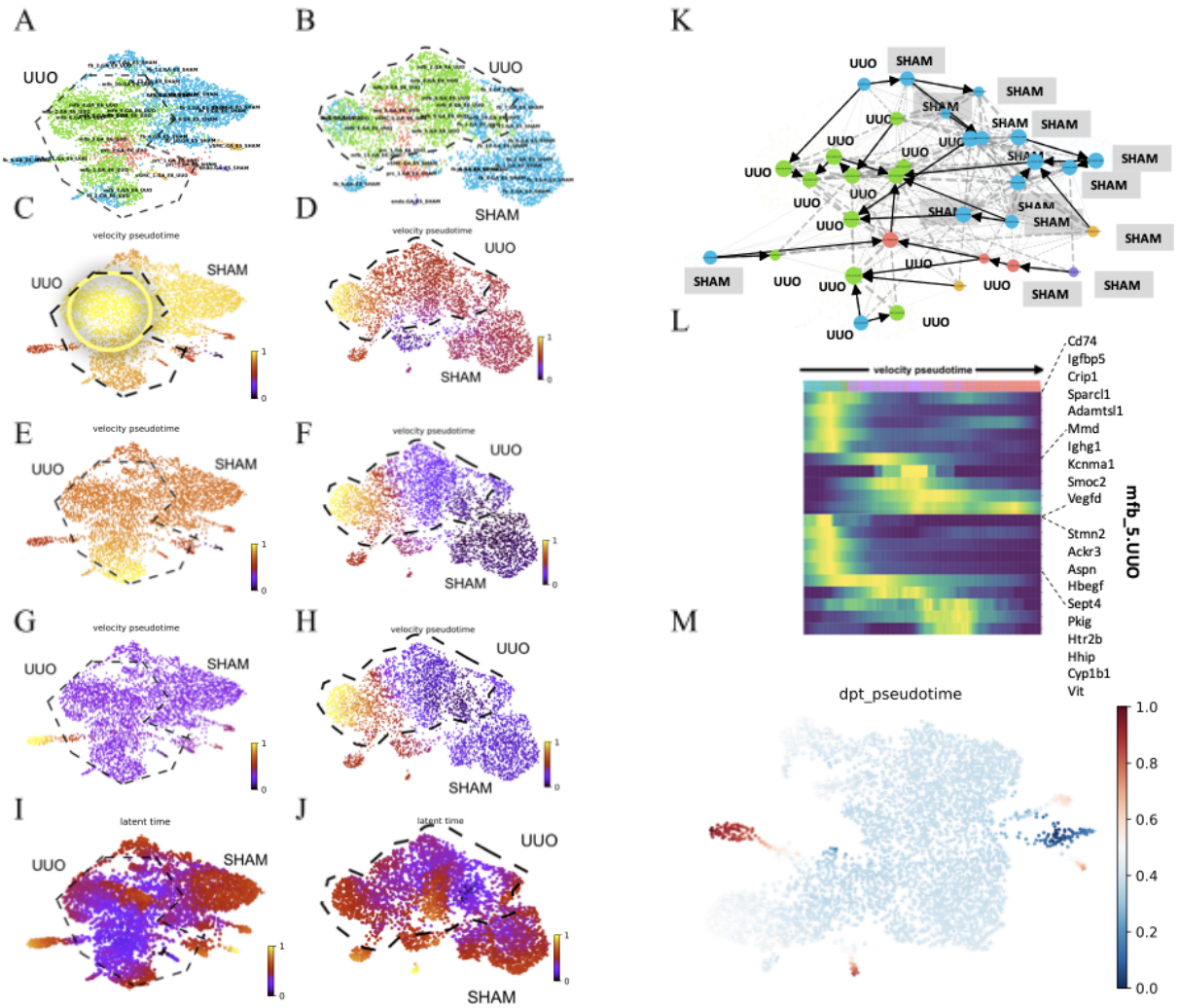


Figure 3.1.26. The scVelo's pseudotime of Gli1 integrated data. Figure A and B are two UMAP integrated by

Harmony and non-integration approach, respectively. It had only fibroblasts (blue), myofibroblasts (green), pericytes (red), vSMC (yellow) and endothelial cells (purple). In Figure C-J, pseudo times were inferred by deterministic (Figure C,D), stochastic (Figure E,F), dynamical (Figure G,H) and latent time (Figure I,J) by dynamical mode. The Figure on the left side in C-J is from data integrated by Harmony, right is from the merged without any integration one. It only included fibroblasts, myofibroblasts, pericytes and vSMC. The most differentiated cells are colored by yellow and the opposite has black and purple colors. In Figure M, the top figure shows the short lineages inferred by scVelo. In Figure L, for the cell type myofibroblasts type 5, top 20 velocity genes are listed with a gene expression matrix in which single cells are ordered by inferred by pseudo time (Time moves from pericytes of SHAM in blue, pericyte type1 of UUO in green, pericyte type 2 of UUO in purple to myofibroblasts type 5 of UUO in pink). Figure M is the diffusion map of PAGA.

I studied the pseudotime of Myh11 data in the same way as Gli1. Pseudotime was inferred by all different models, deterministic, stochastic and dynamical models. Like Gli1 data, the deterministic model (not shown in figures) and stochastic model (Figure 3.1.27.C) inferred pseudotime with the right biological point. Even though there were minor differences between the two outputs from deterministic and stochastic models, the stochastic model performed better than the deterministic model biologically (Figure 3.1.27.C). Figure 3.1.27.E shows the directions inferred by PAGA based on the pseudotime by stochastic model. The top 20 differentially expressed velocity genes for myofibroblasts type 1 were displayed in Figure 3.1.27.F. Several genes were found as interesting, *Adam19, Aldh1a2, Tpm1, Nkd2, Palld* in the top 20 differentially expressed velocity genes for myofibroblasts type 1 in Myh11 UUO mice. *Adam19* (ADAM metallopeptidase domain 19) gene has been studied to be a therapeutic target for renal fibrosis [124]. *Aldh1a2* (aldehyde dehydrogenase 1 family member A2) is an enzyme which has been known to be important in nephrogenesis and is upregulated almost 50-fold in myofibroblasts [125]. *Tpm1* (tropomyosin 1) has recently been revealed that the MicroRNA MiR-29c can lead to renal fibrosis attenuation and inhibit myofibroblasts formation by suppressing *Tpm1 [126]*. *Nkd2* (NKD inhibitor of WNT signaling pathway 2) has been identified as one of myofibroblast-specific genes in human kidney fibrosis [71]. Like *Aldh1a2, Fap* (fibroblast activation protein alpha) is also known as one of upregulated genes in the myofibroblasts [125]. *Palld* (Palladin, cytoskeletal associated protein) is over-represented in both tumor-related fibroblasts and kidney disease [127]. As a last step, I used the PAGA, and the output from PAGA was not suitable for the further analysis in Figure 3.1.27.G.

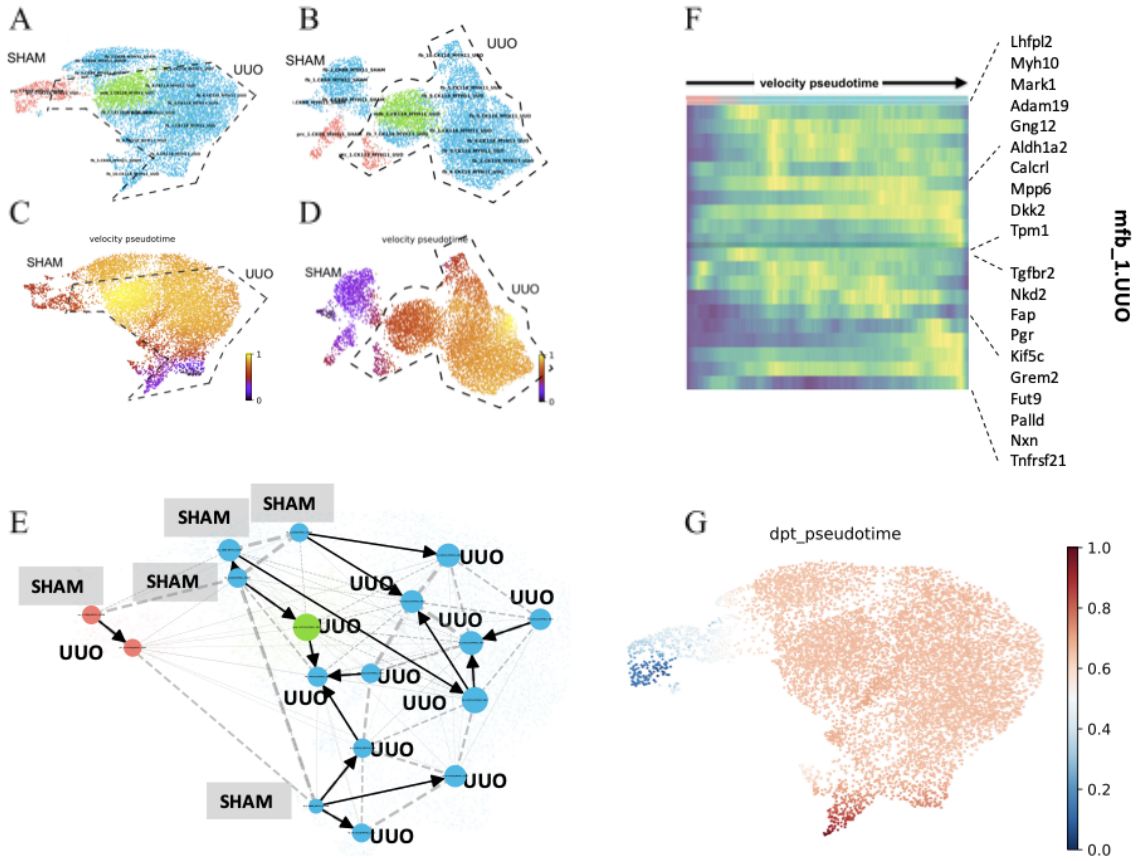


Figure 3.1.27. The scVelo's pseudotime of Myh11 integrated data. Figure A and B are two UMAP integrated by Harmony and non-integration approach, respectively. It only includes fibroblasts (blue), myofibroblasts (green), pericytes (red), vSMC (yellow) and endothelial cells (purple). In Figure C-D, pseudo-times were inferred by a stochastic model (Figure C (harmony), D (non-integration tool)). The most differentiated cells are colored by yellow and the opposite has black and purple colors. Figure E shows the short lineages inferred by scVelo based on the stochastic model. In Figure F, for the cell type myofibroblasts type 1, top 20 velocity genes are listed with a gene expression matrix in which single cells are ordered by inferred by pseudo time (Time flows from fibroblasts type 4 of UUO in pink to myofibroblasts type 1 of UUO in blue). Figure G is the diffusion map of PAGA.

Regarding Ng2 pseudotime analysis, I applied the same approach as Gli and Myh11. The pseudotime inferred by the deterministic model fitted biological knowledge (Figure 3.1.28.C (integrated by Harmony), D (integrated by non-integration tool)). In fibroblasts type 3 in Ng2 UUO, I found *Dapk2, Fhl2, Eln* and *Sorbs2* as interesting velocity genes (Figure 3.1.28.F). *Dapk2* is interesting in the top 20 genes in fibroblasts type 3 in Ng2 UUO. *Dapk2* (death associated protein kinase 2) knocked-out mice have some resistance to the accumulation of extracellular matrix in experimental renal fibrosis indicating that *Dapk2* plays a crucial role in profibrotic kidney injury [128]. *Fhl2* (four and a half LIM domains 2) is one of the genes which lead to the TGF-β1-induced tubular epithelial-to-mesenchymal transition through interacting with Wnt/β-catenin signaling [129]. Regarding *Eln*, *Eln* (elastin) was downregulated in *Txndc5*

(thioredoxin domain containing 5) null mice alleviating the over-expression of ECM protein genes in the mouse kidneys injury [130]. The silencing of *Sorbs2* (sorbin and SH3 domain containing 2) with the knockdown of *Kcnq1ot1* inhibits proliferation and fibrosis in diabetic nephropathy cells in humans [131]. I used the PAGA, and then reviewed the output in Figure 3.1.28.G. Two pseudo times inferred by scVelo and PAGA have similar time flow, but the output from PAGA had less time differences across cells in UUO so it was not suitable for further analysis.

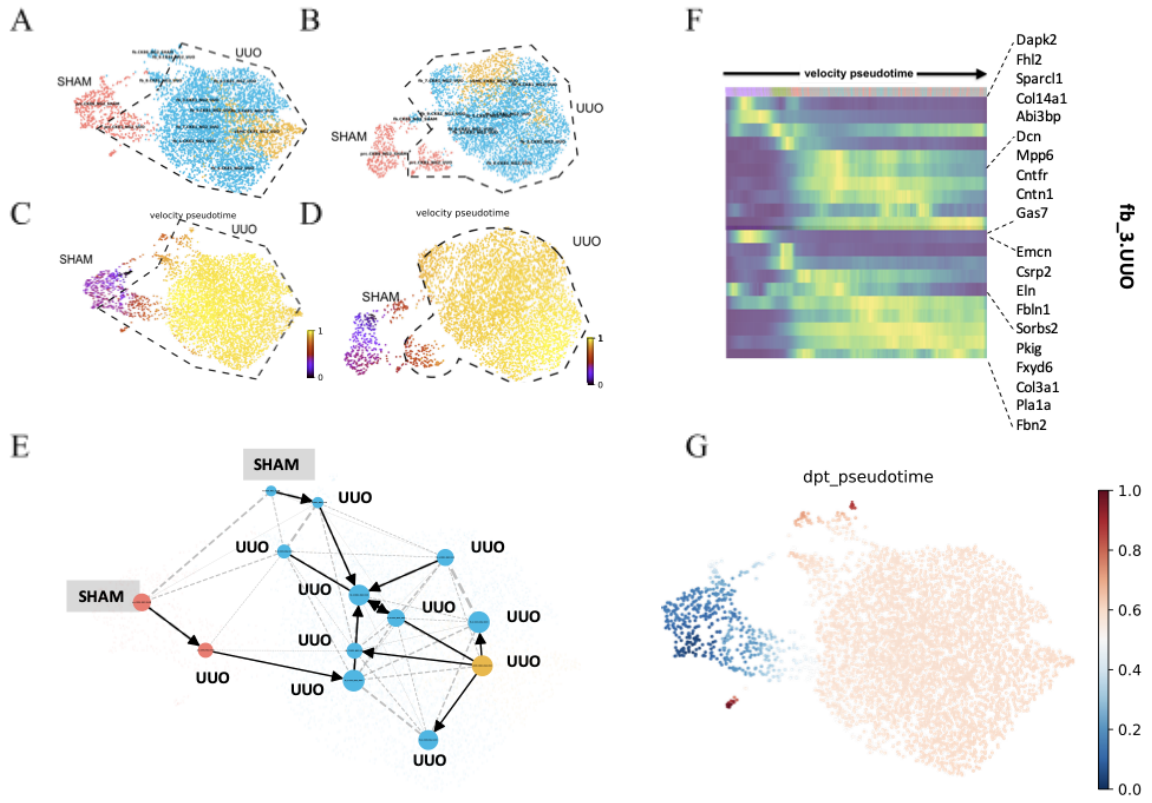


Figure 3.1.28. The scVelo's pseudotime of Ng2 integrated data. Figure.A and B are two UMAP integrated by Harmony and non-integration approach, respectively. It only includes fibroblasts (blue), myofibroblasts (green), pericytes (red), vSMC (yellow) and endothelial cells (purple). In Figure C-D, pseudo times were inferred by a deterministic model (Figure C (integrated by Harmony), D (integrated by non-integration tool)). The most differentiated cells are colored by yellow and the opposite has black and purple colors. Figure.E shows the short lineages inferred by scVelo based on the determinist model. In Figure F, for the cell type fibroblasts type 3, top 20 velocity genes are listed with a gene expression matrix in which single cells are ordered by inferred by pseudo time (Time flows from pericytes of SHAM in purple, pericytes of UUO in green, fibroblasts type 1 of UUO in pink to fibroblasts type 3 of UUO in blue). Figure.G is the diffusion map of PAGA.

Figure 3.1.29, I reviewed the both pseudotime and directions for the Pdgfrb integrated data, and then chose the deterministic mode for the further analysis. In myofibroblasts type 1 of Pdgfrb UUO mice, there were interesting velocity genes, *Il34, Sorbs2, Calcrl, Fit1, Dcn, Gucy1A1. Il34*

(interleukin 34) has been famous to stimulate macrophages-mediated tubular epithelial cells destruction during acute kidney disease, which worsens chronic kidney disease and increases chemokines reactions [132]. In the case of *Calcrl* (calcitonin receptor-like), *RAMP1 (receptor activity modifying protein 1), RAMP2 (receptor activity modifying protein 2)* and *CRLR* (Alias *Calcrl*) were upregulated in a rat model of renal injury, especially, upon induction of fibrosis for obstructive nephropathy [133]. In diabetic nephropathy (DN), *KCNQ1OT1* (KCNQ1 opposite strand/antisense transcript 1) and *SORBS2* (sorbin and SH3 domain containing 2) were overexpressed. Both knockdowns of these two genes, *KCNQ1OT1* and *SORBS2* suppressed proliferation, fibrosis and increased apoptosis in DN cells by repressing NF-κB pathway [131]. *Fit-1* (interleukin 1 receptor-like 1) is the vascular endothelial growth factor (VEGF) receptor 1 and soluble Flt-1 antagonizes VEGF [134]. High soluble *Flt-1* concentrations in patients with chronic kidney disease (CKD) displayed that it was directly associated with dysfunction of renal function [134]. *Dcn* (decorin) is one of the negative feedback loops of TGF-β which is implicated in renal fibrosis [135]. sGCα1 (*Gucy1A1*) has less literature which showed the direct relation between renal fibrosis and sGCα1 (*Gucy1A1*), but there is a report that this gene was overexpressed in fibrotic livers 2 times more than healthy livers [136]. I used the PAGA and then reviewed the output in Figure 3.1.29.G. The PAGA output was not matched with biological knowledge of disease models.

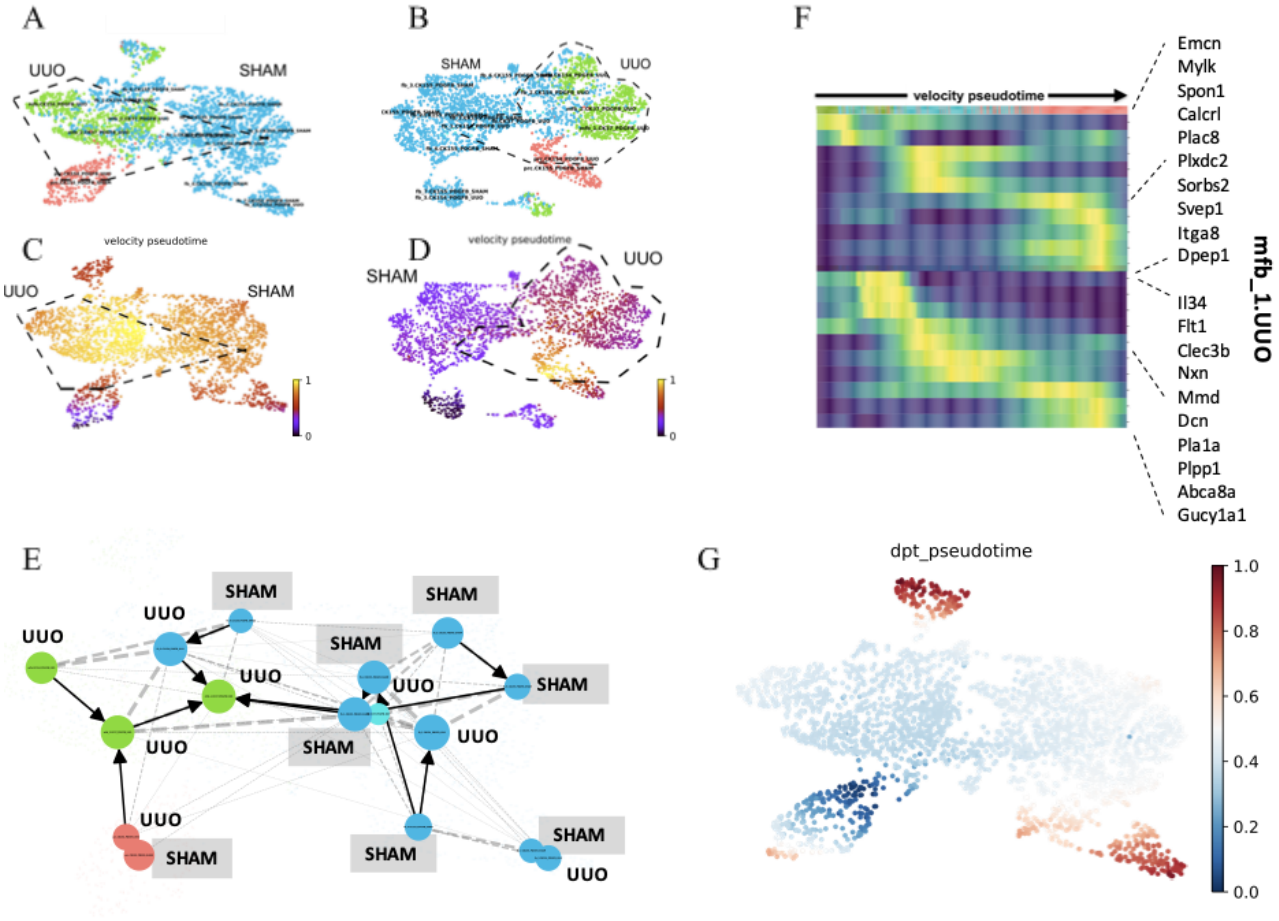


Figure 3.1.29. The scVelo's pseudotime of Pdgfrb integrated data. Figure.A and B are two UMAP integrated by Harmony and non-integration approach, respectively. It only includes fibroblasts (blue), myofibroblasts (green), pericytes (red), vSMC (yellow) and endothelial cells (purple). In Figure C-D, pseudo times were inferred by a deterministic model (Figure C (integrated by Harmony), D (integrated by non-integration tool)). The most differentiated cells are colored by yellow and the opposite has black and purple colors. Figure.E shows the short lineages inferred by scVelo based on the determinist model. Figure E shows the short lineages inferred by scVelo. In Figure F, for the cell type myofibroblasts type 1, top 20 velocity genes are listed with a gene expression matrix in which single cells are ordered by inferred by pseudo time (Time flows from pericytes of UUO in green, myofibroblasts type 2 of UUO in blue to myofibroblasts type 1 of UUO in red). Figure G is the diffusion map of PAGA.

In summary, I reviewed top 20 velocity genes per cell type of interest across all the renal fibrosis mice data, and 40 genes were already studied as renal fibrosis driver genes or fibrosis-related genes in other prior researches in kidneys, livers or hearts (Table 3.1.2). It indicates that this analysis was trustworthy, and other genes with insufficient literature evidence could be novel genes which can have high association with renal fibrosis.

		Gli1		Myh11	Ng2				Pdgfrb				
		mfb 5	prc 2	mfb 1	prc	fb 1	fb 3	fb 9	prc	mfb 1	mfb2	fb 2	Shared
1	*Tgfbi*												1
2	*Vegfd*												2
3	*Smoc2*												1
4	*Kcnma1*												1
5	*Hhip*												1
6	*Sept4.*												1
7	*Ramp1*												1
8	*Adam19*												1
9	*Aldh1a2*												2
10	*Tpm1*												1
11	*Nkd2*												2
12	*Palld*												1
13	*Il34*												2
14	*Mef2c*												2
15	*Slit3*												1
16	*Heph*												2
17	*Col12a1*												1
18	*Lpar1*												1
19	*Sema5a*												1
20	*Pdgfra*												1
21	*Gas6*												1
22	*Dapk2*												1
23	*Fhl2*												1
24	*Sorbs2*												3
25	*Eln*												2
26	*Bmper*												1
27	*Sfrp1*												1
28	*Plk2*												1
29	*Rbm24*												1
30	*Ednrb*												1
31	*Jag1*												1
32	*Pcsk6*												1
33	*Postn*												1
34	*Kitl*												1
35	*Calcrl*												2
36	*Flt1*												1
37	*Dcn*												2
38	*Gucy1a1*												1
39	*Cd86*												1
40	*Sulf2*												1
41	*Tgfb1*												1
42	*Spp1*												1
43	*Fn1*												2
44	*Epas1*												1
	Literatures	5	3	5	4	5	4	5	7	6	5	6	

Table 3.1.2. The 44 velocity genes with prior studies in renal or heart fibrosis. This table shows the 44 velocity genes were identified by cell differentiation studies on Gli1 UUO, Ng2 UUO, Myh11 UUO and Pdgfrb UUO with literature studies. Blue color corresponds to the fibrosis-related genes, whereas the pink color indicates the repair-related genes. The most right column shows how many times the same genes were found across data sets, the lowest column displays the number of papers which explains the relation between the gene and fibrosis.

3.2 Drug repositioning on chronic kidney disease microarray data

I performed drug repositioning for human chronic kidney disease. For bulk-level data analysis, 5 different public data sets were used, GSE20602, GSE32591, GSE37460, GSE47183 and GSE50469 [89]. The 9 different chronic kidney diseases include diabetic nephropathy (DN), hypertensive nephropathy (HN), lupus nephritis (LN), IgA glomerulonephritides (IgAN), membranous glomerulonephritis (MGN), minimal change disease (MCD), focal segmental glomerulosclerosis (FSGS), focal segmental glomerulosclerosis with minimal change disease (FSGS_MCD) and rapidly progressive glomerulonephritis (RPGN).

As a first step, I calculated the cosine similarity between a characteric direction of disease and a gene by the public tool (http://www.maayanlab.net/CD/) [85]. The characteristic direction was a linear line which had 90 degrees to the hyperplane which separated disease and health controls by LDA (Linear Discriminant Analysis) in gene-wise vector space. Secondly, I added the cosine profile of disease to the L1000CDS2 webtool (maayanlab.cloud/l1000cds2/). For each 9 chronic kidney diseases, the webtool provided top 50 ranked drug candidates which were reversely matched to the disease transcriptomics profiles of 9 CKD separately. Figure 3.2.2 has 9 circular graphs including the name of drug candidates, the number of cell lines, how many hours and how much the drugs were treated to that cell line. In order to simplify the drug candidates, I chose the drug candidates found in at least over 3 different chronic kidney diseases as below:

Small molecule	Name of Small molecule	The list of subtypes of CKD reversely matched to the signature of small molecule	The number of subtypes of CKD
BRD-K04853698	LDN-193189	mcd,fsgs,rpgn,ht,igan,mgn,dn	7
BRD-A75409952	wortmannin	fsgs_mcd,fsgs,rpgn,ln,igan,mgn	6
BRD-K71199328	YL-55	fsgs_mcd,mcd,ht,igan,dn	5
BRD-A34205397	suloctidil	mcd,fsgs,ln,ht,dn	5
BRD-K81528515	nilotinib	mcd,ht,igan,dn	4

BRD-K80786583	BRD-K80786583	mcd,ht,igan,dn	4
BRD-A80502530	cinobufagin	mcd,fsgs,ht,igan	4
BRD-A75144621	digoxin	mcd,ht,mgn,dn	4
BRD-K99252563	QL-XII-47	fsgs_mcd,ln,dn	3
BRD-K93034159	cladribine	fsgs_mcd,mcd,ht	3
BRD-K76703230	YM-155	mcd,ht,dn	3
BRD-K64517075	heliomycin	fsgs,ht,dn	3
BRD-K36740062	GSK-1070916	fsgs,ln,mgn	3
BRD-K16533489	I-606051	mcd,ln,dn	3
BRD-K06792661	narciclasine	ht,igan,dn	3
BRD-A77467113	EMF-sumo1-11	fsgs,ht,mgn	3
BRD-A67788537	salermide	mcd,ht,dn	3
BRD-A62184259	cycloheximide	fsgs,rpgn,ln	3
BRD-A52650764	ingenol	fsgs,ln,mgn	3
BRD-A19633847	perhexiline	ht,mgn,dn	3

Table 3.2.1. The table of the collapsed list of drug (small molecule) candidates across all 9 chronic kidney diseases. This table shows the information of 20 small molecules including a name, which & how many CKD subtypes belong to each drug / small molecule candidate.

I performed literature curations for the above 20 small molecules in Table 3.2.1. The 4 of 20 small molecules were identified to have prior studies which support the possibility that they could be used as a treatment for CKD. LDN-193189 (BRD-K04853698) is a BMP signaling

inhibitor. In 2015, Kajimoto et al. found out that this molecule prevented endothelial cells from dysfunction in chronic kidney disease mice [137]. Wortmannin as a PI3K inhibitor was studied to protect streptozotocin (STZ)-induced proteinuric renal disease in mice which had early diabetic nephropathy as well [138]. Narciclasine (BRD-K06792661) has an anti-inflammatory component which showed that this reduced the macrophage infiltration in UUO mice [139]. Interestingly, I found a drug approved by the FDA, Nilotinib (BRD-K81528515) [140]. This study showed the possibility that nilotinib could be used as a drug candidate for chronic kidney disease by attenuating renal disease progression and inflammation [140].

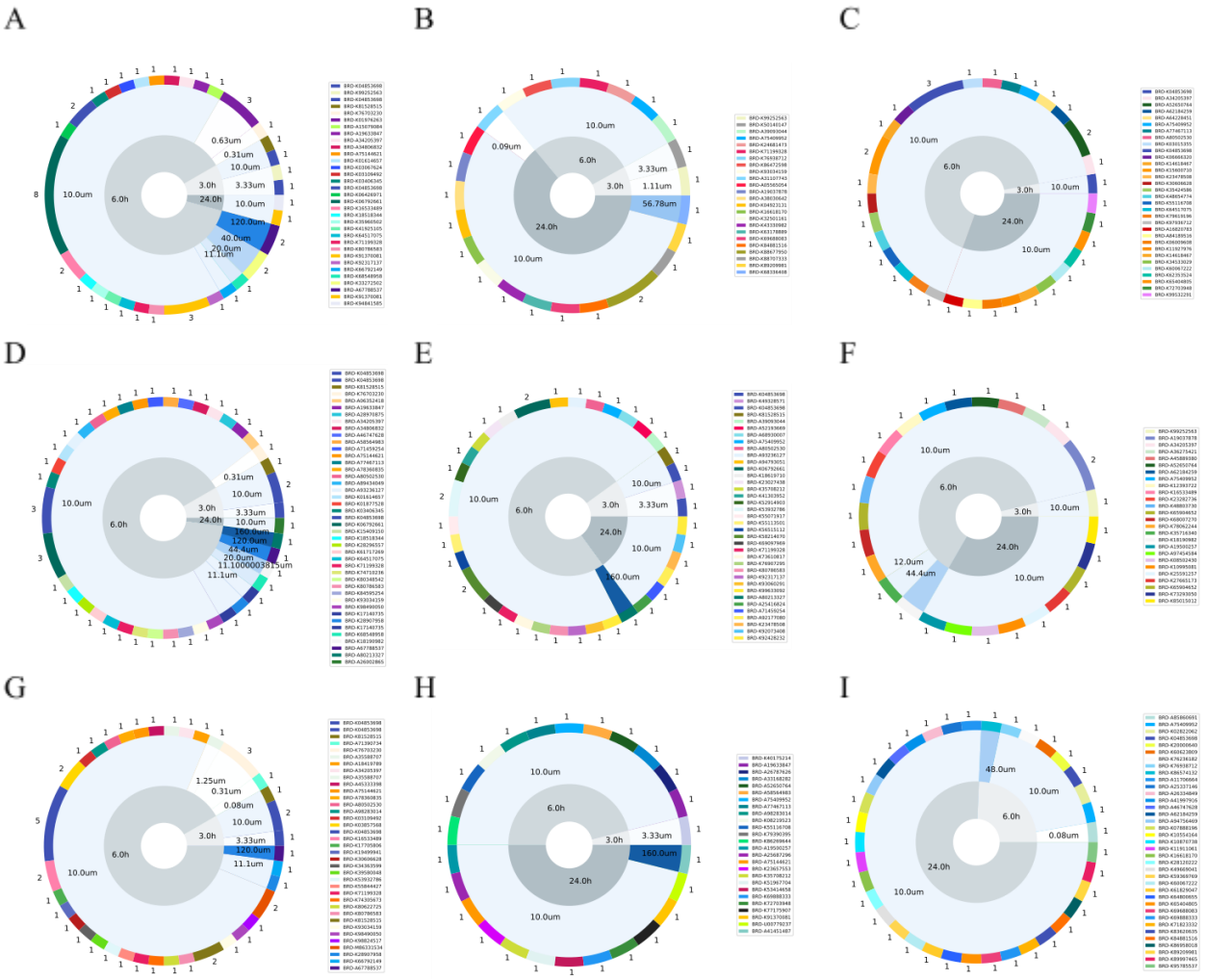


Figure 3.2.1. The 9 different circular graphs of top 50 ranked drug candidates from L1000CDS2 for the 9 different chronic kidney diseases. There are 9 circular graphs of top 50 drug (small molecules) candidates from the 9 different chronic kidney disease, (A) diabetic nephropathy (DN), (B) hypertensive nephropathy (HN), (C) lupus nephritis (LN), (D) IgA glomerulonephritides (IgAN), (E) membranous glomerulonephritis (MGN), (F) minimal change disease (MCD), (G) focal segmental glomerulosclerosis (FSGS), (H) focal segmental glomerulosclerosis with minimal change disease (FSGS_MCD) and (I) rapidly progressive glomerulonephritis (RPGN). Each graph shows the meta information of top 50 drug (small molecules) candidates for each disease. Inner (grey) circle area represents how

many hours small molecules were treated for in a range of 3 hours to 24 hours. Middle (blue) circle area captures how much of the small molecules were treated (from 0.31 um to 120.0 um). The outer circle area is how many times small molecules were inferred as candidates. In the case of the BRD-K0679266 (Figure A), it was mentioned 8 times in the list of top 50 candidates, which 8 corresponds to the number of different cell lines used for the drug treatment.

4. Supplements

ATAACTTCGTATAATGTATGCTATACGAAGTTATTAGGTCCCTCGACCTGCAGCCCAAGCTAGATCGAATT
CGGCCGGCCTTGTACGCGTTAAGTGCAACACGATCCCGCCACCATGGTGAGCAAGGGCGAGGAGGTCA
TCAAAGAGTTCATGCGCTTCAAGGTGCGCATGGAGGGCTCCATGAACGGCCACGAGTTCGAGATCGAG
GGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGCGGCC
CCCTGCCCTTCGCCTGGGACATCCTGTCCCCCCAGTTCATGTACGGCTCCAAGGCGTACGTGAAGCACC
CCGCCGACATCCCCGATTACAAGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACT
TCGAGGACGGCGGTCTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCACGCTGATCTACAAG
GTGAAGATGCGCGGCACCAACTTCCCCCCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGA
GGCCTCCACCGAGCGCCTGTACCCCCGCGACGGCGTGCTGAAGGGCGAGATCCACCAGGCCCTGAAGC
TGAAGGACGGCGGCCACTACCTGGTGGAGTTCAAGACCATCTACATGGCCAAGAAGCCCGTGCAACTG
CCCGGCTACTACTACGTGGACACCAAGCTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAA
CAGTACGAGCGCTCCGAGGGCCGCCACCACCTGTTCCTGGGGCATGGCACCGGCAGCACCGGCAGCGG
CAGCTCCGGCACCGCCTCCTCCGAGGACAACAACATGGCCGTCATCAAAGAGTTCATGCGCTTCAAGGT
GCGCATGGAGGGCTCCATGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACG
AGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGCGGCCCCCTGCCCTTCGCCTGGGACATCCTG
TCCCCCCAGTTCATGTACGGCTCCAAGGCGTACGTGAAGCACCCCGCCGACATCCCCGATTACAAGAAG
CTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGTCTGGTGACCGT
GACCCAGGACTCCTCCCTGCAGGACGGCACGCTGATCTACAAGGTGAAGATGCGCGGCACCAACTTCC
CCCCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCACCGAGCGCCTGTACCCC
CGCGACGGCGTGCTGAAGGGCGAGATCCACCAGGCCCTGAAGCTGAAGGACGGCGGCCACTACCTGG
TGGAGTTCAAGACCATCTACATGGCCAAGAAGCCCGTGCAACTGCCCGGCTACTACTACGTGGACACCA
AGCTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAGCGCTCCGAGGGCCGC
CACCACCTGTTCCTGTACGGCATGGACGAGCTGTACAAGTAAGAATTGTGTTGCACTTAACGCGTACAA
GGCCGGCCCTGCAGGAATTCGATATCAAGCTTATCGATAATCAACCTCTGGATTACAAAATTTGTGAAAG
ATTGACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCAT
GCTATTGCTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTTATGAGGAG
TTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCACTGTGTTTGCTGACGCAACCCCCACTGGTTGG
GGCATTGCCACCACCTGTCAGCTCCTTTCCGGGACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAAC
TCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGT
TGTCGGGGAAATCATCGTCCTTTCCTTGGCTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTC
CTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCCGGCTCTGCGG
CCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCATCGAT
ACCGTCGACCTCGACCT

Supplementary Table 4.1.1. The sequence of tdTomato used when running CellRanger.

	Pathway	p.value	p.adj	p.adj.signif
Gli1	TGFb	0	0	****
	p53	3.46E-157	4.84E-156	****
	TNFa	2.60E-126	3.64E-125	****
	VEGF	3.52E-70	4.93E-69	****
	WNT	4.81E-69	6.73E-68	****
	NFkB	4.06E-59	5.68E-58	****
	JAK-STAT	1.96E-36	2.74E-35	****
	Androgen	5.38E-36	7.53E-35	****
	Hypoxia	.007879630696	.1103148297	ns
	EGFR	1	1	ns
	Estrogen	1	1	ns
	MAPK	1	1	ns
	PI3K	0.4134238982	1	ns
	Trail	1	1	ns

	Pathway	p	p.adj	p.adj.signif
Ng2	Androgen	0	0	****
	Estrogen	0	0	****
	Hypoxia	0	0	****
	NFkB	0	0	****
	TGFb	0	0	****
	TNFa	0	0	****
	VEGF	0	0	****
	WNT	0	0	****
	p53	7.43E-212	1.04E-210	****
	JAK-STAT	4.76E-132	6.66E-131	****
	MAPK	.001199130841	.0167878318	*
	EGFR	1	1	ns
	PI3K	1	1	ns
	Trail	1	1	ns

	Pathway	p	p.adj	p.adj.signif
Myh11	Androgen	0	0	****
	Estrogen	0	0	****
	Hypoxia	0	0	****
	MAPK	0	0	****
	NFkB	0	0	****
	TGFb	0.00E+00	0.00E+00	****
	TNFa	0.00E+00	0.00E+00	****
	VEGF	0	0	****
	JAK-STAT	1.14E-283	1.60E-282	****
	WNT	5.58E-08	7.81E-07	****
	EGFR	1	1	ns
	PI3K	1	1	ns
	Trail	1	1	ns
	p53	1	1	ns

	Pathway	p	p.adj	p.adj.signif
Pdgfrb	VEGF	1.14E-108	1.60E-107	****
	PI3K	1.78E-98	2.49E-97	****
	MAPK	1.02E-71	1.42E-70	****
	NFkB	9.80E-27	1.37E-25	****
	WNT	1.11E-15	1.55E-14	****
	TNFa	2.96E-09	4.15E-08	****
	Androgen	1	1	ns
	EGFR	0.1620519494	1	ns
	Estrogen	1	1	ns
	Hypoxia	1	1	ns
	JAK-STAT	1	1	ns
	TGFb	0.9710717144	1	ns
	Trail	1	1	ns
	p53	1	1	ns

	Pathway	p	p.adj	p.adj.signif
Cd31	Androgen	0	0	****
	EGFR	0	0	****
	Hypoxia	0	0	****
	MAPK	0	0	****
	NFkB	0	0	****
	TGFb	0	0	****
	TNFa	0	0	****
	VEGF	0	0	****
	JAK-STAT	1.52E-243	2.12E-242	****
	Estrogen	1.51E-203	2.11E-202	****
	p53	3.93E-56	5.50E-55	****
	WNT	3.15E-19	4.41E-18	****
	PI3K	1	1	ns
	Trail	1	1	ns

Supplementary Table 4.1.2. The table of p-value, adjusted p-value for the PROGENy score in UUO mice compared to sham mice. PROGENy were measured for 5 different integrated data sets. For each 14 pathways, the significance was measured by a one sample t-test in which the null hypothesis was that PROGENy scores in UUO were not greater than sham, and the p-value was corrected by bonferroni. The t-tests were performed after reviewing density plots, Q-Q plots to check if it's normal distribution.

	TF	p-value	p.adj	p.adj.signif
Gli1	Ssrp1	0	0	****
	Myc	2.07E-286	2.07E-285	****
	Smad3	7.45E-251	7.45E-250	****
	Nr2c2	3.24E-206	3.24E-205	****
	Klf6	8.78E-183	8.78E-182	****
	E2f4	1.88E-153	1.88E-152	****
	Atf2	1.06E-85	1.06E-84	****
	Elk1	7.28E-15	7.28E-14	****
	Nfkb1	1	1	ns
	Thap11	1	1	ns

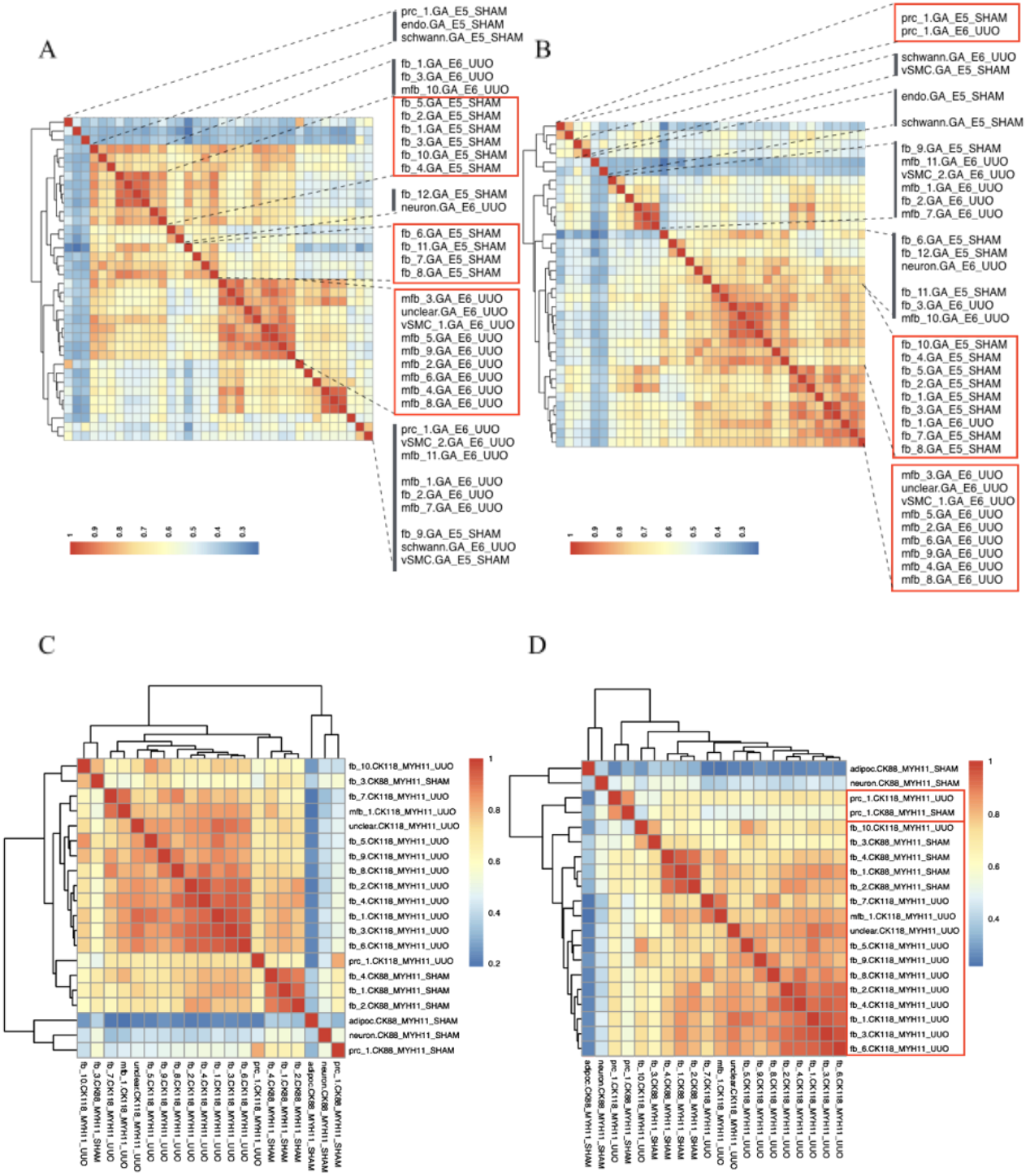
	TF	p-value	p.adj	p.adj.signif
NG2	Atf2	0	0	****
	Bcl3	0	0	****
	Klf6	0	0	****
	Myc	0	0	****
	Nr2c2	0	0	****
	Smad3	0	0	****
	Srf	0	0	****
	Ssrp1	0	0	****
	Tead1	0	0	****
	Rfx5	5.88E-14	5.88E-13	****

	TF	p-value	p.adj	p.adj.signif
MYH11	Klf6	0	0	****
	Mef2a	0	0	****
	Smad3	0	0	****
	Taf1	0	0	****
	Tcf12	0	0	****
	Tead1	0	0	****
	Zfp384	0	0	****
	Myc	1.03E-255	1.03E-254	****
	Ssrp1	2.11E-102	2.11E-101	****
	Foxp1	1	1	ns

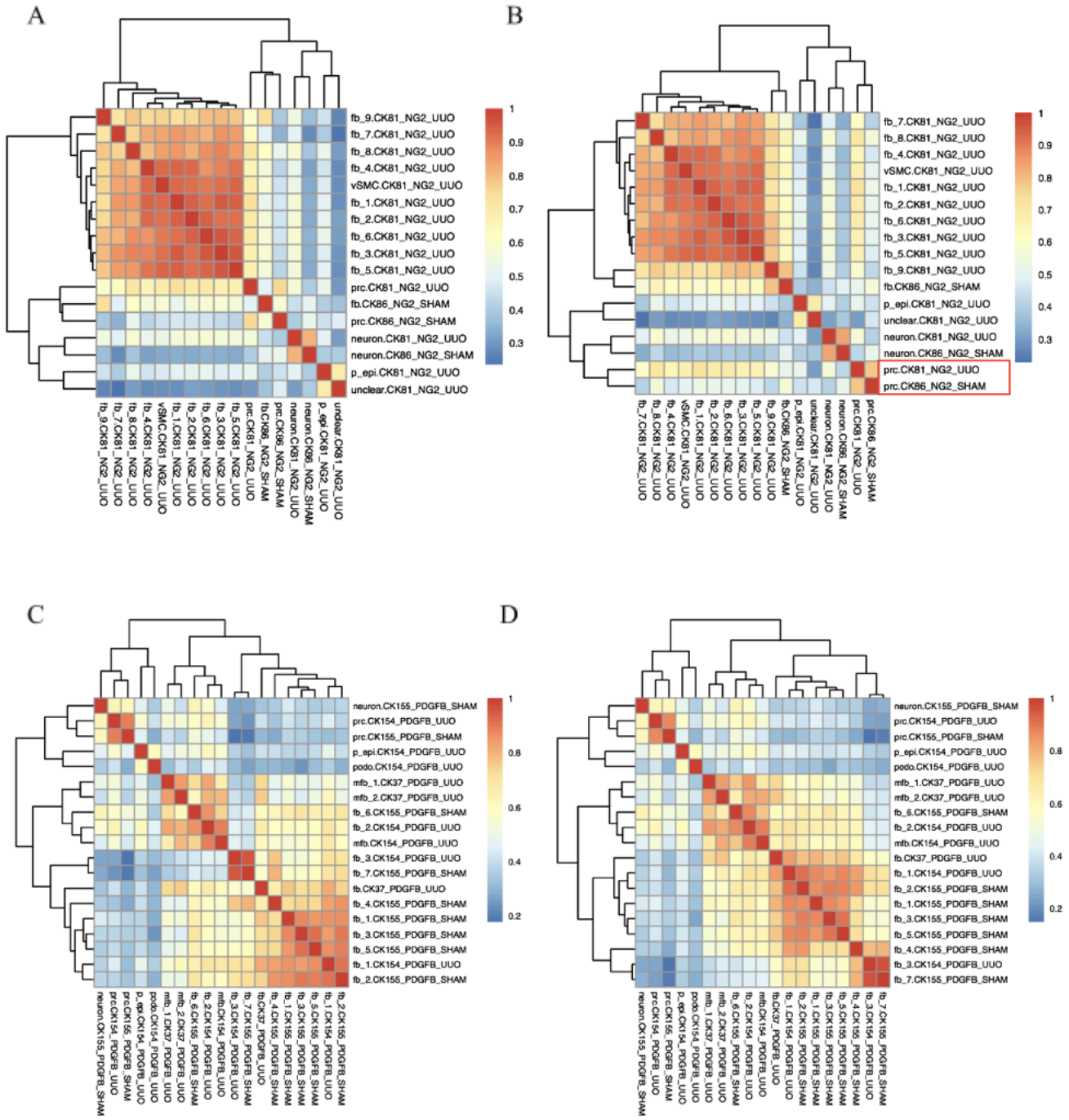
	TF	p-value	p.adj	p.adj.signif
PDGFRB	Atf2	4.93E-64	4.93E-63	****
	Tead1	7.24E-07	7.24E-06	****
	Mef2a	1.05E-06	1.05E-05	****
	Smad3	4.50E-05	.00045045451	***
	Taf1	8.07E-05	000806878515	***
	Esr1	1	1	ns
	Myb	1	1	ns
	Myc	0.9494993372	1	ns
	Ssrp1	0.1816815405	1	ns
	Tcf12	0.756776542	1	ns

	TF	p-value	p.adj	p.adj.signif
CD31	Klf6	0	0	****
	Kmt2a	0	0	****
	Mef2a	0	0	****
	Stat2	0	0	****
	Taf1	0	0	****
	Tcf12	0	0	****
	Tead1	0	0	****
	Tead4	0	0	****
	Myc	9.92E-76	9.92E-75	****
	Foxp1	1	1	ns

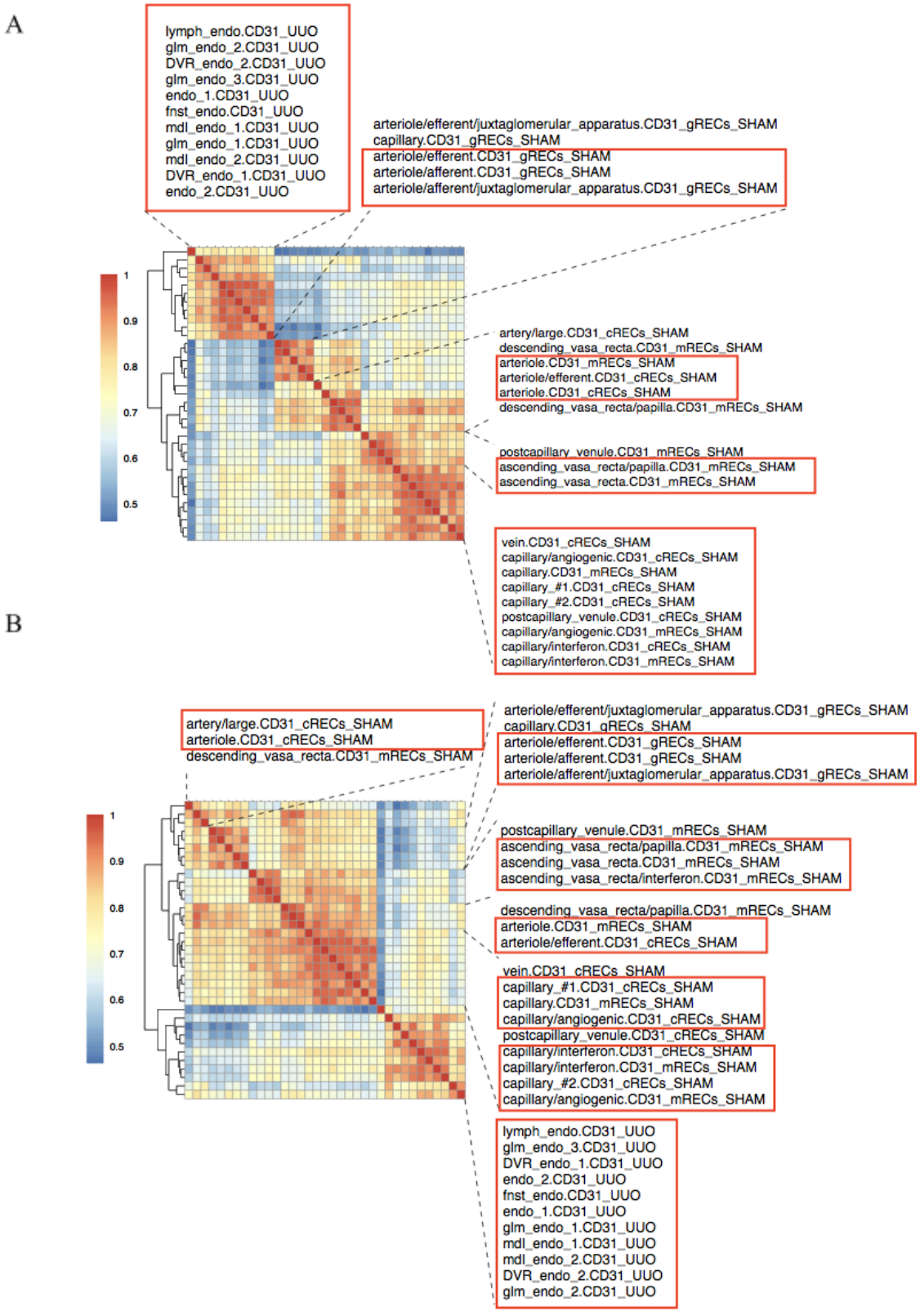
Supplementary Table 4.1.3. The table of p-value, adjusted p-value for the 10 different transcription factor activity scores in UUO mice compared to sham mice. DoRothEA and VIPER were used to measure transcription factor activity scores for the top 10 highly variable transcription factors across cell types. For those transcription factors, the significance was measured by a one sample t-test in which the null hypothesis was that transcription factor activity scores in UUO were not greater than sham, and the p-value was corrected by bonferroni. The t-tests were performed after reviewing density plots, Q-Q plots to check if it's normal distribution.



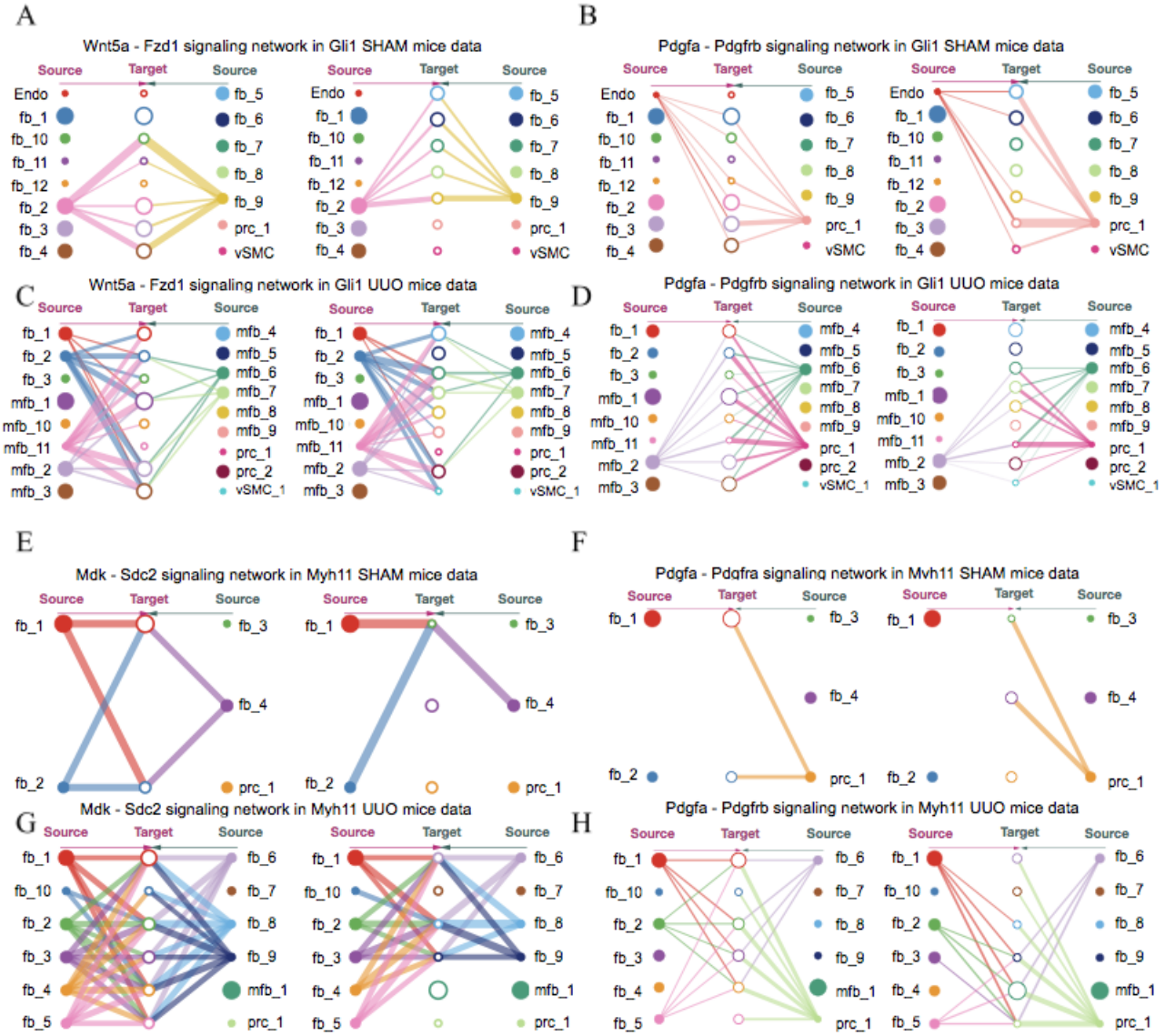
Supplementary Figure 4.1.1. The cell type correlation on the integrated matrix of Gli1 and Myh11. The upper panel (A and B) is from Gli1 integrated data sets combined without any integration tool (A) and Harmony (B). The lower panel (C and D) from Myh11 integrated data set combined without any integration tool (C) and Harmony (D). Cell type-specific correlation was measured by Pearson correlation on "CondGeneProb" matrix computed by genesorteR. All sub cell type numbers ("_1", "_2"…) are independent per each cell type such as fibroblasts, myofibroblasts, pericytes, etc. For example, sub-celltype 2 of fibroblasts ("fb_2") in SHAM is not related to the sub-cell type 2 of fibroblasts ("fb_2") in UUO.



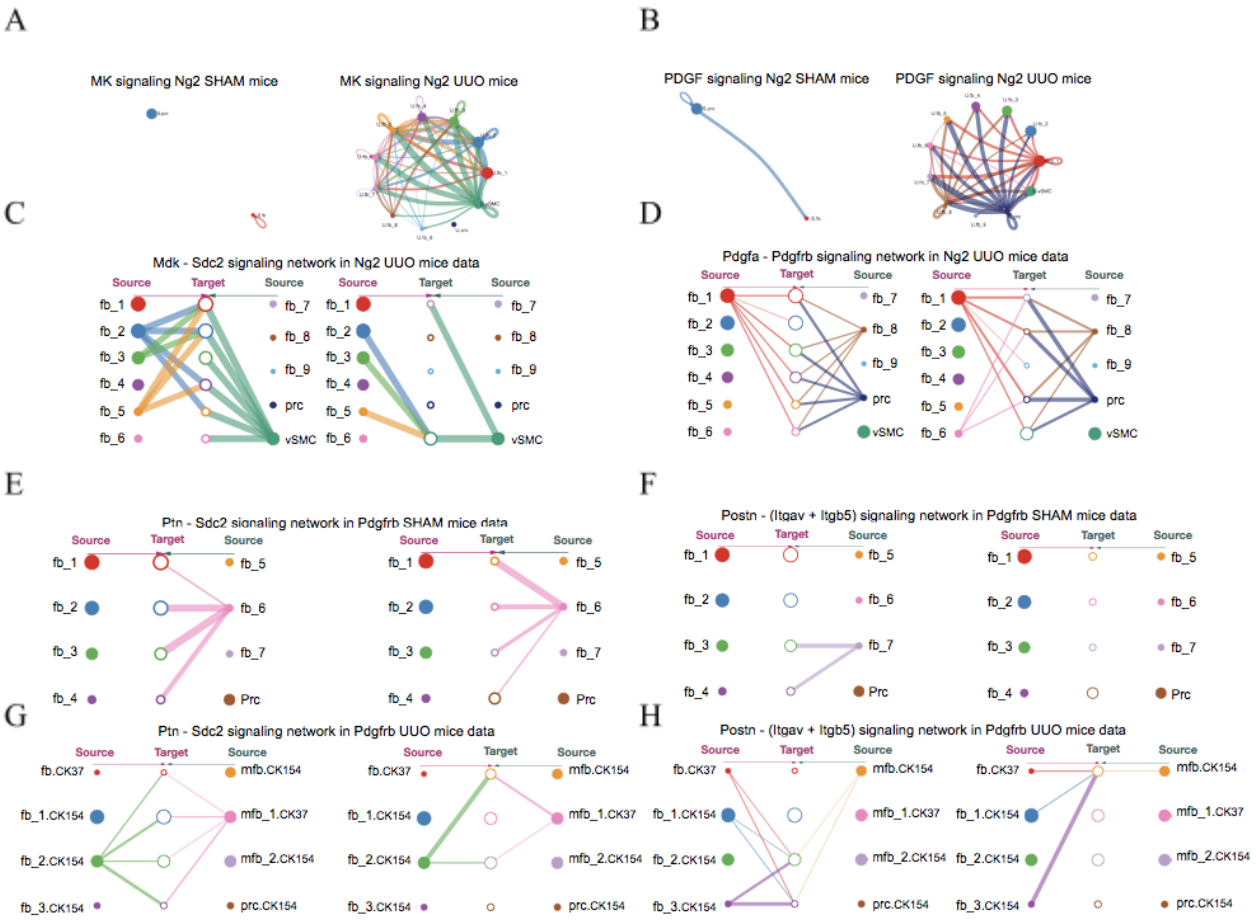
Supplementary Figure 4.1.2. The cell type correlation on the integrated matrix of Ng2 and Pdgfrb. The upper panel (A and B) is from Ng2 integrated data sets combined without any integration tool (A) and Harmony (B). The lower panel (C and D) from Pdgfrb integrated data set combined without any integration tool (C) and Harmony (D). Cell type-specific correlation was measured by Pearson correlation on "CondGeneProb" matrix computed by genesorteR. All sub cell type numbers ("_1", "_2"…) are independent per each cell type such as fibroblasts, myofibroblasts, pericytes, etc. For example, sub-celltype 2 of fibroblasts ("fb_2") in SHAM is not related to the sub-cell type 2 of fibroblasts ("fb_2") in UUO.



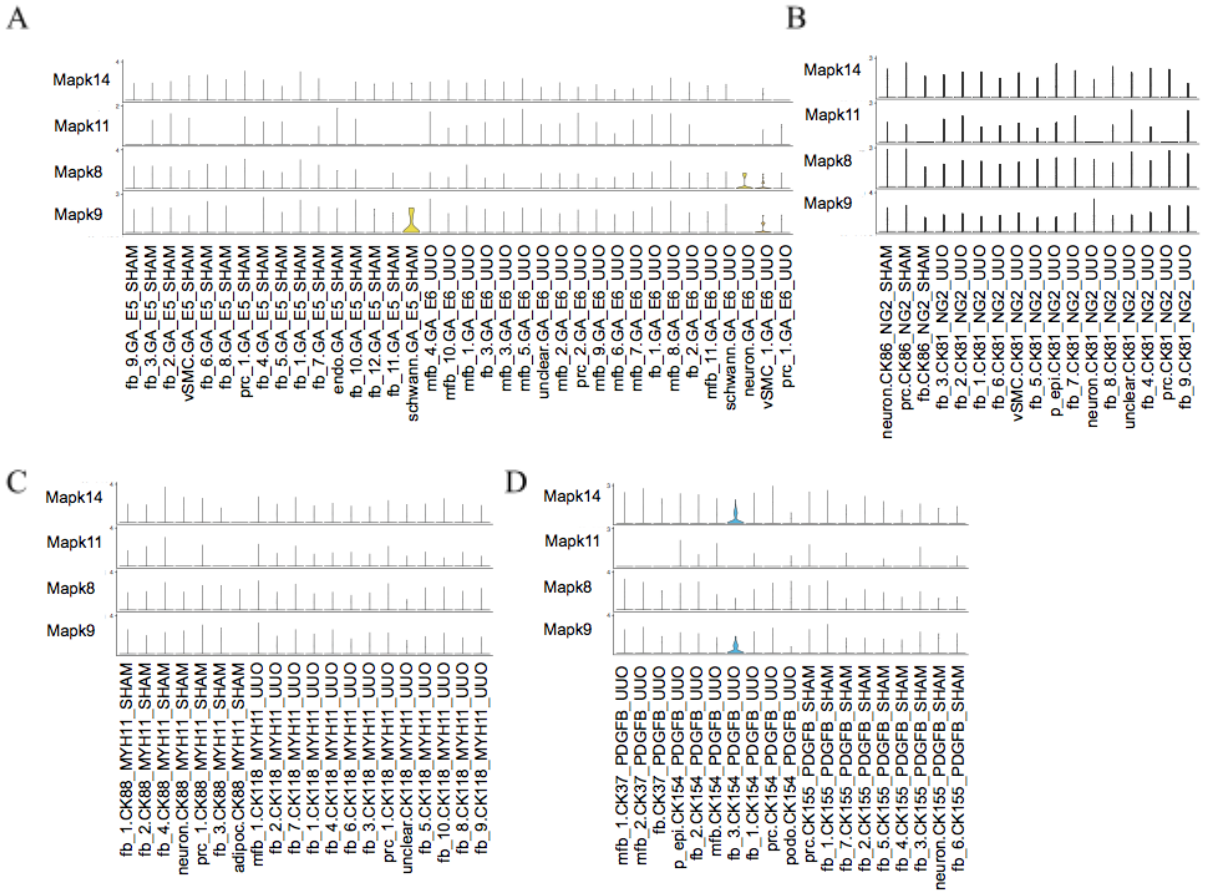
Supplementary Figure 4.1.3. The cell type correlation on the integrated matrix of Cd31. The cell-ctype correlation matrices of Cd31 combined by (A) the non-integration tool and (B) Harmony. Cell type-specific correlation was measured by Pearson correlation on "CondGeneProb" matrix computed by genesorteR.



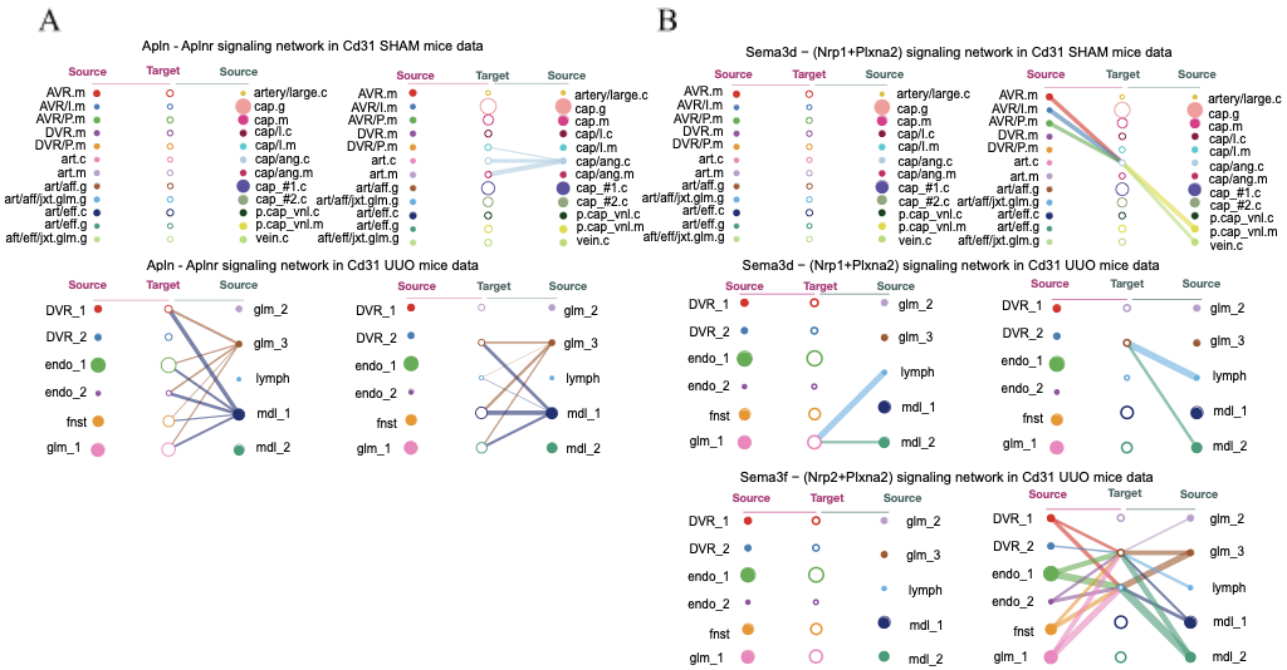
Supplementary Figure 4.1.4. The CellChat intercellular communication of the Gli1 and Myh11 data. There are 3 different signalings, *Wnt5a-Fzd1, Pdgfa-Pdgfrb* and *Mdk-Sdc2*. The interaction width represents communication probabilities measured by the law of mass action with the amounts of the modeled ligand-receptor and each node corresponds to the cell type. (A) the intercellular interactions of *Wnt5a-Fzd1* from Gli1 sham mice data, (C) from Gli1 UUO mice data. (B) the intercellular interactions of *Pdgfa-Pdgfrb* from Gli1 sham mice data, (D) from Gli1 UUO mice data. (E) the intercellular interactions of *Mdk-Sdc2* from Myh11 sham mice data, (G) from Myh11 UUO mice data. (F) the intercellular interactions of *Pdgfa-Pdgfra* from Myh11 sham mice data, (H) *Pdgfa-Pdgfrb* from Myh11 UUO mice data.



Supplementary Figure 4.1.5. The CellChat intercellular communication of the Ng2 and Pdgfrb data. There are 4 different signals, *Mdk-Sdc2, Pdgfa-Pdgfrb, Ptn-Sdc2* and *Postn-(Itgav+Itgb5)*. The interaction width represents communication probabilities measured by the law of mass action with the amounts of the modeled ligand-receptor and each node corresponds to the cell type. (A) the intercellular interactions of *Mdk-Sdc* from Ng2 sham mice data, (C) same interaction from Ng2 UUO mice data. (B) the intercellular interactions of *Pdgfa-Pdgfrb* from Ng2 sham mice data, (D) from Ng2 UUO mice data. (E) the intercellular interactions of *Ptn-Sdc2* from Pdgfrb sham mice data, (G) from Pdgfrb UUO mice data. (F) the intercellular interactions of *Postn-(Itgav+Itgb5)* from Pdgfrb sham mice data, (H) from Pdgfrb UUO mice data.



Supplementary Figure 4.1.6. The gene expression of p38 and JNK across data. Mapk14 and 11 are known as p38, and Mapk8 and 9 are c-Jun N-terminal kinases (JNK). (A) Gli sham and UUO mice data, (B) Ng2, (C) Myh11 and (D) Pdgfrb. p38 and JNK are highly expressed in fibroblasts type3 in the Pdgfrb UUO mice data set compared to other data sets.



Supplementary Figure 4.1.7. The CellChat intercellular communication of the CD31 sham and UUO data set. There are (A) The upper panel displays the intercellular interactions of APELIN from Cd31 sham mice data, the lower one from Cd31 UUO mice data. (B) The upper panel shows the intercellular interactions of *Sema3d-(Nrp1, Plxna2)* from Cd31 sham mice data, the middle one from Cd31 UUO mice data and *Sema3f-(Nrp2, Plxna2)* the lower one from Cd31 UUO mice data.

5. Discussion

From June 2017 to December 2020, I worked on single-cell RNA-seq data analysis from mice renal fibrosis and bulk RNA-seq data analysis from human chronic kidney disease. In the mice renal fibrosis studies at a single cell level, I used scRNA-seq data from *Gli1-CreER-tdTomato*, *Ng2-CreER-tdTomato*, *Myh11-CreER-tdTomato*, and *Pdgfrb-CreER-tdTomato* mice and Cd31 sorted mouse cells to trace the pathology of renal fibrosis. *Gli1*, *Ng2*, *Myh11*, and *Pdgfrb* are expressed in fibroblasts, myofibroblasts, pericytes, neurons, and, rarely, endothelial cells around the perivascular niche of the kidney. In these mouse lines, the genes of interest were genetically tagged with tdTomato, allowing the cells to be sorted by color (fluorescent orange) and computationally by sequence. Cd31 is used as an endothelial cell marker without the requirement for a fluorescent tag. Dr. Kuppe performed these experiments in normal and renal fibrosis mice, generating 9 data sets. The 9 different data sets were labeled by Dr. Kuppe. Additionally, I used public data for Cd31 sham mice (normal mice).

I used CellRanger for aligning reads and Seurat for preprocessing and clustering. I found Harmony to perform better than Seurat for data integration, resulting in fewer artificial effects from the integration algorithms as compared to data merged by non-integration approaches. However, the batch-corrected, normalized data from Harmony produced negative values (e.g., 0 had meaning) and was further preprocessed by scaling before being used for GSEA and cell differentiation studies. The GSEA output fully represented the biological features of cell types. For other downstream analyses, such as functional and intercellular communication studies, I used data merged without any integration as they produced stronger and clearer signals than batch-corrected, positively scaled data (Figure 5.1.1.A).

Downstream analysis consisted of 3 major parts: pathway/transcription-factor functional studies, intercellular communication, and cell differentiation. The functional studies were subdivided into 3 additional parts: 1) pathway studies with PROGENy, 2) Enrichr using gene sets from GO Biological Process (2018), and 3) transcription factor studies with DoRothEA and VIPER. PROGENy has its own gene-by-pathway weight matrix for its 14 built-in pathways to compute pathway activities [30]. Enrichr calls a gene set from GO Biological Process (2018) and ranks biological pathways or terms from that gene set [26]. DoRothEA provides transcription factors with their targets (genes), and VIPER calculates transcription factor activity based on the expression of target genes. I used 2 different tools, CellChat and ICELLNET database with genesorteR, to analyze intercellular communication. CellChat is useful for inferring intercellular

communication using differentially expressed genes, mass actions, and agonists and antagonists [43]. As ICELLNET [40] with genesorteR [20] simply queries the gene set for known ligand-receptor interactions identified in CellChat, it was used as a complementary approach to CellChat. scVelo was used for the cell differentiation study [36].

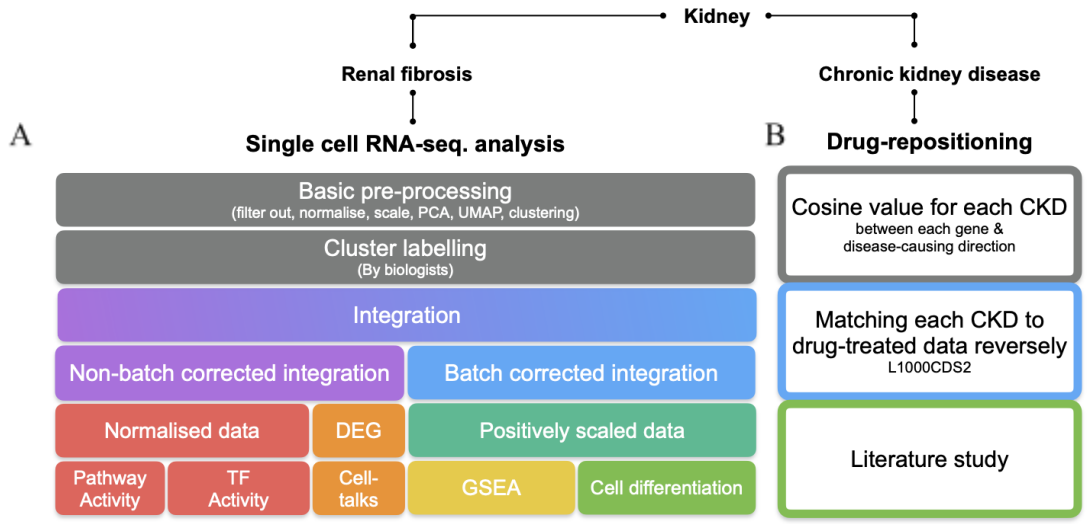


Figure 5.1.1. The graphical overview of data-analysis pipeline on single cell RNA-seq. data (left) and drug repositioning (right). This thesis has two main parts, the single-cell RNA-seq data analysis from mice renal fibrosis (Figure A) and drug-repositioning on 9 different chronic kidney disease microarray data sets (Figure B). In Figure A, single-cell RNA-seq data analysis consists of basic preprocessing which includes filtering out cells & genes, normalizing count-matrix datasets, reducing data by PCA & UMAP, and then clustering. For the preprocessed data set, Chrisotph Kuppe labelled each cluster to the cell type for 9 different data sets (the 10th data set from public data with its annotation). For the labelled data sets, I integrated datasets in 2 different ways while removing batch effects and remaining batches (non-batch corrected integration), separately. For the non-batch corrected integrated data set, pathway activities, transcription factor (TF) activities were measured, and intercellular communication were performed. For the batch-corrected normalized data set, it had minus values as minimum. So, I scaled the data, and then took only positively scaled data. Gene set enrichment test was performed on this data, and the outputs represented the main features of cell types fully. With this biological point, cell differentiation was also conducted on the same batch-corrected positively scaled data. The right panel, Figure B shows the drug-repositioning analysis on the microarray data from 9 different chronic kidney diseases (CKD). Data sets were already preprocessed by Ferenc Tajti and Mahmoud M. Ibrahim. The preprocessed data were used to measure the disease-causing direction in gene-wise vector space between disease and healthy controls for each CKD. Cosine similarities between the direction and each gene were measured. The 9 different cosine profiles for each CKD were used as an input for a tool, L1000CDS2 in order to match the disease profile to drug-treated data of LINCS-L1000 reversely. For the 20 drug candidates found in at least over 3 different CKD, literature study was conducted in order to know which drugs were already proved experimentally by other researchers.

I interpreted the functional studies (pathway/transcription factor activities) along with

intercellular communication results. Interestingly, I connected several ligand-receptor pairs with biologically matched signaling pathways (or transcription factors), consistent with previous renal fibrosis research. Specifically, I found a greater number of *Pdgfa-Pdgfrb* interactions in renal fibrosis mice than normal mice, which was connected to upregulated JAK-STAT signaling in renal fibrosis mice [105], [107]. Additionally, *MIF-CD74-CXCR7* (*Ackr3* in mice) interactions were associated with activated NFkB pathways in the renal fibrosis mice [108]. I also discovered interactions between *Postn* (ligands) and *Igtav* + *Itgb5* (2 receptors) with p38 MAPK pathways in the renal fibrosis mice [109]. These 3 pathways appear to function in the pathology of renal fibrosis, consistent with previous work (Figure 5.1.2). PDGFR was also identified as a target in the drug repositioning study, as nilotinib, identified as a drug repurposing candidate for CKD, is known to inhibit the progression of CKD by inactivating PDGFR [140]. Hypoxia, which induces endothelial to mesenchymal transition (EndMT), was upregulated in the PROGENy outputs in endothelial cells of renal fibrosis, particularly when *Myc* is activated (Figure 5.1.2). *Myc* suppression is reported to alleviate fibrosis in tubular epithelial cells [96]. Another transcription factor, *Klf6*, was implicated owing to its role in upregulating TGF‑β, which in turn induces epithelial‑mesenchymal transition (EMT) [96]. Another transcription factor, *Klf6*, was implicated owing to its role in upregulating TGF‑β, which in turn induces epithelial‑mesenchymal transition (EMT) [97]. TGF‑β activates TNF-α and NF-κB, promoting EndMT and leading to renal fibrosis. Additionally, *Sema3f* and *Sema3a* were found to be co-expressed in the endothelial cells of renal fibrosis mice with upregulated VEGF. Interestingly, these semaphorins are known to compete with VEGF, inhibiting VEGF-promoting angiogenesis [112] and potentially leading to renal fibrosis.

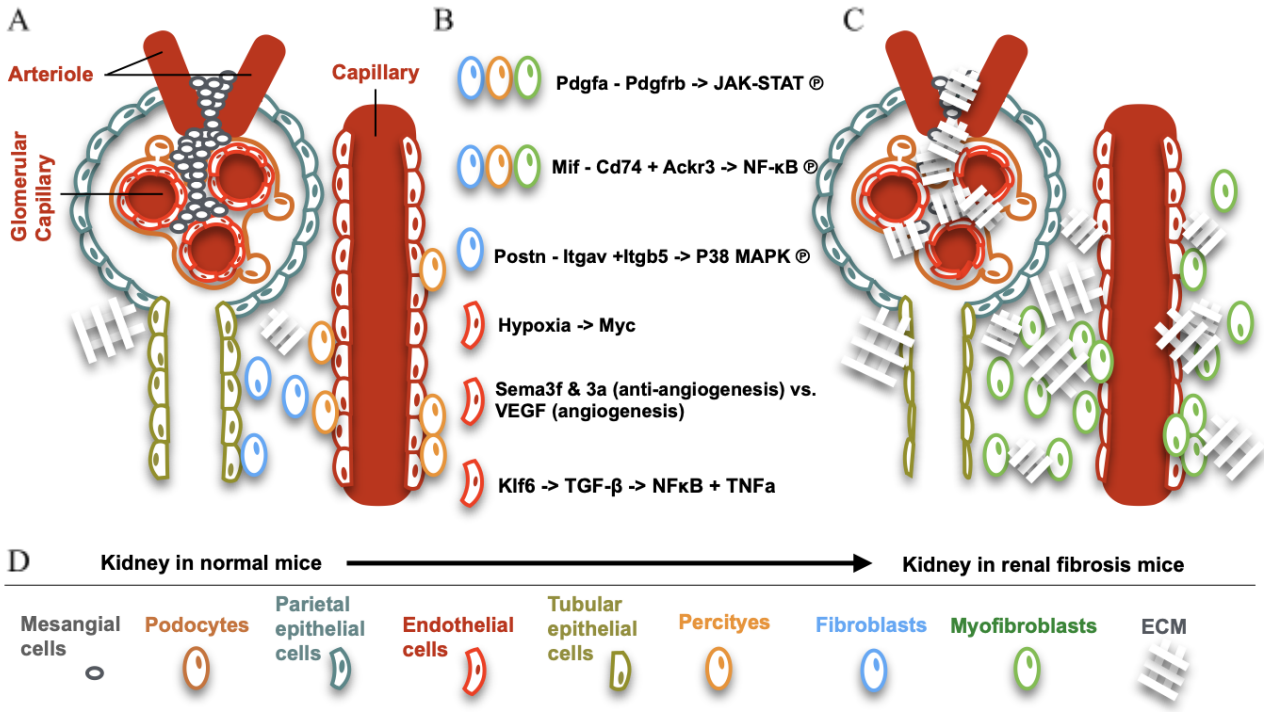


Figure 5.1.2. The graphical summary of renal fibrosis-related ligand-receptor interactions and pathways shared with other prior studies. Left figure (Figure A) represents the normal kidney in mice and the right (Figure.) is renal fibrosis in mice. Two kidneys figures include glomerular and proximal tubules with blood vessels (Arteriole, glomerular capillary and another capillary). Figure D has a legend about cell types with colors, fibroblasts with blue, myofibroblasts with green, percityes with orange, parietal epithelial cells with olive, endothelial cells with red, tubular epithelial cells with dark cyan, mesangial cells with grey, and extracellular matrix (ECM) with a white sharp shape. In Figure.B, there are 6 different renal fibrosis-related ligand-receptor interactions with matched pathways. These 6 different pathological stories were identified in this study, and proved by other prior studies.

I performed cell differentiation studies with scVelo, which uses 3 different models to infer pseudo-time: deterministic, stochastic, and dynamic. These models make different assumptions about the splicing rate/gene [36]. I used all 3 models to fit the data sets and reviewed the pseudo-time outputs to select the model that best fit the biological knowledge. Typically, the deterministic model was optimal, with the stochastic model occasionally employed. Positively scaled data sets generally gave more biologically meaningful outputs than the normalized data merged without integration. Using the selected model, scVelo provided the top 100 differentiation-related genes that were transcriptionally upregulated in each cell type [36]. I curated the top 20 genes found in the specific cell types that represent biologically meaningful lineages for each data set (*Gli1*, *Ng2*, *Myh11*, and *Pdgfrb*). There were 220 genes (117 unique genes) in the 11 selected cell types across the 4 data sets, 40 of which had already been validated as driver genes or upregulated genes in kidney, liver, or heart fibrosis. This suggests that the genes with lack of prior studies, but identified in this work could be novel genes involved in renal fibrosis, which will be validated by Dr. Kuppe.

I also conducted drug repositioning studies using data processed by Tajti et al. from 9 CKD microarray data sets [89]. I matched the CKD expression profiles to the LINCS-L1000 drug-treated data sets using the L1000CDS2 tool [84]. Cosine similarities between each gene and disease-causing direction were used as input for L1000CDS2 instead of gene expression values, as the direction of gene expression is assumed to be more important than the magnitude of the values. The disease-causing direction is represented as a single line in gene vector space 90° to the hyperplane separating the disease and healthy data sets. I found 20 common drug candidates with reversely matching profiles to at least 3 different CKDs. Four of these had already been experimentally validated, and nilotinib is FDA-approved.

Several challenges arose during these single-cell and bulk-level renal fibrosis studies. First, the lack of original raw files from the normal mouse endothelial cell data (Cd31 sham data) limited the cell differentiation studies. Second, single-cell level drug repurposing could not be performed due to the time requirements for data production and unexpected delays caused by COVID-19.

Third, there were combined technical and lineage effects when generating data sets. Each reporter mouse was generated in a different batch. So, it was difficult to identify the right biological stories from each reporter mouse after integrating data sets. Fourth, Cd31 sham and UUO mice data had different levels of cell-type annotations because the Cd31 sham mice came from a public database with its annotation.

Bibliography

[1] “Encyclopedia of Cell Biology.” 2016. doi: 10.1016/c2011-1-06109-x.

[2] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[3] N. Azad, A. K. V. Iyer, and Y. Rojanasakul, “DNA Microarrays in Drug Discovery and Development,” *Biopharmaceutical Drug Design and Development*. pp. 47–66, 2008. doi: 10.1007/978-1-59745-532-9_4.

[4] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, “Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells,” *PLoS One*, vol. 9, no. 1, p. e78644, Jan. 2014.

[5] A. M. Newman *et al.*, “Robust enumeration of cell subsets from tissue expression profiles,” *Nat. Methods*, vol. 12, no. 5, pp. 453–457, May 2015.

[6] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat. Methods*, vol. 6, no. 5, pp. 377–382, May 2009.

[7] A. Crinier *et al.*, “High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice,” *Immunity*, vol. 49, no. 5, pp. 971–986.e5, Nov. 2018.

[8] A. Noé, T. N. Cargill, C. M. Nielsen, A. J. C. Russell, and E. Barnes, “The Application of Single-Cell RNA Sequencing in Vaccinology,” *J Immunol Res*, vol. 2020, p. 8624963, Aug. 2020.

[9] A. Nguyen, W. H. Khoo, I. Moran, P. I. Croucher, and T. G. Phan, “Single Cell RNA Sequencing of Rare Immune Cell Populations,” *Front. Immunol.*, vol. 9, p. 1553, Jul. 2018.

[10] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications,” *Genome Med.*, vol. 9, no. 1, p. 75, Aug. 2017.

[11] V. Svensson *et al.*, “Power analysis of single-cell RNA-sequencing experiments,” *Nat. Methods*, vol. 14, no. 4, pp. 381–387, Apr. 2017.

[12] X. Zhang *et al.*, “Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems,” *Mol. Cell*, vol. 73, no. 1, pp. 130–142.e5, Jan. 2019.

[13] J. A. Briggs *et al.*, “The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution,” *Science*, vol. 360, no. 6392, Jun. 2018, doi: 10.1126/science.aar5780.

[14] A. M. Klein *et al.*, “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells,” *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015.

[15] D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, and A. M. Klein, “Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo,” *Science*, vol. 360, no. 6392, pp. 981–987, Jun. 2018.

[16] R. Zilionis *et al.*, “Single-cell barcoding and sequencing using droplet microfluidics,” *Nat. Protoc.*, vol. 12, no. 1, pp. 44–73, Jan. 2017.

[17] E. Z. Macosko *et al.*, “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015.

[18] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells,” *Nat. Commun.*, vol. 8, p. 14049, Jan. 2017.

[19] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nat. Biotechnol.*, vol. 36, no. 5, pp. 411–420, Jun. 2018.

[20] N. Ye, “K-Means Clustering and Density-Based Clustering,” *Data Mining*. pp. 153–166, 2013. doi: 10.1201/b15288-9.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10. p. P10008, 2008. doi: 10.1088/1742-5468/2008/10/p10008.
[22] I. Korsunsky *et al.*, “Fast, sensitive and accurate integration of single-cell data with Harmony,” *Nat. Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019.
[23] B. Simhachalam and G. Ganesan, “Performance comparison of fuzzy and non-fuzzy classification methods,” *Egyptian Informatics Journal*, vol. 17, no. 2. pp. 183–188, 2016. doi: 10.1016/j.eij.2015.10.004.
[24] M. M. Ibrahim and R. Kramann, “genesorteR: Feature Ranking in Clustered Single Cell Data.” doi: 10.1101/676379.
[25] A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.
[26] E. Y. Chen *et al.*, “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,” *BMC Bioinformatics*, vol. 14, p. 128, Apr. 2013.
[27] M. Ashburner *et al.*, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.
[28] S. Aibar *et al.*, “SCENIC: single-cell regulatory network inference and clustering,” *Nat. Methods*, vol. 14, no. 11, pp. 1083–1086, Nov. 2017.
[29] I. Tirosh *et al.*, “Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq,” *Science*, vol. 352, no. 6282, pp. 189–196, Apr. 2016.
[30] C. H. Holland *et al.*, “Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data,” *Genome Biol.*, vol. 21, no. 1, p. 36, Feb. 2020.
[31] M. Ung, D. Mattox, G. Wang, and C. Cheng, “Abstract 3497: Network-based systematic inference of lncRNA activity in breast cancer,” *Molecular and Cellular Biology, Genetics*. 2017. doi: 10.1158/1538-7445.am2017-3497.
[32] L. Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez, “Benchmark and integration of resources for the estimation of human transcription factor activities,” *Genome Res.*, vol. 29, no. 8, pp. 1363–1375, Aug. 2019.
[33] L. Haghverdi, F. Buettner, and F. J. Theis, “Diffusion maps for high-dimensional single-cell analysis of differentiation data,” *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, Sep. 2015.
[34] K. Street *et al.*, “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics,” *BMC Genomics*, vol. 19, no. 1, p. 477, Jun. 2018.
[35] G. La Manno *et al.*, “RNA velocity of single cells,” *Nature*, vol. 560, no. 7719, pp. 494–498, Aug. 2018.
[36] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, “Generalizing RNA velocity to transient cell states through dynamical modeling,” *Nat. Biotechnol.*, vol. 38, no. 12, pp. 1408–1414, Dec. 2020.
[37] F. A. Wolf *et al.*, “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells,” *Genome Biol.*, vol. 20, no. 1, p. 59, Mar. 2019.
[38] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, “A comparison of single-cell trajectory inference methods,” *Nat. Biotechnol.*, vol. 37, no. 5, pp. 547–554, May 2019.
[39] J. A. Ramilowski *et al.*, “A draft network of ligand–receptor-mediated multicellular signalling in human,” *Nature Communications*, vol. 6, no. 1. 2015. doi: 10.1038/ncomms8866.
[40] F. Noël *et al.*, “ICELLNET: a transcriptome-based framework to dissect intercellular communication.” doi: 10.1101/2020.03.05.976878.
[41] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, “OmniPath: guidelines and gateway for literature-curated signaling pathway resources,” *Nat. Methods*, vol. 13, no. 12, pp. 966–967, Nov. 2016.
[42] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo, “CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes.” doi: 10.1101/680926.
[43] S. Jin *et al.*, “Inference and analysis of cell-cell communication using CellChat,” *Nat. Commun.*, vol. 12, no. 1, p. 1088, Feb. 2021.

[44] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.
[45] N. R. Hill *et al.*, “Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis,” *PLOS ONE*, vol. 11, no. 7. p. e0158765, 2016. doi: 10.1371/journal.pone.0158765.
[46] M. Neovius, S. H. Jacobson, J. K. Eriksson, C.-G. Elinder, and B. Hylander, “Mortality in chronic kidney disease and renal replacement therapy: a population-based cohort study,” *BMJ Open*, vol. 4, no. 2. p. e004251, 2014. doi: 10.1136/bmjopen-2013-004251.
[47] M. Edeling, G. Ragi, S. Huang, H. Pavenstädt, and K. Susztak, “Developmental signalling pathways in renal fibrosis: the roles of Notch, Wnt and Hedgehog,” *Nat. Rev. Nephrol.*, vol. 12, no. 7, pp. 426–439, Jul. 2016.
[48] X.-M. Meng, “Inflammatory Mediators and Renal Fibrosis,” *Advances in Experimental Medicine and Biology*. pp. 381–406, 2019. doi: 10.1007/978-981-13-8871-2_18.
[49] X.-M. Meng, D. J. Nikolic-Paterson, and H. Y. Lan, “Inflammatory processes in renal fibrosis,” *Nature Reviews Nephrology*, vol. 10, no. 9. pp. 493–503, 2014. doi: 10.1038/nrneph.2014.114.
[50] R. D. Bülow and P. Boor, “Extracellular Matrix in Kidney Fibrosis: More Than Just a Scaffold,” *J. Histochem. Cytochem.*, vol. 67, no. 9, pp. 643–661, Sep. 2019.
[51] “Tumor Vascularization.” 2020. doi: 10.1016/c2019-0-00115-x.
[52] I. Shaw, S. Rider, J. Mullins, J. Hughes, and B. Péault, “Pericytes in the renal vasculature: roles in health and disease,” *Nat. Rev. Nephrol.*, vol. 14, no. 8, pp. 521–534, Aug. 2018.
[53] L. S. Brown, C. G. Foster, J.-M. Courtney, N. E. King, D. W. Howells, and B. A. Sutherland, “Pericytes and Neurovascular Function in the Healthy and Diseased Brain,” *Front. Cell. Neurosci.*, vol. 13, p. 282, Jun. 2019.
[54] T. L. Pannabecker, “Renal vascular pericytes: long overlooked and poorly understood, but clearly important, and what about those regulatory pathways?,” *Am. J. Physiol. Renal Physiol.*, vol. 314, no. 1, pp. F67–F69, Jan. 2018.
[55] J. Yang *et al.*, “Putative endothelial progenitor cells do not promote vascular repair but attenuate pericyte–myofibroblast transition in UUO-induced renal fibrosis,” *Stem Cell Research & Therapy*, vol. 10, no. 1. 2019. doi: 10.1186/s13287-019-1201-5.
[56] L. M. Black, J. M. Lever, and A. Agarwal, “Renal Inflammation and Fibrosis: A Double-edged Sword,” *J. Histochem. Cytochem.*, vol. 67, no. 9, pp. 663–681, Sep. 2019.
[57] M. Hammouda, A. Ford, Y. Liu, and J. Zhang, “The JNK Signaling Pathway in Inflammatory Skin Disorders and Cancer,” *Cells*, vol. 9, no. 4. p. 857, 2020. doi: 10.3390/cells9040857.
[58] A. Sureshbabu, S. A. Muhsin, and M. E. Choi, “TGF-β signaling in the kidney: profibrotic and protective effects,” *Am. J. Physiol. Renal Physiol.*, vol. 310, no. 7, pp. F596–F606, Apr. 2016.
[59] S. P. Higgins *et al.*, “TGF-β1/p53 signaling in renal fibrogenesis,” *Cell. Signal.*, vol. 43, pp. 1–10, Mar. 2018.
[60] Y. Wang, C. J. Zhou, and Y. Liu, “Wnt Signaling in Kidney Development and Disease,” *Progress in Molecular Biology and Translational Science*. pp. 181–207, 2018. doi: 10.1016/bs.pmbts.2017.11.019.
[61] B. T. MacDonald, K. Tamai, and X. He, “Wnt/beta-catenin signaling: components, mechanisms, and diseases,” *Dev. Cell*, vol. 17, no. 1, pp. 9–26, Jul. 2009.
[62] W. He, C. Dai, Y. Li, G. Zeng, S. P. Monga, and Y. Liu, “Wnt/beta-catenin signaling promotes renal interstitial fibrosis,” *J. Am. Soc. Nephrol.*, vol. 20, no. 4, pp. 765–776, Apr. 2009.
[63] J. Pace, P. Paladugu, B. Das, J. C. He, and S. K. Mallipattu, “Targeting STAT3 signaling in kidney disease,” *Am. J. Physiol. Renal Physiol.*, vol. 316, no. 6, pp. F1151–F1161, Jun. 2019.
[64] S. Tanaka, T. Tanaka, and M. Nangaku, “Hypoxia and Dysregulated Angiogenesis in Kidney Disease,” *Kidney Dis (Basel)*, vol. 1, no. 1, pp. 80–89, May 2015.
[65] C. Braicu *et al.*, “A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer,” *Cancers* , vol. 11, no. 10, Oct. 2019, doi: 10.3390/cancers11101618.
[66] E. F. Wagner and A. R. Nebreda, “Signal integration by JNK and p38 MAPK pathways in cancer development,” *Nat. Rev. Cancer*, vol. 9, no. 8, pp. 537–549, Aug. 2009.
[67] J. Lee, J. N. An, J. H. Hwang, H. Lee, J. P. Lee, and S. G. Kim, “p38 MAPK activity is associated with the histological degree of interstitial fibrosis in IgA nephropathy patients,”

PLoS One, vol. 14, no. 3, p. e0213981, Mar. 2019.
[68] J. Park *et al.*, “Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease,” *Science*, vol. 360, no. 6390, pp. 758–763, May 2018.
[69] R. Hoess, K. Abremski, S. Irwin, M. Kendall, and A. Mack, “DNA specificity of the Cre recombinase resides in the 25 kDa carboxyl domain of the protein,” *J. Mol. Biol.*, vol. 216, no. 4, pp. 873–882, Dec. 1990.
[70] “Gene Knockout Protocols,” *Methods in Molecular Biology*. 2009. doi: 10.1007/978-1-59745-471-1.
[71] C. Kuppe *et al.*, “Decoding myofibroblast origins in human kidney fibrosis,” *Nature*, Nov. 2020, doi: 10.1038/s41586-020-2941-1.
[72] R. Kramann *et al.*, “Perivascular Gli1+ progenitors are key contributors to injury-induced organ fibrosis,” *Cell Stem Cell*, vol. 16, no. 1, pp. 51–66, Jan. 2015.
[73] S. J. Dumas *et al.*, “Single-Cell RNA Sequencing Reveals Renal Endothelium Heterogeneity and Metabolic Adaptation to Water Deprivation,” *J. Am. Soc. Nephrol.*, vol. 31, no. 1, pp. 118–138, Jan. 2020.
[74] C. Simoes-Pires *et al.*, “Reverse pharmacology for developing an anti-malarial phytomedicine. The example of Argemone mexicana,” *Int. J. Parasitol. Drugs Drug Resist.*, vol. 4, no. 3, pp. 338–346, Dec. 2014.
[75] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, “Principles of early drug discovery,” *British Journal of Pharmacology*, vol. 162, no. 6. pp. 1239–1249, 2011. doi: 10.1111/j.1476-5381.2010.01127.x.
[76] T. N. Jarada, J. G. Rokne, and R. Alhajj, “A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions,” *Journal of Cheminformatics*, vol. 12, no. 1. 2020. doi: 10.1186/s13321-020-00450-7.
[77] E. P. Sampaio, E. N. Sarno, R. Galilly, Z. A. Cohn, and G. Kaplan, “Thalidomide selectively inhibits tumor necrosis factor alpha production by stimulated human monocytes,” *J. Exp. Med.*, vol. 173, no. 3, pp. 699–703, Mar. 1991.
[78] T. T. Ashburn and K. B. Thor, “Drug repositioning: identifying and developing new uses for existing drugs,” *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, Aug. 2004.
[79] J. D. Kranzler and R. M. Gendreau, “Role and rationale for the use of milnacipran in the management of fibromyalgia,” *Neuropsychiatr. Dis. Treat.*, vol. 6, pp. 197–208, May 2010.
[80] J. Lamb *et al.*, “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease,” *Science*, vol. 313, no. 5795, pp. 1929–1935, Sep. 2006.
[81] J. M. Ramsey *et al.*, “Entinostat Prevents Leukemia Maintenance in a Collaborating Oncogene-Dependent Model of Cytogenetically Normal Acute Myeloid Leukemia,” *STEM CELLS*, vol. 31, no. 7. pp. 1434–1445, 2013. doi: 10.1002/stem.1398.
[82] A. Musa *et al.*, “A review of connectivity map and computational approaches in pharmacogenomics,” *Brief. Bioinform.*, vol. 18, no. 5, p. 903, Sep. 2017.
[83] D. Vidović, A. Koleti, and S. C. Schürer, “Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action,” *Front. Genet.*, vol. 5, p. 342, Sep. 2014.
[84] Q. Duan *et al.*, “L1000CDS: LINCS L1000 characteristic direction signatures search engine,” *NPJ Syst Biol Appl*, vol. 2, Aug. 2016, doi: 10.1038/npjsba.2016.15.
[85] N. R. Clark *et al.*, “The characteristic direction: a geometrical approach to identify differentially expressed genes,” *BMC Bioinformatics*, vol. 15, p. 79, Mar. 2014.
[86] Z. Zhang, P. Wu, J. Zhang, S. Wang, and G. Zhang, “The effect of statins on microalbuminuria, proteinuria, progression of kidney function, and all-cause mortality in patients with non-end stage chronic kidney disease: A meta-analysis,” *Pharmacol. Res.*, vol. 105, pp. 74–83, Mar. 2016.
[87] X. Xie *et al.*, “Renin-Angiotensin System Inhibitors and Kidney and Cardiovascular Outcomes in Patients With CKD: A Bayesian Network Meta-analysis of Randomized Clinical Trials,” *Am. J. Kidney Dis.*, vol. 67, no. 5, pp. 728–741, May 2016.
[88] T. Papadopoulos *et al.*, “Omics databases on kidney disease: where they can be found and how to benefit from them,” *Clin. Kidney J.*, vol. 9, no. 3, pp. 343–352, Jun. 2016.

[89] F. Tajti *et al.*, “A Functional Landscape of CKD Entities From Public Transcriptomic Data,” *Kidney Int Rep*, vol. 5, no. 2, pp. 211–224, Feb. 2020.
[90] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.
[91] A. T. L. Lun *et al.*, “EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data,” *Genome Biol.*, vol. 20, no. 1, p. 63, Mar. 2019.
[92] L. Chen, J. Z. Clark, J. W. Nelson, B. Kaissling, D. H. Ellison, and M. A. Knepper, “Renal-Tubule Epithelial Cell Nomenclature for Single-Cell RNA-Sequencing Studies,” *J. Am. Soc. Nephrol.*, vol. 30, no. 8, pp. 1358–1364, Aug. 2019.
[93] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7. pp. 1888–1902.e21, 2019. doi: 10.1016/j.cell.2019.05.031.
[94] D. Szklarczyk *et al.*, “STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.
[95] F. Ehsan Elahi and A. Hasan, “A method for estimating Hill function-based dynamic models of gene regulatory networks,” *R Soc Open Sci*, vol. 5, no. 2, p. 171226, Feb. 2018.
[96] G. J. Mahler, E. J. Farrar, and J. T. Butcher, “Inflammatory cytokines promote mesenchymal transformation in embryonic and adult valve endothelial cells,” *Arterioscler. Thromb. Vasc. Biol.*, vol. 33, no. 1, pp. 121–130, Jan. 2013.
[97] S.-J. Chen, L.-L. Lv, B.-C. Liu, and R.-N. Tang, “Crosstalk between tubular epithelial cells and glomerular endothelial cells in diabetic kidney disease,” *Cell Prolif.*, vol. 53, no. 3, p. e12763, Mar. 2020.
[98] N. Sacilotto *et al.*, “MEF2 transcription factors are key regulators of sprouting angiogenesis,” *Genes Dev.*, vol. 30, no. 20, pp. 2297–2309, Oct. 2016.
[99] F. Noël *et al.*, “Dissection of intercellular communication using the transcriptome-based framework ICELLNET,” *Nat. Commun.*, vol. 12, no. 1, p. 1089, Feb. 2021.
[100] J. Li *et al.*, “Blocking Macrophage Migration Inhibitory Factor Protects Against Cisplatin-Induced Acute Kidney Injury in Mice,” *Mol. Ther.*, vol. 26, no. 10, pp. 2523–2532, Oct. 2018.
[101] P. Boor, T. Ostendorf, and J. Floege, “PDGF and the progression of renal disease,” *Nephrology Dialysis Transplantation*, vol. 29, no. suppl 1. pp. i45–i54, 2014. doi: 10.1093/ndt/gft273.
[102] W. Sato and Y. Sato, “Midkine in nephrogenesis, hypertension and kidney diseases,” *Br. J. Pharmacol.*, vol. 171, no. 4, pp. 879–887, Feb. 2014.
[103] R. J. Tan, D. Zhou, L. Zhou, and Y. Liu, “Wnt/β-catenin signaling and kidney fibrosis,” *Kidney Int. Suppl.*, vol. 4, no. 1, pp. 84–90, Nov. 2014.
[104] M. Liang *et al.*, “Protective Role of Insulin-Like Growth Factor-1 Receptor in Endothelial Cells against Unilateral Ureteral Obstruction–Induced Renal Fibrosis,” *The American Journal of Pathology*, vol. 185, no. 5. pp. 1234–1250, 2015. doi: 10.1016/j.ajpath.2015.01.027.
[105] E. M. Buhl *et al.*, “Dysregulated mesenchymal PDGFR-β drives kidney fibrosis,” *EMBO Molecular Medicine*, vol. 12, no. 3. 2020. doi: 10.15252/emmm.201911021.
[106] E. M. Buhl *et al.*, “The role of PDGF-D in healthy and fibrotic kidneys,” *Kidney Int.*, vol. 89, no. 4, pp. 848–861, Apr. 2016.
[107] T. Ostendorf, F. Eitner, and J. Floege, “The PDGF family in renal fibrosis,” *Pediatric Nephrology*, vol. 27, no. 7. pp. 1041–1050, 2012. doi: 10.1007/s00467-011-1892-z.
[108] L. Valiño-Rivas, C. Baeza-Bermejillo, L. Gonzalez-Lafuente, A. B. Sanz, A. Ortiz, and M. D. Sanchez-Niño, “CD74 in Kidney Disease,” *Frontiers in Immunology*, vol. 6. 2015. doi: 10.3389/fimmu.2015.00483.
[109] J. N. An *et al.*, “Periostin induces kidney fibrosis after acute kidney injury via the p38 MAPK pathway,” *Am. J. Physiol. Renal Physiol.*, vol. 316, no. 3, pp. F426–F437, Mar. 2019.
[110] R. del Toro *et al.*, “Identification and functional analysis of endothelial tip cell–enriched genes,” *Blood*, vol. 116, no. 19. pp. 4025–4033, 2010. doi: 10.1182/blood-2010-02-270819.
[111] C. S. M. Helker *et al.*, “Apelin signaling drives vascular endothelial cells toward a pro-angiogenic state,” *eLife*, vol. 9. 2020. doi: 10.7554/elife.55589.
[112] V. Iragavarapu-Charyulu, E. Wojcikiewicz, and A. Urdaneta, “Semaphorins in Angiogenesis

and Autoimmune Diseases: Therapeutic Targets?," *Front. Immunol.*, vol. 11, p. 346, Mar. 2020.
[113] L. O. Lerman and A. R. Chade, "Angiogenesis in the kidney: a new therapeutic target?," *Curr. Opin. Nephrol. Hypertens.*, vol. 18, no. 2, pp. 160–165, Mar. 2009.
[114] M. E. J. Reinders, T. J. Rabelink, and D. M. Briscoe, "Angiogenesis and endothelial cell repair in renal disease and allograft rejection," *J. Am. Soc. Nephrol.*, vol. 17, no. 4, pp. 932–942, Apr. 2006.
[115] A. Goodarzi, S. Mozafarpoor, M. Dodangeh, F. Seirafianpour, and M. H. Shahverdi, "The role of topical timolol in wound healing and the treatment of vascular lesions: A narrative review," *Dermatol. Ther.*, p. e14847, Feb. 2021.
[116] R. T. Chitturi, A. M. Balasubramaniam, R. A. Parameswar, G. Kesavan, K. T. M. Haris, and K. Mohideen, "The role of myofibroblasts in wound healing, contraction and its clinical implications in cleft palate repair," *J Int Oral Health*, vol. 7, no. 3, pp. 75–80, Mar. 2015.
[117] Q. Yuan, R. J. Tan, and Y. Liu, "Myofibroblast in Kidney Fibrosis: Origin, Activation, and Regulation," *Adv. Exp. Med. Biol.*, vol. 1165, pp. 253–283, 2019.
[118] B. D. Humphreys, "Targeting pericyte differentiation as a strategy to modulate kidney fibrosis in diabetic nephropathy," *Semin. Nephrol.*, vol. 32, no. 5, pp. 463–470, Sep. 2012.
[119] T. Zhao *et al.*, "Vascular endothelial growth factor-D mediates fibrogenic response in myofibroblasts," *Mol. Cell. Biochem.*, vol. 413, no. 1–2, pp. 127–135, Feb. 2016.
[120] C. Gerarduzzi *et al.*, "Silencing SMOC2 ameliorates kidney fibrosis by inhibiting fibroblast to myofibroblast transformation," *JCI Insight*, vol. 2, no. 8, Apr. 2017, doi: 10.1172/jci.insight.90299.
[121] A. M. Scruggs, G. Grabauskas, and S. K. Huang, "The Role of KCNMB1 and BK Channels in Myofibroblast Differentiation and Pulmonary Fibrosis," *Am. J. Respir. Cell Mol. Biol.*, vol. 62, no. 2, pp. 191–203, Feb. 2020.
[122] X.-P. Zhao *et al.*, "Hedgehog Interacting Protein Promotes Fibrosis and Apoptosis in Glomerular Endothelial Cells in Murine Diabetes," *Sci. Rep.*, vol. 8, no. 1, p. 5958, Apr. 2018.
[123] K. Neubauer, B. Neubauer, M. Seidl, and B. Zieger, "Characterization of septin expression in normal and fibrotic kidneys," *Cytoskeleton* , vol. 76, no. 1, pp. 143–153, Jan. 2019.
[124] V. Ramdas, M. McBride, L. Denby, and A. H. Baker, "Canonical transforming growth factor-β signaling regulates disintegrin metalloprotease expression in experimental renal fibrosis via miR-29," *Am. J. Pathol.*, vol. 183, no. 6, pp. 1885–1896, Dec. 2013.
[125] I. Grgic *et al.*, "Translational profiles of medullary myofibroblasts during kidney fibrosis," *J. Am. Soc. Nephrol.*, vol. 25, no. 9, pp. 1979–1990, Sep. 2014.
[126] H. Huang *et al.*, "The MicroRNA MiR-29c Alleviates Renal Fibrosis via TPM1-Mediated Suppression of the Wnt/β-Catenin Pathway," *Front. Physiol.*, vol. 11, p. 331, Apr. 2020.
[127] H.-M. Sun *et al.*, "PALLD Regulates Phagocytosis by Enabling Timely Actin Polymerization and Depolymerization," *J. Immunol.*, vol. 199, no. 5, pp. 1817–1826, Sep. 2017.
[128] J. A. Guay, D. M. Wojchowski, J. Fang, and L. Oxburgh, "Death associated protein kinase 2 is expressed in cortical interstitial cells of the mouse kidney," *BMC Res. Notes*, vol. 7, p. 345, Jun. 2014.
[129] Y. Duan, Y. Qiu, X. Huang, C. Dai, J. Yang, and W. He, "Deletion of FHL2 in fibroblasts attenuates fibroblasts activation and kidney fibrosis via restraining TGF-β1-induced Wnt/β-catenin signaling," *J. Mol. Med.* , vol. 98, no. 2, pp. 291–307, Feb. 2020.
[130] Y.-T. Chen *et al.*, "Endoplasmic reticulum protein TXNDC5 promotes renal fibrosis by enforcing TGF-β signaling in kidney fibroblasts," *J. Clin. Invest.*, vol. 131, no. 5, Mar. 2021, doi: 10.1172/JCI143645.
[131] R. Jie, P. Zhu, J. Zhong, Y. Zhang, and H. Wu, "LncRNA KCNQ1OT1 affects cell proliferation, apoptosis and fibrosis through regulating miR-18b-5p/SORBS2 axis and NF-κB pathway in diabetic nephropathy," *Diabetol. Metab. Syndr.*, vol. 12, p. 77, Sep. 2020.
[132] J.-H. Baek *et al.*, "IL-34 mediates acute kidney injury and worsens subsequent chronic kidney disease," *J. Clin. Invest.*, vol. 125, no. 8, pp. 3198–3214, Aug. 2015.
[133] M. Mukoyama *et al.*, "Role of adrenomedullin and its receptor system in renal pathophysiology," *Peptides*, vol. 22, no. 11, pp. 1925–1931, Nov. 2001.
[134] G. S. Di Marco *et al.*, "Soluble Flt-1 links microvascular disease with heart failure in CKD," *Basic Res. Cardiol.*, vol. 110, no. 3, p. 30, May 2015.

[135] D. Holmes, “Decorin has role in differentiation,” *Nature Reviews Nephrology*, vol. 10, no. 2. pp. 65–65, 2014. doi: 10.1038/nrneph.2013.271.
[136] K. C. Hall *et al.*, “sGC stimulator praliciguat suppresses stellate cell fibrotic transformation and inhibits fibrosis and inflammation in models of NASH,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 22, pp. 11057–11062, May 2019.
[137] H. Kajimoto *et al.*, “BMP type I receptor inhibition attenuates endothelial dysfunction in mice with chronic kidney disease,” *Kidney Int.*, vol. 87, no. 1, pp. 128–136, Jan. 2015.
[138] S. H. Kim, Y. W. Jang, P. Hwang, H. J. Kim, G. Y. Han, and C. W. Kim, “The reno-protective effect of a phosphoinositide 3-kinase inhibitor wortmannin on streptozotocin-induced proteinuric renal disease rats,” *Exp. Mol. Med.*, vol. 44, no. 1, pp. 45–51, Jan. 2012.
[139] S. Fuchs *et al.*, “Haemanthus coccineus extract and its main bioactive component narciclasine display profound anti-inflammatory activities in vitro and in vivo,” *J. Cell. Mol. Med.*, vol. 19, no. 5, pp. 1021–1032, May 2015.
[140] M. Iyoda, T. Shibata, Y. Hirai, Y. Kuno, and T. Akizawa, “Nilotinib attenuates renal injury and prolongs survival in chronic kidney disease,” *J. Am. Soc. Nephrol.*, vol. 22, no. 8, pp. 1486–1496, Aug. 2011.

Acknowledgements

I would like to thank Prof.Dr. Julio Saez-Rodriguez and Prof.Dr. Rafael Kramann for providing me the opportunity to work on this project with their supervision.

I would like to thank Prof.Dr. Benedikt Brors for giving me meaningful comments at every annual meeting.

I would like to thank Dr. Christoph Kuppe who suggested the single-cell RNA-seq project.

I would like to thank my colleagues, Dr. Javier Perales-Patón, Ricardo O. Ramirez-Flores for providing support and assistance in interpreting the statistical and technical nuances of single-cell RNA-seq.

I would like to thank Oliver group members, especially Dr. Hana Susak, Dr. R. Gonzalo Parra and Jan Gleixner who taught me how to use CellRanger, to start pseudotime analysis and to analyze demultiplexing hashtag data, separately.

I would like to thank other Ph.D. candidates from my year, Nicolas Palacio-Escat, Aurélien Dugourd, Christian Holland, Olga Ivanova for useful discussions regarding the coursework and for the fun times outside of work.

I would like to thank all of Julio group members and alumni, especially, Dr. Vigneshwari Subramanian, Dr. Panuwat Trairatphisan and Dr. Enio Gjerga for their support during trying times throughout my Ph.D.

I would like to thank both Erika Schulz and Sabine Blum for their help with administrative tasks, which kept things running smoothly.

I would like to thank Dr. Rebecca Terrall Levinson and Dr. Steven Pauff for their help with proofreading my resume and the part of this thesis.

I would like to thank Konrad Hoeft for translating my English abstract to German.

I would like to dedicate my thesis to my family in South Korea, who provided a nurturing environment and support to fulfill all my goals. I have always counted myself lucky to know such incredible people as my family and oldest friends. I would not be the person I am without them.