Towards Building a Smart Kidney Atlas: Network-based integration of multimodal transcriptomic, proteomic, metabolomic and imaging data in the Kidney Precision Medicine Project

Jens Hansen[1,*], Rachel Sealfon[2,*], Rajasree Menon[3,*], Michael T. Eadon[4], Blue B. Lake[5], Becky Steck[3], Dejan Dobi[6], Samir Parikh[7], Tara K. Sidgel[6], Theodore Alexandrov[8], Andrew Schroeder[6], Edgar A. Otto[3], Christopher R. Anderton[9,10], Daria Barwinska[4], Guanshi Zheng[10], Michael P. Rose[3], John P. Shapiro[7], Dusan Velickovic[9], Annapurna Pamreddy[10], Seth Winfree[4], Yongqun He[3], Ian H. de Boer[11], Jeffrey B. Hodgin[3], Abhijit Nair[3], Kumar Sharma[10], Minnie Sarwal[6], Kun Zhang[5], Jonathan Himmelfarb[11], Zoltan Laszik[6], Brad Rovin[7], Pierre C. Dagher[4], John Cijiang He[1], Tarek M. El-Achkar[4], Sanjay Jain[12], Olga G. Troyanskaya[2,#], Matthias Kretzler[3,#], Ravi Iyengar[1,#], Evren U. Azeloglu[1,#] **for the Kidney Precision Medicine Project Consortium**

* Contributed equally, joint first authors

Affiliations:

1. Icahn School of Medicine at Mount Sinai, New York, New York
2. Princeton University, Princeton, New Jersey and Flatiron Institute, New York, New York
3. University of Michigan School of Medicine, Ann Arbor, Michigan
4. Indiana University School of Medicine, Indianapolis, Indiana
5. University of California San Diego, Jacobs School of Engineering, San Diego, California
6. University of California San Francisco School of Medicine, San Francisco, California
7. Ohio State University College of Medicine, Columbus, Ohio
8. European Molecular Biology Laboratory, Heidelberg, Germany
9. Pacific Northwest National Laboratory, Richland, Washington
10. UT-Health San Antonio School of Medicine, San Antonio, Texas
11. University of Washington, Schools of Medicine and Public Health, Seattle, Washington
12. Washington University in Saint Louis School of Medicine, St. Louis, Missouri

[#] Corresponding Authors, joint senior authors:

Evren U. Azeloglu, Ph.D.
Assistant Professor of Medicine, Nephrology
Icahn School of Medicine at Mount Sinai, New York, NY
Email: evren.azeloglu@mssm.edu
Twitter: @azeloglu

Ravi Iyengar, Ph.D.
Dorothy H and Lewis H Rosenstiel Professor of Pharmacological Sciences
Icahn School of Medicine at Mount Sinai, New York, NY
Email: ravi.iyengar@mssm.edu

Matthias Kretzler, M.D.
Professor of Medicine, Nephrology
University of Michigan School of Medicine, Ann Arbor, MI
Email: kretzler@med.umich.edu

Olga Troyanskaya, Ph.D.
Professor of Computer Science
Princeton University, Princeton, NJ
Email: ogt@genomics.princeton.edu

ABSTRACT

The Kidney Precision Medicine Project (KPMP) plans to construct a spatially specified tissue atlas of the human kidney at a cellular resolution with near comprehensive molecular details. The atlas will have maps of healthy, acute kidney injury and chronic kidney disease tissues. To construct such maps, we integrate different data sets that profile mRNAs, proteins and metabolites collected by five KPMP Tissue Interrogation Sites. Here, we describe a set of hierarchical analytical methods to process, combine, and harmonize single-cell, single-nucleus and subsegmental laser microdissection (LMD) transcriptomics with LMD and near single-cell proteomics, 3-D nondestructive and immunofluorescence-based Codex imaging and spatial metabolomics datasets. We use nephrectomy, healthy living donor and surveillance transplant biopsy tissues to create a harmonized reference tissue map. Our results demonstrate that different assays produce reliable and coherent identification of cell types and tissue subsegments. They further show that the molecular profiles and pathways are partially overlapping yet complementary for cell type-specific and subsegmental physiological processes. Focusing on the proximal tubules, we find that our integrated systems biology-based analyses identify different subtypes of tubular cells with potential for different levels of lipid oxidation and energy generation. Integration of our omics data with pathways from the literature, enables us to construct predictive computational models to develop a smart kidney atlas. These integrated models can describe physiological capabilities of the tissues based on the underlying cell types and pathways in health and disease.

INTRODUCTION

The kidney is one of the most diverse organs in the human body in terms of its cellular heterogeneity, and possibly second only to the brain in its spatial complexity. Accordingly, decoding the functional and pathogenic mechanisms of kidney disease has been challenging; as such, nephrology has consistently ranked behind all other subspecialties of medicine in terms of the drug discovery pipeline [1]. Delineating the cell types and subtypes in different regions of the kidney during health and disease will help identify the tissue-level, cellular and subcellular pathways and processes involved in disease initiation and progression, and aid in drug discovery.

The Kidney Precision Medicine Project (KPMP) is a consortium funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that aims to ethically and safely obtain kidney biopsies from participants with chronic kidney disease (CKD) or acute kidney injury (AKI); create a reference kidney atlas; characterize disease subgroups to stratify patients based on molecular features of disease; and identify critical cells, pathways, and targets for novel therapies and preventive strategies. The KPMP features an expanding set of complementary set of high throughput assays for molecular entities that span transcriptomic, proteomic, metabolomic profiles and spatial/structural properties of kidney tissue. These assays, described here for the five initially funded Tissue Interrogation Sites (TISes), will be integrated to create a comprehensive knowledge environment for the human kidney. This knowledge environment will be compiled by the KPMP Central Hub to serve as a foundation for a spatially specified interactive smart tissue atlas that will include molecular and physiological information on healthy and diseased states of all individual cell types within the adult human kidney.

The KPMP envisions that harmonization and integration of different types of molecular data from omics assays, combined with state-of-the-art pathological and clinical descriptors, will allow us to classify different disease subtypes and states for diagnostic and therapeutic purposes. Numerous groups have proposed the use of integrated multiomics analysis to characterize disease phenotypes using tools that include Bayesian, correlative, network-based and machine learning-based clustering algorithms [2-4]. The goals of these approaches include prediction of clinical outcomes, identification of underlying disease mechanisms and stratification of patients [5]. KPMP further envisions that the final integrated analytical environment will serve as a knowledge base for the entire field that will empower a molecular anchored outcome prediction and development of targeted treatments.

Here, we present an overview of KPMP's strategies to harmonize and integrate multiple data types through identification of subcellular pathways and functions that delineate cell-level biochemical and physiological functions. Using reference kidney pilot tissue samples, we have performed data harmonization and integration to investigate the complementarity of different data types and develop a pipeline for the generation of tissue maps.

RESULTS

Outline of KPMP Data Types

In these analyses, there were four transcriptomic, two proteomic, one imaging-based, and one spatial metabolomics tissue interrogation assays that consisted of 3 to 48 different datasets obtained from 3 to 22 participants (Supplementary Table 1). These assays and their detailed tissue pre-analytical, tissue processing, data acquisition and analytical data processing pipelines are outlined in Figure 1. We also summarize the steps whereby the data sets were integrated and harmonized in the upper right side of this descriptive map view of the KPMP data integration paradigm.

Pathway- and network-level integration of multiple molecular interrogation techniques reveals cell- and tissue-specific biological processes that are critical for renal physiology

To overcome the inherent challenges of multiomics integration and assay dependent divergence, we employed dynamic enrichment analysis [6] and network mapping [7]. We evaluated the convergence of subcellular processes (SCPs) and pathways that are over-represented in different cell types or subsegments within the kidney (in comparison to the other cell types or subsegments), using single cell RNASeq data from PREMIERE TIS (Michigan, Princeton, Broad) [8], single nucleus RNASeq data from UCSD/WU TIS [9], Laser microdissected (LMD) bulk RNASeq (Supplementary Table 2) and LMD proteomics (Supplementary Table 3) from the OSU/IU TIS, Near Single Cell (NSC) proteomics from the UCSF TIS (Supplementary Table 4) and spatial metabolomics from the UTHSA-PNNL-EMBL TIS (Supplementary Table 5A/B/C from 3 different participants).

Single-cell [8] and -nucleus [9] RNASeq analysis resulted in the grouping of multiple cells or nuclei into clusters that were assigned to a particular cell type based on the expression of essential genes. The top 300 most significantly differentially expressed genes (DEGs) and proteins (DEPs) of each cluster or subsegment compared to all other clusters or subsegments as well as the metabolites assigned to glomerular and non-glomerular kidney regions (Supplementary Table 6) were subjected to enrichment analysis to create pathway maps (Supplementary Table 7) for the three representative cell types contributing diverse function to kidney physiology: proximal tubular epithelial cells (Figure 2A, Supplementary Figure 1A for nonspecific pathways), podocytes (Supplementary Figure 1B) and principal cells of the collecting ducts (Supplementary Figure 1C). The final maps revealed highly interrelated SCPs that are intimately linked to the physiological function of the respective cell types. Furthermore, these SCPs are highly overlapping between assays and datasets with up to 74% of them being repeatedly enriched in two or more assays, confirming the inherent agreement among these different assays. While the individual significant genes or gene products coming from multiple assays were not necessarily the same, placement of these gene products into an interconnected pathway map showed innate congruence between the assays. The key subcellular processes (SCPs) for the different cell types differed significantly.

Cell-type specific SCP networks predict overlapping and complementary pathways that accurately support each cell type's whole cell function. Proximal tubule networks predict a high metabolic activity and describe ion reabsorption and ion-triggered glucose reabsorption pathways as well as ammonia metabolism and detoxification pathways (Figure 2A). The predictions are in agreement with the energy intensive ion, glucose and other small molecule reabsorption by the proximal tubule cells [10] and their predominant function in ammonium excretion and renal drug clearance [11]. The identification of cellular iron homeostasis pathways documents the iron storage capacity of proximal tubule cells [12] that among other functions, also mitigates kidney damage during acute kidney injury [13]. Podocyte/glomerular networks focus on cell-cell/cell-matrix adhesion, glomerular basement membrane/extracellular matrix (ECM) and actin dynamics (Supplementary Figure 1B), all pathways fundamental for barrier generation and consequently for glomerular filtration. Principal cell/collecting duct networks concentrate on ion reabsorption (Supplementary Figure 1C), emphasizing the important role of the collecting duct in fine-tuning these mechanisms, thereby regulating systemic electrolyte and water balance.

These networks document that 13% (principal cells/collecting duct), 27% (proximal tubule cells/tubulointerstitium) and 74% (podocytes/glomerulus) of all predicted SCPs were

discovered by at least two different technologies. A closer investigation of the SCPs further highlights that the overlap is even higher, if only the SCPs that describe cell type specific functions are considered. Furthermore, the different datasets describe complementary subfunctions of the same physiological processes. For example, both proteomic datasets of the proximal tubule subsegments describe fatty acid transport via carnitine shuttling into the mitochondrial matrix, where the enzymes for mitochondrial beta oxidation are localized (Figure 2A). The PREMIERE SC RNASeq dataset predicts carnitine biosynthesis, i.e. synthesis of the central molecule of the carnitine shuttle.

Integration of pathways that were predicted based on the tubulointerstitial metabolites, such as 'Glycolysis and Gluconeogenesis' and 'D-Arginine and D-ornithine metabolism' (Supplementary Figure 1D), into the Molecular Biology of the Cell Ontology (MBCO) SCP-networks (Figure 1A) further underline the predicted high metabolic activity of the proximal tubule cells. Glomerular metabolites enrich for pathways (Supplementary Figure 1C), such as sphingolipid and arachidonic acid metabolism, that support cell-matrix/cell-cell adhesion and gap junctions, respectively [14]. Dynamic enrichment analysis of both single-cell RNA-seq datasets predicts the involvement of another metabolic pathway, i.e. retinol metabolism, in podocyte function, in particular as a regulator of tight junctions (Supplementary Figure 1B). Retinoic acid has a regulatory effect on tight junctions [15, 16] and plays a significant role in mitigating podocyte apoptosis and dedifferentiation during podocyte injury [17].

The enrichment results suggest that proximal tubular cells have the capacity to meet the high energy demand by not only fueling the citric acid cycle via beta oxidation, but also via glucose and glutamine catabolism. Nevertheless, beta oxidation is most consistently predicted, in agreement with previous studies documenting lipid metabolism as the preferential energy source in proximal tubule cells [18, 19]. Investigation of the pathway components of these SCPs documents that the different omics technologies identify different components of these pathways that integrate into a comprehensive description of the relevant biochemical pathways (Figure 2B). Each technology contributes genes, proteins and metabolites for a fuller description of the pathways than would be obtained by a single technology. Tubulointerstitial metabolites, for example, contain glucose, cofactors of the pyruvate dehydrogenase complex and multiple adenosine nucleotides/nucleosides (i.e. metabolites of the energy carrier ATP). In agreement with the results of the pathway predictions, network mapping [7] revealed that cell-type specific DEGs and DEPs lie within the same area of the human interactome (Supplementary Figure 1E), indicative of close functional relationships.

In parallel, we identified modules in a kidney-specific functional network using the top ranked 300 marker genes and proteins across all datatypes in order to detect sets of cell-type specific, functionally related genes [20, 21]. The module detection algorithm finds groups of genes that form tightly connected communities within a kidney-specific functional network, which is constructed using a data-driven approach from gene-gene relationships across thousands of experimental assays. After module detection, gene enrichment analysis is performed within each module to understand the key functions of the genes in each module. As with dynamic enrichment analysis, the modules display clear cell-type specific functional enrichments (Supplementary Table 8). For example, the network of proximal tubule marker genes includes modules enriched in anion transport and cellular response to metal ions (Figure 2C), the network of podocyte marker genes includes modules enriched in glomerulus development and cell-cell adhesion (Supplementary Figure 1F), and the network of principal cell marker genes includes modules enriched in sodium ion transport (Supplementary Figure 1G)

Integration of diverse transcriptomic interrogation techniques shows strong agreement and technological synergy between assays

While single-cell RNASeq has the ability to identify different cell types and map individual transcriptomic profiles of cells, the technique does not provide spatial context or adequate coverage of transcriptomic depth that laser microdissection (LMD) combined with bulk RNASeq could offer. LMD transcriptomics can identify subsegment-specific differential gene expression (i.e., over- or under-representation of a specific gene within a subsegment as compared to all other subsegments) for nearly the entire human transcriptome (Supplementary Figure 1H). An idealized integration scenario would combine these assays synergistically such that they could complement the shortcomings of each other, improve quality control metrics across technologies, and increase rigor and reproducibility of the overall study. These pilot studies combined four distinct transcriptomic assays: single-cell RNASeq, multiplexed single-cell RNASeq, single-nucleus RNASeq, and LMD RNASeq. Our results show that the combination of approaches provide a fuller description of spatial organization and molecular components within each cell type than any one approach by itself.

Multiscale integration of the three different transcriptomic interrogation techniques shows strong agreement and technological synergy between assays. KPMP Tissue Interrogation Sites incorporate a diverse portfolio of transcriptomic assays that have different advantages and shortcomings. Three TISes, PREMIERE (U. Michigan, Princeton, Broad), UCSF and UCSD/Washington U, performed two separate single-cell RNA-seq assays and one single-nucleus RNASeq assay, respectively. From these dissociated cell/nuclei-based transcriptomic assays, we identified a total of 22,268, 27,757, and 17,659 cells. Our results showed that the 3 disparate transcriptomic assays, regardless of their spatial specificity, can be hierarchically harmonized (Figure 3A). First, we see that different single-cell and single-nucleus transcriptomic assays in KPMP align favorably for most of the cell types, whereby top cluster descriptors, i.e., marker gene candidates for each cell type, agree across different assays. The integrated analysis of the single cell and single nuclei transcriptomic data yielded 16 clusters at a resolution of 0.6 from 45,536 cells/nuclei. Cells from the three TISes were included in all clusters indicating successful integration of the data (Figure 3B). Differential expression analysis of the genes expressed in each cluster versus all other clusters generated cell-type specific gene sets. Each gene set included known kidney cell-type specific markers which enabled assignment of clusters to specific cell types and established the validity of these measurements for generation of a tissue map. The cell types identified included intrinsic kidney cell types as well as immune cell types.

We also performed bulk transcriptomic assays on LMD subsegments, which are shown in Figure 3A, and mapped to the associated single cell clusters. We integrated single-cell sequencing, single-nucleus sequencing, and LMD transcriptomics datasets by assigning each cell in the integrated single-cell datasets to the nearest LMD subsegment, based on the Pearson correlation between the expression profiles for each cell and each subsegment. We find that there is strong concordance across the datasets in case of the low (Figure 3C) and integrated single cell/nucleus clustering, whereby the majority of the cells from each cluster were assigned to the correct corresponding LMD subsegment in an unbiased manner (for example, proximal tubule cells are assigned to the proximal tubule subsegment, while podocytes are assigned to the glomerular subsegment).

Post-hoc power analysis reveals that 16-25 participants are sufficient for consistent classification of cell types

We evaluated how many reference participant samples need to be processed to generate RNASeq libraries that will be subjected to whole genome transcriptomic sequencing to obtain consistent reproducible results. This allows an estimation of how many participants need to be recruited for the identification of reliable disease-related mechanisms. We used the PREMIERE single-cell [8] and the USCD/WU single-nucleus [9] RNASeq datasets that were obtained from 24 and 47 libraries (Supplementary Table 1) (17,532 cells and 12,100 nuclei after quality control), respectively. We separately subjected both RNASeq datasets, with and without random and progressive removal of libraries, to a standardized single-cell sequencing analysis pipeline (Supplementary Figure 2A) followed by comparison of the results obtained for the down-sampled datasets with the complete datasets (Supplementary Figure 2B). Similarly, we analyzed the LMD proteomic dataset obtained from 11 nephrectomy samples and subjected the full or down-sampled LMD datasets to a standardized analysis pipeline, followed by comparison of the results. Our results indicate that for a consistent detection of podocytes (i.e. in more than 95% of all down sampled datasets with the same library counts), at least 16 (~11,727 cells) or 7 libraries (1,837 nuclei) are needed if subjected to single-cell RNASeq (Figure 4A) or single-nucleus RNASeq (Figure 4B), respectively. Consistent detection of proximal tubule and principal cells depended on only 2 and 3 libraries (~1,367 or ~2184 cells), respectively, in the case of single-cell RNASeq or 4 and 12 libraries (~1,005 or ~3,100 nuclei), respectively, in the case of single-nucleus RNASeq. Additionally, our results suggest that the accuracy of single-cell or single-nucleus assignments to the selected cell types is relatively stable (with the exception of podocytes in the single-cell dataset) as documented by the low number of cells that are assigned as different cell types or mapped to an unrelated tissue subsegment.

The Pearson correlations between the fold changes of the identified DEGs of the down-sampled and the complete datasets also depend on the analyzed libraries. For example, a Pearson correlation of 0.8 or higher is achieved with 11, 4 and 9 libraries (~7,961, ~2,866 and ~ 6,708 cells) for podocytes, proximal tubule cells and principal cells, respectively, based on the single-cell dataset; and 7, 29 and 47 libraries (~1,837, ~7,428 and 12,100 nuclei) based on the single-nucleus dataset. The slow increase in the correlation coefficient of DEGs for proximal tubule cells likely results from the continuous increase in identified proximal tubule clusters, each adding new proximal tubule related DEGs. The Pearson correlation between the p-values of predicted GO BPs or level-3 MBCO SCPs already falls below 0.8, if less than 19 to 24 libraries (~14,096 to 17,532 cells) or 16 to 47 libraries (~4,047 to 12,100 nuclei) are used in the case of single-cell and single-nucleus RNASeq, respectively (Supplementary Figure 2C, 3D, respectively). This is unsurprising given the small size of the SCPs and pathways (GO processes contain 35 ± 71 genes, level-3 MBCO SCPs 11 ± 11 genes). Comparisons of the top predicted GO and MBCO processes document that one would still be able to get a similar amount of information from a downsampled dataset (even if SCP p-values correlation coefficients are below 0.8). An increase of the cutoff rank from 10 up to 20, on average, allows reidentification of 70% of the top predictions, if single cell RNASeq data of only two to 14 libraries (~1,367 to ~10,119 cells) is subjected to MBCO enrichment analysis (Figure 4A) and reidentification of 50% of the top predictions, if data from only two to four libraries (~1,367 to ~2,866 cells) are used (Supplementary Figure 2C). Similar results were obtained for the single-nucleus RNASeq predictions. On average 12 and 15 libraries (~3,100 and 3,835 nuclei) allowed reidentification of seven of the top 10 predicted podocyte and proximal tubule MBCO SCPs, respectively, while 21 libraries (~5,462 nuclei) were sufficient to reidentify five of the top

10 principal cell GO BPs (Figure 4B). Five of the top 10 predicted podocyte and proximal tubule MBCO SCPs could even be reidentified based on eight libraries or ~2,066 nuclei (Supplementary Figure 2D). For the LMD proteomics dataset, six to eight samples were sufficient to reproduce the results obtained for the full datasets with only minor variations in the correlation of identified DEGs (Figure 4C) and SCPs (Supplementary Figure 2E) or SCP rankings (Figure 4C).

Proteomic and transcriptomic assays produce non-overlapping and biologically complementary assessment of subsegmental molecular composition

It has been shown that transcriptomic and proteomic assays do not always produce overlapping results [22]. This discrepancy had been hypothesized to stem from the differences in RNA and protein turnover as well as differences in assay methodologies [23]. We have two subsegmental proteomic and four transcriptomic assays within KPMP that can be cross correlated to assess consistent co-expression of mRNA and protein levels across different cell types or tissue subsegments. Both proteomic datasets identified protein expression in two kidney sub-segments: glomeruli and tubulointerstitium (for LMD) or proximal tubule (for NSC). To allow for an unbiased cross-platform comparison, we focused on podocytes (glomeruli in LMD) and proximal tubular cells and subsegments in the single-nuclei/cell and LMD bulk RNASeq datasets as well. We identified the gene and protein expression values of each participant within each cell type cluster or subsegment, followed by the calculation of all pairwise correlation coefficients between these samples. Hierarchical clustering of the correlation coefficients documented that the absolute gene and protein expression values are specific for a particular platform and not for their anatomical origin (Supplementary Figure 3A). Such clustering behavior could arise from the inconsistent detection of genes and proteins by the different technologies, many genes or proteins are only detected by a particular technology, but not by another. However, even if we remove those genes and proteins that were not consistently detected across all six technologies (Supplementary Figure 3B), the clustering still groups samples obtained by the same technology (and not same anatomical origin) into the same group. Since platform related biases should influence gene or protein expression values independently of their tissue origin, the calculation of expression ratios between two values obtained by the same platform might reduce such biases. In agreement with this statement, the logarithmized ratios of gene and protein expression values between the glomerular and tubular cell types or subsegments (and vice versa) were specific for the anatomical region, as revealed by the clustering results that are obtained with (Figure 5A) or without (Supplementary Figure 3C) removal of those genes and proteins that were consistently detected by all 6 technologies. The clustering results document that such a normalization approach shifts the focus from the platform to the biology. Nevertheless, each platform can provide additional information, as documented by the technology related sub-clusters and the sets of genes and proteins uniquely identified by each technology. Similar results are obtained, if the samples are clustered based on the top 50 predicted GO BPs or MBCO level-3 SCPs (Supplementary Figure 3D and 4E, respectively). Correlation analysis of averaged $\log_2$ fold changes between all RNASeq and proteomic platforms further supports high consistency of results (Figure 5B) and indicates that the consideration of multiple datasets increases accuracy of the results, since averaged RNASeq and proteomic datasets show a higher correlation with each other than any individual RNASeq and proteomic datasets (Figure 5C).

Imaging-based molecular data and non-spatial proteomic and transcriptomic assays together produce cohesive marker expression signatures.

Imaging assays are critical in providing spatial specification of the deep mapping by techniques such as single-cell RNASeq or bulk proteomics [24]. Imaging and single-cell RNASeq assays provide complementary information at the single cell level. While imaging assays identify the spatial localization of individual cells together with their expression signatures for a limited number of proteins, single-cell RNASeq assays provide more extensive transcriptomic profiles for individual cells. Through integrating imaging and single-cell transcriptomic datasets, we can infer extensive molecular profiles for spatially localized cells.

A first step towards this integration is to develop approaches to integrate maps of cell types identified using imaging and single-cell transcriptomic technologies. Here, we construct a mapping matrix to transform the cell-type specific protein expression profiles measured using the imaging assay to cell-type specific gene expression profiles measured using single-cell transcriptomic assays (Figure 6). An entry in the mapping matrix is high if the corresponding imaging cell type is highly weighted in the linear combination of imaging cell type expression profiles that approximate the expression profile of a cell type in the single-cell transcriptomic dataset. We find that this mapping approach performs particularly well for cell types characterized by a smaller number of unique markers in the imaging dataset (for example, endothelial cells or podocytes). Despite differences between protein and gene expression across cell types, the mapping demonstrates congruence across cell type-specific expression profiles measured using imaging and transcriptomic assays.

Future integration will explore mapping of cell types classified by this approach to our large-scale 3-D imaging with common markers including monolithic cell types such as neutrophils, T-cells and proximal tubule epithelial cells. In doing so, we will leverage 3-D relationships (e.g., densities and distance to anatomical structures of interest, etc.) between immune cells, the endothelium and tubular epithelial cells to better understand the biology and enrich the imaging signatures of health and disease.

Integration of gene expression profiles with metabolic pathways show cell subtype-specific metabolic signatures

Single-cell/nucleus RNASeq technologies allow for the classification of cellular subtypes, each might specialize on a few of the cell type specific functions. One example of cellular subtypes are the seven different proximal tubule cell subtypes that were identified in the PREMIERE dataset [8]. At its current resolution, spatial metabolomics allows for distinguishing glomerular and non-glomerular metabolites only. Due to the abundance of proximal tubules in the kidney cortex, a reasonable hypothesis is that most of the nonglomerular metabolites are generated by proximal tubule cells. This hypothesis is supported by our enrichment analysis. The pathways that are predicted for non-glomerular metabolites either overlapped with or were closely related to the pathways that are predicted for proximal tubule cells and subsegments based on the other datasets (Figure 2A).

One noteworthy feature of the differential gene expression pattern in the subtypes of the proximal tubule cells is the variations in the levels of expression of mRNAs encoding enzymes in the beta-oxidation pathways. Due to the high energy demand, lipid metabolism is critical for the physiological activities of the proximal tubule cells [18, 19]. The data from the single cell transcriptomic studies suggest that we may be able to predict the biochemical activity of proximal tubule tissue from the single cell expression data (Figure 7A). As a step towards construction of such predictive dynamical models we generated a canonical pathway map of mitochondrial beta-oxidation pathways from prior knowledge (www.genome.jp/kegg)

(Supplementary Figure 4A). Mitochondrial beta-oxidation describes the iterative generation of acetyl-CoA by sequential removal of two-carbon units from the fatty acid acyl chain. The first cycle during the oxidation of palmitoyl-CoA contains most of the enzymes that participate in its complete oxidation (Figure 7B). We specified the relative expression levels for the beta oxidation enzymes in the different subtypes of proximal tubule cells after assigning subtype specific gene expression values to the participating enzymes (Figure 7B). In Figure 7C we show the relative percentages of each enzyme in the different proximal tubule subtypes, in figure 7D the tissue collection types of the PT cells in each cluster. These percentages are calculated from mean read counts that are given in supplementary Figure 4B. Our pathway map suggests that a subpopulation of about 26% of all PT cells (clusters PT-1 and PT-5) over proportionally contributes to total beta oxidation activity in the proximal tubule.

Decrease in fatty acid oxidation, resulting in a loss of ATP generation, has been shown to be a significant contributor to tubulointerstitial fibrosis [19]. Our results suggest that upon general injury, a subset of proximal tubule cells could - because of decreased beta oxidation - contribute disproportionately to fibrosis and decreases in the physiological capability of the proximal tubule as a whole. This hypothesis that needs to be experimentally tested demonstrates that the integration of multiple omics technologies can give rise to new assumptions of cellular mechanisms involved in kidney physiology and pathophysiology.

Discussion

The advances in transcriptomic technologies along with other omics and imaging assays offer unprecedented insights into the organization of tissues at cellular resolution and the molecular constituents of the different cell types and their subtypes. Here, we have integrated data generated by multiple omics technologies as well as imaging assays to provide a detailed view of the molecular machinery present in kidney tissue. Our integrative analytics allows us to conclude that integration of multiple omics technologies permits identification of pathways and functional modules across cell types with greater confidence than any individual omics technology.

This is partly because the different technologies identify different members of the same functional pathway (i.e. SCP), giving greater confidence that the pathway is likely to be operational in the cell type of interest. Furthermore, these technologies offer differing perspectives into the physiological function or role of a given cell type, and they can act as complementary quality control steps for each other. For example, given the transcriptomic profile of each single-cell cluster and the corresponding subsegments, we show that we can successfully predict the spatial localization of each cell type. In addition, we show that our approach allows a detailed look at a validated pathway within proximal tubular cells, suggesting that there could be multiple subtypes of proximal tubule cells in humans. We predict that these subtypes differ in their potential for lipid metabolism, which is critically important for the physiological function of proximal tubule cells, as cellular energetics have been shown to be critical for reabsorptive activity. Here, it is important to note that mitochondrial content is potentially a driver of cellular subtype identity, and as KPMP researchers have noted before [8], using a universal percent mitochondrial transcript cutoff may not be optimal for processing of these data.

Though the post-hoc power analysis documents how many participants are needed to recover known biology, we cannot conclude from these results how many patients are needed to recover unknown biology that might be associated with disease development. Patients may manifest diverse molecular signatures even within diagnostic categories and sampling many patients may be crucial for elucidating disease subtypes and heterogeneous disease

processes and mechanisms across the patient population. Furthermore, critical disease signatures might be captured by rare cell types or subtle expression differences across individuals. Nevertheless, our post-hoc power analysis can help to estimate the reliability of an identified cell subtype or predicted disease mechanism by documenting that it is consistently recovered using down-sampled datasets.

Our approach is amendable to future computational modeling studies that can further improve the proposed tissue atlas. For example, mapping the beta-oxidation pathway for lipid metabolism to a set of coupled reactions that are specified as differential equations could allow us to develop dynamical models of this pathway in the different subtypes of proximal tubular cells. Simulations using such models can predict how changes in the number of cells in the relevant subtypes or changes in enzyme levels in particular subtypes can change the metabolic capability of the proximal tubules, thus affecting the physiological processes needed for reabsorption. These types of predictive capabilities of tissue maps of interacting molecular entities at cellular resolution will allow us to build a smart kidney atlas, where we do not only describe the different cell types, subtypes and location and levels of molecular entities, but also use this information to predict physiological capability of the tissue. While our integrated computational approaches allow single-cell transcriptomics to identify cellular subtypes with higher confidence, we note that these findings still require experimental confirmation with spatial assays, such as multiplexed *in situ* hybridization, which are being incorporated into the KPMP. Final maps in the atlas will always need to incorporate experimental verification of computationally derived information.

Information becomes knowledge only when it is deliberately and systematically catalogued such that new cohesive insights can readily be drawn. Ontology is an ideal tool that can logically represent the data and metadata in a human- and computer-interpretable manner and to enable the generation of new knowledge, especially when such knowledge involves multiscale relationships between molecules, cell types and their subtypes and tissue level physiological function. In addition to the integrated analytics presented here, the KPMP is also building a community-based Kidney Tissue Atlas Ontology (KTAO), which will systematically integrate different types information (such as clinical, pathological, cell and molecular) into a logically defined tissue atlas, which can then be further utilized to support various applications [34]. Taken together, the final interactive knowledge environment and the smart kidney tissue atlas constructed by KPMP should be able to help redefine cellular types and subtypes in the kidney; improve patient care by providing new disease classifications; and may ultimately lead to new patient-specific novel therapeutic approaches.

METHODS

Omics and imaging assays used within KPMP target different types of molecular components with different resolution, sensitivity and precision. An important function of the KPMP Central Hub is to integrate the different types of data using a set of analytical techniques. This process is summarized in Figure 1. Throughout the paper, we consistently use the same continuous color-code to identify different assays or cell types. The experimental assays that generate the raw data and all their technical details including standard operating procedures are detailed under the supplementary information and publicly released with all their technical details and version-controlled release dates on the KPMP protocols.io page (https://www.protocols.io/groups/kpmp/publications).

Identification of differentially expressed genes, proteins and metabolites

We analyzed data from four types of transcriptomic, two proteomic, one imaging-based and one metabolomic tissue interrogation assays. The pilot data presented for each assay comprises 3 to 48 different datasets that are obtained from 3 to 22 participants (Supplementary Table 1). Participants kidne tissue was procured from a spectrum of tissue resources including from unaffected parts of tumor nephrectomy specimen (n=38), living donor preperfusion biopsies (n=3), diseased donor nephrectomies (n=5), and normal surveillance transplant (n=5) and native kidney biopsies (n=4). Within each assay we generated lists of differentially expressed genes (DEGs), proteins (DEPs) and metabolites that describe those genes, proteins or metabolites that are upregulated or enriched in a particular single cell cluster, single nucleus cluster or kidney subsegment, if compared to all other clusters or subsegments.

For pathway enrichment analysis and module identification, cluster-specific differentially expressed genes (DEGs) were obtained from published analyses from PREMIERE TIS (Michigan, Princeton, Broad) single-cell RNA sequencing (RNASeq) [8] and UCSD/WU TIS single-nucleus RNASeq [9] datasets. We excluded the clusters proximal tubular cells-3 and principal cells-2 from the single-nucleus RNASeq dataset, since these clusters showed an inflammatory or a stress response. Laser microdissected (LMD) RNASeq and proteomics, near-single-cell (NSC) proteomics and spatial metabolomics datasets were individually processed as described in supplementary methods.

Ranking of Differentially Expressed Genes and Proteins

In the case of the DEGs and DEPs that were used for dynamic enrichment analysis, [6] module identification, [21] and post-hoc power analysis, single nucleus and single cell DEGs were first ranked by p-value and then by decreasing fold changes (i.e., fold changes were used as a tiebreaker). Top ranked 300 entities were subjected to downstream analysis. Similarly, DEGs and DEPs obtained for each kidney subsegment based on LMD bulk RNASeq, or LMD and NSC proteomics, were ranked first by p-value and decreasing fold changes and the top ranked 300 DEGs and DEPs subjected to pathway enrichment analysis or module detection (see below).

For the cross-platform comparisons, we did not combine sequencing and proteomic results of multiple participants to generate DEGs and DEPs, but compared the results obtained for each individual person. Therefore, we could not calculate p-values for the LMD and NSC technologies. Furthermore, since both proteomic technologies only generated results for 2 subsegments, i.e. the glomerular and proximal tubule segments for NSC proteomics and the glomerular and tubulointerstitial subsegments for the LMD proteomics, we collectively calculated the fold changes between podocyte/glomeruli and proximal tubules/tubulointerstitial cells or subsegments for each individual participant (see below). DEGs and DEPs were then

ranked by fold change, before submitting the top ranked 300 DEGs and DEPs for downstream analysis.

Standard and Dynamic Enrichment Analysis

Top DEGs and DEPs for each podocyte cluster/glomerulus, proximal tubule cell cluster/tubulointerstitium and principal cell cluster/collecting duct subsegment were separately subjected to standard enrichment analysis using Gene Ontology Biological Processes (GO BPs) or the Molecular Biology of the Cell Ontology (MBCO) level-3 subcellular processes (SCPs) [6] and Fisher's Exact Test. An open access ontological framework for this analysis can be found at https://github.com/SBCNY/Molecular-Biology-of-the-Cell.

The right-tailed Fisher's Exact test calculates the likelihood of obtaining the observed or a higher overlap between a list of DEGs/DEPs and a list of genes/proteins annotated to a particular SCP. To calculate this likelihood, we consider which genes or proteins have a chance to be identified as differentially expressed. Only genes/proteins that are detected by this method and statistically analyzed for differential expression can be identified as DEGs/DEPs and only these genes/proteins are considered as the background set for the Fisher's Exact test. Similarly, the background set only contains genes that have a chance to be assigned to a given SCP in a given ontology. In the case of the single cell [8] and nucleus [9] RNASeq datasets, all genes that are part of the UMI (Unique Molecular Identifier) read count matrices comprise the experimental background genes. In the case of the LMD bulk RNASeq, and the LMD and NSC proteomics datasets, the experimental background genes/proteins were all genes/proteins that were statistically analyzed by the TISs for differential expression (Supplementary Tables 2, 3 and 4, respectively). Ontological background genes/proteins were all genes that are annotated to at least one pathway within that particular ontology. The intersection of the experimental and ontological background genes/proteins is called background genes/proteins and is different for every assay and ontology combination. For additional statistical accuracy we removed all genes and proteins that were not part of the background genes/proteins from the lists of DEGs, DEPs and SCP genes before each enrichment analysis.

Additionally, we subjected the top DEGs and DEPs to dynamic enrichment analysis using MBCO [6]. The annotated interconnected hierarchy of MBCO is enriched using a unique algorithm that infers relationships between functionally related SCPs. Dynamic enrichment analysis uses these relationships to generate context-specific higher-level processes by merging functionally related SCPs that contain at least one DEG or DEP. The context specific higher-level SCPs contain all genes of the original SCPs and are added to the annotated ontology to generate a context specific ontology. The context specific ontology at this point contains single or merged SCPs. This list is then used for enrichment analysis of the DEPs or DEGs using Fisher's Exact test. The top five predicted SCPs or merged SCPs are connected based on the inferred relationships, and all networks for a particular cell type/segment merged, whereby each SCP was color-coded according to the source assay(s) that initiated its dynamic enrichment.

Module Detection

In parallel to enrichment analyses, we also performed another network-based pathway enrichment technique, identifying modules of cell-type specific marker genes within the kidney-specific functional network using the HumanBase interface (hb.flatironinstitute.org). For each cell type (proximal tubule, podocyte and principal cell), module detection was performed using the top 300 DEGs from each cell-type specific transcriptomic dataset cluster and the top 300

DEPs from each proteomics dataset. Module detection is a network-based approach described in Krishnan et al., and construction of the functional networks is described in Greene *et al* [20, 21]. In contrast to the prior knowledge-based MBCO networks, the kidney-specific functional network is constructed using a data-driven regularized Bayesian framework based on the information in thousands of datasets, which include co-expression, transcription factor binding, protein-protein interactions, and other data types. Modules are detected using a community clustering algorithm based on connectivity between genes in the kidney-specific functional network, and enrichment analysis is subsequently performed to identify functional enrichments in each module.

Enrichment Analysis of Metabolites

All glomerular and nonglomerular metabolites that were identified for the three participants were merged and subjected to pathway enrichment analysis using MetaboAnalyst [25]. The top six predicted metabolic pathways were mapped onto MBCO pathways whenever possible; if they did not have a corresponding pathway, the original pathway names were preserved.

Integration of Single-Cell/Single-Nucleus Transcriptomics

In contrast to bulk mRNA sequencing, where the gene expression measurements reflect an average across all captured cell types, single-cell or single-nucleus mRNA sequencing allows the measurement and comparison of comprehensive gene sets obtained from individual cells. This approach enables mapping of cellular heterogeneity with high throughput. In the first phase of the project, three KPMP tissue interrogation sites (TISes) performed this approach to generate single/single nucleus expression data from normal adult kidney tissue. In addition to locally acquired kidney tissue samples, each TIS also used a set of common KPMP pilot tumor nephrectomy tissue samples to generate the expression data. Single-cell transcriptomic data was produced by PREMIERE (24 libraries from 22 participants) [8] and UCSF (10 libraries from 10 participants), whereas the single-nucleus data was made by UCSD (47 libraries from 15 participants) [9]. Following is the brief description of the integration of the data from the three sites.

Data from each site were first processed using the Seurat 3.0 R package [26]. As a quality control step, nuclei/cells with less than 500 and more than 5,000 features and more than 20% mitochondrial genes were removed. The processing steps included normalization and identification of highly variable genes. We then removed potential doublets using DoubletFinder [27] from each dataset. Next, we used the integration algorithm embedded in the Seurat R package to perform combined analysis of single-cell/single-nucleus transcriptomic data. The integration algorithm first identified a set of anchor genes in each processed dataset. These anchor genes were then used to harmonize the datasets. The downstream process included scaling, principal component analysis, batch integration using harmony, dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP), and unsupervised clustering. The clustering was performed at a low resolutions (0.5). Enriched genes for each cluster compared to all other clusters were identified by running the Wilcoxon rank sum test.

Integration of Single-cell, Single-nucleus and Laser Capture Microdissection Bulk Transcriptomics

To integrate single-cell sequencing, single-nucleus sequencing, and LMD bulk transcriptomic datasets, we first determined the overlap between genes identified both in the LMD dataset and in the corresponding single-cell transcriptomic dataset. From this set of

shared genes, we restricted further analyses to a subset of genes showing variable expression in the single-cell dataset. We then computed the Pearson correlation between each individual cell in a scaled single-cell dataset and the LMD transcriptomic dataset for the same participant. For this correlation, we used the logarithm of a "mean ratio vector" which is composed of relative expression of a gene within a subsegment divided by the average expression of the same gene within all other subsegments within the dataset for the same participant, averaged across all participants. Using this approach, we can assign each cell to the appropriate LMD segment that shows the highest correlation value. To evaluate the overall segment assignments for individual cell clusters, we examine the normalized distribution of cells assigned to each LMD segment within a given single-cell cluster and present this as a normalized heatmap that represents overlap between different transcriptomic assays.

Post-hoc power analysis

The PREMIERE single-cell RNASeq [8] and the UCSD/WU single-nucleus RNASeq [9] datasets were obtained from 22 and 15 participants, respectively, whose samples were sequenced in 24 and 47 libraries. We used these datasets to assess the reproducibility and reliability of both assays in a post-hoc power analysis. This analysis compares results of the full datasets with the results from down-sampled datasets where libraries are randomly and systematically removed from the full data. Both datasets were separately subjected to a standardized Seurat pipeline for the identification of single-cell (or -nucleus) clusters and DEGs. Nuclei and cells with less than 500 and more than 5,000 features as well as more than 20% mitochondrial genes were removed. 'SCTransform' was used for data normalization and scaling (based on top 2,000 features), followed by principal component analysis. We used jackstraw analysis to identify the last significant principal component (alpha = 0.01) among the top 20 components. The last and all earlier principal components (independent of their significance) were used for dimensionality reduction before identifying single nucleus/cell clusters (resolution = 0.4).

DEGs of each cluster were identified (adjusted p-value: 0.05) and compared with literature-curated cell-type specific essential genes (Supplementary Table 9) using Fisher's Exact test to assign a kidney cell type to each cluster. The assigned cell type is that cell type whose essential genes had the most significant enrichment among the DEGs of that cluster. To document the reliability of that cell type assignment we compared its p-value to the p-value of the second prediction (that cell type whose essential genes had the second most significant enrichment among the DEGs of that cluster). The larger the distance between both p-values, the more reliable the cell type assignment. The number of clusters that were assigned to each cell type was documented. Nuclei and cells that were assigned to a particular cell type and map or do not map to the corresponding tissue subsegment were counted as well, based on the subsegmental correlation analysis as described above. The top 300 DEGs were subjected to standard and dynamic enrichment analysis using gene ontology and MBCO.

We progressively and randomly removed libraries from the full (reference) datasets to generate 100 non-overlapping downsampled datasets for each number of remaining participants. Downsampled data was subjected to our analysis pipeline and results were compared with the reference results. We calculated the percentage of downsampled datasets for each number of remaining participants that identified a particular cell type. If a particular cell type was identified in a down-sampled dataset we counted how many nuclei/cells were assigned to the same and a different cell type in the reference analysis and how many nuclei/cells of a particular cell type mapped and did not map to the correct tissue subsegment. We calculated the Pearson correlation between the DEGs and the pathways predicted for that

particular cell type in the downsampled datasets, and the reference datasets based on $\log_2$(fold changes) and negative $\log_{10}$(p-values), respectively. Since in the case of dynamic enrichment analysis the same SCP might be part of multiple SCP combinations that are predicted with different p-values, we assigned the most significant -$\log_{10}$(p-value) to each SCP.

Pathway enrichment analysis normally involves identification of the most significant pathways irrespective of their p-values. A down-sampled dataset may still contain the same information as the full dataset, but less data might decrease reliability of the predictions and distort the ranking of the predicted pathways. As such, one might have to investigate lower ranked pathways. Consequently, we determined how many SCPs have to be considered in a down-sampled analysis to re-identify at least 70% (or 50%) of the top 10 or seven predictions obtained from standard or dynamic enrichment analysis with the full dataset, respectively. Using the full datasets, we identified the top 10 predicted GO BPs, the top 10 predicted MBCO level-3 SCPs based on standard enrichment analysis and the top seven predicted level-3 SCPs or SCP-combinations based on dynamic enrichment analysis. We then analyzed for each down-sampled dataset, how many predictions are needed to reidentify at least 70% or 50% of the identified reference SCPs. Notice that the top seven predictions based on dynamic enrichment analysis can contain more than seven SCPs, since each prediction is either a single SCP or a unique combination of two or three SCPs.

Similarly, we randomly and progressively removed participants from the full LMD proteomics dataset, followed by the recalculation of fold changes and significance between the protein expression values in the glomerular and tubulointerstitial subsegments. All significant DEPs (FDR ≤ 0.05) were compared between the full and the downsampled datasets. Additionally, the top 300 significant DEPs of each subsegment were subjected to enrichment analysis and predicted pathways compared as described above.

Proteomic-Transcriptomic Co-expression Analysis

LMD and NSC proteomic datasets identified protein expression in two kidney subsegments: glomeruli and tubulointerstitium for LMD and glomeruli and proximal tubule for NSC. For an unbiased cross-platform comparison we focused on podocytes (glomeruli in LMD) and proximal tubular cells (tubular subsegments in LMD) in the single-nuclei/cell and LMD bulk RNASeq datasets. We identified technology and participant specific cluster gene expression, using the “Average Expression” functionality embedded in Seurat R package (RNA assay, counts slot) on the cells/nuclei assigned to the same clusters in the integrated PREMIERE, UCSF and UCSD/WU data analysis described above. The gene lists of all proximal tubule clusters of an individual participant and technology were merged. If a gene was identified by more than one cluster, we defined the highest expression value as the merged expression value for that gene. For each technology we characterized all genes/proteins that were identified in at least one cluster or subsegment of at least one participant and defined these genes/proteins as a technology specific background set. The intersection of all background sets was defined as the set of common genes. Participant-specific podocyte or glomerular gene and protein expression was calculated by dividing gene and protein expression in podocytes, or glomeruli, by gene and protein expression in proximal tubule cells or proximal tubule/tubulointerstitial subsegments, after adding 1 to prevent division by 0. Ratios were inverted to describe proximal tubule/tubulointerstitial specific gene expression. Log_{10} absolute expression values and $\log_2$(ratios) of all genes/proteins or all common genes/proteins were subjected to pairwise correlation, followed by hierarchical clustering. Log_2 ratios were averaged over each participant within each technology and pairwise Pearson correlation coefficients were determined between the different technologies using the set of common

genes. Mean $\log_2$ ratios were averaged across the four RNASeq platforms and the two proteomic platforms, followed by determination of the Pearson correlation coefficient using the set of common genes.

Comparison of Cell Type-specific Imaging and Transcriptomic Expression Data

To integrate cell type-specific imaging and transcriptomic data, we first constructed matrices with average expression values for each gene in each cell type cluster for both the set of 16 normalized integrated transcriptomic clusters and the CODEX clusters. We normalized each gene in both transcriptomic and CODEX matrices to have a mean of 0 and standard deviation of 1. We then filtered both datasets to include only genes represented in both the transcriptomic and the imaging datasets and computed the average expression of each gene/protein in each cell type. We next considered the problem of constructing a matrix to computationally map transcriptomic cell clusters to the imaging cell clusters. Specifically, let A be the N x k1 matrix of average protein expression values by imaging data clusters, C be the N x k2 matrix of average gene expression values by transcriptomic clusters, and M be the k1 X k2 matrix that maps A to C. We want to find M such that AM ≈ C. We can approximate M by taking the Moore-Penrose pseudoinverse of A, denoted A+, with M ≈ (A+)(C). M then provides a set of weights that map the imaging cell types to the transcriptomic cell types, with a large value for an entry in M in position (i, j) indicating that the imaging cell type i makes a large contribution to approximating the expression vector of transcriptomic cell type j as a linear combination of imaging cell types. Before visualizing matrix M as a heatmap, we first normalized each row to have mean of 0 and standard deviation of 1 in order to identify the transcriptomic cell types that are weighted most heavily in the mapping to each imaging cell type.

Generating Pathway Maps for Beta-oxidation Network from Single-cell RNASeq Clusters

To better understand one of the most significantly enriched pathways in our integrated analytics of proximal tubules, reactions involved in fatty acid beta oxidation were extracted from KEGG (www.genome.jp/kegg). KEGG enzyme and compound IDs were replaced by their human annotated genes and compound names, respectively. Subcellular localization of each gene was identified using the jensenlab human compartment database based on a jensenlab confidence of at least four (i.e. 80% of maximum confidence in the database) [28]. Gene products that were annotated to the mitochondrial matrix or inner mitochondrial membrane were kept, generating a canonical network of mitochondrial beta oxidation. Mean read counts of the 7 proximal tubule subclusters within the PREMIERE dataset were assigned to the enzymes using the average expression functionaly of the Seurat R package (RNA assay, counts slot). Relative levels of the enzymes were visualized as pie charts, where the size of each slice represents the fraction of that enzyme in a particular subcluster. For each enzyme the average read counts in each subcluster was normalized towards the sum of the enzyme read counts in all clusters. The diameter of each enzyme pie chart is an arbitrary minimum diameter (30 arbitrary units) plus the binary logarithm of the sum of the mean UMI counts for that enzyme normalized to the mean UMI counts of the least expressed enzyme.

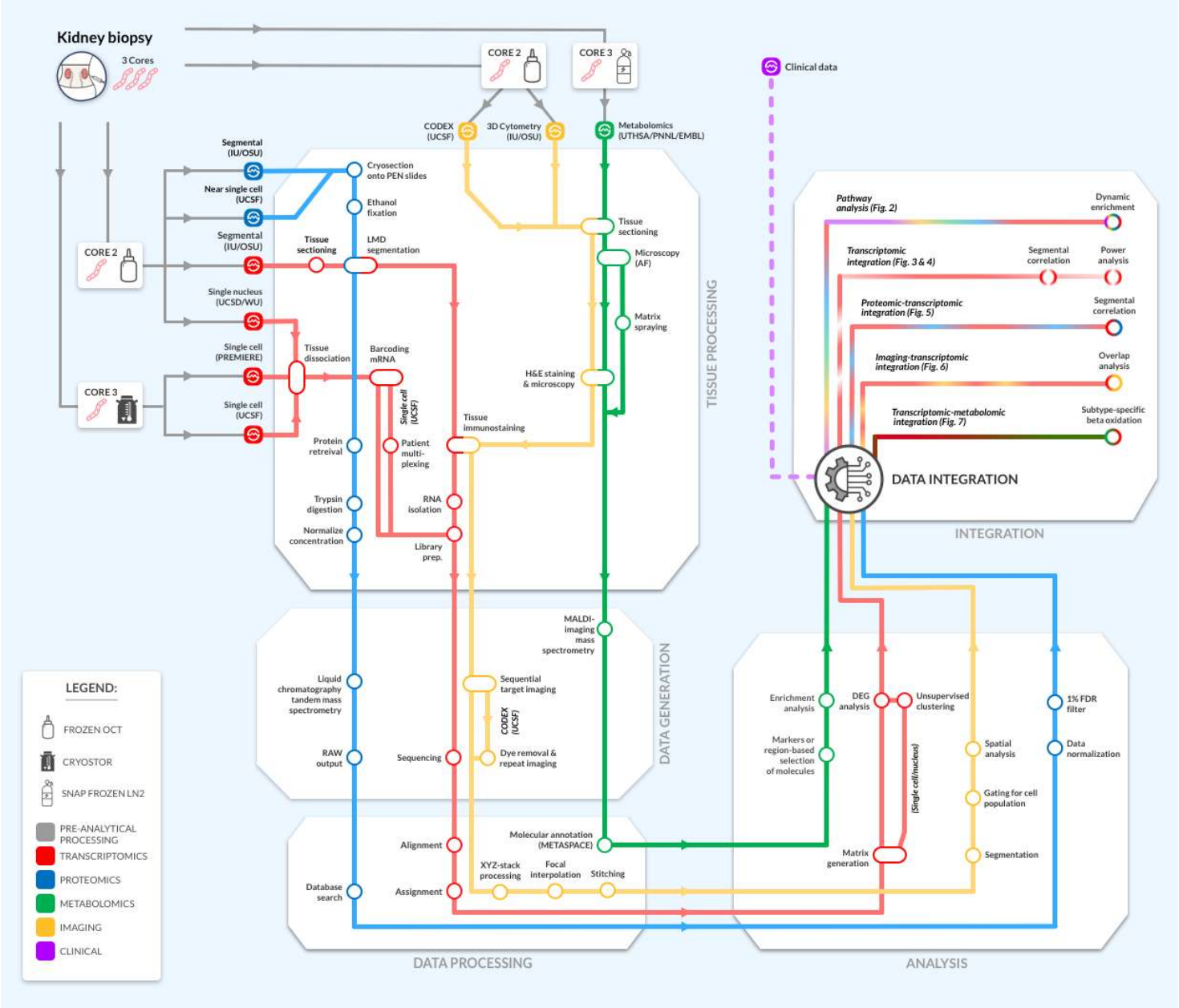
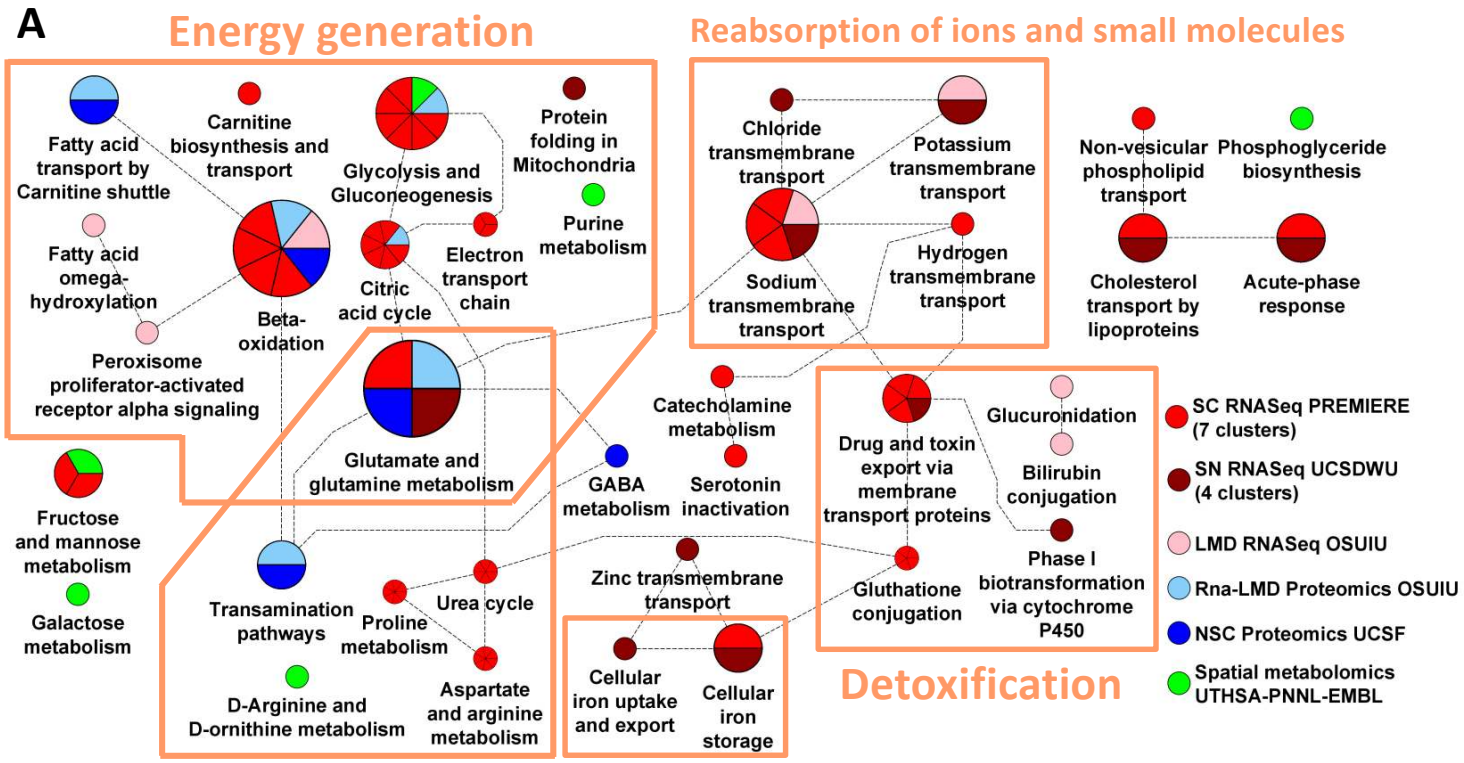
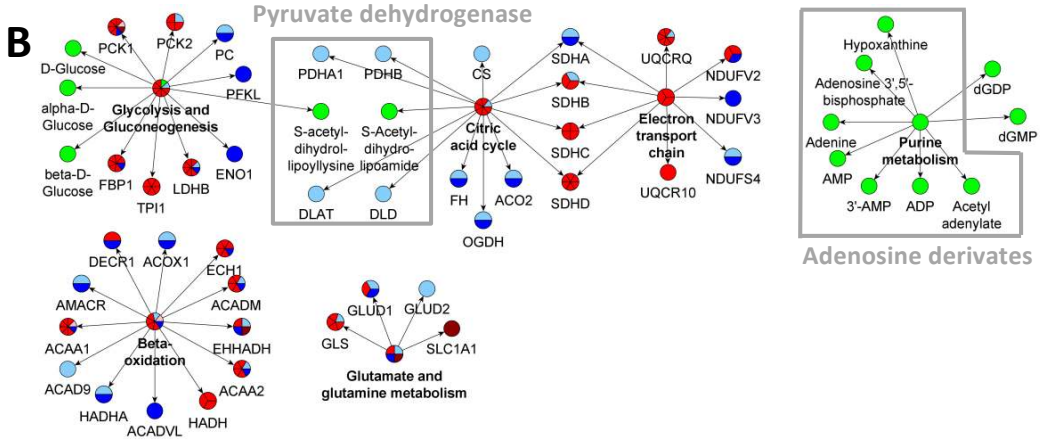


Figure 1

Figure 1. Graphic outline of KPMP data integration and harmonization procedures. The "subway map" representation of the experimental and analytical protocols used within KPMP is shown in operational flow from kidney biopsy to the integrated multimodal data represented in this manuscript. The kidney biopsy, which is processed through three different tissue processing methods, is shared among TISes that generate the data. Four key modalities of molecular data are generated: transcriptomic (red), proteomic (blue), imaging (yellow) and metabolomic (green). Biopsy core 2 and 3 are used for the molecular analysis, biopsy core 1 (not depicted) is used for histological analysis.





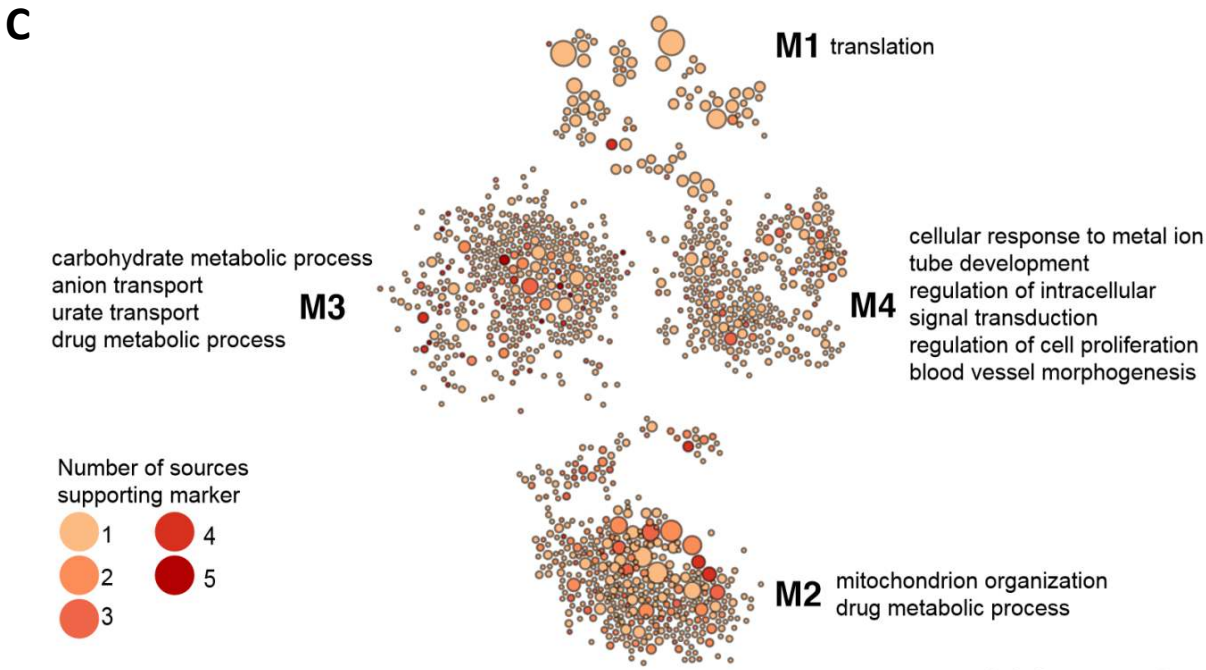
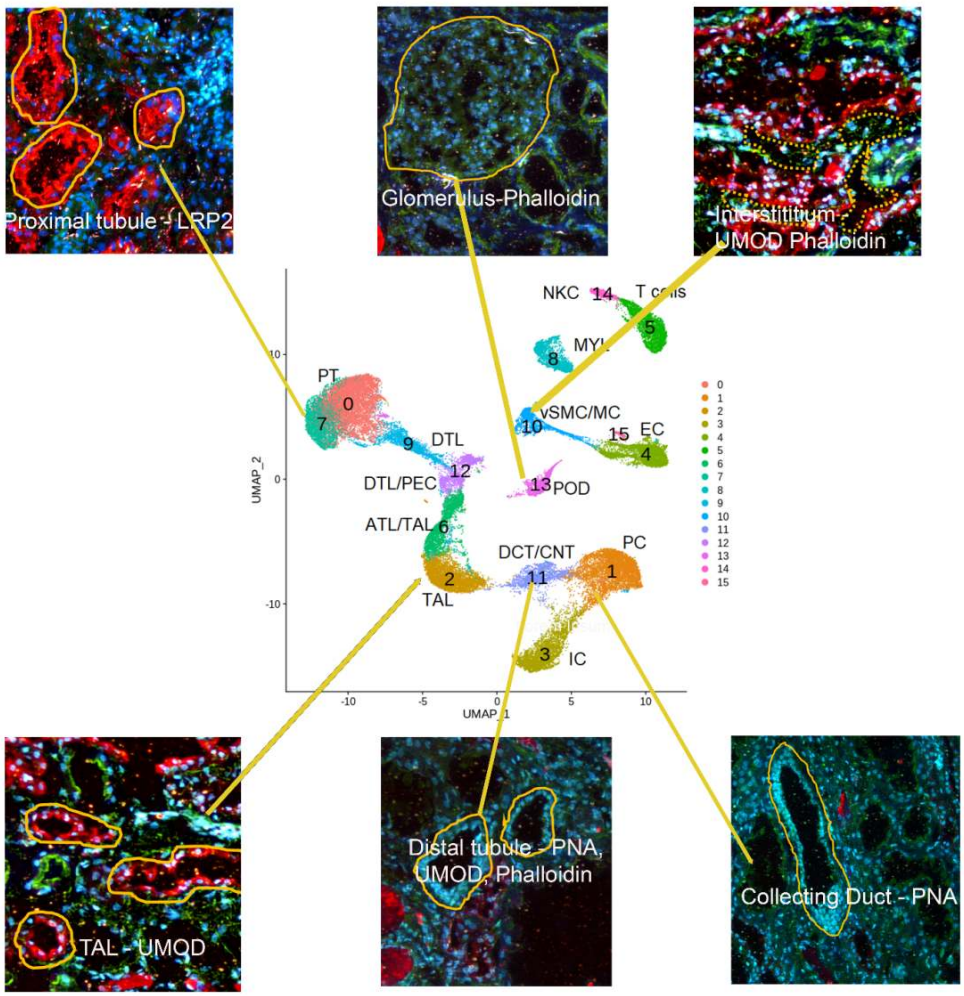


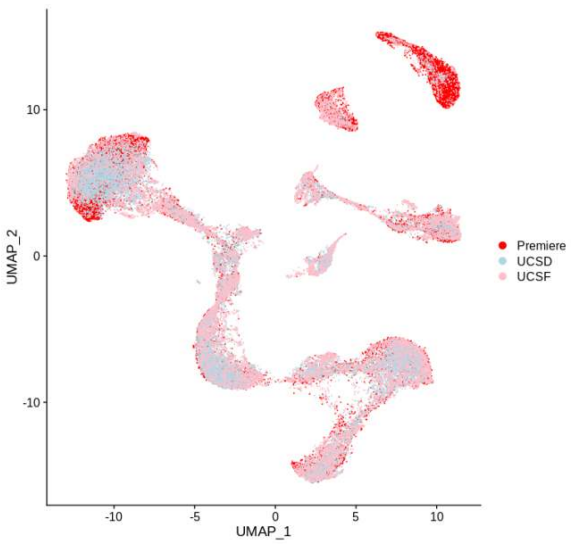
Figure 2

Figure 2. Downstream analysis of differentially expressed genes and proteins documents coherence and complementary results obtained by seven different technologies/centers. Nephrectomy tissues were subjected to single-nucleus (SN) and single-cell (SC) RNASeqq, laser microdissected (LMD) RNASeq and proteomics, near single cell (NSC) proteomics and spatial metabolomics. **(A)** Top 300 differentially expressed genes (DEGs) and differentially expressed proteins (DEPs) of each proximal tubule/tubulointerstitium cluster or subsegment were subjected to dynamic enrichment analysis using the Molecular Biology of the Cell Ontology (MBCO). Subcellular process (SCP)-networks were generated from the top seven most significant predictions. Metabolites associated with non-glomerular compartments were subjected to MetaboAnalyst enrichment analysis and the top six predicted pathways were integrated into the SCP networks after mapping to MBCO SCPs. Predicted SCPs that describe canonical cell biological functions were removed from this figure and are instead shown in Supplementary Figure 1A. **(B)** All six datasets were screened for genes, proteins and metabolites that are expressed or generated by proximal tubule cells or subsegments and are associated with metabolic SCPs selected from the SCP-networks. **(C)** The top 300 DEGs/DEPs for each proximal tubule cell type or subsegment were subjected to module identification in a kidney-specific functional network, and gene ontology enrichment analysis was performed for each module. The size of each node represents its degree, and its color the number of data sources supporting that gene.

A



B



C

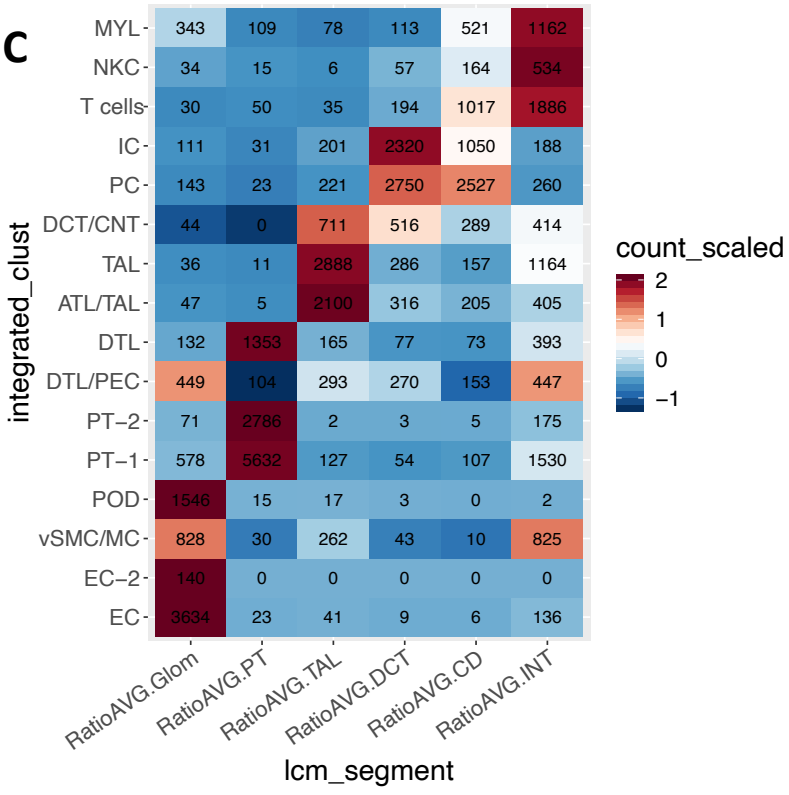


Figure 3

Figure 3. Integrated transcriptomic analysis reveals coherent cell-specific signatures. **(A)** Integrated UMAP of single-cell and single-nucleus technologies with associated LMD images of nephron subsegments. The corresponding segments shown include the markers used to identify each: Phalloidin – FITC labeled phalloidin for dissection of glomeruli and other structures; LRP2 – Megalin with AlexaFluor 568 secondary (red); UMOD – directly conjugated AlexaFluor 546 Ab to uromodulin (red); fluorescein labeled PNA – Peanut Antigen labels collecting ducts (green); DAPI included for nuclei (blue). **(B)** Each cluster contains cells and nuclei obtained by each one of the three technologies. **(C)** Each cell in the integrated transcriptomic single-cell analysis is mapped to the closest subsegment (subsegment with highest Pearson correlation of gene expression) in the laser capture microdissection data. To compute the Pearson correlation between the gene expression profiles of cells and LCM segments, the gene profiles were restricted to genes shared between the two datasets and showing variable expression in the single-cell dataset and correlations were computed between the logarithm of the mean ratio vector for each LCM segment and the scaled expression profile of each cell in the single cell dataset. For each single-cell cluster and LCM subsegment, the number of cells from that cluster assigned to the corresponding segment is displayed in the heatmap. The heatmap is colored according to the number of cells assigned to each LCM subsegment, scaled so each row has mean of 0 and standard deviation of 1.

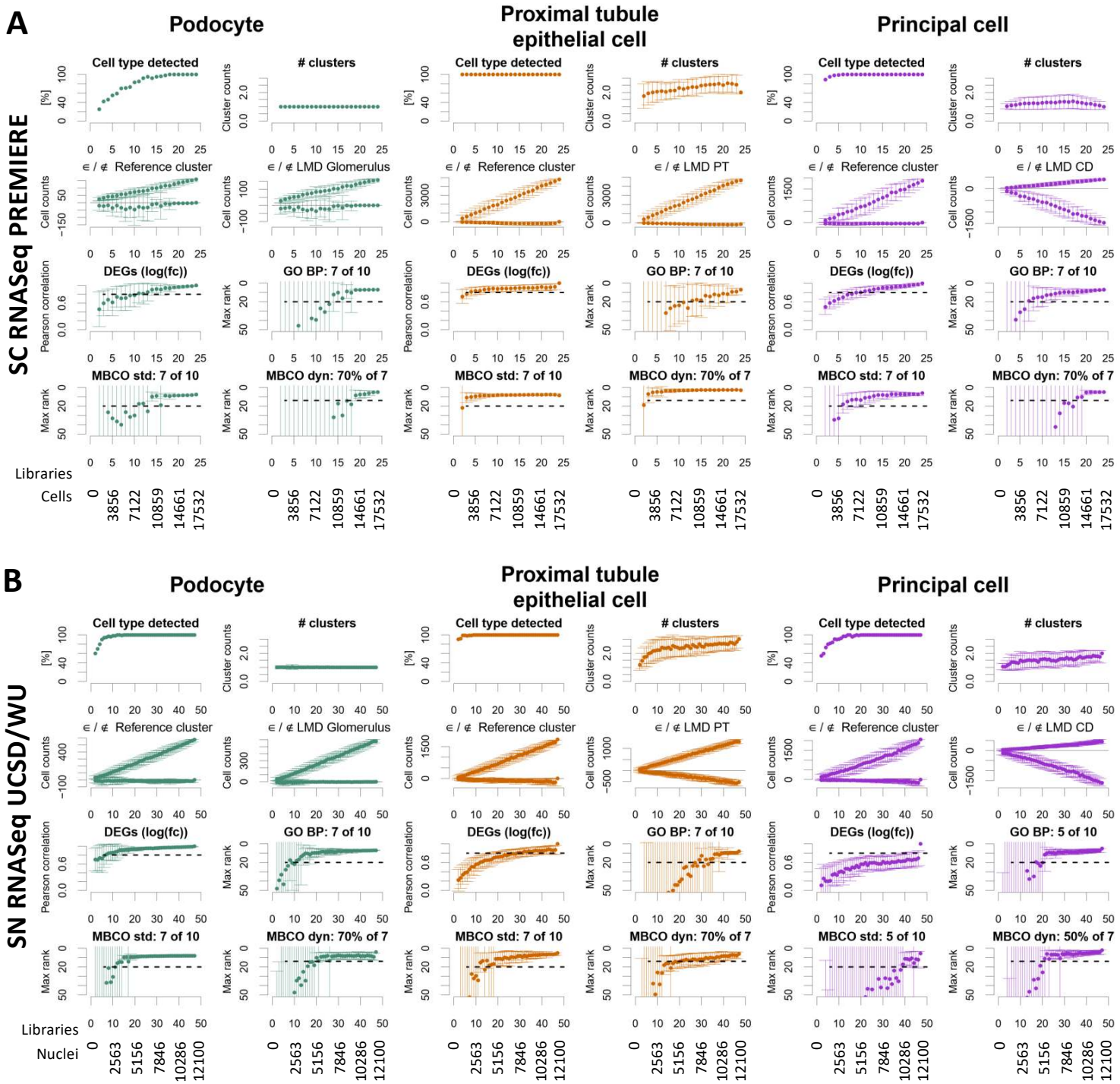
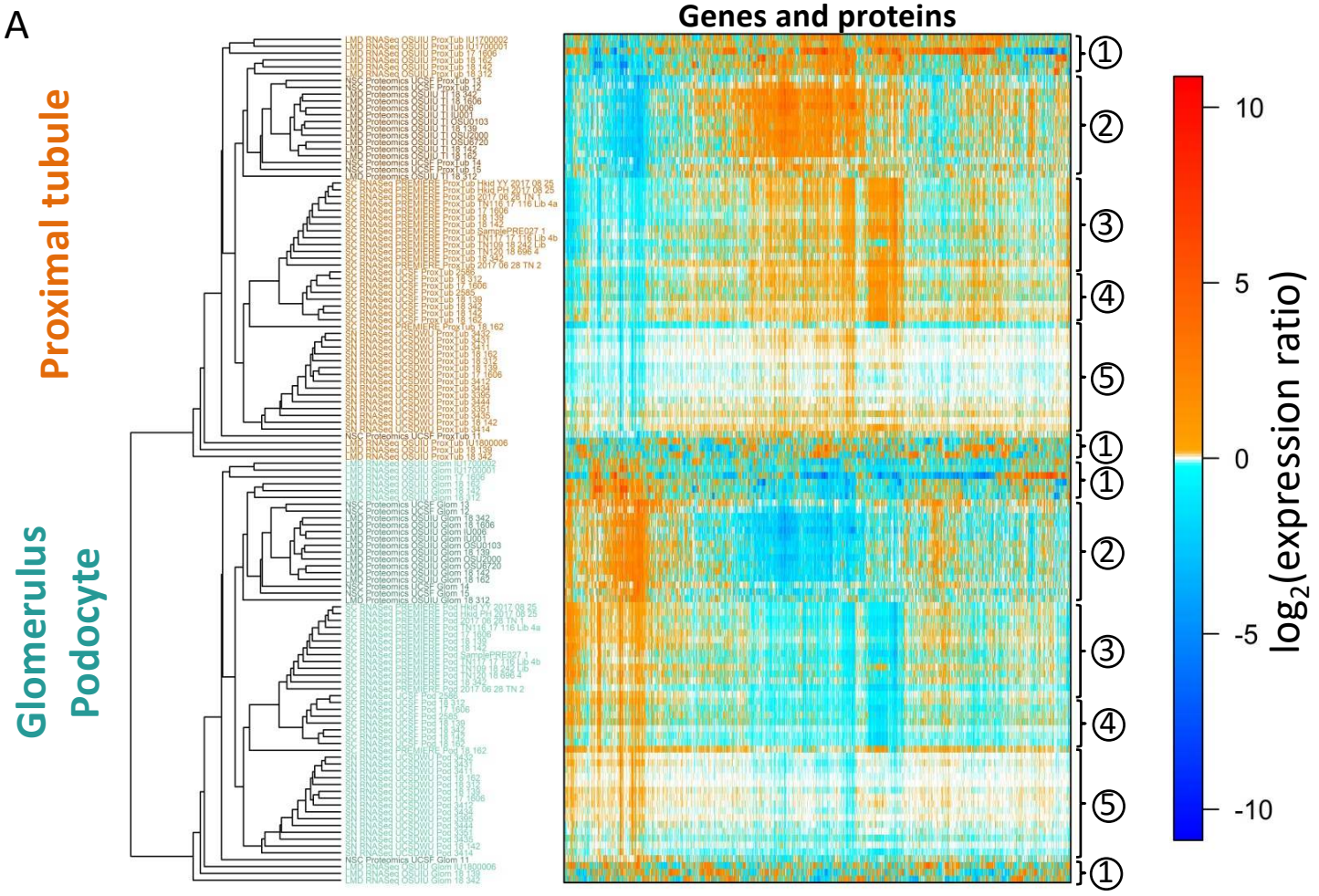
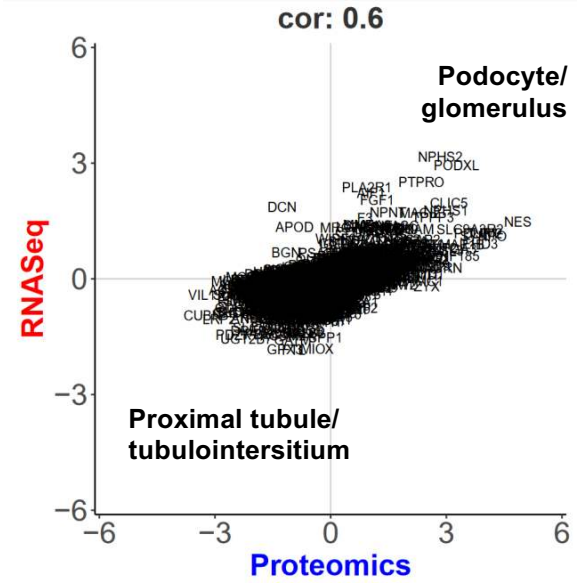


Figure 4

Figure 4. Post-hoc power analysis suggests that 16-25 participants are sufficient to reproducibly characterize most major cell types in single-cell transcriptomics. Participant libraries or samples were randomly and progressively removed from the **(A)** PREMIERE single-cell (24 libraries), **(B)** UCSD single-nucleus (47 libraries) RNASeq, and **(C)** LMD proteomics (11 samples) datasets to generate at max 100 non-overlapping random groups for each number of remaining participants. Datasets were subjected to an automated single-cell/nucleus and proteomic data analysis pipeline and results compared between the downsampled and complete reference datasets. 'Cluster count' documents how many clusters were assigned to a particular cell type. ∈/∉ reference cluster counts how many cells that were assigned to a particular cell type were assigned to the same cell type (above abscissa, positive cell counts) or a different cell type (below abscissa, negative cell counts) in the reference analysis. ∈/∉ LMD subsegment counts how many cells assigned to a particular cell type have gene expression values that correlated the most with the named (above abscissa, positive value) or a different (below abscissa, negative value) LMD subsegment. The Pearson correlation of the $\log_2$(fold changes) of cell type specific DEGs or subsegment specific DEPs was determined between each down-sampled dataset and the reference dataset. To quantify how reliably biological information can be reproduced, we identified to which rank a researcher has to lower the cutoff to re-identify at least seven or five of the top ten pathways that were predicted based on the full dataset and standard enrichment analysis and to re-identify 70% or 50% of all SCPs that were part of the top seven predictions based on dynamic enrichment analysis. Libraries label the number of used sequencing libraries for each down-sampled dataset, cells the average number of total cells that were obtained from those libraries. See Supplementary Figure 2 for additional quality control measures and cell types.





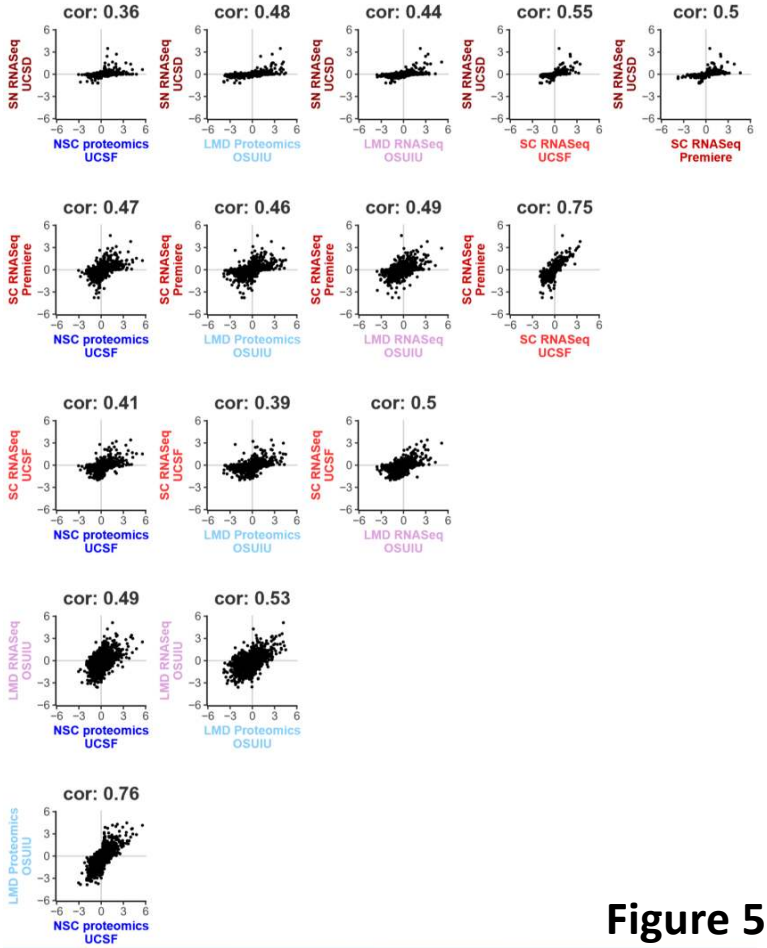


Figure 5

Figure 5. Correlation analyses demonstrate concordance across different omics technologies. Log_2(fold changes) between podocyte (or glomerulus) and proximal tubule cells (or tubulointerstitium) were calculated for each participant based on each technology and common genes/proteins identified by each technology subjected to comparative analysis. **(A)** Hierarchical clustering of pairwise correlation coefficients between the log_2(fold changes). Heatmap shows up- and downregulated genes/proteins of each sample in red and blue, respectively. Genes and proteins were rearranged according to the clustering results. White spots indicate undetected genes or no expression differences. Labels describing podocyte/glomerular and proximal tubule/tubulointerstitium RNASeq and proteomic datasets are colored aquamarine and orange, respectively.Curly brackets group samples obtained by the same technology: 1: LMD RNASeq, 2: NSC/LMD Proteomics, 3: SC RNASeq PREMIERE, 4: SC RNASeq UCSF, 5: SN RNASeq UCSD/WU. **(B)** Log_2(fold changes) obtained by the same platform were averaged across all participants, followed by averaging of the results across all four transcriptomics and two proteomics platforms. Positive (negative) log_2(fold changes) indicate podocyte/glomerular (proximal tubule/tubulointerstitial) expression, respectively. **(C)** Pairwise correlations between the single cell/nucleus RNASeq and proteomic datasets.

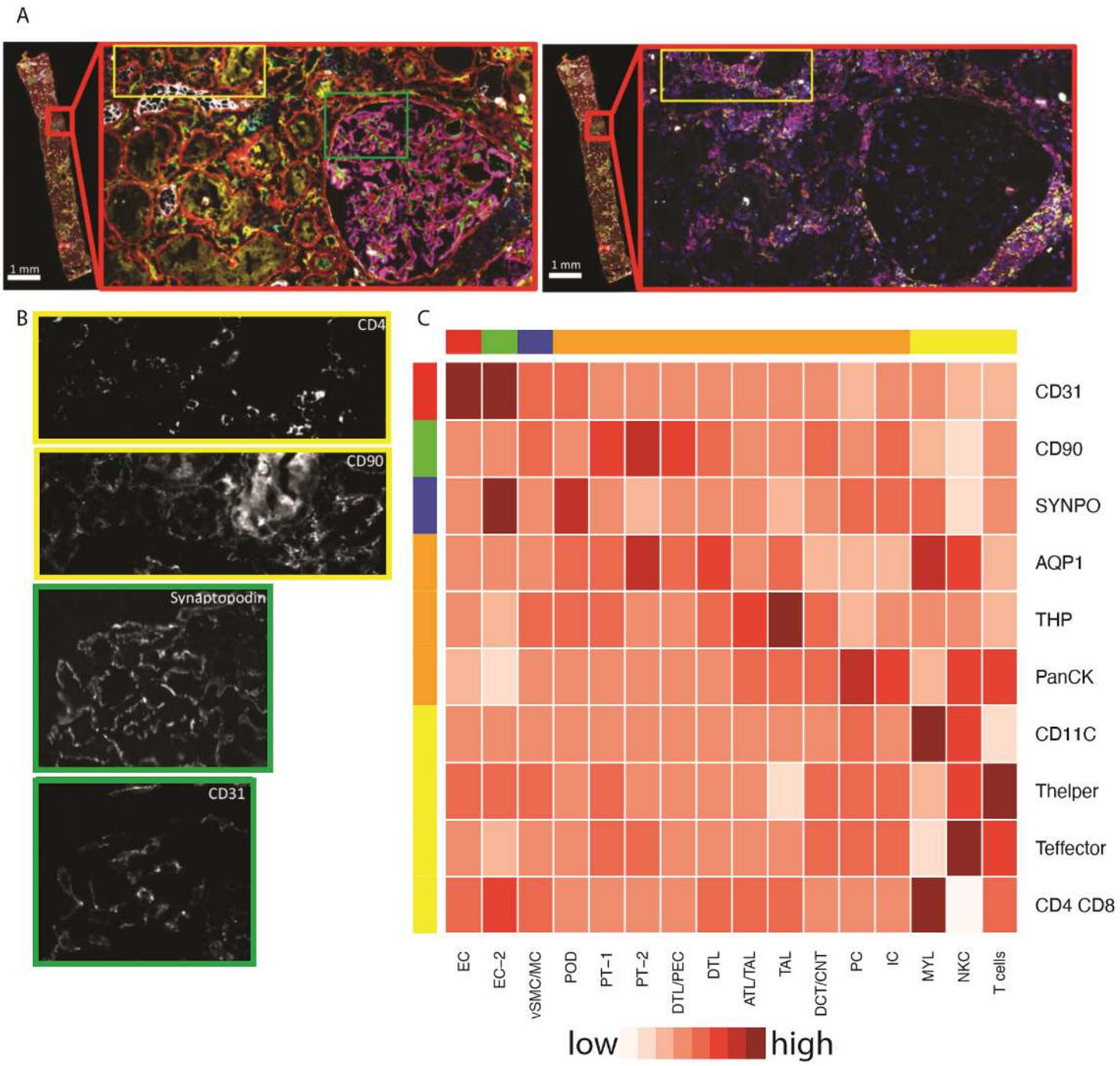
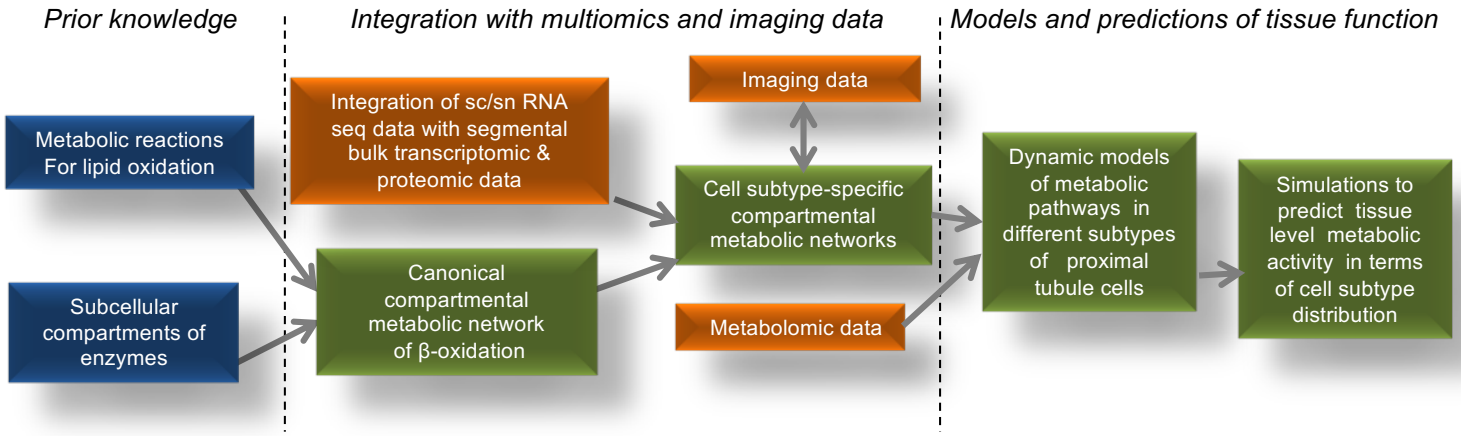


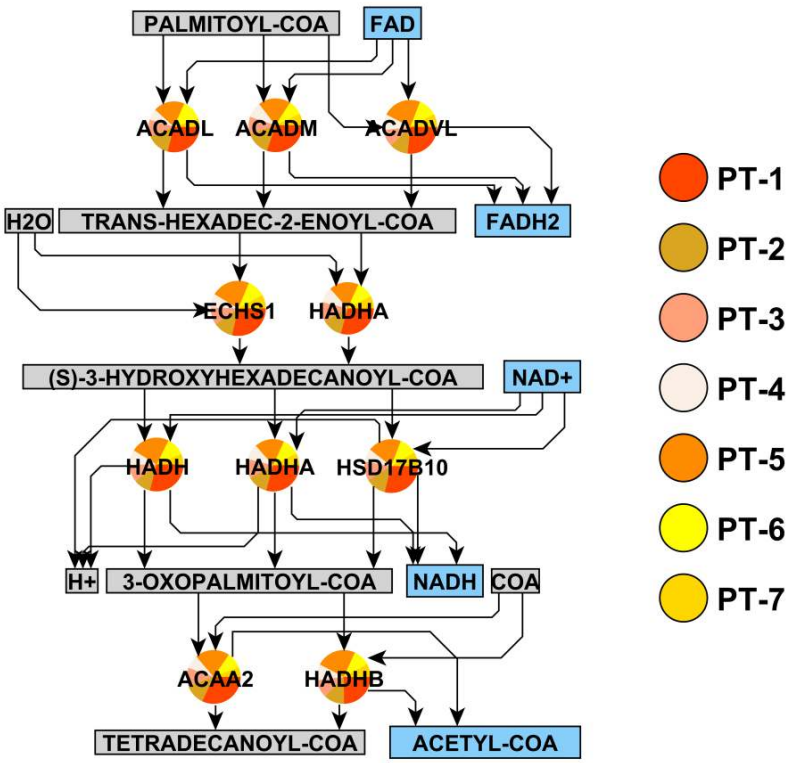
Figure 6

Figure 6. Imaging-based and transcriptomic assays show consistent cell-type-specific marker signatures. (A) CODEX multiplexed immunofluorescent image of an exemplary biopsy section (renal transplant biopsy with acute cellular rejection), showing identical areas with two different sets of markers. Upper panel: collagen IV (red), CD31 (green), synaptopodin (magenta), Pan-CK (gray), CD278 (cyan), CD90 (yellow). Lower panel: CD45 (red), CD4 (green), CD11c (gray), CD8 (cyan), CD38 (magenta), CD279 (yellow). Cell nuclei are blue (stained with Hoechst dye). Original magnification 200x. **(B)** Selected individual channels of the subregions highlighted by yellow and green brackets above. **(C)** Mapping matrix showing relationships between markers characterizing CODEX cell-type clusters and transcriptomic cell-type clusters. Colorbars to the top and left of the heatmap show broad cell-type categories (red = endothelial, green = fibroblast/mesangial, blue = podocyte, orange = tubular, yellow = immune).

A



B



C

	PT-1	PT-2	PT-3	PT-4	PT-5	PT-6	PT-7	Total
ACADL	29	15	11	7	20	13	5	100
ACADM	30	14	10	9	19	9	7	100
ACADVL	27	11	10	10	23	11	8	100
ACADS	28	12	9	10	22	10	9	100
ECHS1	29	12	10	8	23	11	8	100
HADH	29	10	11	9	24	11	6	100
HADHA	29	13	12	10	18	11	8	100
HSD17B10	29	12	11	8	20	11	9	100
ACAA2	32	13	11	8	21	10	6	100
HADHB	25	12	11	10	24	10	8	100
ACAT1	31	13	11	9	21	9	7	100

D

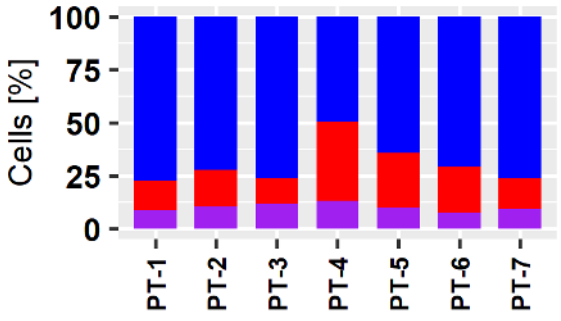


Figure 7

Figure 7. Proximal tubule cell subtype-specific beta oxidation pathways. (A) Simplified flow chart for the generation of cell subtype specific metabolic reaction networks that can form the basis for dynamic models. **(B)** Proximal tubule (PT) cell subtype specific mean enzyme expression values were added to a canonical reaction network of mitochondrial beta oxidation (using the PREMIERE single cell data). Enzymes are visualized as pie charts, where the sizes of the slices represent the relative enzyme expression (i.e. averaged read counts) in each cell subtype. Total pie sizes logarithmically increase with increasing summed enzyme expression of all cell subtypes. The prominence of the PT-1 and PT-2 slices suggests that most of the beta oxidation activity arises from two particular PT cell subtypes. Notify that we used directed arrows to indicate which reactants are on the same side of the equation, although all reactions are reversible. **(C)** Mean UMI counts of each enzyme in percent of the summed mean UMI counts of that particular enzyme. Colors shift from light to dark orange with increasing percentage. Enzyme names are official NCBI gene symbols. **(D)** Cell counts of the proximal tubule clusters documents that all clusters contain cells from each tissue collection procedure.

SUPPLEMENTARY INFORMATION

Supplementary Methods

Seven different RNASeq, proteomics, metabolomics and imaging datasets were generated and analyzed by five different TISes. The PREMIERE TIS (composed of Michigan, Princeton, Broad) generated single cell RNASeq data, the USCD/WashU TIS generated single-nucleus data, the UCSF TIS generated single-cell RNASeq, near-single-cell proteomics and Codex imaging data, the IU/OSU TIS generated laser microcapture dissection (LMD) RNASeq and LMD proteomics data and the UTHSA-PNNL-EMBL TIS generated spatial metabolomics data.

Single-nucleus RNASeq (UCSD/WashU) and Single-cell RNASeq (PREMIERE)

UMI count matrixes and list of differentially expressed genes were downloaded from published analyses for the PREMIERE TIS (composed of Michigan, Princeton, Broad) single-cell RNA sequencing (RNASeq) [8] and UCSD/WashU TIS Single-nucleus RNASeq [9] datasets. We excluded the proximal tubular cells-3 and principal cells-2 clusters from the single-nucleus RNASeq dataset, since these clusters showed an inflammatory or a stress response.

Subsegmental LMD Transcriptomics (IU/OSU)

A comprehensive Laser MicroDissection (LMD) protocol is published on protocols.io (https://www.protocols.io/view/laser-microdissection-8rkhv4w). Briefly, 12 μm frozen sections are obtained from an Optimal Cutting Temperature (OCT) preserved tissue block and adhered to LMD membrane slides (Leica, Buffalo Grove, IL). Tissue undergoes a rapid staining protocol involving acetone fixation, washes with RNAse-free PBS, and antibody incubation in 10% bovine serum albumin. Slides undergo dissection with a Leica LMD6500 system with pulsed UV laser. After collecting a minimum tissue area of 500,000 μm^2 in an RNAse-free micro-centrifuge tube, RNA is isolated using the PicoPure RNA IsolationKit according to manufacturer's instructions (Applied Biosystems, Cat# KIT0204). RNA quality is assessed by bioanalyzer, ribosomal RNA is depleted, and cDNA libraries are prepared using the SMARTer Universal Low Input RNA Kit (Takara, No. 634938). Sequencing was conducted on an Illumina HiSeq4000. Mapping was performed using STAR (v2.5.2b) and read counts were quantified with featureCounts (subread v.1.5.0). Total read counts mapping to each gene were generated with edgeR, normalized, and converted to expression ratios.

Segment specific gene expression was compared to the gene expression in all other subsegments using an unpaired ttest with equal variance. Subsegment specific gene expression ratios were calculated similarly.

Subsegmental LMD Proteomics (IU/OSU)

A comprehensive Laser MicroDissection (LMD) proteomics protocol is published on protocols.io https://www.protocols.io/view/laser-microdissection-for-regional-transcriptomics-8rkhv4w?version_warning=no. Our LMD proteomic methods have also been previously published in detail [29, 30]. Briefly, 10 μm frozen sections are obtained from an OCT preserved tissue block and adhered to polyethylene naphthalate (*PEN*) membrane *slides* for LMD. Frozen sections are fixed in 70% ethanol, incubated in H_2O to remove OCT, briefly stained with hematoxylin, and dehydrated in ethanol. LMD is performed and glomeruli and tubulointerstitial samples are collected separately in 0.5% Rapigest/50 mM NH_3HCO_3 solution. The collected samples are then boiled for 20 minutes for protein retrieval and digested overnight with trypsin. Peptides are dried, re-suspended in acetonitrile/formic acid and analyzed using liquid chromatography tandem-mass spectrometry (LC-MS/MS) analysis using

an Easy-nLC 1000 HPLC coupled to an Orbitrap Fusion mass spectrometer (Thermo Scientific, Waltham, MA). Data is searched using Proteome Discoverer 2.1 (Thermo Scientific) and searched against a human Uniprot database (version 05/26/18). Data are analyzed following global normalization of spectral counts.

Glomerular gene expression was compared to the tubulointerstitial gene expression using an unpaired t-test with equal variance. Glomerular to tubular specific gene expression ratios were calculated similarly.

3-D Immunofluorescence Imaging and Tissue Cytometry (IU/OSU)

The entire 3-D fluorescence imaging and tissue cytometry protocol is published on protocols.io (dx.doi.org/10.17504/protocols.io.9avh2e6). Briefly, frozen cores are sectioned at 50 μm using a cryostat and fixed using 4% paraformaldehyde. A panel of up to 8 antibodies was incubated to identify renal and immune cell types. Images were acquired in up to 8 channels using a Leica SP8 Confocal Microscope. Volume stacks spanning the whole thickness of the tissue were taken using a 20× NA 0.75 or 40× NA 1.3 objectives with 0.5- to 1.0-μm spacing. Large scale confocal imaging of overlapping volumes was performed with an automated stage and stitched using Leica LASX software (Germany). 3-D image rendering was done using Voxx v2.09d. The 3-D tissue cytometry was performed on image volumes using VTEA, which was developed as a plugin for ImageJ/FIJI as previously described [31].

CODEX Imaging (UCSF)

The CODEX system is the combination of an (1) oligo-nucleotide based antibody labeling-detection technique, (2) a microfluidics instrument coupled with an inverted microscope capable of whole slide scanning, and a (3) software suite that consists of an image processor and an ImageJ-based image analysis solution [32]. First, a section from an optimal cutting temperature compound-embedded tissue block is cut and incubated manually in a single step, with a set of antibodies each tagged with a unique oligonucleotide sequence. The following phase consists of iterative cycles of detection, imaging, and dye removal. In each cycle, a maximum of three targets are revealed by spectrally distinct dyes (AF488, Atto 550, and Cy5) tagged with oligonucleotides complementary to the oligonucleotide tag of a given antibody.

The acquired images are processed by the CODEX processor in a set sequence of steps: shading correction, tile registration, deconvolution, drift compensation, overlap cropping, background subtraction, best focus detection/interpolation, stitching, cell segmentation, and spillover compensation.

The output of the cell segmentation step of image processing is an .fcs file (similarly to flow-cytometry solutions). This file contains the individual fluorescent intensity values (can range from 0 to 65k) of each cell for each marker. Fluorescent intensity values allow the definition of cell populations by manual gating of the segmented cells using visual assessment of the image and previous literature data on the expression pattern of our marker set in human kidney.

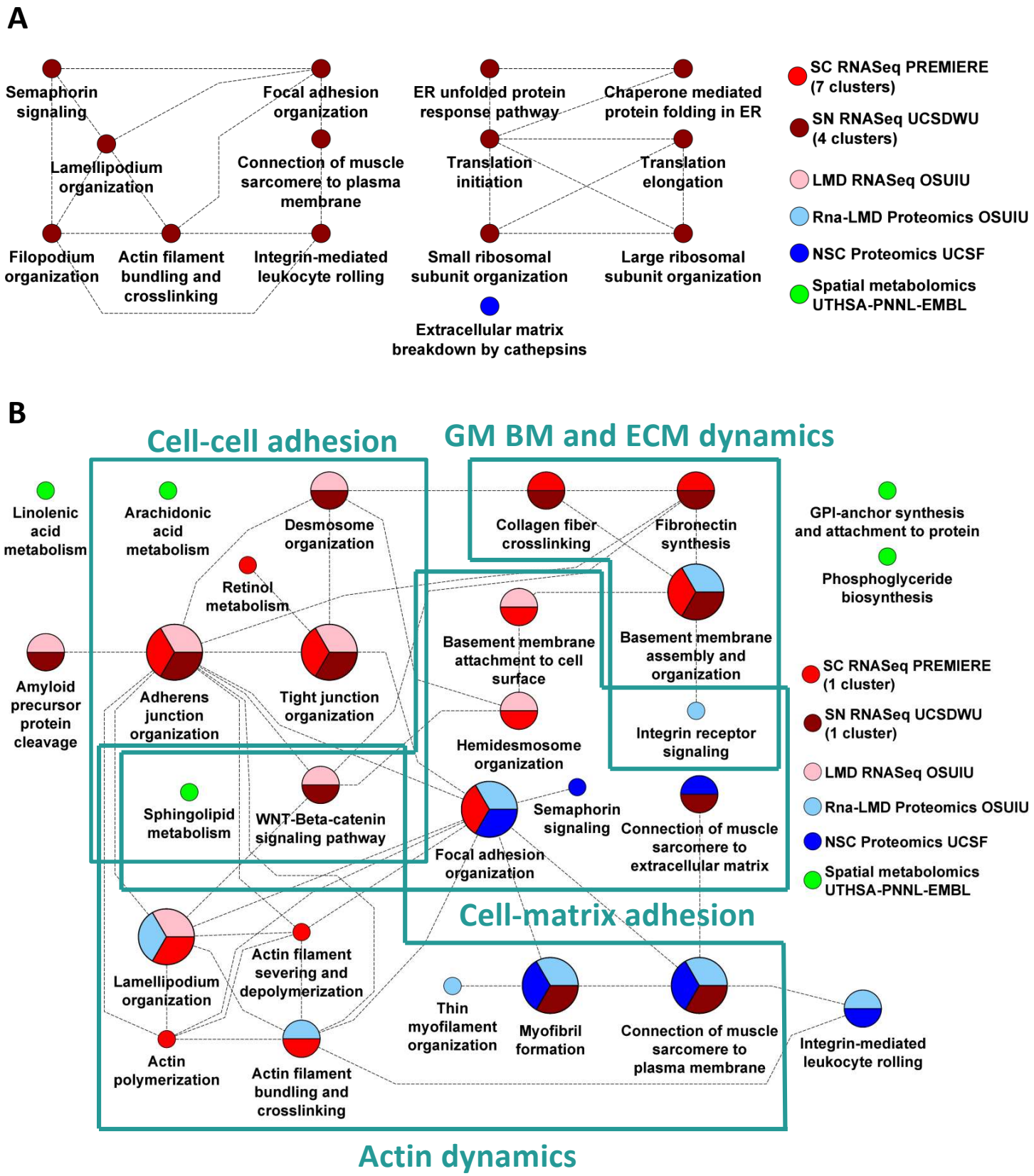
Native renal biopsies taken at University of California, San Francisco from patients with minimal change disease (n = 3), thin-basement membrane disease (n = 1), and post-surgical biopsies from tumor nephrectomies (n = 2) were used. In addition, case 18-162 from KPMP pilot sample pool was also processed (Supplementary Table 1).

Spatial Metabolomics (UTHSA-PNNL-EMBL)

10 μm thick renal cortical tissues were sectioned on a cryostat (Leica Microsystems) and prepared for matrix assisted laser deposition imaging (MALDI) mass spectrometry by spraying

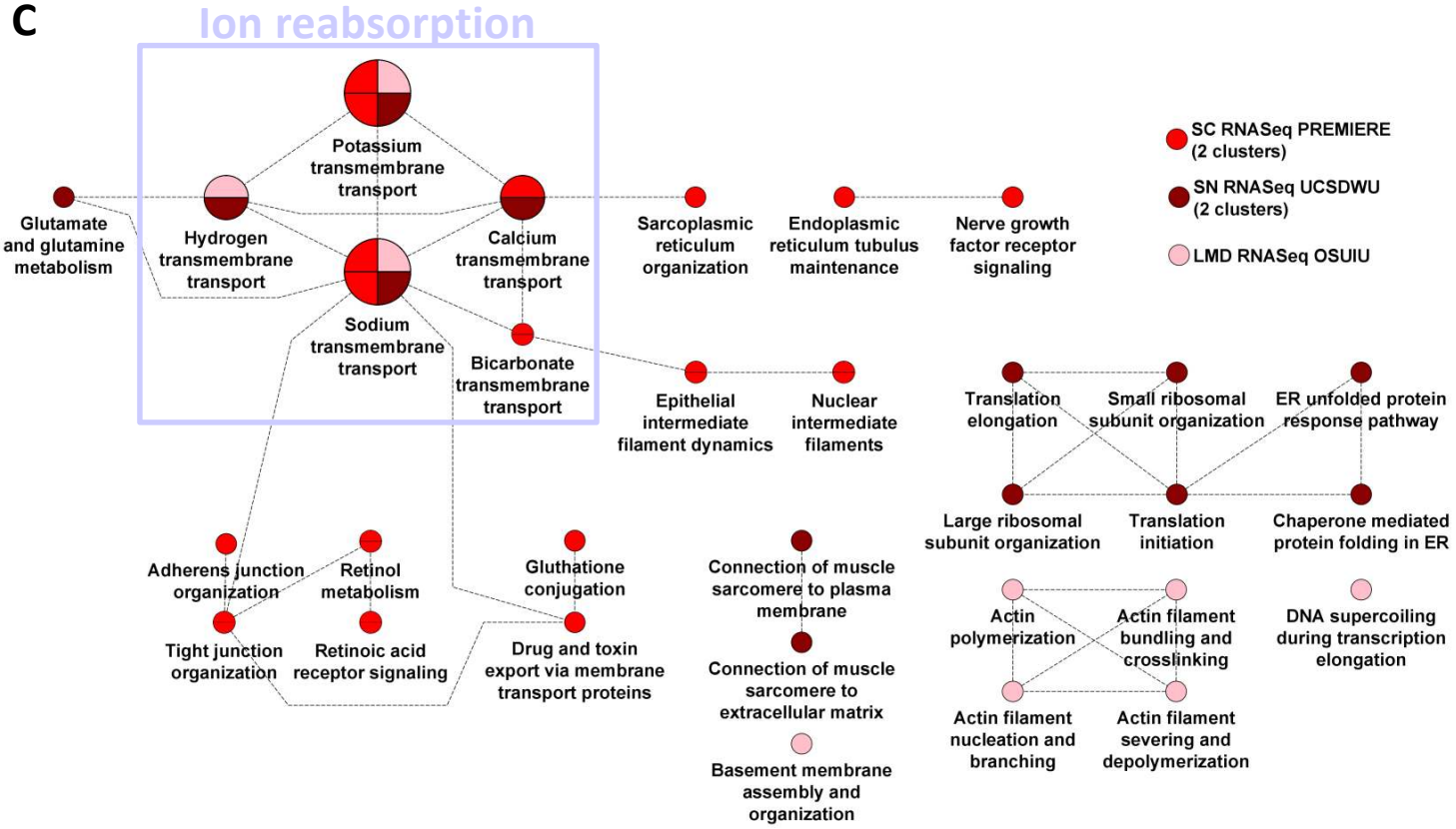
with the norharmane MALDI matrix (Sigma) using the TM-Sprayer automated spraying robot (HTX Technology). Imaging was performed using a MALDI-FTICR imaging mass spectrometer (Bruker Daltonics) set at a 120,000 resolving power at m/z 400 or a MALDI-Orbitrap mass spectrometer (Thermo Scientific) set at the 120,000 resolving power at m/z 200. The data was inspected following the quality control guidelines as developed within KPMP and converted into the imzML centroided format using the SCiLS software (Bruker Daltonics) or ImageInsight software (Spectroglyph, LLC), followed by the submission to METASPACE and annotation against the SwissLipids and HMDB molecular databases with the false discovery rate of 20%, as described in [33].

We have developed an approach to find glomeruli markers in MALDI imaging mass spectrometry data by using METASPACE and co-localization analysis. First, we have selected a template marker that was localized within the glomerular regions, as confirmed by the histology. This ion was annotated by METASPACE as ceramide phosphate CerP(d34:1) [33]. Then, we performed a spatial co-localization analysis by calculating for all other detected metabolites and lipids their spatial correlation with CerP(d34:1) using the cosine score. The molecules with the correlation above 0.2 were considered and manually curated to show the co-localization with the glomeruli regions by overlaying every ion image with the histological image. The resulting 30 markers were uploaded to the KPMP DataLake and were used for the multiomics integration analysis.



Supplementary Figure 1

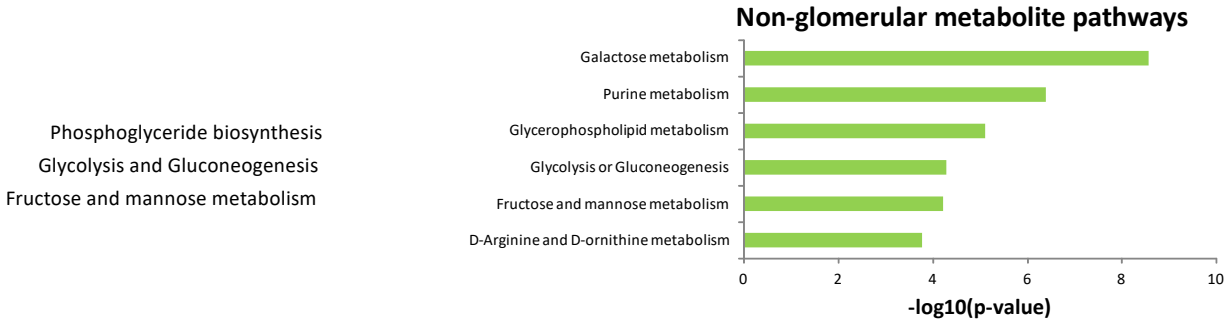
C

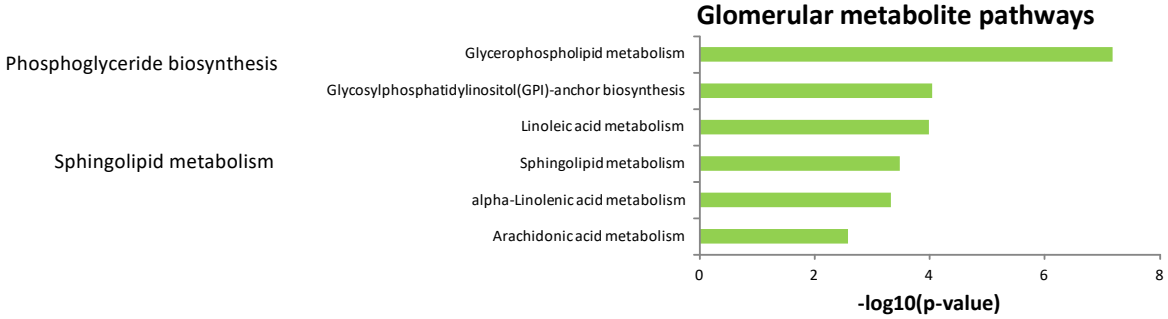


D

MetaboAnalyst – Enrichment of metabolites

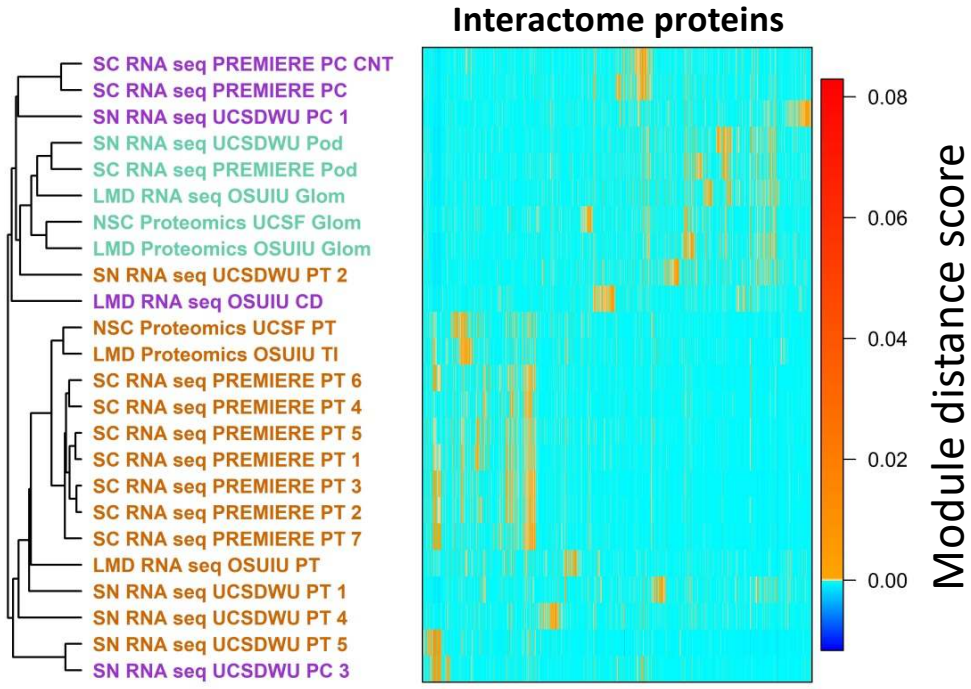
Equivalent MBCO SCPs





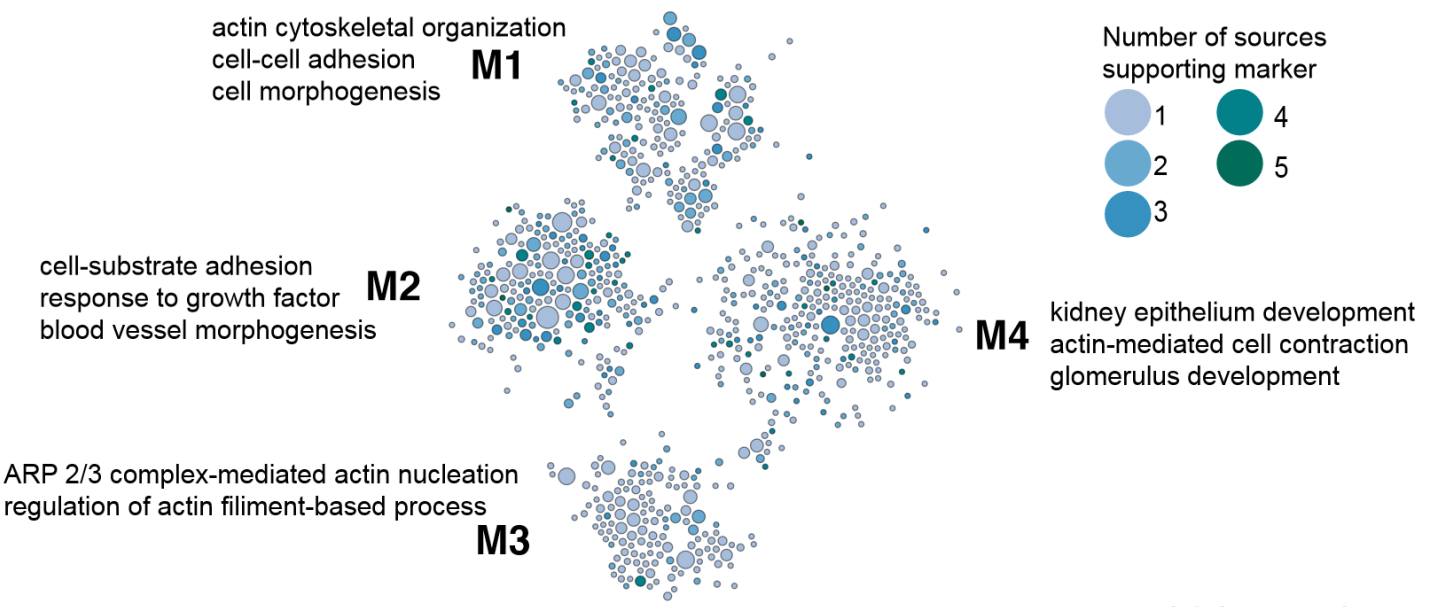
Supplementary Figure 1

E

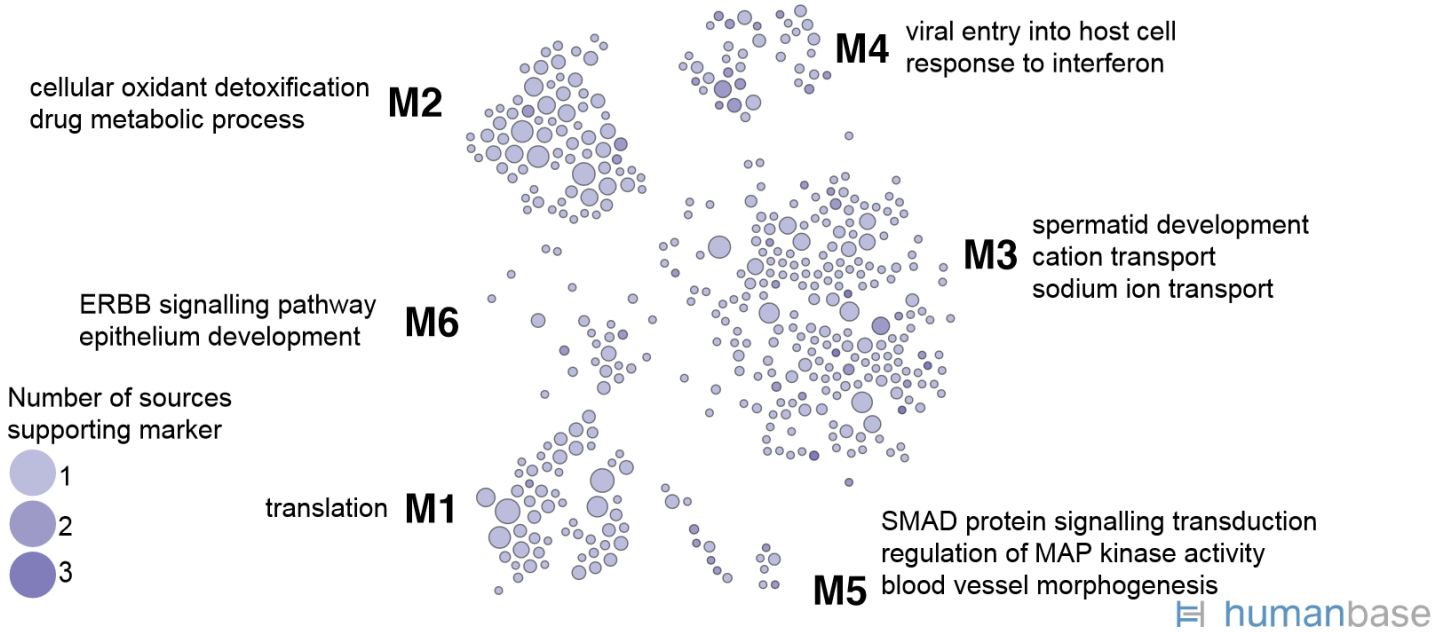


Supplementary Figure 1

F

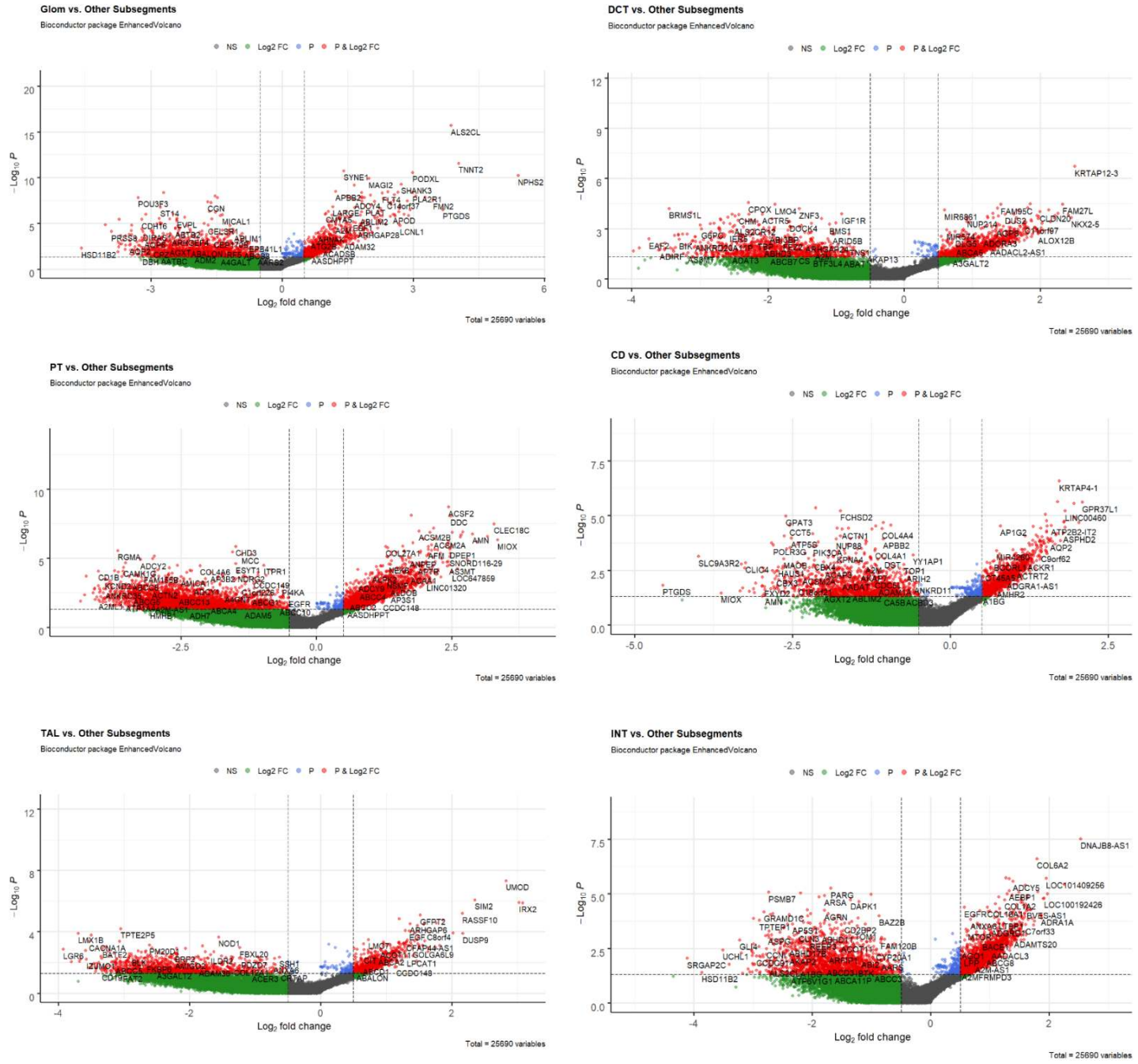


G



Supplementary Figure 1

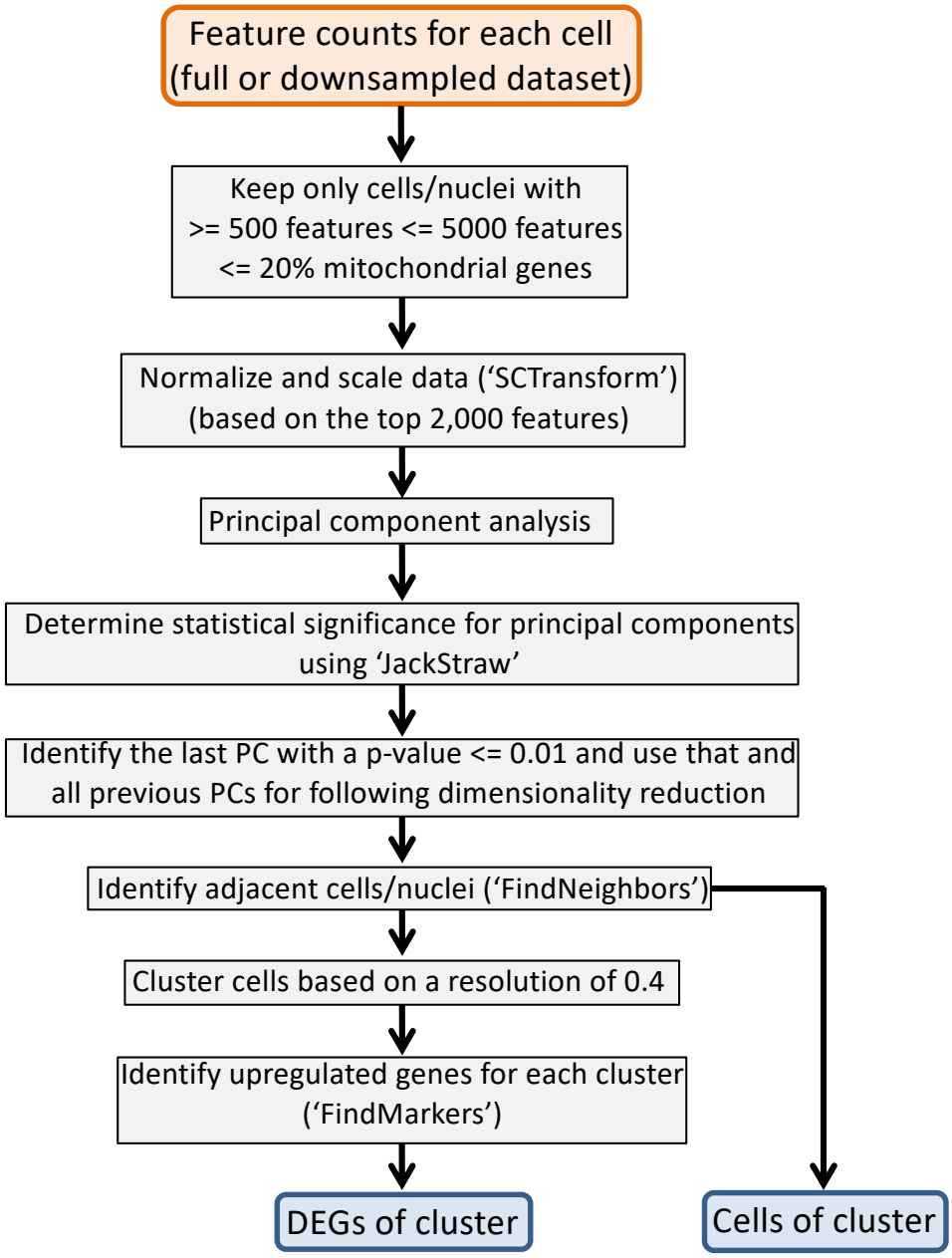
H



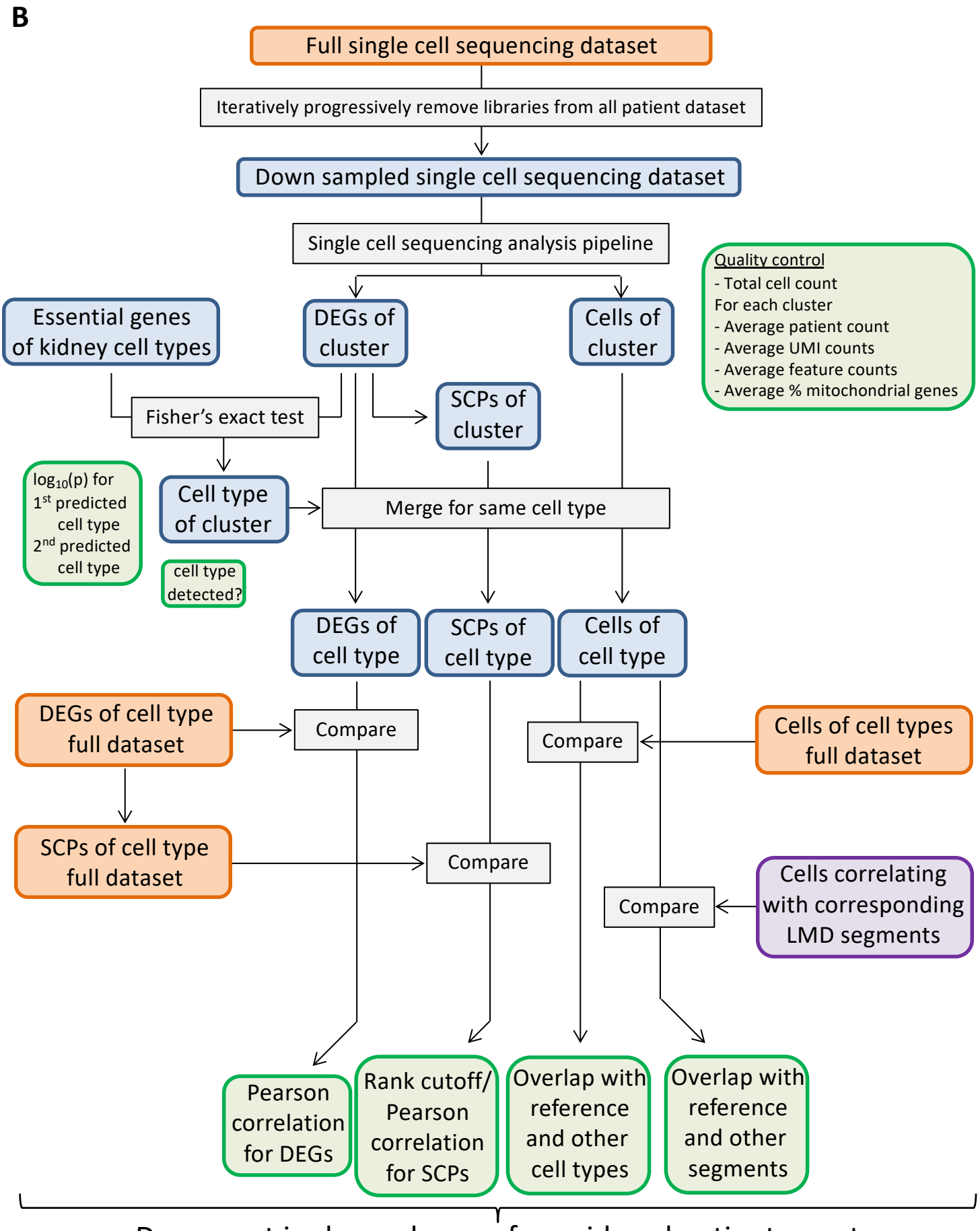
Supplementary Figure 1

Supplementary Figure 1. Top 300 differentially expressed genes (DEGs) and differentially expressed proteins (DEPs) of each proximal tubule/tubulointerstitium, podocyte/glomerular and principal cell/collecting duct cluster or tissue subsegment were subjected to dynamic enrichment analysis using the Molecular Biology of the Cell Ontology (MBCO). SCP-networks were generated from the top seven most significant predictions. **(A)** The predicted proximal tubule SCP-networks contained SCPs that describe basic cell biological functions. SCP-networks that are related to specific proximal tubule cell functions are visualized in figure 2A. **(B)** Podocyte/glomerular SCP-networks almost exclusively focus on SCPs involved in barrier formation, a fundamental prerequisite for glomerular filtration. **(C)** Principal cell/collecting duct networks describe ion reabsorption pathways that document collecting duct function in the fine tuning of ion and water reabsorption for whole body water homeostasis. **(D)** All glomerular and non-glomerular metabolites obtained from the three nephrectomy samples were subjected to pathway enrichment analysis using MetaboAnalyst. Top six predictions were mapped to MBCO level-3 SCPs, if possible. **(E)** The human interactome is a network of ~15,000 proteins that are connected, if they physically interact with each other. To identify if the DEGs and DEPs of each cell type/tissue subsegment map to the same area in the human interactome and consequently interact with the same proteins, we used the top 300 DEGs and DEPs as seed nodes and calculated the Module Distance Score (MDS) for each interactome protein [7]. The MDS is higher the closer a protein is to the seed nodes when compared to all other non-seed nodes. Hierarchical clustering of pairwise correlations between MDS documents that the different technologies identify cell type specific DEGs and DEPs that map to neighboring areas in the human interactome. Red and blue indicate positive and negative module distance scores, respectively. Sample rows and interactome protein columns were rearranged according to the hierarchical clustering results. Modules were generated for proximal tubule cells and subsegments **(F)** and for principal cells and **(G)** collecting duct subsegment, as described in Figure 2C. **(H)** Volcano plots illustrating differentially expressed genes in laser micro-dissected sub-segments for the glomerulus, proximal tubule, thick ascending loop of Henle, distal convoluted tubule, collecting duct, and interstitium without tubules. Comparisons are made between the sub-segment of interest and all other subsegments. The p-value threshold depicted is an unadjusted $p < 0.05$.

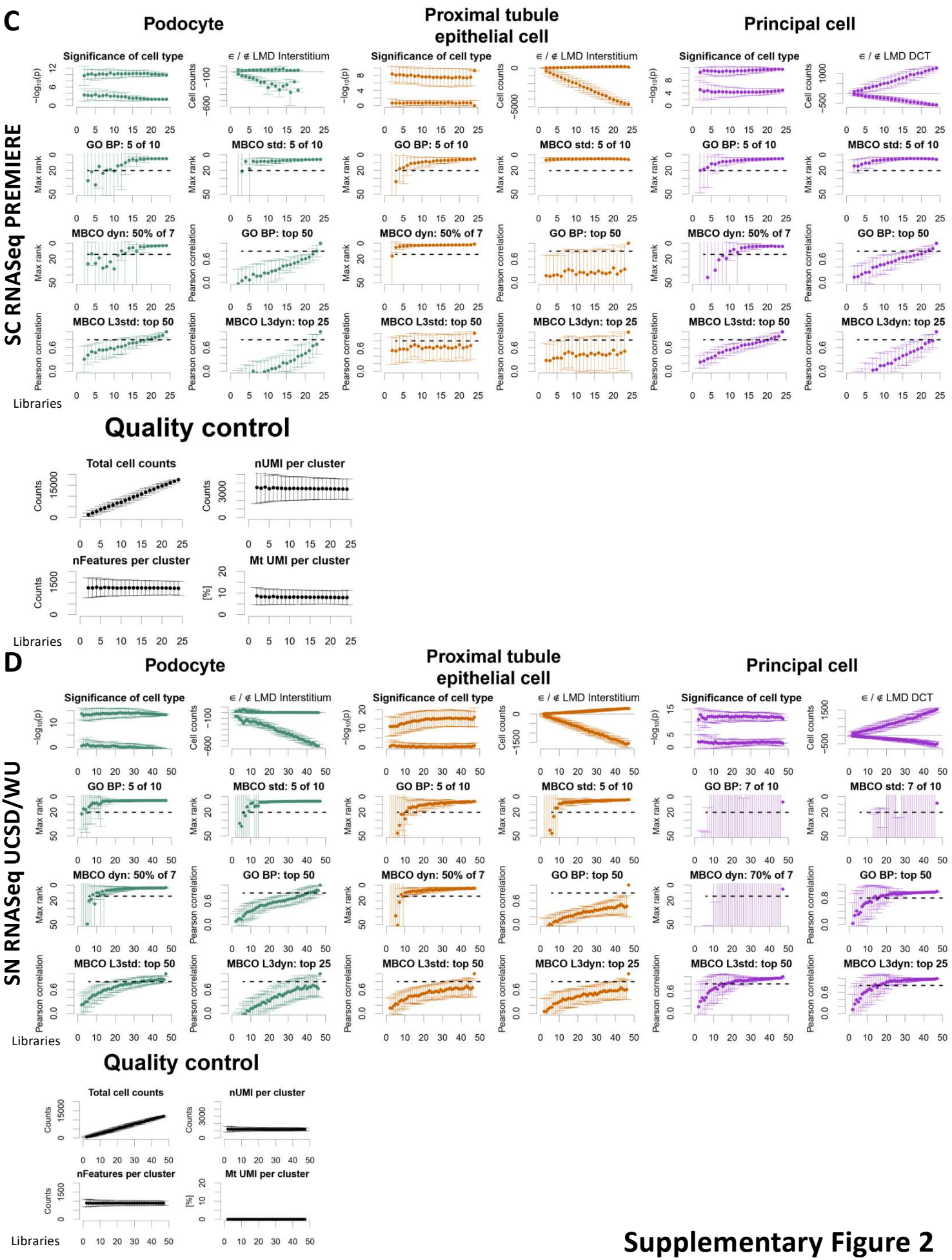
A



Supplementary Figure 2

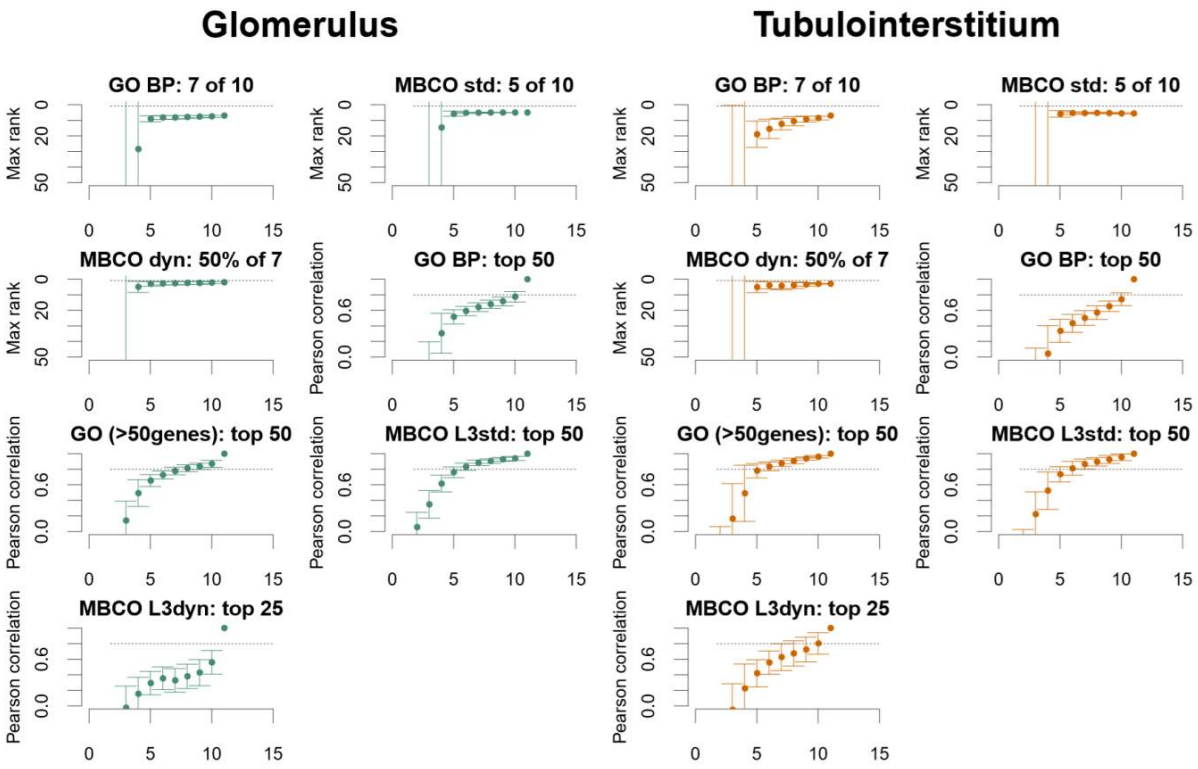


Supplementary Figure 2



Supplementary Figure 2

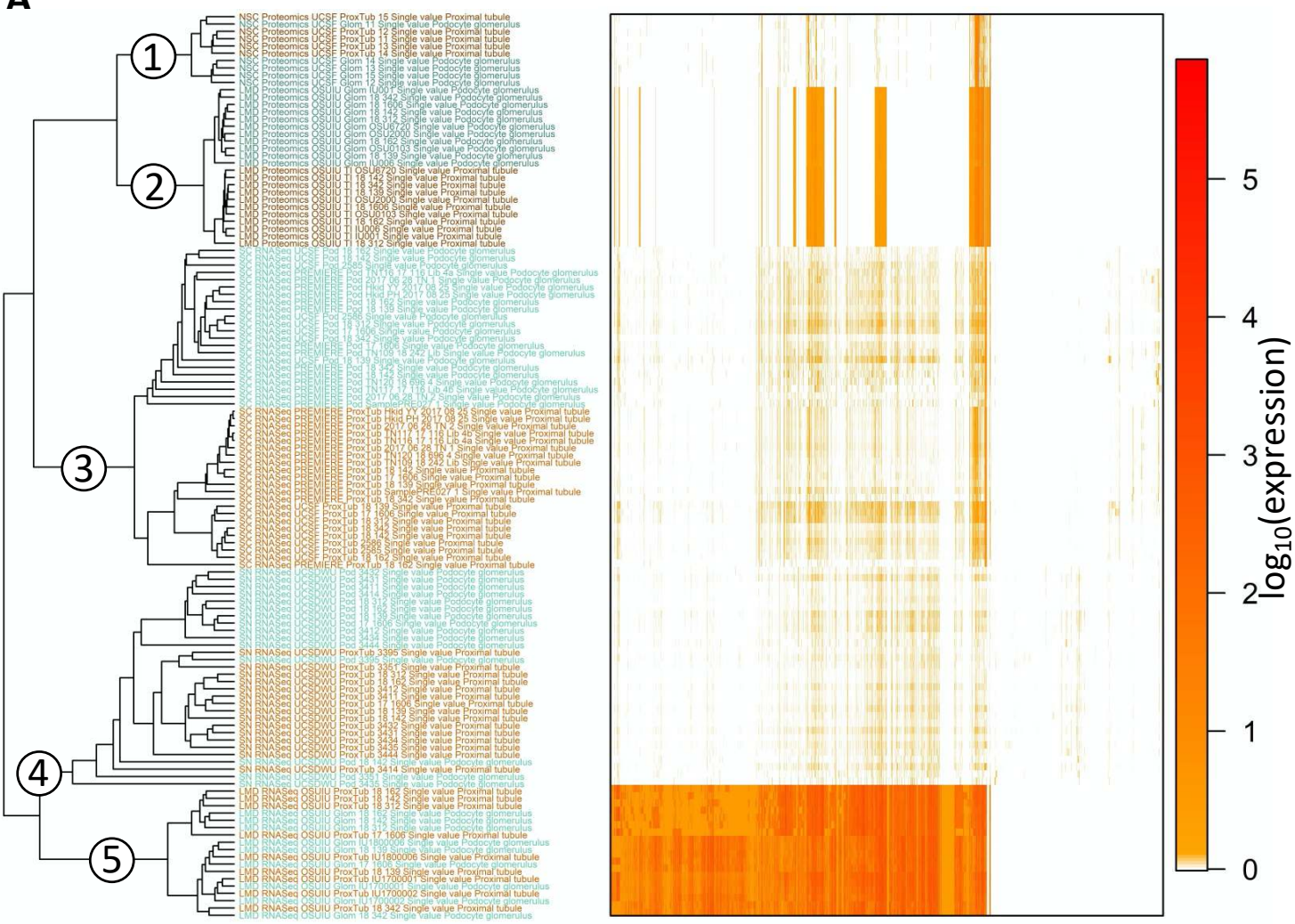
E



Supplementary Figure 2

Supplementary Figure 2. Single-cell transcriptomic post-hoc power analysis. **(A)** Single-cell and single-nucleus sequencing analytical pipeline **(B)** and downstream post-processing analytical pipeline used for post-hoc power analysis. Single-cell clusters were assigned to cell types based on the significance of the overlap between cluster-specific DEGs and cell type specific essential genes as determined by Fisher's exact test. Complete post-hoc power analysis profile for **(C)** PREMIERE single-cell, **(D)** UCSD single-nucleus, and **(E)** LMD proteomics. Differences between the minus $\log_{10}$(p-values) of the most significant and second most significant overlaps document high certainty of cell type assignments. $\in/\notin$ LMD subsegment shows a how many cells map (above abscissa, positive values) or do not map (below abscissa, negative values) to different subsegments than the ones shown in figure 4. Rank cutoffs were generated as described within Figure 4. Pearson correlations between down-sampled datasets and complete (reference) datasets based on the top-50 GO Biological Processes (with or without removal of all processes with less than 50 annotated genes) and the top-50 MBCO level-3 SCPs. For dynamic enrichment analysis all SCPs among the top 25 predictions were compared. Since an SCP can be predicted as part of multiple SCP combinations or as a single SCP, its p-value was assigned to be the p-value of the most significant prediction. RNASeq quality control measures include the number of total cells in each analysis as well as the number of unique molecular identifiers (UMIs), features and the percentage of expressed mitochondrial (MT) genes in each cluster.

A



1) NSC Proteomics USCF
2) LMD Proteomics OSUIU
3) SC RNASeq PREMIERE/UCSF
4) SN RNASeq UCSD\WU
5) LMD RNASeq OSUIU

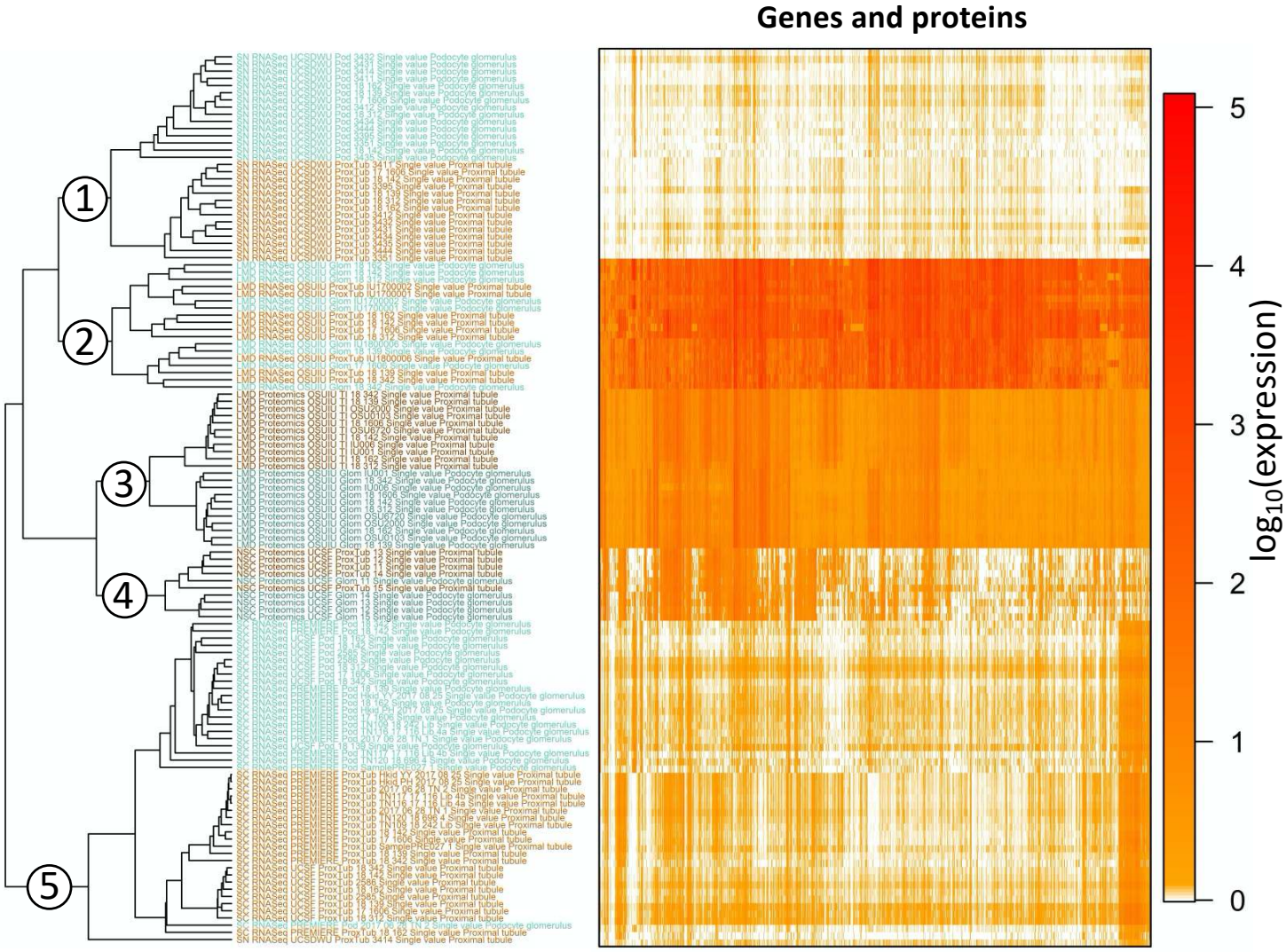
Proximal tubule

Glomerulus

Podocyte

Supplementary Figure 3

B



1) SN RNASeq UCSD\WU
2) LMD RNASeq OSUIU
3) LMD Proteomics OSUIU
4) NSC Proteomics USCF
5) SC RNASeq PREMIERE/UCSF

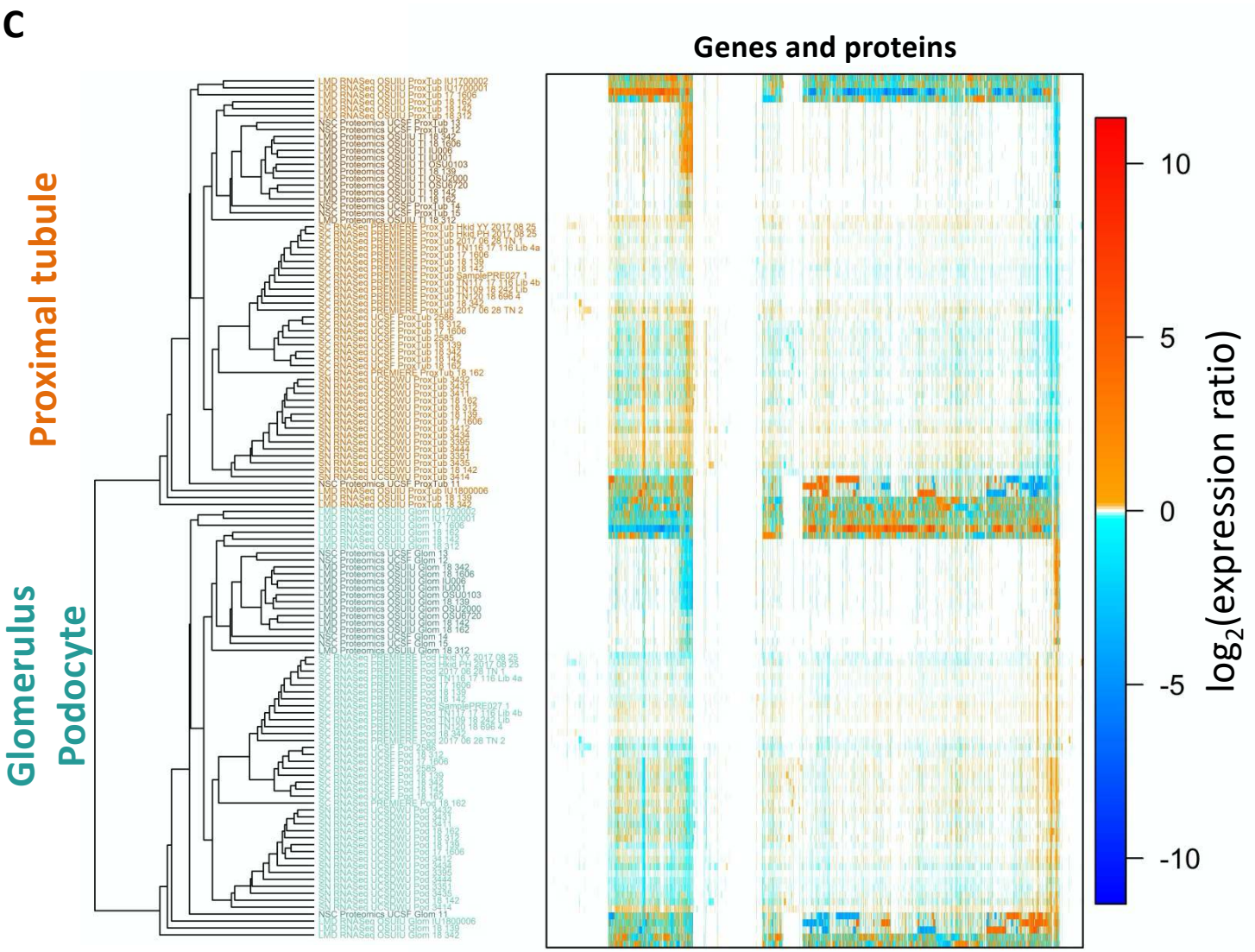
Proximal tubule

Glomerulus

Podocyte

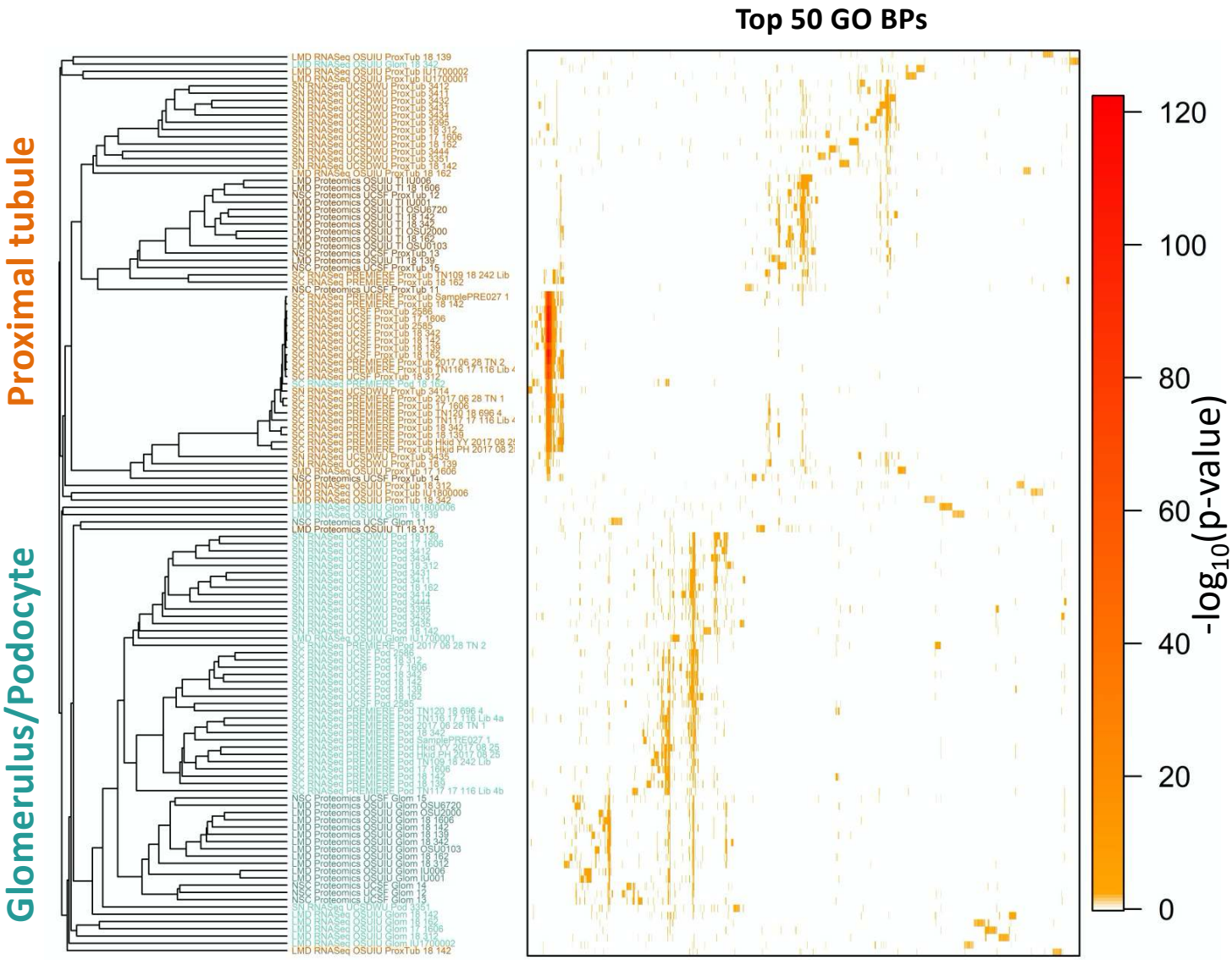
Supplementary Figure 3

C



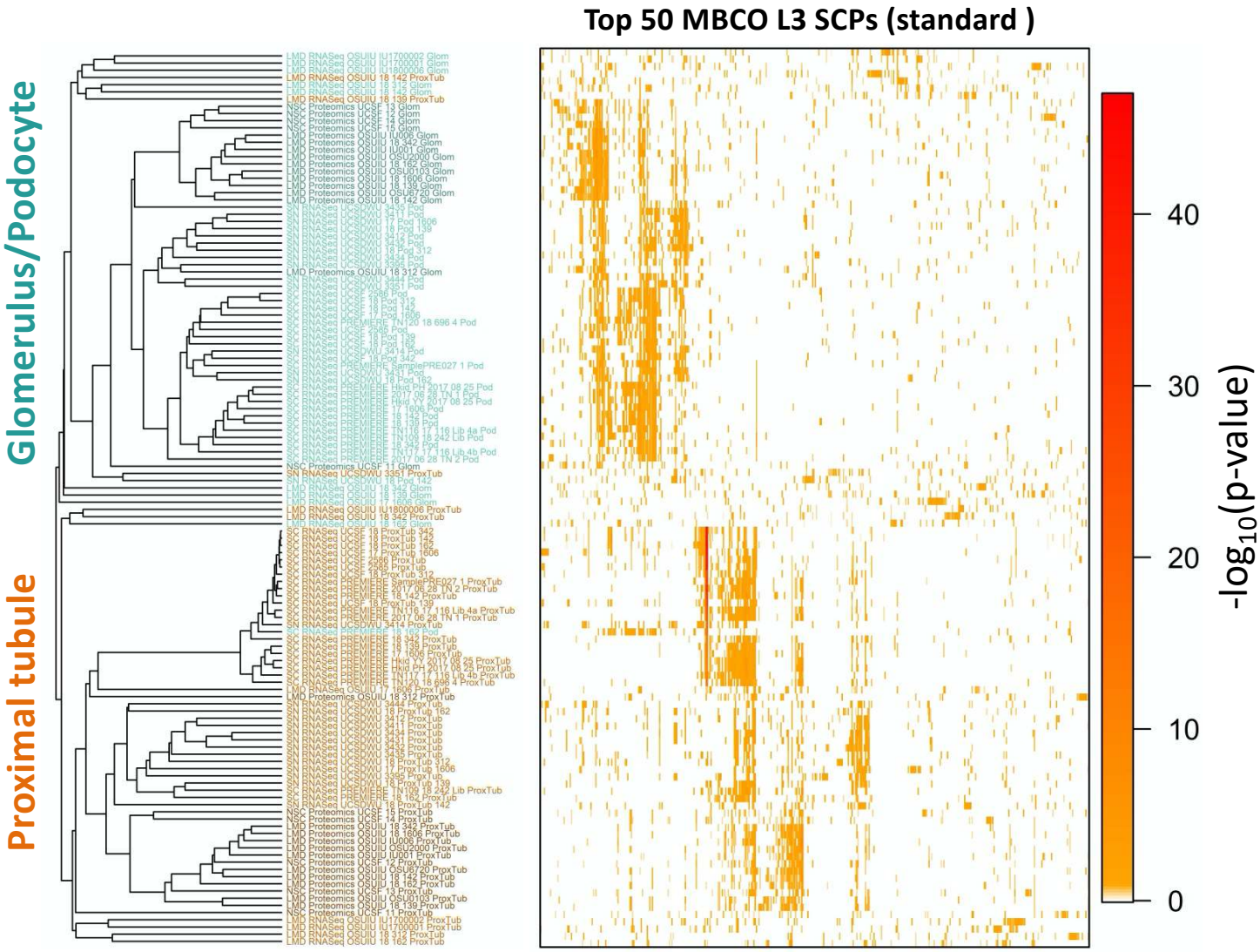
Supplementary Figure 3

D



Supplementary Figure 3

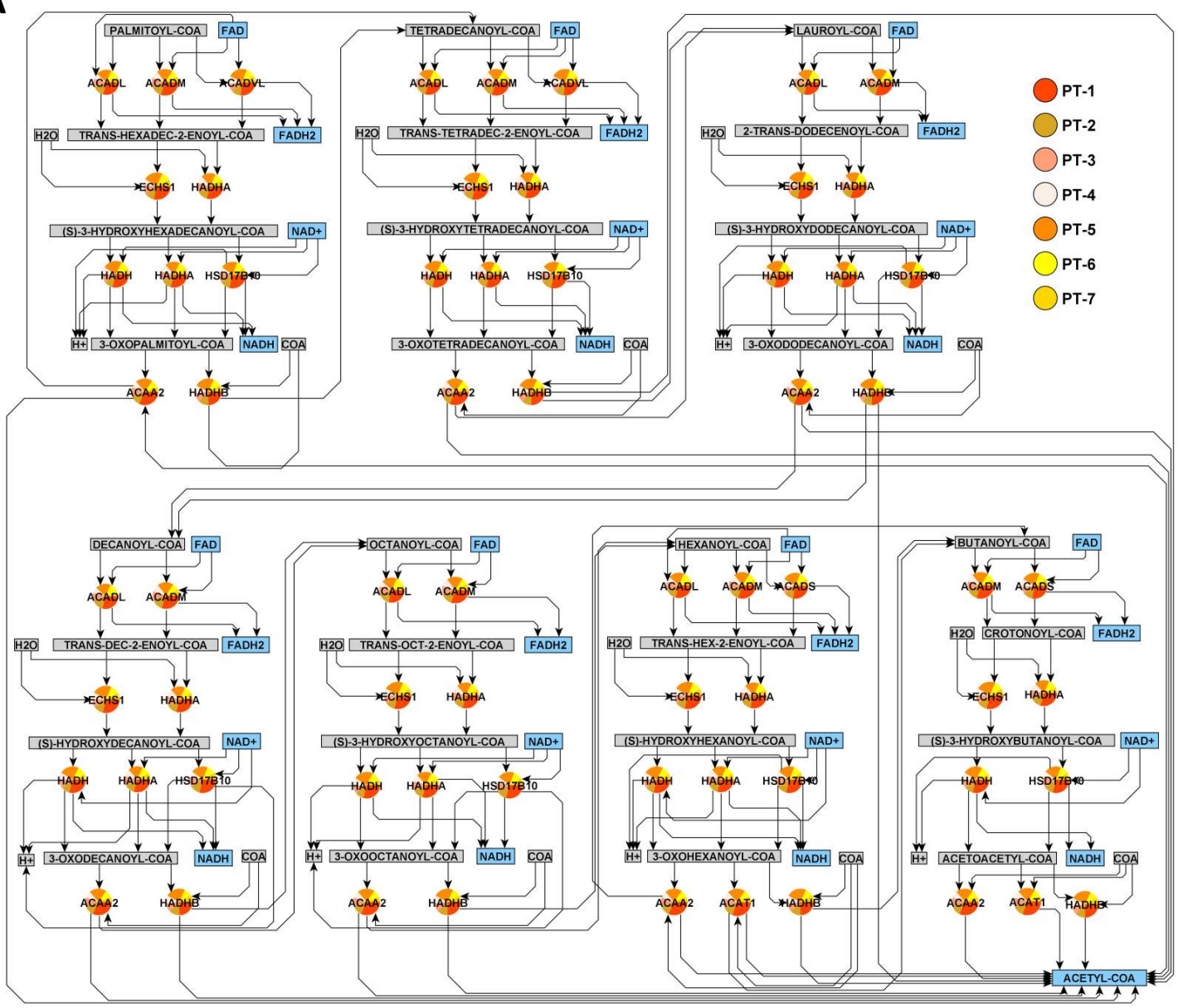
E



Supplementary Figure 3

Supplementary Figure 3. Correlation analysis between the different omics technologies. Pairwise correlation coefficients of **(A)** absolute gene/protein expression in podocyte/glomerulus and proximal tubule/tubulointerstitial hierarchically cluster by technology (and not by anatomical region), **(B)** independently if the full dataset or **(C)** a dataset that contains only shared genes and proteins. However, considering the expression ratio of each gene/protein versus the other anatomical region enables clustering by anatomical region, as shown in Figure 5A, even if all genes/proteins are considered. Top 300 DEGs or DEPs were subjected to pathway enrichment analysis and **(D)** the top-50 GO BPs and **(E)** MBCO level-3 SCPs subjected to hierarchical clustering based on pairwise correlation coefficients between $-\log_{10}$(p-values). All heatmaps show either logarithmized absolute gene/protein expression values of each sample, $\log_2$(fold changes) or minus $\log_{10}$(p-values). White spots indicate undetected genes/proteins/SCPs or genes and proteins that are equally expressed in both cell types/kidney subsegments. Sample rows and gene/protein columns were rearranged according to the hierarchical clustering results.

A



B

	PT-2	PT-1	PT-6	PT-3	PT-5	PT-4	PT-7
ACADL	0.04	0.09	0.04	0.03	0.06	0.02	0.02
ACADM	1.03	2.15	0.64	0.75	1.39	0.67	0.5
ACADVL	0.35	0.85	0.36	0.33	0.74	0.31	0.25
ACADS	0.13	0.31	0.11	0.1	0.25	0.11	0.1
ECHS1	2.42	5.96	2.19	2.14	4.74	1.68	1.69
HADH	0.49	1.4	0.54	0.51	1.13	0.43	0.31
HADHA	0.24	0.56	0.2	0.23	0.35	0.18	0.15
HSD17B10	0.5	1.17	0.43	0.46	0.82	0.32	0.35
ACAA2	0.63	1.61	0.48	0.56	1.05	0.42	0.3
HADHB	0.37	0.74	0.3	0.33	0.72	0.29	0.24
ACAT1	1.45	3.48	1.05	1.23	2.35	1.01	0.74

Supplementary Figure 4

Supplementary Figure 4. Proximal tubule cell subtype specific beta oxidation. **(A)** Full beta oxidation pathway map containing the relative enzyme expression in each proximal tubule cell subtype. **(B)** Mean UMI counts of each enzyme in each cell subtype.

SUPPLEMENTARY TABLES

Supplementary Table 1. Samples used for different analytical pipelines.

Supplementary Table 2. Laser microdissected (LMD) RNASeq gene expression

Supplementary Table 3. Laser microdissected (LMD) Proteomics protein expression

Supplementary Table 4. Near Single Cell (NSC) Proteomics protein expression

Supplementary Table 5. Spatial metabolomics metabolite correlations for participants 18-139 (A), 18-142 (B) and 18-342 (C).

Supplementary Table 6. Top 300 differentially expressed genes (DEGs) and proteins (DEPs) predicted by each assay for each analyzed cell type/tissue subsegment.

Supplementary Table 7. Dynamic enrichment analysis results of the top 300 DEGs and DEPs.

Supplementary Table 8. Gene Ontology enrichments for modules in the kidney-specific functional network of top DEGs and DEPs in proximal tubules, podocytes, and principal cells.

Supplementary Table 9. Literature curated cell-type specific essential genes used for cell type identification.

Data Availability

All raw and processed data described in this manuscript is available through the KPMP Data Portal at kpmp.org.

Acknowledgements

Kidney Precision Medicine Project acknowledges the all the participants, patients, and the scientific officers from National Institute of Diabetes and Digestive and Kidney Diseases. KPMP was supported by NIH grants UH3 DK114923, UH3 DK114920, UH3 DK114933, UH3 DK114937, UH3 DK114907 and U2C DK114886. A complete list of all KPMP members can be found at kpmp.org. We thank Joseph Goldfarb for critically reading of the manuscript.

Author contributions

Integrated analysis and interpretation: Jens Hansen[1,*], Rachel Sealfon[2,*], Rajasree Menon[3,*], John Cijiang He[1], Jonathan Himmelfarb[11], Olga G. Troyanskaya[2,#], Matthias Kretzler[3,#], Ravi Iyengar[1,#], Evren U. Azeloglu[1,#]

Pilot tissue procurement, data coordination and metadata curation: Becky Steck[3], Abhijit Nair[3], Jeffrey B. Hodgin[3], Matthias Kretzler[3,#], Evren U. Azeloglu[1,#]

SC/SN RNASeq data generation and processing: Rajasree Menon[3,*], Blue B. Lake[5], Andrew Schroeder[6], Edgar A. Otto[3], Jeffrey B. Hodgin[3], Minnie Sarwal[6], Kun Zhang[5], Sanjay Jain[12], Matthias Kretzler[3,#], Jens Hansen[1,*]

LMD transcriptomics data generation and processing: Michael T. Eadon[4], Daria Barwinska[4], Pierre C. Dagher[4]

LMD and NSC proteomic data generation and processing: Samir Parikh[7], John P. Shapiro[7], Tara K. Sidgel[6], Minnie Sarwal[6], Brad Rovin[7]

Imaging data generation and processing: Dejan Dobi[6], Seth Winfree[4], Tarek M. El-Achkar[4], Zoltan Laszik[6]

Spatial metabolomics data generation and processing: Theodore Alexandrov[8], Dusan Velickovic[9], Christopher R. Anderton[9,10], Guanshi Zheng[10], Annapurna Pamreddy[10], Kumar Sharma[10]

Manuscript preparation: Jens Hansen[1,*], Rachel Sealfon[2,*], Rajasree Menon[3,*], Michael P. Rose[3], Yongqun He[3], Ian H. de Boer[11], Olga G. Troyanskaya[2,#], Matthias Kretzler[3,#], Ravi Iyengar[1,#], Evren U. Azeloglu[1,#]. All authors commented and edited the manuscript and assisted in the assembly of the final version.

Disclosure of Financial Interests

None.

REFERENCES

1. Strippoli GF, Craig JC, Schena FP. The number, quality, and coverage of randomized controlled trials in nephrology. *Journal of the American Society of Nephrology : JASN* 2004; **15:** 411-419.

2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017; **18:** 83.

3. Yan J, Risacher SL, Shen L*, et al.* Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform* 2018; **19:** 1370-1381.

4. Shi WJ, Zhuang Y, Russell PH*, et al.* Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics* 2019; **35:** 4336-4343.

5. Bersanelli M, Mosca E, Remondini D*, et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016; **17 Suppl 2:** 15.

6. Hansen J, Meretzky D, Woldesenbet S*, et al.* A flexible ontology for inference of emergent whole cell function from relationships between subcellular processes. *Sci Rep* 2017; **7:** 17689.

7. Berger SI, Ma'ayan A, Iyengar R. Systems pharmacology of arrhythmias. *Sci Signal* 2010; **3:** ra30.

8. Menon R, Otto EA, Hoover P*, et al.* Single cell transcriptomics identifies focal segmental glomerulosclerosis remission endothelial biomarker. *JCI Insight* 2020; **5**.

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

10. Bhargava P, Schnellmann RG. Mitochondrial energetics in the kidney. *Nat Rev Nephrol* 2017; **13:** 629-646.

11. Wang K, Kestenbaum B. Proximal Tubular Secretory Clearance: A Neglected Partner of Kidney Function. *Clin J Am Soc Nephrol* 2018; **13:** 1291-1296.

12. van Swelm RPL, Wetzels JFM, Swinkels DW. The multifaceted role of iron in renal health and disease. *Nat Rev Nephrol* 2020; **16:** 77-98.

13. Zarjou A, Bolisetty S, Joseph R*, et al.* Proximal tubule H-ferritin mediates iron trafficking in acute kidney injury. *J Clin Invest* 2013; **123:** 4423-4434.

14. Eich C, Manzo C, de Keijzer S*, et al.* Changes in membrane sphingolipid composition modulate dynamics and adhesion of integrin nanoclusters. *Sci Rep* 2016; **6:** 20693.

15. Dar MI, Jan S, Reddy GL*, et al.* Differentiation of human neuroblastoma cell line IMR-32 by sildenafil and its newly discovered analogue IS00384. *Cell Signal* 2020; **65:** 109425.

16. Li J, Li Q, Geng S. All-trans retinoic acid alters the expression of the tight junction proteins Claudin-1 and -4 and epidermal barrier function-associated genes in the epidermis. *Int J Mol Med* 2019; **43:** 1789-1805.

17. Mallipattu SK, He JC. The beneficial role of retinoids in glomerular disease. *Front Med (Lausanne)* 2015; **2:** 16.

18. Meyer C, Nadkarni V, Stumvoll M*, et al.* Human kidney free fatty acid and glucose uptake: evidence for a renal glucose-fatty acid cycle. *Am J Physiol* 1997; **273:** E650-654.

19. Kang HM, Ahn SH, Choi P*, et al.* Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. *Nature medicine* 2015; **21:** 37-46.

20. Greene CS, Krishnan A, Wong AK*, et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015; **47:** 569-576.

21. Krishnan A, Zhang R, Yao V*, et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* 2016; **19:** 1454-1462.

22. Koplev S, Lin K, Dohlman AB*, et al.* Integration of pan-cancer transcriptomics with RPPA proteomics reveals mechanisms of epithelial-mesenchymal transition. *PLoS Comput Biol* 2018; **14:** e1005911.

23. Schwanhausser B, Busse D, Li N*, et al.* Global quantification of mammalian gene expression control. *Nature* 2011; **473:** 337-342.

24. Calizo RC, Bhattacharya S, van Hasselt JGC*, et al.* Disruption of podocyte cytoskeletal biomechanics by dasatinib leads to nephrotoxicity. *Nat Commun* 2019; **10:** 2061.

25. Xia J, Psychogios N, Young N*, et al.* MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* 2009; **37:** W652-660.

26. Stuart T, Butler A, Hoffman P*, et al.* Comprehensive Integration of Single-Cell Data. *Cell* 2019; **177:** 1888-1902.e1821.

27. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* 2019; **8:** 329-337.e324.

28. Binder JX, Pletscher-Frankild S, Tsafou K*, et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014; **2014:** bau012.

29. Satoskar AA, Shapiro JP, Bott CN*, et al.* Characterization of glomerular diseases using proteomic analysis of laser capture microdissected glomeruli. *Mod Pathol* 2012; **25:** 709-721.

30. Shapiro JP, Biswas S, Merchant AS*, et al.* A quantitative proteomic workflow for characterization of frozen clinical biopsies: laser capture microdissection coupled with label-free mass spectrometry. *J Proteomics* 2012; **77:** 433-440.

31. Winfree S, Khan S, Micanovic R*, et al.* Quantitative Three-Dimensional Tissue Cytometry to Study Kidney Tissue and Resident Immune Cells. *J Am Soc Nephrol* 2017; **28:** 2108-2118.

32. Goltsev Y, Samusik N, Kennedy-Darling J*, et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 2018; **174:** 968-981 e915.

33. Palmer A, Phapale P, Chernyavsky I*, et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Methods* 2017; **14:** 57-60.

34. Edison Ong, Lucy L. Wang, Jennifer Schaub, John F. O'Toole, Becky Steck, Avi Z. Rosenberg, Frederick Dowd, Jens Hansen, Laura Barisoni, Sanjay Jain, Ian H. de Boer, M. Todd Valerius, Sushrut S. Waikar, Christopher Park, Dana C. Crawford, Theodore Alexandrov, Christopher R. Anderton, Christian Stoeckert, Chunhua Weng, Alexander D. Diehl, Christopher J. Mungall, Melissa Haendel, Peter N. Robinson, Jonathan Himmelfarb, Ravi Iyengar, Matthias Kretzler, Sean Mooney, Yongqun He and the Kidney Precision Medicine Project. Modeling Kidney Disease Using Ontology: Perspectives from the KPMP. *Nature Nephrology in press.*