Defining cell-type specificity at the transcriptional level in human disease

Wenjun Ju, Casey S Greene, Felix Eichinger, et al.

Genome Res. published online August 15, 2013
Access the most recent version at doi:10.1101/gr.155697.113

Defining cell-type specificity at the transcriptional level in human disease

Wenjun Ju[1,2#], Casey S. Greene[3,4,5#], Felix Eichinger[1], Viji Nair[1], Jeffery B. Hodgin[1], Markus Bitzer[1], Young-suk Lee[3,4], Qian Zhu[3,4], Masami Kehata[6], Min Li[6], Song Jiang[1], Maria Pia Rastaldi[6], Clemens D. Cohen[7], Olga G. Troyanskaya[3,4]*, and Matthias Kretzler[1,2]*

[1]Division of Nephrology, Department of Internal Medicine, [2]Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48105, USA. [3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, 08544, USA. [4]Department of Computer Science, Princeton University, Princeton, NJ, 08540, USA. [5]Department of Genetics, The Geisel School of Medicine at Dartmouth, Hanover, NH, 03755, USA. [6]Laboratorio di Ricerca Nefrologica, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy. [7]Division of Nephrology, University of Zurich, Zurich, Switzerland.

[#]These authors contributed equally to the work

*To whom correspondence should be addressed. Email: kretzler@umich.edu (M.K.); ogt@genomics.princeton.edu (O.G.T.)

Abstract

Cell-lineage-specific transcripts, i.e. those with expression restricted to a limited set of cell types, are essential for differentiated tissue function, mediate acquired chronic diseases, and are implicated in hereditary organ failure. However, experimental identification of cell-lineage specific genes in a genome-scale manner is infeasible for most solid human tissues. To address this challenge, we developed the first genome-scale computational method to identify genes with cell-lineage-specific expression, even in lineages not separable by experimental micro-dissection. Our machine learning-based approach leverages high-throughput functional-genomics data from tissue homogenates in a novel iterative statistical framework. We applied this method to chronic kidney disease and identified transcripts specific to podocytes, key cells in the glomerular filter responsible for hereditary and most acquired glomerular kidney disease. We systematically evaluated our predictions by immunohistochemistry, verifying selective tissue-distribution of podocyte-specific genes within the kidney. Our in silico approach was significantly more accurate (65% accuracy in human) than predictions based on direct measurement of in vivo fluorescence-tagged murine podocytes (23% accuracy in human). In agreement with the hypothesis that cell-lineage specific transcripts are likely to be associated with disease, our method identified genes implicated as causal in hereditary glomerular disease and involved in molecular pathways of acquired and chronic renal diseases. Furthermore, based on expression analysis of human kidney disease biopsy samples, we demonstrated that expression of the podocyte genes identified by our approach is significantly related to the degree of renal impairment in patients.

Our approach is general and broadly applicable to define lineage specificity in both cell physiology and human disease contexts. We provide a user-friendly website that enables researchers to easily apply this method to any cell-lineage or tissue of interest. Identified cell-lineage specific transcripts are expected to play essential tissue-specific roles in organogenesis and disease and can provide starting points for the development of organ-specific diagnostics and therapies.

Introduction

Cell-lineage differentiation plays a defining role in biology. Impairment of differentiated cell functions is responsible for the organ specific manifestation of acquired chronic degenerating diseases including Alzheimer's disease, diabetes, and chronic kidney disease (CKD). Defining lineage-specific cellular function in human physiology and disease remains challenging, as it is frequently impossible to physically isolate a specific cell lineage from the heterogeneous lineages that make up many solid human tissues. This inability to obtain pure cell preparation from human tissue *in vivo* and to identify the functional context of these cell lineages on a genome-scale is a significant barrier to developing an understanding of molecular interactions in complex tissues and diseases.

Here we develop a computational approach (Fig. 1) that identifies genes specifically expressed in a cell lineage from high-throughput expression data of complex solid tissue biopsies. This problem is of significant biological and clinical relevance; obtaining a pure *ex vivo* cell population of sufficient size from the lineage of interest for direct assay is often technically infeasible, particularly when the lineage of interest is a component of solid tissues. This challenge imposes severe limitations on researchers' ability to account for the cell-lineage specific expression and function of most human genes. This problem is distinct from the task of identifying the fractional composition of a heterogeneous sample (e.g. whole blood), and methods to address such problems require whole-genome expression measurements for each underlying cell type, which are unavailable for most solid human cell lineages (Shen-Orr et al. 2010). Our iterative machine-learning-based approach leverages heterogeneous expression data from human tissue homogenates. We term this approach "*in silico* nano-dissection" because it is a computational approach that can analyze lineages that are not separable by experimental micro-dissection.

Chronic diseases like diabetes and hypertension cause morbidity and mortality via alteration of differentiated organ function in a wide range of tissues. Here we use kidney disease as a proof of concept application, focusing on the podocytes, the highly differentiated glomerular epithelial cells responsible for most hereditary and acquired glomerular disease (Groop et al. 2009; Gerstein 2001; D'Agati et al. 2011; Niewold 2011; Roselli et al. 2004; Kim et al. 2003). As with

most other differentiated cell lineages in solid tissues, discovering human podocyte-specific genes on a whole-genome-scale has remained infeasible due to the challenge of obtaining pure *ex vivo* populations of sufficient size for high-throughput evaluation, making this important cell lineage an ideal proof of concept application for nano-dissection. Restricted gene expression is defined in this study as podocyte 'specific' within the renal context if it shows gene expression limited to podocytes within the kidney, and podocyte specific within the renal glomerulus if it is expressed only in podocytes within the glomeruli, but detectable in other extra-glomerular cell lineages in the kidney. Previous high-throughput strategies have relied on mouse (Endlich et al. 2002; Jain et al. 2011; Brunskill et al. 2011) or human (Saleem et al. 2008) immortalized glomerular visceral epithelial cells, but *in vitro* culture leads to rapid loss of both lineage-specific phenotypes and lineage-specific gene expression. Whole tissue based molecular profiles of renal disease are attainable (Woroniecka et al. 2011; Schmid et al. 2006; Lindenmeyer et al. 2010; Ju et al. 2009; Hodgin et al. 2010; Higgins et al. 2004; Henger et al. 2004; Bennett et al. 2007) since human renal tissue is routinely obtained by diagnostic fine needle biopsy, but whole-tissue expression profiles have not previously been capable of identifying gene expression at the cell lineage level. This difficulty is not unique to renal disease. Similar challenges exist for other clinically important lineages, e.g. neuronal-cell-lineage specific markers in neurodegenerative diseases like Alzheimer's disease or Multiple Sclerosis. Employing a computational approach to identify cell-lineage-specific molecules for non-invasive monitoring of neuronal functional status would help to address one of the key challenges pursued in the study of such diseases (Reddy et al. 2011).

When applied to renal gene expression data sets, our nano-dissection method predicts 136 genes not previously known to be podocyte specific. Through systematic immunohistochemistry-based evaluation we show that our iterative *in silico* method significantly outperforms experimental strategies using fluorescence-activated cell sorting (FACS) separated GFP-tagged murine cells for identification of cell-lineage-enriched transcripts. We further demonstrate that expression of the nano-dissection predicted podocyte-specific genes significantly correlate with kidney function, as measured by the glomerular filtration rate (GFR), in patients with CKD. The nano-dissection method also predicts the most recently identified gene responsible for hereditary nephrotic syndrome (Mele et al. 2011). These findings reinforce the concept that defining cell-

lineage-specific genes can provide important insights into the pathogenesis of and targeted therapies for degenerative human disease of the kidney, central nervous system and other highly differentiated tissues. Our approach is freely available in a user-friendly website that allows researchers to easily explore any cell lineage or tissue of interest (http://nano.princeton.edu) and through an open-source C++ library (http://libsleipnir.bitbucket.org).

Results

In silico nano-dissection approach discovers cell-lineage-specific genes

To discover cell-lineage-specific genes, we developed *in silico* nano-dissection, an iterative computational approach that predicts cell-lineage-specific expression of human genes using high-throughput genomic expression data derived from tissue homogenates. This method uses an iterative machine-learning framework that makes robust predictions, even when only limited prior knowledge about cell-lineage-specific markers is available (detailed description of the approach is provided in the Methods section below). Intuitively, our approach discovers patterns of co-expression of the cell-lineage markers in whole-tissue homogenates from a variety of genetic backgrounds, physiological and pathophysiological states. The approach leverages human curated markers of the cell lineage of interest (podocyte in this case) (SI Table 1 & 2) to identify the genetic or pathophysiological perturbations in which the expression patterns of these markers are predictive of their cell-lineage specificity. These patterns of informative conditions are identified from comprehensive transcriptional datasets derived from tissue homogenates, often represent only a small fraction of all data, and are likely reflective of the markers' biological functions. The condition-specific patterns are then used to identify additional cell-lineage-specific genes. Our approach uses an iterative algorithm to refine the weighting of informative perturbations in a manner robust to the limited availability of curated markers (gold standards). Each gold standard provides differential specificity (e.g. those based on double immunofluorescence, immunohistochemical staining, or RNA abundance) and quality. With our strategy, standards are assessed without the need for genome-scale measurements from a pure sample of the cell-lineage of interest. Our method is robust to variable standard quality, and the

machine-learning component of our approach by itself is not sufficient for this robustness (SI Fig 5).

We applied this nano-dissection strategy to a dataset of 452 microarray measurements for micro-dissected human kidney biopsies and predicted 136 genes with novel podocyte-specific expression in the renal context (SI Table 3). These represented all non-gold-standard genes among the top 150 predictions. The selected genes are the set with the maximum F-measure (SI Methods) as assessed by cross validation where precision was weighted five times as much as recall and resulted in a number of genes practical for systematic verification and validation. *In silico* nano-dissection separated known podocyte genes from genes specific to the other glomerular cell-lineages and tubular cells (Fig. 2), while a simple correlation-based approach failed to do so (SI Fig. 1 & SI Methods).

The applicability of our nano-dissection strategy is not limited to the podocyte – it can accurately separate genes from tissues as diverse as skin (skin fibroblast genes from melanocyte and keratinocyte genes) and neuronal tissue (astrocyte genes from other glial-cell-specific genes) from publicly available expression data for the corresponding tissue homogenates (SI Fig. 2 & 3).

Confirmation of the specificity of *in silico* podocyte predictions

In addition to the systematic evaluation by immunohistochemical staining (below), we found that nano-dissection identified genes that were previously reported to have podocyte-specific expression patterns such as *PLA2R1* (Beck et al. 2009) and *GJA1* (Yaoita et al. 2002; Sawai et al. 2006) (SI Table 3 & 4), but which were withheld during the expert curation. In addition to recapitulating past literature, this strategy predicted concurrent discoveries. While this manuscript was being prepared, two genes predicted by nano-dissection, myosin IE (*MYO1E*) and PDZ and LIM domain 2 (*PDLIM2*), were shown to display podocyte-specific expression and play a role in renal function and hereditary and acquired glomerular disease (Mele et al. 2011; Sistani et al. 2011). *MYO1E* mutations were shown to cause childhood-onset, glucocorticoid-resistant focal segmental glomerulosclerosis (FSGS) (Mele et al. 2011; Ingelfinger 2011) and *PDLIM2* exhibited a reduced expression in patients with minimal change disease (MCD) and membranous nephropathy (MN) (Sistani et al. 2011).

We used high-throughput immunohistochemical (IHC) stainings from the Human Protein Atlas (HPA) (www.proteinatlas.org) to systematically validate the podocyte-specific expression of genes identified by nano-dissection. Although genes with staining data available in the HPA were not annotated to the level of cell-lineage localization, intra-glomerular cell types were identified based on their localization pattern inside of the glomerular tuft by three investigators with expertise in renal histopathology independently in a blinded manner (see Methods). This enabled us to systematically evaluate our predictions by immunohistochemistry and to compare the performance of *in silico* nano-dissection with experimental predictions from *in vivo* fluorescence-tagged murine podocytes (Fig. 3).

As the first step of the validation strategy, a blinded evaluation of the localization of predicted podocyte proteins in IHC staining images from HPA was performed for predicted podocyte-specific transcripts and an equivalently sized set of randomly selected genes (included as control). Of the predicted 136 podocyte-specific proteins in HPA, 31 were found to have sufficient expression pattern for evaluation at the time of our study (using HPA version 7.0 - 2010.11.15). Of these 31 proteins, 20 (65%) were found to be podocyte-specific within the renal context, with staining for 7 (23%) exclusively attributed to the podocyte within kidney tissue and an additional 13 proteins (42%) stained exclusively to the podocyte within the glomerulus. The other 11 proteins (35%) were stained to other renal cells (Fig. 3b, SI Table 4). The nano-dissection-identified gene set significantly (Fisher's exact p-value = 3.256e-06) outperformed the control (background) genes of which 1 (2%) showed podocyte-specific staining within the kidney and 5 (10%) showed podocyte-specific staining within the glomerulus (Fig. 3b). Forty-two (88%) background genes were stained to other renal cells.

In contrast to humans, where podocyte-specific micro-dissection is technically infeasible, murine transgenic model systems have been developed by the GenitoUrinary Development Molecular Anatomy Project (GUDMAP) consortium specifically to define cell-lineage-specific genes (McMahon et al. 2008). Lineage tracing was established by GFP-expression using cell type-specific promoters, followed by FACS and genome-wide expression profiling of GFP-positive single cell suspension (Brunskill et al. 2011). Using the podocyte, mesangial and endothelial cell-lineage-specific GUDMAP expression data sets (SI Table 5), we subtracted endothelial and

mesangial gene expression profiles from the transcriptome obtained from the podocyte preparation and identified 102 podocyte-specific transcripts (McMahon et al. 2008) (Methods & SI Table 6). These transcripts underwent the same cell-lineage-specific evaluation in HPA as the *in silico* nano-dissected human transcripts. Thirty of the 102 murine experimental approach derived podocyte-specific transcripts had staining patterns identifiable in HPA, of which staining for 2 was exclusively attributed to the podocyte within human kidney tissue and 5 proteins were stained to the podocyte within the glomerulus (Fig. 3b). Thus the *in silico* nano-dissection approach exhibited a significantly higher accuracy (Fisher's exact p-value = 0.0059) than the murine experimental strategy for discovering transcripts with podocyte-specific expression (nano-dissection's 65% vs murine experimental approach's 23% of predictions confirmed as podocyte-specific). The *in silico* prediction accuracy of cell-lineage enrichment using human tissue homogenate exceeded that obtained from *in vivo* fluorescence-tagged and sorted cells in a murine model system.

Disease specific regulation of the nano-dissection gene sets

To test the hypothesis that the discovered podocyte-specific genes were associated with human renal disease, the transcript with the highest podocyte-specific score and positive HPA validation, procollagen C-endopeptidase enhancer 2 (*PCOLCE2*), was selected for further characterization. PCOLCE2 protein modulates binding of procollagen C-proteases to collagen in a BMP1 dependent and cell-lineage-restricted manner (Steiglitz et al. 2002), a process with significant relevance for the development and function of the glomerular basement membrane (Tanaka et al. 2010). We investigated the disease specific transcriptional regulation of *PCOLCE2*. The steady-state mRNA level of *PCOLCE2* in glomeruli from human renal biopsies was significantly repressed in patients with FSGS (n=17), a glomerular disease with podocyte damage and end-stage renal disease (ESRD), compared to controls (n=39, $p<0.05$) (Fig. 4a). In contrast, in glomeruli from patients with MCD (n=12), a proteinuric disease without progression to ESRD, *PCOLCE2* transcript levels were not significantly altered. In a cohort of CKD patients with heterogeneous glomerular pathophysiology (n=139) loss of *PCOLCE2* glomerular gene expression was significantly correlated with loss of renal function (R=0.32, p=1.17E-04).

Disease-specific *PCOLCE2* regulation was further validated in human kidneys affected by glomerular disease using immunohistochemical staining in an independent biopsy cohort. In concordance to the IHC staining patterns reported in HPA (Fig.3a VII), the podocyte-specific localization of PCOLCE2 protein was confirmed. In contrast to the nuclear and perinuclear PCOLCE2 signal seen in IHC in glomeruli of 5 healthy kidneys (Fig. 4b I), PCOLCE2 staining was not detectable in glomeruli from 8 FSGS patients (Fig. 4b II), demonstrating the ability of the nano-dissection strategy to detect genes with both cell-lineage-specific expression and disease-specific alteration in glomerular failure.

Glomerular disease stratification by the *de novo* predicted podocyte-specific transcripts

Podocyte damage leads to progressive loss of kidney function and the need for dialysis and renal transplantation. To test the association with glomerular function, the glomerular regulation of the podocyte gene set discovered by nano-dissection was compared between 39 controls and 17 patients with FSGS, a renal disease caused by severe podocyte damage (Kriz et al. 1994; Pavenstädt 2000) and the leading cause of glomerular failure in children. Using significance analysis of microarrays (SAM, Tusher et al. 2001), 60 of the 136 genes identified by nano-dissection were significantly repressed in glomeruli from FSGS patients versus controls (q value < 0.05).

Next, the regulation of predicted transcripts in chronic renal disease was evaluated in a cohort of 139 patients with glomerular diseases, including FSGS, diabetic nephropathy (DN), IgA nephropathy (IgAN), Membranous Nephropathy (MN), lupus nephritis (LN) and MCD (SI Table 7). Steady state mRNA expression measurements of nano-dissection predicted podocyte-specific transcripts were correlated with glomerular filtration rate (GFR) at the time of biopsy, currently the best overall index of kidney function used to classify the stages of CKD patients by the Kidney Disease Outcomes Quality Initiative (KDOQI). Expression of the set of 136 *de novo* predicted podocyte-specific genes was significantly ($p < 0.01$) more correlated with GFR than observed in permuted gene-GFR associations (Fig.4c). These results demonstrate the potential for predicted podocyte-specific genes to be candidate markers for disease progression. This finding has significant clinical utility, as the cell-lineage specific and disease associated genes can provide superior specificity for biomarker testing in heterogeneous biofluids like urine or

blood compared to ubiquitously expressed disease markers. The GFR correlation of the podocyte-specific gene set predicted by nano-dissection supports the tight link of podocyte differentiation and function with renal impairment irrespective of initiation of renal disease by genetic or environmental causes.

Application of nano-dissection to non-podocyte lineages

Mesangial cells are one of the three major cell types in kidney glomeruli, and mesangial expansion is a hallmark of diabetic nephropathy (DN). We investigated the expression profile of the top 52 mesangial cell-specific genes predicted by nano-dissection (SI Fig 6, cutoff based on the same F-measure criterion) in an independent DN dataset (by Woroniecka et al. 2011, data include glomerular gene expression profile of 13 healthy donors and 9 patients with DN). In this dataset, 50 of the 52 predicted mesangial cell-specific genes showed robust expression in the micro-dissected glomeruli. 44% of these genes (22 out of 50) exhibited increased steady state mRNA levels in patients with DN compared to controls (SAM analysis $q<0.01$); no transcript showed significantly reduced mRNA levels. This demonstrates both nano-dissection's applicability to other renal cell-lineages, as well as its ability to identify lineage-specific genes with increased mRNA levels.

We further evaluated nano-dissection on non-renal cell lineages. For this analysis, we used tissue annotations from the Human Protein Reference Database (HPRD, Keshava Prasad et al. 2009) as gold standards for tissue-specific expression. For example, we have applied nano-dissection to identify genes specifically expressed in skin fibroblasts (density estimate cross-validation based evaluation, SI Fig 3). We evaluated genes above the maximum F-measure criterion for significant enrichment in disease annotations from OMIM. The genes above the maximum F-measure criterion showed significant enrichment of genes involved in two collagen disorders with known fibroblast expression: Ehlers-Danlos syndrome and Osteogenesis imperfecta. Six of the ten genes associated with Ehlers-Danlos syndrome were above this threshold (FDR corrected q-value = 0.00038), as were six of the eight genes associated with Osteogenesis imperfecta (FDR corrected q-value = 0.00011). None of these genes were included in the fibroblast gold standard from HPRD used by nano-dissection; all were new predictions of fibroblast-specific genes. This

demonstrates both the potential of nano-dissection to identify cell-lineage specific genes as well as the potential for those genes to be associated with cell-lineage specific diseases.

Discussion

Organ specific transcriptional programs define the final stages in tissue development and the mature function of metazoan organisms. Alterations in the functions of genes with cell-lineage restricted expression patterns are widely believed to lead to tissue specific disease manifestations. Furthermore, inherited diseases are frequently caused by mutations in genes with restricted expression patterns (D'Agati 2008; Cai and Petrov 2010; Winter et al. 2004). Mutations in such genes often do not cause early embryonic lethality, but rather manifest disease at the time when the function of these genes becomes critical for a specific tissue and subsequently for organismal survival (D'Agati 2008). In acquired disease like diabetes or hypertension, the vulnerability of a specific organ to the systemic disease is defined by the expression of tissue-specific genes (Woroniecka et al. 2011; Koop et al. 2003; Doublier et al. 2003). Defining cell-lineage-specific transcripts therefore has immediate clinical implications for such cell-lineages. However, a major challenge to define a specific cellular transcriptome has been the inability to obtain pure cell preparation from human tissue *in vivo* (e.g. as recently summarized by Lindenmeyer et al. (2010) for renal cell lineages).

To identify cell-lineage specific transcripts on a genome-wide scale even when direct experimental assays are infeasible, as is the case for most solid human tissues, we developed *in silico* nano-dissection. This iterative machine-learning-based approach robustly leverages existing knowledge about the cell-lineage of interest to identify transcripts with similar behavior in heterogeneous transcriptional data sets of tissue homogenates. *In silico* nano-dissection does not require expression data of pure genome-wide profiles from the cell lineage of interest and is robust to small numbers and varying specificity of available cell-lineage markers. Although our strategy uses Support Vector Machines (SVM) as the machine-learning component of the nano-dissection method (Figure 1), in principal any machine learning approach that leverages positive and negative examples for training can be integrated in place of the SVM.

This study represents the first high-throughput approach for identification of cell-lineage-specific genes for any cell lineage from *in vivo* human data. The approach is general -- we found that our predictions remained robust (significantly overrepresented by podocyte-specific genes) even when the directly targeted expression data (renal glomeruli) constitutes only 5% of the total data sets (the rest being diverse human expression data from the Gene Expression Omnibus). *In silico* nano-dissection applied to public gene expression data from tissue homogenates (implementation available through our nano.princeton.edu website) is also capable of accurately separating cell-lineage-specific genes in skin (skin fibroblast genes from melanocyte and keratinocyte genes) and neuronal tissue (astrocyte genes from other glial cell specific genes). Furthermore this approach is effective even with a limited number of positive marker genes or in situations when some provided lineage specific genes are inaccurate / irrelevant to the cell lineage (SI Fig. 4 & 5)). This makes *in silico* nano-dissection a promising approach to identify cell-lineage-specific genes that might be potentially associated with other acquired or inherited diseases, for which targeted data may not be available.

Cell-type specific markers predicted by nano-dissection could be used to extend the applicability of methods that define the fractional composition of a mixture. These methods are then capable of deconvoluting an expression signal to perform tests of gene expression within individual lineages. While some approaches require knowledge of the mixture's fractional composition, which limits their application to cell types measurable in cell sorter experiments (Shen-Orr et al. 2010), others can start with either pure *ex vivo* samples of each mixture component to estimate the sample composition or a high quality set of known markers (Kuhn et al. 2011). These requirements currently limit the applicability of these methods, as they require information not available in studies of most complex solid human tissues. Markers identified by nano-dissection provide a promising starting point for the application of such deconvolution approaches to many more human cell lineages, including those not amenable to experimental micro-dissections.

Nano-dissection is the first genome-scale method that identifies cell-lineage-specific genes important for renal disease in humans. In addition to the *de novo* identified podocyte-specific genes, genes that have been shown by genetic or functional studies to be causally involved in hereditary glomerular diseases and chronic progressive renal failure, including *DACH1* (nano-

dissection rank 184 (Köttgen et al. 2010)), *APOL1* (rank 185 (Kopp et al. 2011; Genovese et al. 2010)), *VEGFA* (rank 247 (Köttgen et al. 2010; Eremina et al. 2008)) and *MYH9* (rank 260 (Kao et al. 2008; Kopp et al. 2008)) were ranked highly by our method. During the preparation of this manuscript, *MYO1E* (rank 75), was reported to be associated with autosomal-recessive glucocorticoid resistant nephrotic syndrome (Mele et al. 2011). *MYO1E* was found to exhibit, as predicted by our study, podocyte-specific expression and appears to interact with other cell-lineage-specific genes in podocyte cytoskeletal dynamics. *PDLIM2* (rank 121), another gene identified by nano-dissection, was recently reported to show podocyte-specific expression and repression in acquired glomerular disease (Mele et al. 2011; Sistani et al. 2011). Interestingly, neither of these genes is present in the list of 102 genes (SI Table 6) identified as podocyte-specific using the murine *ex vivo* cell lineage separation in the GUDMAP datasets. Brunskill and coworkers (Brunskill et al. 2011) recently generated a transcriptional data set (144 genes) regulated during murine podocyte development and enriched in adult podocytes in comparison with the renal cortex. Analysis in HPA of the human orthologues exhibited a similar enrichment to the GUDMAP *ex vivo* cell-lineage data set used in Figure 3 (2 podocyte-specific in kidney (5%) and 13 podocyte-specific in glomeruli (30%)), but the murine data did not reach the specificity of the *in silico* nano-dissection approach (65%).

Cell lineage specific strategies capable of identifying genes associated with a disease provide additional value compared to unbiased genome wide approaches that identify genes with expression correlation to a specific phenotype (like GFR). The latter captures a different pool of transcripts: abundantly expressed non-cell lineage specific genes constitute the majority of the transcripts correlated with renal function, but these transcripts do not necessarily perform cell lineage specific functions and may not be associated with hereditary disease. For example, expression levels of *MYO1E* and *APOL1* are not strongly correlated to GFR (ranked 1236 and 7430 respectively by GFR-expression correlation), yet these two genes were identified by nano-dissection to be cell type specific and have been shown experimentally (independent of and parallel to our work) to cause hereditary glomerular disease (Genovese et al. 2010; Kopp et al. 2011; Mele et al. 2011) Furthermore, a systematic analysis focusing on literature-curated hereditary FSGS-associated genes that do not overlap with our podocyte gold standard, (see SI

Methods) also demonstrates that a genome-wide assessment of GFR-transcript expression correlation alone could not identify genes associated with this hereditary renal disease (SI Fig 8A). In contrast, nano-dissection can identify genes associated with disease, with FSGS genes receiving significantly higher podocyte nano-dissection scores than those without known FSGS association (SI Fig 8B). Thus, nano-dissection's ability to identify cell lineage specificity is important for identifying genes potentially associated with such diseases and clinical phenotypes. Beyond simply addressing issues of statistical power, methods that consider cell lineage specificity provide additional utility because they address targeted biological questions that are tightly coupled to the disease etiology.

Our findings have significant potential for clinical utility. In the study of hereditary diseases, next generation exome sequencing technologies are now widely applied across hereditary diseases and are capable of identifying putative causal genetic variants in very small pedigrees. However, these studies often result in multiple candidate genes in need of further prioritization. As hereditary diseases are often caused by cell-lineage-specific transcripts (see above and Hinkes et al. (Hinkes et al. 2006)), the systematic scoring system for cell-lineage-specific enrichment provided by the nano-dissection approach can become a crucial tool to prioritize candidate genes for further validation using their cell-lineage enrichment scores. Vice versa, several hundreds of tissue specific genes identified by nano-dissection can be screened comprehensively in families with a hereditary disease of the organ of interest using targeted exon sequencing strategies as currently is pursued by our group in a rare disease cohort (Gadegbeku et al. 2013; Halbritter et al. 2012).

For acquired chronic disease, the search is still ongoing to define specific and robust biomarkers of differentiated organ function. Unbiased molecular screening approaches have been largely disappointing in this context. Our data strongly supports the close association of cell-lineage-specific transcripts and loss of end organ function in complex, chronic human kidney disease. Proteins encoded by these genes may be detected in plasma and urine and provide a non-invasive means to measure organ function in a cell-lineage-specific manner. In contrast to ubiquitously expressed molecules involved in fibrosis and inflammation, which are currently the most common source of candidate biomarkers for chronic diseases, cell-lineage-specific biomarkers

are less likely to be confounded by extra-renal processes and should provide superior diagnostic specificity (Fukuda et al. 2012). This has been demonstrated in the context of podocyte failure in model systems and human disease (Sato et al. 2009) and nano-dissection provides an opportunity to expand the scope of podocyte specific transcripts analyzed in complex mixture of urinary cells by these approaches. Finally, functional studies of the cell-lineage-specific genes identified in human disease tissue offer the opportunity to develop a targeted therapeutic approach for chronic disease. Targeting a disease-specific molecular mechanism selectively in the tissue manifesting the disease has the potential to significantly increase efficacy while reducing off-target effects.

In summary, nano-dissection is a novel computational approach for defining the specificity of cell types at the transcriptional level. As demonstrated for glomerular disease, but applicable across all organs with large scale transcriptional data sets available, nano-dissection can reveal novel transcripts with essential tissue-specific function in organogenesis and hereditary human disease. In chronic progressive diseases, the nano-dissection-identified transcripts can serve as highly specific markers of disease stages and provide a starting point for the development of organ specific targeted therapies. While we have shown that HPRD annotations can guide successful nano-dissection analyses, we believe the method is most powerful when combined with high-quality user constructed standards, which can be easily accomplished using our nano-dissection web server. Nano-dissection can be performed on user-curated tissue-specific gene expression compendia via the user-friendly nano-dissection webserver at http://nano.princeton.edu. This web-server includes 452 microarrays from micro-dissected kidney biopsy samples from this study, as well as 7539 samples across 28 diverse human tissue collections manually curated from the Gene Expression Omnibus (GEO). Nano-dissection of investigator-specific gene-expression datasets can be performed with the Sleipnir library for functional genomics (version 3 or higher) available for Windows, OS X, and Linux systems from http://libsleipnir.bitbucket.org/ (Huttenhower et al. 2008).

Materials and Methods

Patient characteristics

Human renal biopsy specimens were procured through an international multicenter study, the European Renal cDNA Bank-Kroener-Fresenius biopsy bank. Biopsies were obtained from patients after informed consent and with approval of the local ethics committees. All biopsies were stratified by the reference pathologist of the ERCB according to their histological diagnoses. Histology reports, clinical data and gene expression information were stored in a de-identified manner. A total of 452 microarrays from kidney biopsies were used for nano-dissection, of which 139 patients were used for kidney function correlation analysis. Demographic data of these 139 patients are provided in the SI Table 7.

Micro-dissected human kidney biopsy data

Micro-dissection into glomerular and tubule-interstitial compartments and Affymetrix based gene expression profiling were performed as previously reported (Ju et al. 2009). Affymetrix GeneChip Human Genome U133A 2.0 and U133 Plus 2.0 Array were used in this study. For this analysis we restricted ourselves to only the probesets present on both platforms. Normalized data files are uploaded on the GEO (Edgar 2002) website and accessible under reference numbers GSE32591 (Berthier et al. 2012), GSE35488 (Reich et al. 2010), GSE37455 (Berthier et al. 2012), and GSE37460 (Berthier et al. 2012). For simplicity, we use “in vivo” to refer to these assays of genes measuring gene-expression in human biopsies of complex tissues.

***In silico* nano-dissection for the prediction of cell-lineage specific gene expression**

Our approach uses machine learning within a novel iterative framework to predict genes with cell-lineage specific expression on the whole-genome scale based on gene expression data from tissue homogenates. This problem is especially challenging because, in order to work for cell lineages that are infeasible to micro-dissect experimentally such as the podocytes, our approach must function without example expression profiles of the lineage of interest.

Intuitively, our method leverages patterns of expression of cell-lineage specific genes that it discovers from whole-genome expression compendia not resolved to the cell-lineage of interest. These patterns are specific for each cell-lineage and generally only found in a small subset of experimental conditions, which may include genetic, physiological, pathophysiological, environmental or experimental states/perturbation (for example biopsy specimens from different patients). To discover these cell-lineage-specific expression patterns as well as the subsets of

conditions that are informative for a given cell lineage, our approach uses a machine learning approach in an iterative probabilistic framework to combine an expert-provided standard of known cell-lineage specific genes (positives) as well as example genes that are expressed in other cell-lineages (negatives). However, most solid-tissue cell-lineages cannot be studied experimentally in high-throughput, and thus only few cell-lineage-specific genes are often known with high accuracy (e.g. from immunohistochemistry). The additional challenge here is that these standards are often limited in size (especially for cell lineages not amenable to experimental micro dissection), and can be of varying specificity (e.g. specific to cell lineage within the immediate structure or whole organ or defined by different experimental approaches). This paucity of high-quality standards and need to effectively leverage lower-quality or less specific examples severely limits direct application of traditional machine learning approaches, for example a Support Vector Machine classifier (SVM) (Supplemental Figure 5).

Because it is experimentally infeasible to obtain pure example expression profiles for cell lineages from solid human tissues, our method must perform well even while available standards are often very limited in size and can be of highly varying specificity. This paucity of high-quality standards and the need to effectively leverage lower quality or less specific examples severely limits the direct application of traditional machine learning approaches (e.g. Support Vector Machine (SVM) performance outside of the iterative framework is shown in SI Fig. 5).

To address these challenges, we developed an iterative classification approach that continually refines both the predictive cell-lineage specific patterns and informative conditions based on statistical scoring and refinement (through informative subset selection) of the provided standard. This iterative approach allows the user to provide tiered standards, that is, the investigator identifies only the relative specificity of evidence tiers (i.e. low-throughput high specificity approaches are more reliable as compared to high-throughput experimental platforms with lower specificity). The *in silico* nano-dissection method is then able to make high-accuracy predictions of cell-lineage specific genes on the whole-genome scale and, within the tiered standard constraint, is robust to variable specificity of example cell-lineage specific genes.The iterative strategy is necessary to allow investigators to add standards of questionable quality without dramatically compromising the quality of cell-lineage predictions. A linear SVM without this

iterative approach fails when standards of lower quality are added to high quality standards (SI Fig. 5).

The researcher defines standards within tiers. Tiers represent levels of specificity (i.e. in descending order: double immunofluorescence, annotated in literature curated database, high-throughput protein expression). For each tier, nano-dissection calculates the sum of the ranks of genes from the classifier (for the case of SVM this is the ranked distance from the SVM hyperplane) for each positive example, R_i,(here podocyte genes) against each of M negative standards, j, (e.g. glomerular, mesangial, tubular) as $SR_j = \sum_{i=1}^{n_p} R_i$, where n_p represented the number of positives and ranks were calculated from only the positive examples and the negative examples from standard j. It then computes a test statistic for this individual separation, U_j for each negative standard as $U_j = \max(C_j, n_p n_j - C_j)$, where $C_j = n_p n_j + \frac{n_p(n_p+1)}{2} - SR_j$. This is normalized by converting it to a z-score by using the mean and standard deviation through $z_j = \frac{U_j - \frac{n_j n_p}{2}}{\sqrt{\frac{n_j n_p (n_j + n_p + 1)}{12}}}$. The scores for the individual separations are then combined to provide a final score for this tier of standards $p = \sqrt[M]{\prod_{j=1}^{M} (1 - cdf(z_j))}$. Nano-dissection automatically selects the standards resulting in the lowest p (which ranges from zero to one), i.e. that which corresponds to a better separation of positives from each negative standard.

In certain cases, an additional (and optional) external validation gene-set may be available. Because nano-dissection can be applied where experimental micro-dissection was insufficient, these standards may represent both positives and negatives (e.g. in this case where additional microarray measurements of the renal glomerulus were available as validation). We termed genes in this standard as “high-throughput-validating” genes and other genes as “non-validating” genes. Nano-dissection can use this validation set to identify the set of standards providing the

best separation of validating genes by calculating $SR = \sum_{i=1}^{n_v} R(|d_i|)$, where $R(|d_i|)$ is the rank of the absolute value of the distance to the hyperplane of the validating gene i in a list containing the n_v validating genes and the n_{nv} non-validating genes. It then calculates U as $U = \max(C, n_v n_{nv} - C)$, where $C = n_v n_{nv} + \frac{n_v(n_v+1)}{2} - SR$, which is then converted to a z-score $z = \frac{U - \frac{n_v n_{nv}}{2}}{\sqrt{\frac{n_v n_{nv}(n_v + n_{nv} + 1)}{12}}}$. Finally, p for validating versus non-validating is calculated as $p = 1 - cdf(z)$. Selecting the standard tier that provided the lowest p results in the standard where validating genes were most extreme (i.e. best separated from each other). Our results demonstrate that this approach enables us to use a non-cell-lineage specific validation (i.e. glomerular) gene set to grade our separation of putative cell-lineage (podocyte) specific genes by selecting that standard which leads to example genes on the extremes (in our example, this has potential podocytes at the top of the list and potential non-podocyte glomerular genes at the bottom). In the case where there exists a validation standard of high quality specific to our cell-lineage of interest, we instead use $R(d)$ directly instead of $R(|d|)$. In that case, this value would represent the one sided Wilcoxon rank-sum p-value for a comparison of validating and non-validating genes. Because this iterative nano-dissection approach relies on genome-scale data obtained from the surrounding compartment and because this evaluation was used to identify the optimum standards, this p provides a quality measure for the resulting standard. Thus nano-dissection allows us to obtain cell-lineage specific signal from *in vivo* human data.

The nano-dissection algorithm therefore proceeds as follows (see SI Fig. 7 for pseudocode). Given user-supplied standards in tiers of increasing specificity, for each standard-level, *k*, combine standards of that level with all standards of higher specificity levels. Apply the selected classification algorithm (here we applied SVM from the SVM^{perf} package (Joachims 2006) using the Sleipnir library (Huttenhower et al. 2008)) and generate a ranked list of predictions. Score the

predictions for *k* as described above to calculate *p* for the *k*th level of specificity. Select the level of specificity providing the lowest *p*.

In this work, standards were obtained from expert literature review. The positive podocyte-specific standard genes were required to have at least one of the following levels of evidence: immunofluorescence staining, *in situ* hybridization, or electron microscopy image of immuno-gold staining of podocytes *in vivo*. Two levels of specificity were evaluated. The most stringent level contained genes specifically expressed only in podocytes and no other cell types in the human kidney, referred to as podocyte-specific in kidney (as an example see Nephrin staining pattern in Fig. 3a I). For the majority of selected genes, evidence for disease association in human glomerular failure or murine model systems was also available. The less stringent level contained all of the above, as well as genes expressed in podocytes and no other cell types in glomeruli, but did contain genes detected in extra-glomerular cells of the kidney (Synaptopodin (SYNPO), and CD2AP staining in Fig. 3a II & III). For the majority of selected genes evidence for disease association in human glomerular failure or murine model systems was also available. Application of nano-dissection resulted in the use of both tiers of standards, which corresponded to a total of 46 genes that were both podocyte-specific and present in the gene-expression dataset.

Gene expression data extraction from the GUDMAP and data processing

The genome wide expression data of murine podocytes, mesangial and endothelial cells were obtained from http://www.gudmap.org/. The IDs of datasets that have been utilized in our study are listed in Supplementary Table 5. The detailed protocol of data generation is described in recently published paper by Brunskill et al. (2011), and can also be found in our Supplementary Methods. We preprocessed and normalized data as described in "Micro-dissected human kidney data". By comparing the expression level in podocytes versus the other two major cell types in glomeruli, we define a gene to be podocyte-specific if its expression in podocytes is 4.76 fold over mesangial cells and 4.65 fold over glomerular capillary endothelial cells. The cut off values represent three standard deviations of the average difference between podocytes and mesangial / endothelial cell transcripts, respectively. Using HomoloGene from NCBI Entrez (Maglott et al. 2011), 102 murine genes could be mapped to their Homo sapiens ortholog (SI Table 6).

Evaluation of intra-renal protein localization in HPA

We evaluated the intra-renal localization pattern of protein products of predicted genes based on HPA 7.0 - 2010.11.15. Intra-glomerular cell types are identified based on their localization pattern inside of the glomerular tuft by three investigators with expertise in renal histopathology independently in a blinded manner. Conflicts were resolved by a majority vote. The following staining patterns were considered inconclusive and excluded from the analysis: 1) Proteins with negative staining or "data not available"; 2) Proteins with only a single renal histology image available; 3) Proteins with a diffuse non-specific staining pattern. If several antibodies were evaluated for a specific protein, the images from the antibody with the highest degree of specificity were used for evaluation. Tubular brush border staining was considered unspecific. In general, protein localization patterns were classified into 3 groups: 1. Expressed exclusively in podocytes with no other cell types exhibiting staining in kidney section, referred to as "podocyte-specific in kidney" (i.e. Nephrin (NPHS1) in Fig. 3aI); 2. Expressed specifically in podocytes within the glomerulus but with positive staining also observed in tubular-interstitial compartments, referred to as "podocyte-specific in glomerulus" (i.e. Synaptopodin (SYNPO) and CD2-associated protein (CD2AP) in Fig. 3a II and III respectively); 3. All remaining staining patterns as "other renal cell". For clarity we refer to category 1 and 2 in aggregate as 'podocyte-specific' genes.

Immunohistochemistry staining of kidney biopsy tissues

Following a previously described protocol (Lorz et al. 2008), immunohistochemical studies were performed using a PCOLCE2 primary rabbit antibody (HPA013203, Sigma-Aldrich, St. Louis, MO).

Association of RNA expression of cell-lineage-specific genes to kidney function

The expression of the top 136 candidate genes in a cohort of 139 patients with CKD was correlated to square root of GFR, calculated by MDRD equation (Modified diet in renal disease formula) (Levey et al. 2006), using Pearson correlation. The correlations were compared against a randomized set by analyzing their correlation density plots using the sm R package (Bowman

and Azzalini 2010). Randomization was performed by randomly reassigning expression values to GFR 100 times on the given dataset, followed by recalculation of the correlation**.**

Manual tissue-of-origin sample annotation for the webserver

Microarray experiments (Affymetrix U133 Plus 2) were manually annotated to the sample's tissue of origin using the controlled vocabulary in the Brenda Tissue Ontology. To ensure wide coverage of tissue-types, a broad set of candidate samples for each tissue was identified with an initial term matching, corrected for linguistic variations with stemming, for each Brenda term (and its synonyms) on sample descriptions available in GEO. These term-to-experiment matches were then manually curated, verified, or corrected based on the corresponding sample descriptions. Only matches for terms across at least two independent datasets were reviewed. Only the tissue-of-origin information was considered in the manual evaluation, and so tumor-adjacent normal breast biopsy samples were correctly annotated to 'breast,' for example. We excluded tissue mixture samples, reference samples, and non-human samples. We also excluded samples with ambiguous descriptions, as well as cell line and cancer terms. Samples annotated to detailed terms in the controlled vocabulary were propagated up to organ-level annotations, based in the organ of origin. Terms with fewer than ten annotated samples after propagation were excluded at this stage. This procedure resulted in a manually annotated compendium of 7539 samples from 28 tissues that we make available through the nano-dissection webserver for nano-dissection analysis.

Data Access

Normalized gene expression data files of micro-dissected human kidney biopsies are uploaded on the GEO (Edgar 2002) website and accessible under reference number GSE47185.

Acknowledgements

This work was supported in part by R01 GM071966 to OGT and MK, by RO1 HG005998 and DBI0546275 to OGT, by P50 GM071508, and by R01 DK079912 and P30 DK081943 to MK. OGT is a Senior Fellow of the Canadian Institute for Advanced Research.

The authors acknowledge the members of the European Renal cDNA Bank-Kroener-Fresenius biopsy bank at the time of the expression profiling: Clemens David Cohen, Holger Schmid, Michael Fischereder, Lutz Weber, Matthias Kretzler, Detlef Schlöndorff, Munich/Zurich/AnnArbor/New York; Jean Daniel Sraer, Pierre Ronco, Paris; Maria Pia Rastaldi, Giuseppe D'Amico, Milano; Peter Doran, Hugh Brady, Dublin; Detlev Mönks, Christoph Wanner, Würzburg; Andrew Rees, Aberdeen and Vienna; Frank Strutz, Gerhard Anton Müller, Göttingen; Peter Mertens, Jürgen Floege, Aachen; Norbert Braun, Teut Risler, Tübingen; Loreto Gesualdo, Francesco Paolo Schena, Bari; Jens Gerth, Gunter Wolf, Jena; Rainer Oberbauer, Dontscho Kerjaschki, Vienna; Bernhard Banas, Bernhard Krämer, Regensburg; Moin Saleem, Bristol; Rudolf Wüthrich, Zurich; Walter Samtleben, Munich; Harm Peters, Hans-Hellmut Neumayer, Berlin; Mohamed Daha, Leiden; Katrin Ivens, Bernd Grabensee, Düsseldorf; Francisco Mampaso(†), Madrid; Jun Oh, Franz Schaefer, Martin Zeier, Hermann-Joseph Gröne, Heidelberg; Peter Gross, Dresden; Giancarlo Tonolo; Sassari; Vladimir Tesar, Prague; Harald Rupprecht, Bayreuth; Hermann Pavenstädt, Münster; Hans-Peter Marti, Bern; Peter Mertens, Magdeburg.

Disclosure Declaration

The authors do not have any conflicts of interests to declare.

Figure Legends

Fig. 1. Schematic overview of the *in silico* nano-dissection workflow, an iterative approach for cell-lineage-specific gene prediction, validation and functional analysis. Expert curated literature annotations are iteratively combined with gene-expression data to predict genes specific to a cell lineage. These predictions are assessed and the standards are refined. Validation of podocyte-specificity of our predictions used publicly available resources followed by evaluation of intra-renal mRNA and protein expression analysis in correlation with clinical phenotypes to define regulation of predicted gene sets in human disease.

Fig. 2. *In silico* nano-dissection. Distribution of cell type-specific prediction by percentile, estimated using a Gaussian kernel. Genes are ordered on the X-axis from worst (0^{th} percentile) to

best (100^{th} percentile). The dotted line shows *in silico* nano-dissection cutoff for the top 136 genes. Nano-dissection successfully separates (AUC 0.83) podocyte-specific genes (green) from genes specific to other renal cell lineages (glomerular endothelial in dark blue, glomerular mesangial in light blue and tubular in red).

Fig. 3. Evaluation of podocyte-specific genes based on qualified Human Protein Atlas (HPA) staining images. (a) HPA images demonstrate podocyte-specific pattern of positive standard markers and predicted genes. Podocyte positive standard markers with kidney specific pattern (I): Nephrin (NPHS1); and glomerulus specific pattern (II): SYNPO and (III): CD2AP. Exemplary staining patterns for de novo nano-dissection predicted proteins (IV-IX): (IV): FGF1; (V): ARHGAP28; (VI): PRKAR2B; (VII): PCOLCE2; (VIII): GJA1; and (IX): ZDHHC6. (b) HPA based distribution intra-renal protein staining pattern in random gene set, nano-dissection-identified gene set and the murine experimental approach-derived gene set: The *in silico* nano-dissection approach (65%) significantly outperforms a random set of genes (12%) and the *ex vivo* murine experimental approach (23%) for identifying podocyte-specific genes. Grey bars show the proportion of genes with exclusively podoctye-specific staining within the kidney and black bars show the proportion of genes with exclusively podocyte-specific staining within the glomerulus.

Fig. 4. Regulation of predicted podocyte-specific gene set in human disease. (a) Box-and-whisker plot of glomerular mRNA Expression of *PCOLCE2* in biopsies from living donor controls (LD, n=35), MCD patients (n=12) and FSGS patients (n=19). * denotes significantly ($P<0.05$) differentially expressed in FSGS versus in living donors. (b) Immunohistochemical staining of PCOLCE2 on kidney biopsies from controls (I) and FSGS patients (II). In comparison with control kidneys, PCOLCE2 signal disappears in FSGS patients. Images shown are the representative images in the glomerulus of controls (n=5) and FSGS patients (n=8). (c) Density plot of the association (Pearson correlation, X-axis) of the 136 predicted podocyte-specific genes (red) with renal function as quantified by GFR value, compared to density plot of repeatedly (100 times) randomized gene expression-GFR associations (black). The randomized set shows a distribution centered on 0 (meaning no correlation with GFR), whereas the podocyte-specific genes show a skewed distribution towards positive correlation, indicating reduced gene

expression is associated with impaired renal function. Correlation with GFR of the 136 transcripts across all renal diseases analyzed was significantly enriched compared to the permuted sample ($p < 0.01$). Black line indicates the correlation of *PCOLCE2* mRNA level with GFR.

References

Beck LH, Bonegio RGB, Lambeau G, Beck DM, Powell DW, Cummins TD, Klein JB, Salant DJ. 2009. M-type phospholipase A2 receptor as target antigen in idiopathic membranous nephropathy. *The New England Journal of Medicine* **361**: 11–21.

Bennett MR, Czech KA, Arend LJ, Witte DP, Devarajan P, Potter SS. 2007. Laser capture microdissection-microarray analysis of focal segmental glomerulosclerosis glomeruli. *Nephron Experimental Nephrology* **107**: e30–40.

Berthier CC, Bethunaickan R, Gonzalez-Rivera T, Nair V, Ramanujam M, Zhang W, Bottinger EP, Segerer S, Lindenmeyer M, Cohen CD, et al. 2012. Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *Journal of Immunology* **189**: 988–1001.

Bowman AW, Azzalini A. 2010. R package sm: nonparametric smoothing mtehods.

Brunskill EW, Georgas K, Rumballe B, Little MH, Potter SS. 2011. Defining the molecular character of the developing and adult kidney podocyte. ed. K. Stadler. *PLoS ONE* **6**: e24640.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome biology and evolution* **2**: 393–409.

D'Agati VD. 2008. The spectrum of focal segmental glomerulosclerosis: new insights. *Current Opinion in Nephrology and Hypertension* **17**: 271–81.

D'Agati VD, Kaskel FJ, Falk RJ. 2011. Focal segmental glomerulosclerosis. *The New England Journal of Medicine* **365**: 2398–411.

Doublier S, Salvidio G, Lupia E, Ruotsalainen V, Verzola D, Deferrari G, Camussi G. 2003. Nephrin expression is reduced in human diabetic nephropathy: evidence for a distinct role for glycated albumin and angiotensin II. *Diabetes* **52**: 1023–30.

Edgar R. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**: 207–210.

Endlich N, Sunohara M, Nietfeld W, Wolski EW, Schiwek D, Kränzlin B, Gretz N, Kriz W, Eickhoff H, Endlich K. 2002. Analysis of differential gene expression in stretched podocytes: osteopontin enhances adaptation of podocytes to mechanical stress. *FAESB J* **16**: 1850–2.

Eremina V, Jefferson JA, Kowalewska J, Hochster H, Haas M, Weisstuch J, Richardson C, Kopp JB, Kabir MG, Backx PH, et al. 2008. VEGF inhibition and renal thrombotic microangiopathy. *The New England Journal of Medicine* **358**: 1129–36.

Fukuda A, Wickman LT, Venkatareddy MP, Wang SQ, Chowdhury MA, Wiggins JE, Shedden KA, Wiggins RC. 2012. Urine podocin:nephrin mRNA ratio (PNR) as a podocyte stress biomarker. *Nephrology, Dialysis, Transplantation* **27**: 4079–87.

Gadegbeku CA, Gipson DS, Holzman LB, Ojo AO, Song PXK, Barisoni L, Sampson MG, Kopp JB, Lemley K V, Nelson PJ, et al. 2013. Design of the Nephrotic Syndrome Study Network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach. *Kidney International*.

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, et al. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**: 841–5.

Gerstein HC. 2001. Albuminuria and Risk of Cardiovascular Events, Death, and Heart Failure in Diabetic and Nondiabetic Individuals. *The Journal of the American Medical Association* **286**: 421–426.

Groop P-H, Thomas MC, Moran JL, Wadèn J, Thorn LM, Mäkinen V-P, Rosengård-Bärlund M, Saraheimo M, Hietala K, Heikkilä O, et al. 2009. The presence and severity of chronic kidney disease predicts all-cause mortality in type 1 diabetes. *Diabetes* **58**: 1651–8.

Halbritter J, Diaz K, Chaki M, Porath JD, Tarrier B, Fu C, Innis JL, Allen SJ, Lyons RH, Stefanidis CJ, et al. 2012. High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. *Journal of Medical Genetics* **49**: 756–67.

Henger A, Kretzler M, Doran P, Bonrouhi M, Schmid H, Kiss E, Cohen CD, Madden S, Porubsky S, Gröne EF, et al. 2004. Gene expression fingerprints in human tubulointerstitial inflammation and fibrosis as prognostic markers of disease progression. *Kidney International* **65**: 904–17.

Higgins JPT, Wang L, Kambham N, Montgomery K, Mason V, Vogelmann SU, Lemley K V, Brown PO, Brooks JD, van de Rijn M. 2004. Gene expression in the normal adult human kidney assessed by complementary DNA microarray. *Molecular Biology of the Cell* **15**: 649–56.

Hinkes B, Wiggins RC, Gbadegesin R, Vlangos CN, Seelow D, Nürnberg G, Garg P, Verma R, Chaib H, Hoskins BE, et al. 2006. Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nature Genetics* **38**: 1397–405.

Hodgin JB, Borczuk AC, Nasr SH, Markowitz GS, Nair V, Martini S, Eichinger F, Vining C, Berthier CC, Kretzler M, et al. 2010. A molecular profile of focal segmental glomerulosclerosis from formalin-fixed, paraffin-embedded tissue. *The American Journal of Pathology* **177**: 1674–86.

Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG. 2008. The Sleipnir library for computational functional genomics. *Bioinformatics* **24**: 1559–61.

Ingelfinger JR. 2011. MYO1E, focal segmental glomerulosclerosis, and the cytoskeleton. *The New England Journal of Medicine* **365**: 368–9.

Jain S, De Petris L, Hoshi M, Akilesh S, Chatterjee R, Liapis H. 2011. Expression profiles of podocytes exposed to high glucose reveal new insights into early diabetic glomerulopathy. *Laboratory Investigation* **91**: 488–98.

Joachims T. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, p. 217, ACM Press, New York, New York, USA.

Ju W, Eichinger F, Bitzer M, Oh J, McWeeney S, Berthier CC, Shedden K, Cohen CD, Henger A, Krick S, et al. 2009. Renal gene and protein expression signatures for prediction of kidney disease progression. *The American Journal of Pathology* **174**: 2073–85.

Kao WHL, Klag MJ, Meoni LA, Reich D, Berthier-Schaad Y, Li M, Coresh J, Patterson N, Tandon A, Powe NR, et al. 2008. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature Genetics* **40**: 1185–92.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Research* **37**: D767–72.

Kim JM, Wu H, Green G, Winkler CA, Kopp JB, Miner JH, Unanue ER, Shaw AS. 2003. CD2-associated protein haploinsufficiency is linked to glomerular disease susceptibility. *Science* **300**: 1298–300.

Koop K, Eikmans M, Baelde HJ, Kawachi H, De Heer E, Paul LC, Bruijn JA. 2003. Expression of podocyte-associated molecules in acquired human kidney diseases. *Journal of the American Society of Nephrology* **14**: 2063–71.

Kopp JB, Nelson GW, Sampath K, Johnson RC, Genovese G, An P, Friedman D, Briggs W, Dart R, Korbet S, et al. 2011. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. *Journal of the American Society of Nephrology* **22**: 2129–37.

Kopp JB, Smith MW, Nelson GW, Johnson RC, Freedman BI, Bowden DW, Oleksyk T, McKenzie LM, Kajiyama H, Ahuja TS, et al. 2008. MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nature Genetics* **40**: 1175–84.

Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith A V, et al. 2010. New loci associated with kidney function and chronic kidney disease. *Nature Genetics* **42**: 376–84.

Kriz W, Elger M, Nagata M, Kretzler M, Uiker S, Koeppen-Hageman I, Tenschert S, Lemley K V. 1994. The role of podocytes in the development of glomerular sclerosis. *Kidney International Supplement* **45**: S64–72.

Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R. 2011. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature Methods* **8**: 945–7.

Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Kusek JW, Van Lente F. 2006. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Annals of Internal Medicine* **145**: 247–54.

Lindenmeyer MT, Eichinger F, Sen K, Anders H-J, Edenhofer I, Mattinzoli D, Kretzler M, Rastaldi MP, Cohen CD. 2010. Systematic analysis of a novel human renal glomerulus-enriched gene expression dataset. ed. G. Chua. *PLoS ONE* **5**: e11545.

Lorz C, Benito-Martín A, Boucherot A, Ucero AC, Rastaldi MP, Henger A, Armelloni S, Santamaría B, Berthier CC, Kretzler M, et al. 2008. The death ligand TRAIL in diabetic nephropathy. *Journal of the American Society of Nephrology* **19**: 904–14.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **39**: D52–7.

McMahon AP, Aronow BJ, Davidson DR, Davies JA, Gaido KW, Grimmond S, Lessard JL, Little MH, Potter SS, Wilder EL, et al. 2008. GUDMAP: the genitourinary developmental molecular anatomy project. *Journal of the American Society of Nephrology* **19**: 667–71.

Mele C, Iatropoulos P, Donadelli R, Calabria A, Maranta R, Cassis P, Buelli S, Tomasoni S, Piras R, Krendel M, et al. 2011. MYO1E mutations and childhood familial focal segmental glomerulosclerosis. *The New England Journal of Medicine* **365**: 295–306.

Niewold TB. 2011. Stabilizing the Kidney's Skeleton. *Science Translational Medicine* **3**: 95ec128–95ec128.

Pavenstädt H. 2000. Roles of the podocyte in glomerular function. *American Journal of Physiology Renal Physiology* **278**: F173–9.

Reddy MM, Wilson R, Wilson J, Connell S, Gocke A, Hynan L, German D, Kodadek T. 2011. Identification of candidate IgG biomarkers for Alzheimer's disease via combinatorial library screening. *Cell* **144**: 132–42.

Reich HN, Tritchler D, Cattran DC, Herzenberg AM, Eichinger F, Boucherot A, Henger A, Berthier CC, Nair V, Cohen CD, et al. 2010. A molecular signature of proteinuria in glomerulonephritis. ed. A.B. Khodursky. *PLoS ONE* **5**: e13451.

Roselli S, Heidet L, Sich M, Henger A, Kretzler M, Gubler M-C, Antignac C. 2004. Early glomerular filtration defect and severe renal disease in podocin-deficient mice. *Molecular and Cellular Biology* **24**: 550–60.

Saleem MA, Zavadil J, Bailly M, McGee K, Witherden IR, Pavenstadt H, Hsu H, Sanday J, Satchell SC, Lennon R, et al. 2008. The molecular and functional phenotype of glomerular podocytes reveals key features of contractile smooth muscle cells. *American Journal of Physiology Renal Physiology* **295**: F959–70.

Sato Y, Wharram BL, Lee SK, Wickman L, Goyal M, Venkatareddy M, Chang JW, Wiggins JE, Lienczewski C, Kretzler M, et al. 2009. Urine podocyte mRNAs mark progression of renal disease. *Journal of the American Society of Nephrology* **20**: 1041–52.

Sawai K, Mukoyama M, Mori K, Yokoi H, Koshikawa M, Yoshioka T, Takeda R, Sugawara A, Kuwahara T, Saleem MA, et al. 2006. Redistribution of connexin43 expression in glomerular podocytes predicts poor renal prognosis in patients with type 2 diabetes and overt nephropathy. *Nephrology, Dialysis, Transplantation* **21**: 2472–7.

Schmid H, Boucherot A, Yasuda Y, Henger A, Brunner B, Eichinger F, Nitsche A, Kiss E, Bleich M, Gröne H-J, et al. 2006. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* **55**: 2993–3003.

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. 2010. Cell type-specific gene expression differences in complex tissues. *Nature Methods* **7**: 287–9.

Sistani L, Dunér F, Udumala S, Hultenby K, Uhlen M, Betsholtz C, Tryggvason K, Wernerson A, Patrakka J. 2011. Pdlim2 is a novel actin-regulating protein of podocyte foot processes. *Kidney International* **80**: 1045–54.

Steiglitz BM, Keene DR, Greenspan DS. 2002. PCOLCE2 encodes a functional procollagen C-proteinase enhancer (PCPE2) that is a collagen-binding protein differing in distribution of expression and post-translational modification from the previously described PCPE1. *The Journal of Biological Chemistry* **277**: 49820–30.

Tanaka M, Asada M, Higashi AY, Nakamura J, Oguchi A, Tomita M, Yamada S, Asada N, Takase M, Okuda T, et al. 2010. Loss of the BMP antagonist USAG-1 ameliorates disease in a mouse model of the progressive hereditary kidney disease Alport syndrome. *The Journal of Clinical Investigation* **120**: 768–77.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 5116–21.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research* **14**: 54–61.

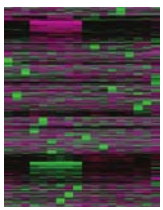
Woroniecka KI, Park ASD, Mohtat D, Thomas DB, Pullman JM, Susztak K. 2011. Transcriptome analysis of human diabetic kidney disease. *Diabetes* **60**: 2354–69.

Yaoita E, Yao J, Yoshida Y, Morioka T, Nameta M, Takata T, Kamiie J, Fujinaka H, Oite T, Yamamoto T. 2002. Up-regulation of connexin43 in glomerular podocytes in response to injury. *The American Journal of Pathology* **161**: 1597–606.

Iterative In Silico Nano-dissection

Lineage Specific Standards

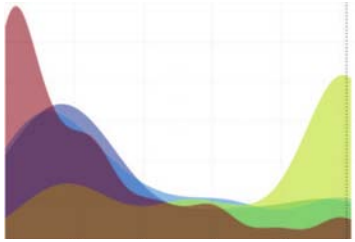
Tissue Biopsy (Gene Expression)



$$z_j = \frac{U_j - \frac{n_j n_p}{2}}{\sqrt{\frac{n_j n_p (n_j + n_p + 1)}{12}}}$$

$$p = \sqrt[M]{\prod_{j=1}^{M} (1 - cdf(z_j))}$$

Refine



Assess

Top Novel Predictions:

Gene Name	Estimated Probability
PCOLCE2	0.97
CHI3L1	0.96
FGF1	0.95
PLA2R1	0.95
SPOCK1	0.94
TPPP3	0.92

Machine Learning Algorithm (SVM)

Predict

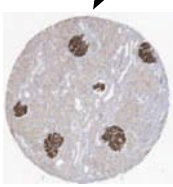
Validate and Compare

Evaluate Clinical Relevance

RNA level

Protein Level

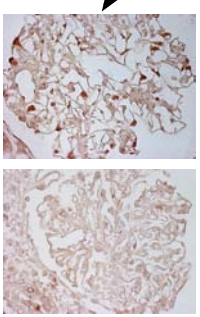
GUDMAP
GenitoUrinary Molecular Anatomy Project



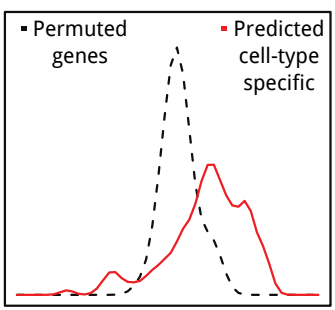
The Human Protein Atlas

Control

FSGS

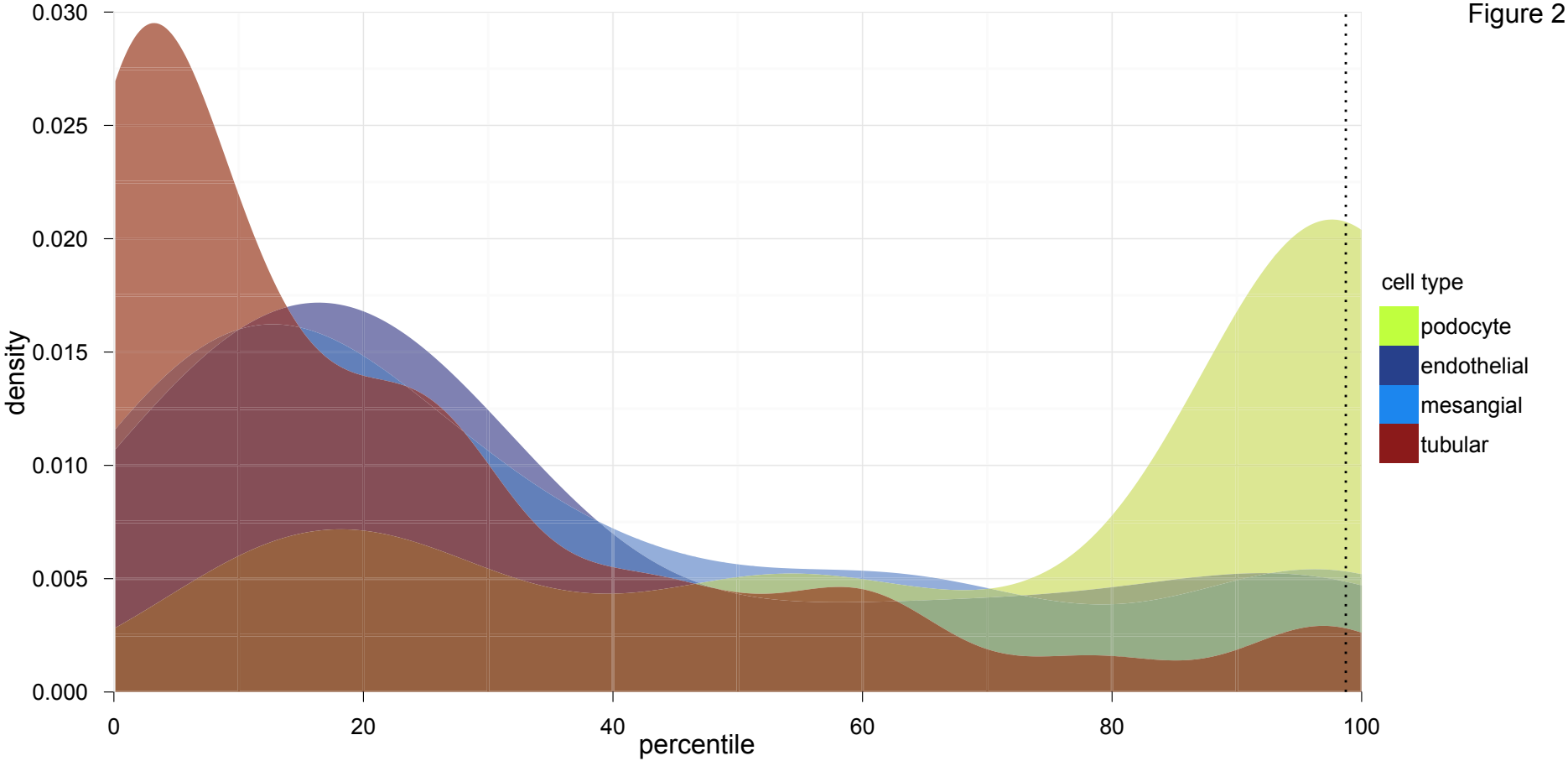


Tissue Expression in Disease

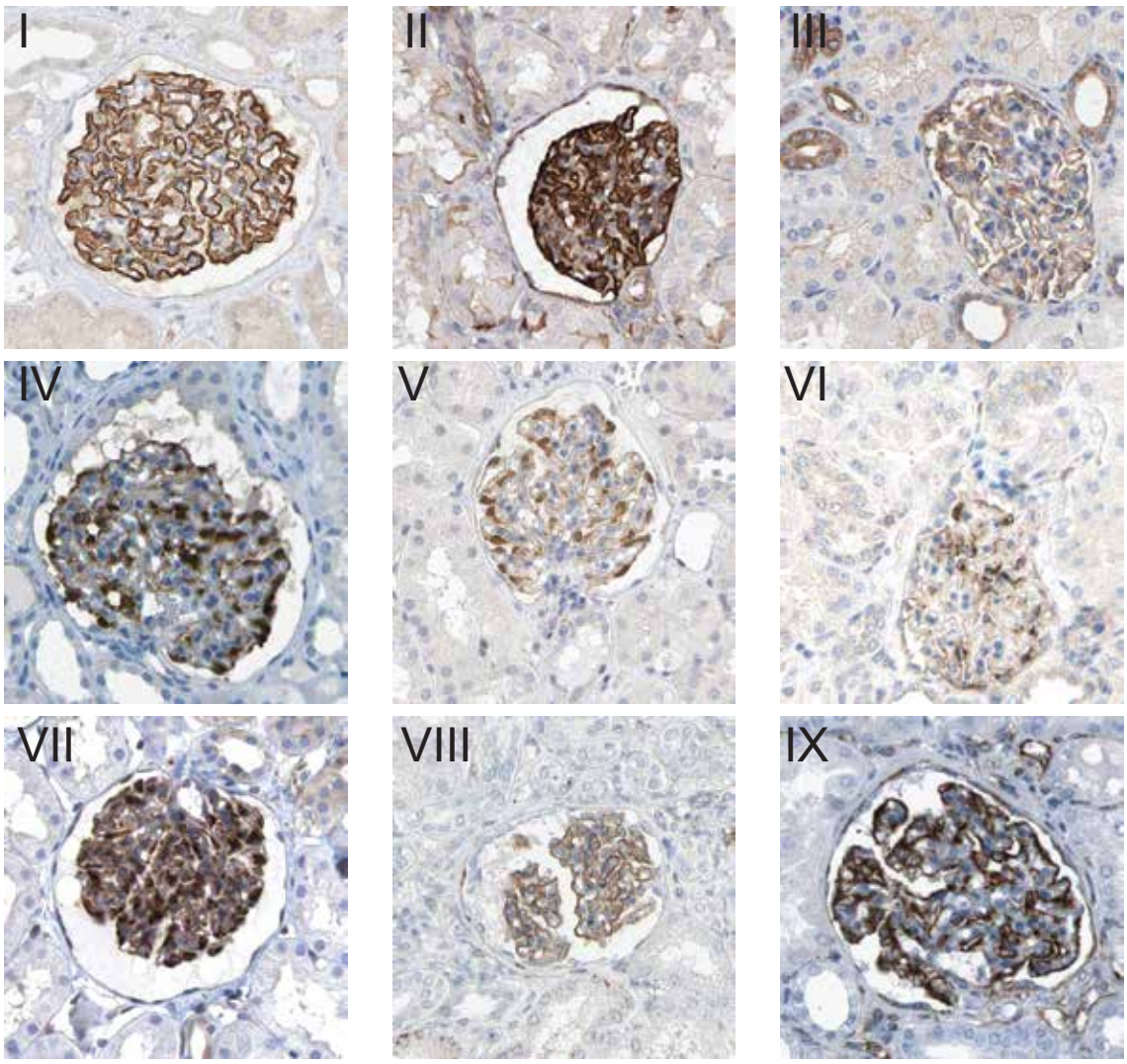


Correlation with Renal Function (CKD) in Disease

Figure 2.


0.030
0.025
0.020
0.015
0.010
0.005
0.000
density
0
20
40
60
80
100
percentile
cell type
podocyte
endothelial
mesangial
tubular

A



B