Integration of GWAS Summary Statistics and Gene Expression Reveals Target Cell Types Underlying Kidney Function Traits

Yong Li,1 Stefan Haug,1 Pascal Schlosser,1 Alexander Teumer,2,3 Adrienne Tin,4 Cristian Pattaro,5 Anna Köttgen,1,4 and Matthias Wuttke1

1Institute of Genetic Epidemiology, Medical Center–University of Freiburg, Freiburg, Germany
2Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
3German Center for Cardiovascular Research, Partner Site Greifswald, Greifswald, Germany
4Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
5Eurac Research, Institute for Biomedicine (affiliated with the University of Lübeck), Bolzano, Italy

ABSTRACT

Background Genetic variants identified in genome-wide association studies (GWAS) are often not specific enough to reveal complex underlying physiology. By integrating RNA-seq data and GWAS summary statistics, novel computational methods allow unbiased identification of trait-relevant tissues and cell types.

Methods The CKDGen consortium provided GWAS summary data for eGFR, urinary albumin-creatinine ratio (UACR), BUN, and serum urate. Genotype-Tissue Expression Project (GTEx) RNA-seq data were used to construct the top 10% specifically expressed genes for each of 53 tissues followed by linkage disequilibrium (LD) score–based enrichment testing for each trait. Similar procedures were performed for five kidney single-cell RNA-seq datasets from humans and mice and for a microdissected tubule RNA-seq dataset from rat. Gene set enrichment analyses were also conducted for genes implicated in Mendelian kidney diseases.

Results Across 53 tissues, genes in kidney function–associated GWAS loci were enriched in kidney ($P = 9.1E-8$ for eGFR; $P = 1.2E-5$ for urate) and liver ($P = 6.8 \cdot 10^{-5}$ for eGFR). In the kidney, proximal tubule was enriched in humans ($P = 8.5E-5$ for eGFR; $P = 7.8E-6$ for urate) and mice ($P = 0.0003$ for eGFR; $P = 0.0002$ for urate) and confirmed as the primary cell type in microdissected tubules and organoids. Gene set enrichment analysis supported this and showed enrichment of genes implicated in monogenic glomerular diseases in podocytes. A systematic approach generated a comprehensive list of GWAS genes prioritized by cell type–specific expression.

Conclusions Integration of GWAS statistics of kidney function traits and gene expression data identified relevant tissues and cell types, as a basis for further mechanistic studies to understand GWAS loci.

JASN 31: 2326–2340, 2020. doi: https://doi.org/10.1681/ASN.2020010051

Genome-wide association studies (GWAS) are a powerful method to identify genetic variants associated with complex traits and diseases, but a systematic translation of associated gene loci into mechanistic insights has been challenging and is still lacking for the majority of loci.1 One reason that has complicated such translational efforts is that target tissues and cell types important in disease pathogenesis are partly unknown, both with respect to the
global genetic architecture of complex diseases and traits as well as on the level of the individual gene.

The analytical integration of findings from GWAS of human traits and diseases with information on gene expression represents an attractive option to link two unbiased sources of evidence to connect genetic variation to the respective phenotypes.1–3 Advances in bulk tissue as well as single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA sequencing (snRNA-seq)4–7 together with the establishment of community gene expression profiling efforts, such as the GTEx Project,8,9 and the availability of murine and human kidney single-cell transcriptomes10,11 offer new opportunities to highlight trait-relevant target tissues and cell types in an unbiased manner.

Such opportunities are of high interest in areas where few specific treatments exist and novel therapeutic options are lacking, such as in CKD. CKD affects 10% of the global adult population,12 is a leading cause of death,13 is treated insufficiently and largely symptomatically,14 and lags behind other disciplines with respect to the number of clinical trials.15 In this project, we therefore capitalize on the recent generation of powerful GWAS summary statistics from the CKDGen Consortium for kidney function–related traits—eGFR,16 urinary albumin-creatinine ratio (UACR),17 BUN,16 and serum urate.18

Our objective was to identify relevant target tissues in which the genes associated with these traits are highly expressed, followed by the subsequent identification of relevant kidney cell types for the respective traits as a whole and also at the level of the individual genes in order to pave the way for translational studies.

**METHODS**

**GWAS Summary Statistics**

GWAS summary statistics of the four kidney function traits eGFR, UACR, BUN, and serum urate16–18 were downloaded from the CKDGen Consortium Data website (https://ckdgen.imbi.uni-freiburg.de). To match the European linkage disequilibrium (LD) score reference in linkage disequilibrium score regression for specifically expressed genes (LDSC-seq), we used the European ancestry–specific GWAS summary statistics. GWAS summary statistics for asthma (ICD10 code J45) from UK Biobank were downloaded from Gene Atlas,19 and those for schizophrenia from the CLOZUK + PGC2 Consortium20 were downloaded from the Walters group data repository (https://walters.psycm.cf.ac.uk/).

**Genetic Correlations of GWAS Traits**

Pairwise genetic correlations and single-nucleotide polymorphism (SNP)–based heritability of the six GWAS traits were estimated by LDSC21 software with the option “–rg.”22 LD scores were calculated for all SNPs that were part of the HapMap Project Phase 323 using the 1000 Genomes Project Phase 3 European population genotypes as the LD reference.24 The MHC region was excluded from the analysis due to its unusual LD pattern as recommended. A Bonferroni correction was applied to control for multiple testing.

**Processing of Gene Expression Datasets**

GTEx V7 RNA-sequencing (RNA-seq) data was downloaded as transcripts per million (TPM) from the GTEx Portal (https://gtexportal.org/home/). Genes for which fewer than four samples had at least one read count per million in the TPM matrix were filtered out. We used 53 tissues grouped by 30 categories as defined in the provided sample annotation file. Sample sizes of the tissue and tissue categories are available in Supplemental Table 1. Age and sex were extracted from the subject phenotype file and used as covariates in linear models.

Human kidney scRNA-seq data (Young et al.)11 This dataset was used as our main dataset and includes single-cell transcriptomes from a number of renal tumor samples as well as samples from healthy fetal, pediatric, adolescent, and adult kidneys. We used the data of 8707 cells in the bar-code group of normal epithelium and vasculature without the bulk of proximal tubule (PT) cells, which originated from five healthy adult kidneys. The authors’ R code (https://github.com/constantAmateur/scKidneyTumors) was adapted to process and normalize the data using the R package Seurat25 v2.3.4. The individual 10× sequencer channel information was corrected for by the batch correction algorithm “ComBat” while preserving zero expression values. Expression data were then normalized by the Seurat function “NormalizeData” with default parameters. Cluster assignment of the cells was taken from Supplemental Table 11.

Mouse kidney scRNA-seq data (Park) were downloaded as an RNA-seq count matrix of 43,745 kidney cells from seven mice from Gene Expression Omnibus10 (GEO; accession no. GSE107585). Data were processed with Seurat25 v2.3.4. Normalization was performed by the function “NormalizeData” with default parameters, and the Seurat function “removeBatchEffect” was used to correct for the batch variable “mouse identifiers.” Cluster assignment information was extracted from the cell identifiers.

Human kidney snRNA-seq data (Lake) were provided by Lake et al.26 This dataset from the Kidney Precision Medicine Project Consortium includes single-nucleus transcriptomes from tumor-free regions of nephrectomies and discarded
deceased donor kidneys from 15 different individuals. The Seurat object file contained gene expression values of 17,659 single nuclei in 30 cell clusters that had passed quality control.

Human kidney snRNA-seq data (Wu) were downloaded as an RNA-seq count matrix of 4524 single nuclei from a healthy human adult from GEO (accession no. GSE118184). Data were processed using Seurat v2.3.4 as described previously. Normalization by the “NormalizeData” function was performed with default parameters. Assignment of cells to 17 cell clusters was provided by the authors.

Human kidney organoid scRNA-seq data originated from Wu et al. and were downloaded as a count matrix from GEO (accession no. GSE118184). Two subsets of the induced pluripotent stem cell organoid data were used: 18,072 cells with the Morizane protocol and 26,890 cells with the Takasato protocol. Processing of the datasets followed the same procedures as in the snRNA-seq data by Wu et al.

Rat microdissected tubule scRNA-seq data (Lee) are available from supplemental dataset 1 of Lee et al. RNA-seq was used to profile gene expression in each of 14 microdissected renal tubule segments and glomeruli of rat kidneys. We imported data from the worksheet “RPMK_withPAseq,” which includes reads per kilobase of exon model per million mapped reads of 105 tubule samples and 5 glomeruli samples.

Mouse glomerulus scRNA-seq data (Karaikos) were downloaded from https://shiny.mdc-berlin.de/mgsca/. Karaikos et al. used scRNA-seq of 12,954 mouse glomerular cells to generate a normalized and filtered gene expression matrix with cluster information. We processed the data containing five major cell clusters using Seurat v2.3.4 and performed normalization using the same procedures as with the other single-cell datasets.

Cell cluster labels in each single-cell dataset were taken from the respective publication, with adjustment according to the nomenclature of Chen et al. For visualization of expression values as heat maps for GWAS-implicated genes, we transformed the average expression values across cell types to z scores. The transformation was applied to the output of the Seurat function “AverageExpression.” Heat maps were created using the Bioconductor R package ComplexHeatmap. We highlighted known monogenic kidney disease genes in heat maps on the basis of their categorization from Rasouly et al.

Heritability Enrichment by LDSC-seg
We used LDSC-seg to define specifically expressed genes in each tissue or kidney cell type. In brief, t statistics for each gene were extracted from a linear model fit of gene expression on tissue or cell-type membership, where in the design matrix, the sample or cell was set to 1 when belonging to the focal tissue or cell type and −1 otherwise.

We used the TPM values for GTEx, normalized expression values of Seurat objects for single-cell datasets, and reads per kilobase of exon model per million mapped reads values for rat tubule data as the dependent variables. In GTEx data, we excluded samples in the same tissue category as the focal tissue when fitting the model (e.g., other brain regions when looking at a specific brain region) and included age and sex as covariates. In the single-cell and rat tubule datasets, each cell type was compared with all other cell types. We then ranked genes by their t statistics and selected the top 10% of genes of each GTEx tissue and kidney cell type to define a set of specifically expressed genes. Sensitivity analyses using the top 5% and 20% to define specifically expressed genes showed robustness of the results to the choice of the 10% threshold. The human homologs of mouse and rat genes were obtained by querying Ensembl Biomart.

After the top 10% specifically expressed genes were obtained, LD score regression was performed by LDSC software for each GWAS summary statistic file in order to test whether trait heritability was enriched in regions surrounding these genes in each evaluated tissue or cell type. Annotation files were first constructed using the specifically expressed genes as input and adding 100 kb on both sides of the transcribed regions. A control annotation file was created for each expression dataset including all genes existing in the dataset. LD scores were then calculated for each annotation. Other parameters were the same as in the LDSC genetic correlation analyses. The magnitude and significance of per-SNP heritability enrichment were then tested for each annotation conditional on all genes’ annotation and the annotations in the baseline model v1.1. Enrichment was defined by the $h^2$ estimate of the individual annotation divided by the proportion of SNPs contained within the annotation. The coefficient $P$ value was calculated from the coefficient ($\tau$) z score.

LDSC-seg analyses were performed for six GWAS traits. Genome-wide correlation analyses demonstrated the four kidney function traits being correlated with each other. Therefore, we defined the significance threshold using a Bonferroni correction for the number of tissues or cell types and the number of trait categories (kidney, asthma, schizophrenia). For example, in the GTEx dataset, the P-value cutoff was 3.1e-4 (0.05 per 53 tissues per three categories). Additionally, we defined a “suggestive” $P$-value threshold correcting for the number of tissues or cell types only (e.g., GTEx: 9.4e-4=0.05 per 53 tissues).

Gene Set Enrichment of Mendelian Kidney Disease Genes
We used the expression-weighted cell-type enrichment (EWCE) method implemented in the R package EWCE to test if the expression of known monogenic kidney disease genes inherited in a Mendelian fashion was enriched in certain kidney cell types. A list of 625 kidney disease genes was obtained, divided into seven subcategories: congenital or developmental, electrolytes or metabolic, glomerular, tubulointerstitial, cancer, genital, and secondary renal diseases. We tested the enrichment using all seven gene sets, although the last three sets do not represent primary Mendelian kidney disease genes. The specificity matrices were generated by the EWCE.
“generate.celltype.data” function in the human (Young) and mouse (Park) single-cell datasets, as well as two human kidney organoid datasets. Subsequently, EWCE was run with 100,000 bootstrap replicates controlling for transcript length and GC content. We performed multiple-tests correction using the Bonferroni procedure, accounted for the number of cell types and gene sets, and controlled the $\alpha$ at 0.05.

**Projecting Cell Types across Four scRNA-Seq Datasets**

The “scmapCluster” function in R package scmap36 (v1.4.1) was used to project cell-type annotations across the four single-cell kidney datasets. We started with a similarity threshold of 0.7 and decreased it in steps of 0.1 until the proportion of “unassigned” cells was <0.3 (lowest similarity threshold was 0.3). Sankey plots were created using the “plot” function in scmap. Because the projection was bidirectional, for each pair we plotted only the projection with the lower “unassigned” cells proportion.

**Cell-Type Enrichment by Multimarker Analysis of GenoMic Annotation**

As a secondary analysis, the R packages EWCE35 and Multi-marker Analysis of GenoMic Annotation (MAGMA) Celltyping37 were used to calculate cell-type specificity of gene expression and its association with GWAS traits. MAGMA v1.0.738 was used to calculate gene-level association statistics using a window of 10 kb upstream and 1.5 kb downstream of each gene, accounting for LD, gene length, gene density, sample size, and minor allele count. Because the top 10% of specifically expressed genes were used for the heritability enrichment analyses in LDSC-seg, the “top 10% mode” of MAGMA was used to match these analyses.

**RESULTS**

This project built on several datasets that have recently been generated and are on the basis of genetic data from up to 1 million individuals. Tables 1 and 2 display an overview of the investigated kidney function traits GWAS and gene expression datasets. Summary statistics from asthma19 and schizophrenia20 as traits that are not known to be connected to kidney function and disease were used as negative controls. Gene expression data of 53 human tissues and cell types were obtained from the GTEx Project (V7),9 as well as from scRNA-seq projects from primary kidney tissue of humans11,26,27 and mice,10,28 kidney organoids,27 and a microdissected tubule RNA-seq dataset from rats.28 The harmonized kidney cell types from human and mouse single-cell datasets are shown in Figure 1.

As a first step, pairwise genome-wide genetic correlations of the six investigated traits were determined (Methods), and they are summarized in Supplemental Table 2. Significant ($P<0.003$) genetic correlations were observed for five of six pairs of kidney function traits. eGFR was negatively correlated with BUN ($r_g = -0.31; P = 3.8E-14$) and serum urate ($r_g = -0.21; P = 1.3E-9$). The correlation with UACR was positive ($r_g = 0.34; P = 2.6E-33$), which might be caused by UACR containing creatinine as its denominator and urinary albumin values in the numerator being generally low in the mostly population-based studies of the CKDGen Consortium. The observed genetic correlations are consistent with estimates in studies of the individual kidney function traits.16–18 Asthma showed a moderate genetic correlation with serum urate ($r_g = 0.11; P = 0.001$), and schizophrenia was not correlated with any other trait.

We next applied LDSC-seg,33 which uses GWAS summary statistics to test for the presence of heritability enrichment of specifically expressed genes across human tissues, to identify kidney function–relevant tissues (Methods, Supplemental Table 3). Significant ($P<2.4E-4$) enrichment of heritability of eGFR was observed in regions around genes specifically expressed in kidney ($P = 9.1E-8; 2.2$-fold enrichment) and liver ($P = 6.8E-8; 2.2$-fold enrichment) (Figure 2). Significant enrichment in kidney tissue was also observed for BUN ($P = 1.2E-4; 1.9$-fold enrichment) and serum urate ($P = 1.2E-5; 2.1$-fold enrichment). There was no significant enrichment for UACR, but nominally significant heritability enrichment was observed in kidney, liver, and aorta (Figure 2). Heritability of schizophrenia, the negative control, was enriched in regions surrounding genes specifically expressed in multiple brain but no other tissues (Figure 2). Asthma did not show significant enrichment in any tissue.

Given the strong enrichment of heritability in the kidney, we next investigated which kidney cell types showed enriched heritability in genomic regions around genes that are highly expressed in specific cell types. Lists of specifically expressed genes as determined by a linear model fit of gene expression on tissue or cell-type membership using both a human scRNA-seq dataset11 and a murine scRNA-seq dataset10 are presented in Supplemental Tables 4 and 5, respectively. The validity of these lists is underscored by the presence of many positive controls, such as NPHS2 and NPHS1 as top genes for podocytes, LRP2 and CUBN for PT cells, and UMOD and several claudins for the loop of Henle. To facilitate linking this

### Table 1. GWAS datasets used in the analyses

<table>
<thead>
<tr>
<th>Source</th>
<th>Phenotype</th>
<th>Sample Size</th>
<th>Reference: PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKDGen</td>
<td>eGFR</td>
<td>567,460</td>
<td>31152163</td>
</tr>
<tr>
<td>CKDGen</td>
<td>BUN</td>
<td>243,031</td>
<td>31152163</td>
</tr>
<tr>
<td>CKDGen</td>
<td>UACR</td>
<td>547,361</td>
<td>31511532</td>
</tr>
<tr>
<td>CKDGen</td>
<td>Urate</td>
<td>288,666</td>
<td>31578528</td>
</tr>
<tr>
<td>UKB</td>
<td>Asthma</td>
<td>28,628 patients; 423,636 controls</td>
<td>30349118</td>
</tr>
<tr>
<td>CLOZUK + PGC2</td>
<td>Schizophrenia</td>
<td>40,675 patients; 64,643 controls</td>
<td>29483656</td>
</tr>
</tbody>
</table>

PMID, PubMed ID; UKB, UK Biobank.
information to existing knowledge about monogenic kidney diseases, we annotated the presence and type of monogenic kidney disease\textsuperscript{32} for each gene and cell type. These annotations also showed cell-type specificity. For example, four of the ten most specifically expressed genes in the human distal tubule are known as monogenic kidney disease genes for electrolyte or metabolic disorders (Supplemental Table 4).

Investigating the four kidney function GWAS for heritability enrichment in the human kidney scRNA-seq dataset,\textsuperscript{11} we found significant ($P<5.2E-4$) heritability enrichment for eGFR and urate in a PT cluster PT1 ($P=8.5E-5$ and $P=7.8E-6$, respectively). Suggestive ($P<2.3E-3$) enrichment was also found for another PT cluster, PT2, for eGFR and urate, as well as in distal tubule cells for eGFR (Figure 3). Of note, a similar enrichment $P$ value despite the considerably smaller sample size of the serum urate GWAS ($n=288,666$) than the eGFR GWAS ($n=567,460$) may suggest that PT has a more specific role in urate metabolism than in regulation of eGFR. Association statistics for BUN also looked similar, but showed no significantly enriched clusters, at a GWAS summary statistic sample size comparable with serum urate ($n=243,031$). Conversely, for UACR, the highest heritability enrichment was observed in glomerulus, podocyte, and PT. None of these were significant, but the finding reflects existing physiologic knowledge about albuminuria, with a role of the glomerulus in the filtration of albumin and the role of the PT in its subsequent receptor-mediated reabsorption.

Single-cell data from mouse kidney\textsuperscript{10} implicated the same cell types (Figure 3). Although only the PT met multiple testing–corrected criteria for significant enrichment, the consistency in implicated cell types indicates high homology of important genes across species. LDSC results for all traits are shown in Supplemental Tables 6 (human) and 7 (mouse). The results were robust to the choice of expression dataset: the use of two publicly available snRNA-seq datasets from humans\textsuperscript{26,27} confirmed very similar patterns of enrichment, with strongest results in the PT (Supplemental Figure 1, Supplemental Tables 8 and 9). In the snRNA-seq dataset by Wu et al.,\textsuperscript{27} the PT cells were separated into three clusters corresponding to segments 1–3 of the PT. Although the strongest enrichment was in segment 2 ($P=1.5E-4$) for eGFR, it was segment 1 for urate ($P=1.3E-5$).

The LDSC-seg results of the microdissected rat tubule data\textsuperscript{28} and a mouse glomerulus single-cell dataset\textsuperscript{29} (Supplemental Figure 2, Supplemental Tables 10 and 11) further confirmed these findings. In the rat tubule data, the PT segments showed significant enrichment for eGFR, thus providing evidence that our main findings can be recapitulated on the single functional unit of the kidney. In the mouse glomerulus data, tubule cells were enriched for BUN, eGFR, and urate, whereas their proportion was only 6%, indicating the robustness of the observed PT enrichment to cell cluster size.

No kidney cell types in any of the four single-cell and single-nucleus datasets showed evidence for heritability enrichment on the basis of the negative control schizophrenia. For asthma, a few suggestive enrichment signals were observed, including distal tubule in the human single-cell dataset ($P=0.002$), T lymphocyte in the mouse single-cell dataset ($P=0.003$), and macrophage in the human single-nucleus dataset by Wu et al.\textsuperscript{27} ($P=5.9E-4$). T lymphocytes and macrophages are important cell types in the etiology of asthma.\textsuperscript{39,40} However, the heritability enrichment in these cell types does not necessarily suggest that these immune cells in the kidney play important roles in asthma. A more reasonable explanation is that these cells have unique gene expression patterns so that the enrichment can be also detected in the kidney.

To further substantiate these results, we applied a complementary enrichment method\textsuperscript{37} that uses an MAGMA\textsuperscript{38}-based approach. Different from the LDSC-seg, this method tests the association between gene-level statistics with cell-type specificity. After computing gene-level association statistics and estimating the cell-type specificity with serum urate was also statistically significant ($P=1E-5$ and $P=1.7E-3$, respectively), further supporting the important and specific role of the PT in urate metabolism. Similar results were observed on the basis of murine gene expression, where PT cell specificity showed a significant association with eGFR and urate association statistics ($P=5.6E-5$ and $P=2.3E-6$, respectively). Altogether, these

### Table 2. Expression datasets used in the analyses

<table>
<thead>
<tr>
<th>Name</th>
<th>Species</th>
<th>Dataset Type</th>
<th>Sample Size</th>
<th>Reference: PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx v7</td>
<td>Human</td>
<td>Bulk RNA-seq</td>
<td>11,688 samples, 714 donors</td>
<td>25954001</td>
</tr>
<tr>
<td>Park</td>
<td>Mouse</td>
<td>scRNA-seq</td>
<td>43,745 cells (7 mice)</td>
<td>29622724</td>
</tr>
<tr>
<td>Young</td>
<td>Human</td>
<td>scRNA-seq</td>
<td>8707 cells (5 individuals)</td>
<td>30093597</td>
</tr>
<tr>
<td>Lake</td>
<td>Human</td>
<td>snRNA-seq</td>
<td>17,659 nuclei (15 individuals)</td>
<td>31249312</td>
</tr>
<tr>
<td>Wu</td>
<td>Human</td>
<td>snRNA-seq</td>
<td>4524 nuclei (1 individual)</td>
<td>30449713</td>
</tr>
<tr>
<td>Lee</td>
<td>Rat</td>
<td>Bulk RNA-seq</td>
<td>110 samples</td>
<td>25817355</td>
</tr>
<tr>
<td>Karaiskos</td>
<td>Mouse</td>
<td>scRNA-seq</td>
<td>12,954 cells (8 mice)</td>
<td>29794128</td>
</tr>
<tr>
<td>Organoid (Morizane)</td>
<td>Human</td>
<td>scRNA-seq</td>
<td>18,072 cells (1 individual)</td>
<td>30449713</td>
</tr>
<tr>
<td>Organoid (Takasato)</td>
<td>Human</td>
<td>scRNA-seq</td>
<td>26,890 cells</td>
<td>30449713</td>
</tr>
</tbody>
</table>

PMID, PubMed ID.
results demonstrate that the significantly enriched cell types can be detected by independent methods. Full MAGMA association results are summarized in Supplemental Figure 3 and Supplemental Tables 12 (human) and 13 (mouse).

GWAS of eGFR, UACR, and serum urate have implicated genes underlying the association signals by functional annotation combined with statistical fine mapping or colocalization analyses. This process resulted in lists of genes most likely to causally influence the studied traits, which may not be the genes closest to the SNPs with the strongest association signals. We therefore examined the expression patterns of these genes in human and murine kidney cell types. A strong expression signal was observed for many eGFR genes in PT cells (Figure 4), in agreement with the findings from the heritability enrichment analyses. UACR-associated genes were highly expressed in podocytes (Figure 4, Supplemental Figure 4). The benefit of examining gene expression in multiple single-cell datasets from different species is illustrated by the example of CACNA1S, which shows high expression in B lymphocytes in mice but not in B lymphocytes or any other cell types in human kidney. Similar to eGFR genes, urate-associated genes showed a strong expression signal in PT cells (Supplemental Figure 5 for human and Supplemental Figure 6 for mice), consistent with the heritability enrichment findings for urate.

To help prioritize genes in the eGFR and UACR GWAS loci, we built candidate lists containing all genes within a 500-kb window around each locus’ index SNP. For these genes, we list the cell types (if any) where the gene appears in the top 10% specifically expressed gene list (Supplemental Table 14). For example, rs10846157 is an index SNP identified in GWAS of eGFR (P=8.3E-21). Although the closest gene is RERG, using this approach prioritizes the neighboring gene PTPRO.
Figure 2. Heritability enrichment for six traits in GTEx tissues. Bar plots of $-\log_{10}$ (P values) for four kidney traits, asthma, and schizophrenia across GTEx tissues. The dashed red lines denote Bonferroni-adjusted significance ($P=0.05/53/3=3.1E-4$), the dotted red lines denote suggestive significance ($P=0.05/53=9.4E-4$), and the dotted gray lines denote nominal significance ($P=0.05$). The color intensity of the bars denotes the enrichment values. SCZ, schizophrenia.
PTPRO is specifically expressed in podocytes and well known as a cause of monogenic steroid-resistant nephrotic syndrome (OMIM #614196). This systematic approach results in a comprehensive list of GWAS genes prioritized by cell type–specific expression.

Complementary to genes implicated by GWAS, we next asked whether certain gene sets known to cause Mendelian forms of kidney disease were expressed at higher levels within a particular cell type than expected by chance, using the EWCE method (Methods). To this end, 625 genes

Figure 3. Heritability enrichment for kidney traits in human and mouse single-cell (sc) datasets. Bar plots of $-\log_{10}(P)$ values for four kidney traits across cell types in human and mouse sc datasets. The dashed red lines denote Bonferroni-adjusted significance ($P=0.05/24/3=6.9E-4$ in human and $P=0.05/16/3=1.0E-3$ in mouse), the dotted red lines denote suggestive significance ($P=0.05/24=2.1E-3$ and $P=0.05/16=3.1E-3$), and the dotted gray lines denote nominal significance ($P=0.05$). The color intensity of the bars denotes the enrichment values. AVR, ascending vasa recta; B-lymph, B lymphocyte; CD-Trans, collecting duct transient cell; DCT, distal convoluted tubule; DVR, descending vasa recta; Endo, endothelial cell; Fib, fibroblast; GE, glomerular epithelium; IC, intercalated cell; IC-A, intercalated cell type A; IC-B, intercalated cell type B; LOH, loop of Henle; Macro, macrophage; MC, mesangial cell; NE, nephron epithelium; Neutro, neutrophil; NK, natural killer cell; novel, novel cell type; PC, principal cell; Podo, podocyte; T-lymph, T lymphocyte.
previously reported to be connected to Mendelian kidney disease were divided into seven subcategories with a focus on four primary kidney disease categories, namely congenital or developmental, electrolytes or metabolic, glomerular, and tubulointerstitial. In the human single-cell gene expression dataset, five enriched cell clusters were identified for genes implicated in monogenic electrolyte or metabolic disease (Figure 5, Supplemental Table 15), including PT, nephron epithelium, and loop of Henle (three of them with \( P = 1 \times 10^{-5} \)). Of these, PT was confirmed in the mouse dataset (Supplemental Figure 7, Supplemental Table 16). For single-gene causes of glomerular diseases, podocyte was highly enriched (\( P = 1 \times 10^{-5} \) in human and mouse). Fibroblast showed an unexpected enrichment in congenital or developmental diseases in human (\( P = 0.001 \)) but not in mouse (\( P = 0.94 \)). In order to identify cell types related to congenital or developmental disease genes without the presence of nonkidney cell types, we additionally performed EWCE using two kidney organoid datasets from Wu et al.\(^{27}\): 18,072 cells obtained with the Morizane protocol and 26,890 cells with the Takasato protocol. For congenital and developmental disease genes, we detected enrichment in clusters belonging to mesenchymal stem cells in both datasets (Supplemental Figures 8 and 9, Supplemental Tables 17 and 18), which is consistent with existing knowledge about developmental diseases of the kidney. In summary, EWCE analyses correctly identified the effector cell type of genes underlying Mendelian diseases of electrolytes or metabolic as well as glomerular disorders, thereby validating the categorization of disease genes.

The six kidney RNA-seq datasets originated from different species (human, mouse, and rat) and different technologies (scRNA-seq, snRNA-seq, and bulk RNA-seq). In order to investigate to what extent the cell-type assignment was consistent across the four kidney single-cell datasets, we systematically projected cell types from one dataset onto the cell types identified in a different dataset on the basis their gene

---

**Figure 4.** Cell type–specific expression of eGFR and UACR genes in the human single-cell dataset. Heat map showing expression of genes implicated in eGFR (upper) and UACR (lower) across cell types (x axis) in the human single-cell dataset. The expression values are z-score transformed. Genes known to underlie primary monogenic kidney diseases are colored in blue. AVR, ascending vasa recta; DCT, distal convoluted tubule; DVR, descending vasa recta; Fib, fibroblast; GE, glomerular epithelium; IC-A, intercalated cell type A; IC-B, intercalated cell type B; LOH, loop of Henle; MC, mesangial cell; NE, nephron epithelium; Podo, podocyte.
expression values36 (Methods). As shown in the Sankey plots (Supplemental Figure 10), cell-type annotations from the human and mouse single-cell datasets aligned well. For instance, PT cells in mice overwhelmingly matched to the three PT clusters (PT1, PT2, and PT3) in humans. On the other hand, the human scRNA-seq11 dataset and the snRNA-seq dataset by Wu et al.27 did not match very well using the default similarity threshold of 0.7 (Methods), as reflected by the proportion of “unassigned” cells. Lowering the similarity threshold to 0.3 resulted in a much better match between cell types.

**Figure 5.** Gene set enrichment analyses for Mendelian kidney genes in the human single-cell dataset. Gene set enrichment analyses for seven Mendelian kidney disease gene sets in the human single-cell dataset. The y axis indicates $-\log_{10}(P$ values). AVR, ascending vasa recta; DCT, distal convoluted tubule; DVR, descending vasa recta; Fib, fibroblast; GE, glomerular epithelium; IC-A, intercalated cell type A; IC-B, intercalated cell type B; LOH, loop of Henle; MC, mesangial cell; NE, nephron epithelium; Podo, podocyte. *Bonferroni significant; **Bonferroni significant and $P<1E-5$. 
such as for PT, collecting duct, and distal tubule. It is worth noting that in this projection, the PT1 cluster in the scRNA-seq dataset predominately matched the segments 1–3 of the PT in the snRNA-seq dataset by Wu et al.,

## DISCUSSION

GWAS of kidney function traits have identified hundreds of associated loci. However, translation of the associated genetic loci into genes, relevant cell types, and ultimately, mechanistic insights is still a challenging task. By leveraging GWAS summary statistics and gene expression data, we identified kidney function trait–related target tissues and kidney cell types. At the tissue level, enriched heritability was found in eGFR, BUN, and serum urate–associated genetic regions highly expressed in the kidney as well as in liver. The enrichment in both kidney and liver may be driven by shared highly expressed genes in both organs. Shared expression patterns may reflect the fact that kidney and liver are both organs with a high proportion of epithelial cells that may share transcription programs. Indeed, Lindgren et al. reported that HNF1A and HNF4A are critical transcription factors regulating gene expression in both liver and PT cells. More generally, many solute carriers and enzymes exhibit shared expression in both organs. Thus, kidney function traits showed tissue-specific signatures of trait–associated genes, with unbiased transcriptomics data recapitulating existing biologic knowledge about the physiology underlying the traits of interest.

Within the kidney, PT cells were implicated as the relevant cell type for eGFR and serum urate. Moreover, gene set enrichment analyses showed that known single-gene causes of glomerular diseases were highly enriched in podocytes, and electrolyte or metabolic disease genes were enriched in PT cells.

One example for target cell-type prioritization of a GWAS-implicated gene is KNG1. Our analyses show that this gene is highly and specifically expressed in the loop of Henle and was in fact the most specifically expressed gene in this cell type in the human dataset. The index SNP rs11919484 identified in GWAS of eGFR is upstream of KNG1. KNG1 encodes for high-molecular-weight kininogen, which is part of the kallikrein–kinin system. This system is the target of multiple pharmacologic interventions to lower BP, underscoring the importance of the implicated cell types in BP control. Kininogen is cleaved to bradykinin, which in turn, is linked to kidney function and influences BP, natriuresis, and diuresis via the renin-angiotensin-aldosterone system. Another interesting example is TCF21, which is highly specifically expressed in podocytes. The eGFR GWAS index SNP rs3822939 maps close to the gene EYA4, but our analyses prioritize TCF21. Although this gene has not been found to be associated with monogenic kidney disease, it has been shown to be important for kidney development in knockout mice.

Hypomorphic variants could result in reduced nephron number and/or function, providing a potential link to low eGFR.

Heritability enrichments from human and murine scRNA-seq datasets showed very good agreement across species. Furthermore, two additional human snRNA-seq datasets as well as a kidney tubule segments RNA-seq dataset from rat demonstrated consistent results. Moreover, the findings were confirmed by a second independent enrichment method using linear regression models to correlate gene expression cell-type specificity and gene-level kidney function trait associations. Lastly, the Bonferroni correction we applied was over-conservative because many of the cell types within the datasets were related to each other. In total, the tissue- and cell type–specific enrichments that we observed were robust to the choice of species, datasets, and methods, and therefore, they likely represent true positive findings.

Many scRNA-seq datasets across organs from humans and mice have become publicly available over the last few years. Together with genome-wide GWAS summary statistics, they allow for the investigation of cell type–specific expression of implicated genes and eventually, for identifying relevant cell types underlying the studied traits. For kidney traits, Park et al. first reported that CKD-associated genes are enriched in PTs by using a z-score enrichment approach. Qiu et al. generated tubular and glomerular eQTL data to identify novel genes for CKD and found enrichment of these genes for PT expression. In two recent kidney function GWAS, the expression of implicated genes was profiled across kidney cell types. The first study used a murine dataset; the second used a dataset derived from a single human kidney. Differences in prioritized genes in this study and previously published studies may be attributable to the better-powered GWAS summary statistics in this study, as well as from differences in causal gene assignment, species, and technical and analytical variability. Still, the results were largely consistent.
Watanabe et al. collected 43 publicly available scRNA-seq datasets from humans and mice and tested association of cell types with 26 well-powered GWAS using a cell-type specificity regression workflow similar to the one used here. In their study, CKD was included as the only kidney function–related trait, and despite the inclusion of kidney cells, no significant enrichment in any cell type was detected. One possible reason is that the included kidney datasets were from the Mouse Cell Atlas and Tabula Muris Consortium, and kidney cells constituted only a small proportion of both datasets. Also, the included GWAS of CKD was not well powered. Thus, the overall power to detect associations may have been limited.

A very recent GWAS of 38 blood and urine biomarkers in 358,072 participants of the UK Biobank identified 1857 associated loci. Included kidney biomarkers were creatinine, cystatin C, UACR, urate, and eGFR. Heritability enrichment was tested on the basis of marker gene lists in specific cell types of kidney, liver, and pancreas. In line with our results, enrichments were found for kidney traits in podocytes and tubule cells. Our study focuses on kidney function traits using additional methods as well as several scRNA-seq datasets across organisms and from primary tissue, microdissected tissues, and organoids.

For UACR, we were able to prioritize kidney tissue of 52 other GTEx tissues. Within the kidney, podocytes and PT cells were nominally enriched for UACR. Among the individual genes prioritized as most likely to be causal in the original publication, PRKCI was found to be highly and specifically expressed in podocytes. The gene product, a serin/threonine protein kinase, is important in actin cytoskeletal regulation in podocytes. A podocyte-specific deletion in mice resulted in nephrotic syndrome, and knockdown of PRKCI orthologs in Drosophila nephrocytes resulted in disrupted architecture of slit diaphragm orthologs. Furthermore, CUBN was the first gene reported in GWAS of UACR, and it was found to be expressed in PT, where the encoded protein is important for the endocytosis of filtered albumin. Both Cubulin-knockout mice and patients with rare disruptive mutations in the gene exhibit proteinuria.

Our findings on the basis of the integration of unbiased data sources are consistent with current biologic knowledge by highlighting the well known role of the PT in urate transport and of glomerulum and particularly, podocytes in albuminuria. The ability to implicate relevant cell types at a single-gene level is illustrated by the examples above and may help to detect the cell-type origin and to inform mechanistic studies for newly discovered kidney disease risk genes with little pre-existing knowledge.

The strengths of our study are in the unbiased input datasets and their examination from different angles: the well-powered and very recent GWAS summary statistics of different important kidney traits on the one hand and bulk as well as single-cell gene expression data across species and sequencing technologies on the other hand. In addition, we used two different enrichment methods and incorporated negative controls. We found consistent results supported by existing biologic knowledge and provide comprehensive lists as a resource to the community that can aid post-GWAS follow-up experiments. The restriction to GWAS of European ancestry individuals limits the extrapolation of our findings to groups of non-European ancestry. Differences in cell cluster assignment between the scRNA-seq and snRNA-seq datasets might limit the comparability of cell-type enrichment, but the projection of cell-type annotations across datasets showed good agreement. Furthermore, the power to detect enrichment in rare kidney cell types is limited. This will be alleviated by the future availability of large-scale single-cell data from global initiatives, such as the Human Cell Atlas. Although it has been shown that increasing the sample size of sequenced cells increases the significance of the enrichment on average, we did not find evidence suggesting that cluster size acted as a confounding factor. Although PT was the dominant cell type in the mouse data, the human data that we used were from a subcompartment that excluded the mass of proximal tubular cells. We cannot, however, exclude that true enrichment could be missed because of limited power for rare cell types. Lastly, the association of fibroblasts detected with congenital and developmental diseases in one of the human scRNA-seq datasets might be the result of misclassification of these cells. Independent experiments with two kidney organoid datasets lacking fibroblasts showed biologically plausible enrichment in mesenchymal stem cells instead.

The presented heritability enrichment analyses assume a predominant genetic contribution from tissue- or cell type–specific genes. However, it is also conceivable that transcripts not confined to a specific tissue or cell type contribute to the traits of interest. Indeed, colocalization analyses of GWAS and eQTLs support the presence of both signals that colocalize very specifically only with tubular or glomerular eQTLs and signals that show colocalization in many tissues and organs. Lastly, epigenetic modifications are known to contribute to cell type–specific expression. The integrative analyses of epi-genetic data are likely to become informative in the future after well powered and importantly—comparable epigenetic datasets across a wide range of human tissues with matching gene expression data become available.

In summary, the integration of GWAS summary statistics and gene expression data systematically linked genes in kidney function–associated GWAS loci to specific tissues and cell types. These findings will help experimental studies subsequent to GWAS by prioritizing the potentially relevant cell type for each trait-associated gene.

DISCLOSURES

All authors have nothing to disclose.
FUNDING

The work of A. Köttgen was supported by Deutsche Forschungsgemeinschaft grants KO 3598/3-1 and KO 3598/5-1. The work of A. Köttgen and P. Schlosser was supported by Deutsche Forschungsgemeinschaft grant CRC 992 (to A. Köttgen). The work of A. Köttgen and M. Wuttke was supported by Deutsche Forschungsgemeinschaft grants CRC 1140 (to A. Köttgen) and 246781735 (to A. Köttgen). Y. Li was supported by Deutsche Forschungsgemeinschaft grant KO 3598/4-1 (to A. Köttgen).

ACKNOWLEDGMENTS

We thank the CKDGen Consortium for making GWAS results publicly available, and our thanks also go to CKDGen Consortium collaborators, all study investigators, personnel, and patients.

Prof. Anna Köttgen, Dr. Yong Li, and Dr. Matthias Wuttke designed research; Dr. Cristiano Pattaro, Dr. Pascal Schlosser, Dr. Alexander Teumer, and Dr. Adrienne Tin provided essential data; Dr. Stefan Haugi, Dr. Yong Li, and Dr. Matthias Wuttke performed data analysis; Prof. Anna Köttgen, Dr. Yong Li, and Dr. Matthias Wuttke wrote the manuscript; and all authors read and approved the final manuscript.

SUPPLEMENTAL MATERIAL

This article contains the following supplemental material online at http://jasn.asnjournals.org/lookup/suppl/doi:10.1681/ASN.2020010051/-/DCSupplemental.

Supplemental Figure 1. Heritability enrichment results of four kidney traits, asthma, and schizophrenia in the human snRNA-seq dataset Lake.

Supplemental Figure 2. Heritability enrichment results of four kidney traits in the rat tubule dataset Lee and the mouse glomerulus scRNA-seq dataset Karaiskos.

Supplemental Figure 3. MAGMA cell-type association results of four kidney traits in human and mouse scRNA-seq datasets.

Supplemental Figure 4. Cell type-specific expression of eGFR- and UACR-associated genes most likely to be causal in the mouse scRNA-seq dataset.

Supplemental Figure 5. Cell type-specific expression of serum urate-associated genes most likely to be causal in the human scRNA-seq dataset.

Supplemental Figure 6. Cell type-specific expression of genes implicated in serum urate in the mouse scRNA-seq dataset.

Supplemental Figure 7. Gene sets enriched results of Mendelian disease genes in the mouse scRNA-seq dataset.

Supplemental Figure 8. Gene sets enriched results of Mendelian disease genes in the human organoid scRNA-seq dataset Morizane protocol.

Supplemental Figure 9. Gene sets enriched results of Mendelian disease genes in the human organoid scRNA-seq dataset Takakosato protocol.

Supplemental Figure 10. Projection of assigned cell-type clusters across four single-cell and single-nucleus datasets by scmap.

Supplemental Table 1. Sample sizes of GTEx tissue and tissue categories.

Supplemental Table 2. Genetic correlations and heritability of six GWAS traits.

Supplemental Table 3. LDSC-seq results of six traits in GTEx tissues.

Supplemental Table 4. Specifically expressed genes in the human scRNA-seq dataset (Young).

Supplemental Table 5. Specifically expressed genes in the mouse scRNA-seq dataset (Park).

Supplemental Table 6. LDSC-seq results of six traits in the human scRNA-seq dataset (Young).

Supplemental Table 7. LDSC-seq results of six traits in the mouse scRNA-seq dataset (Park).

Supplemental Table 8. LDSC-seq results of six traits in the human snRNA-seq dataset (Lake).

Supplemental Table 9. LDSC-seq results of six traits in the human snRNA-seq dataset (Wu).

Supplemental Table 10. LDSC-seq results of four kidney traits in the micro-dissected rat tubule RNA-seq dataset (Lee).

Supplemental Table 11. LDSC-seq results of four kidney traits in the mouse glomerulus scRNA-seq data (Karaiskos).

Supplemental Table 12. MAGMA association results of kidney function traits in the human scRNA-seq dataset (Young).

Supplemental Table 13. MAGMA association results of kidney function traits in the mouse scRNA-seq dataset (Park).

Supplemental Table 14. Prioritization of genes in GWAS loci using expression specificity results.

Supplemental Table 15. Gene set enrichment results of kidney disease genes in the human scRNA-seq dataset (Young).

Supplemental Table 16. Gene set enrichment results of kidney disease genes in the mouse scRNA-seq dataset (Park).

Supplemental Table 17. Gene set enrichment results of kidney disease genes in the human organoid scRNA-seq dataset Morizane protocol.

Supplemental Table 18. Gene set enrichment results of kidney disease genes in the human organoid scRNA-seq dataset Takakosato protocol.

REFERENCES

25. Caminati M, Pham DL, Bagnasco D, Canonica GW: Type 2 immunity in Th2-driven asthma. World Allergy Organ J 11: 13, 2018
54. Tabula Muris Consortium; Overall coordination; Logistical co-ordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators: Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562: 367–372, 2018