# Supplementary Information for

# Epigenomic and transcriptomic analyses define core cell types, genes and targetable mechanisms for kidney disease

Hongbo Liu[1,2,3], Tomohito Doke[1,2,3], Dong Guo[4], Xin Sheng[1,2,3], Ziyuan Ma[1,2,3], Joseph Park[1,3,5], Ha My T. Vy[6,7], Girish N. Nadkarni[6,7,8,9], Amin Abedini[1,2,3], Zhen Miao[1,2,3], Matthew Palmer[10], Benjamin F. Voight[2,3,11,12], Hongzhe Li[13], Christopher D. Brown[3], Marylyn D. Ritchie[3,5], Yan Shu[4] and Katalin Susztak[1,2,3]*

Correspondence: Katalin Susztak (ksusztak@pennmedicine.upenn.edu)

[1]Department of Medicine, Renal Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Institute of Diabetes Obesity and Metabolism, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland at Baltimore, Baltimore, MD 21201, USA
[5]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[6]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[7]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[8]The Hasso Plattner Institute of Digital Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[9]The Mount Sinai Clinical Intelligence Center, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[10]Pathology and Laboratory Medicine at the Hospital of the University of Pennsylvania, Philadelphia, 19104, USA
[11]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA
[12]Institute of Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA
[13]Department of Biostatistics, Epidemiology, and Informatics, and Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

**The supplementary information includes:**

# Supplementary Note

## Data generation

*Genotype data*. Genomic DNA isolated from kidney samples was used for genotyping. 271 samples were genotyped using Axiom Tx SNP GWAS array, and 239 samples were genotyped using Affymetrix Axiom Biobank array (**Supplementary Table 8**). For each dataset, PLINK (v1.9)[1] was utilized for quality control. First, duplicates and variants with genotyping call rate < 95% were removed. Samples with >5% missing values were excluded. Additional samples were excluded because of ambiguous sex. To identify poor DNA quality or sample contamination, heterozygosity test was performed to exclude samples with high heterozygosity (extreme inbreeding coefficient cutoff was determined by heterozygosity rate ± 3-fold standard deviations from the mean). To further identify potential sample contamination, identity-by-descent (IBD) for each pairwise sample combination was computed and all samples passed PI_HAT < 0.2. In total, 267 individuals remained in the dataset by Axiom Tx SNP GWAS array, and 227 individuals remained in the dataset by Affymetrix Axiom Biobank genotyping array, respectively. Finally, variant-level tests were performed, and the following variants were excluded: monomorphic variants (MAF=0), Hardy-Weinberg equilibrium $p < 1\times10^{-6}$, genotype missingness predicted using surrounding haplotypes ($p < 1\times10^{-8}$), association with chemistry plate batch ($p < 1\times10^{-8}$), and variants on sex chromosomes.

To merge the genotypes from Axiom Tx and Axiom Biobank arrays, we extracted the genotype calls of an overlapping subset of 327,366 variants between two platforms. This enabled imputation of the same set of variants in all samples. To ensure inclusion of only high confidence variants, multiple sample and variant QC steps were performed before imputation. First, we excluded variants whose reference and alternative alleles did not align between two platforms and those with a frequency difference larger than 0.15 between two platforms. Two individuals with call rate <95% were excluded. Then, variant-level tests were performed, and the following variants were excluded: Hardy-Weinberg equilibrium $p < 1\times10^{-6}$, genotype missingness predicted using surrounding haplotypes ($p < 1\times10^{-8}$), association with chemistry plate batch ($p < 1\times10^{-8}$). After quality control, genotypes were phased with SHAPEIT2 (v2.17)[2] and imputed by IMPUTE2 (v2.3.2)[3,4], using the multi-ancestry panel reference from 1,000 Genome Phase 3 (NCBI build 37, released in October 2014). To quantify the population structure, genotype-based principal component analysis (PCA) was conducted using EIGENSTRAT (v7.2.1)[5] on 488 individuals, with additional 2,504 samples from the 1,000 Genomes Project Phase 3 (503 EUR, 661 AFR, 347 AMR, 504 EAS, 489 SAS)[6]. This genotype data were used for meQTL mapping (**Supplementary Fig. 5**), eQTM analysis (**Extended Data Fig. 5**) and heritability analysis (**Extended Data Fig. 6**).

*DNA methylation data.* DNA methylation at over 850,000 methylation sites was measured in 506 kidney samples using Infinium Methylation EPIC BeadChip. SeSAMe (v1.5.3)[7] was used for pre-processing and quality control steps including low intensity-based detection calling achieved by pOOBAH, bleed-through correction in background subtraction, nonlinear dye bias correction, stricter non-detection calling and control for bisulfite conversion based on C/T-extension probes. For each sample, residual incomplete bisulfite conversion was quantified using GCT score based on C/T-extension probes. Leukocyte fraction was estimated by cell composition deconvolution using a two-component model. Beta values were defined as methylated signal/(methylated signal + unmethylated signal). 56,552 probes with missing values in >20% samples were excluded. We further masked 107,847 probes: probes with non-unique 30bp 3'-subsequence, low mapping quality (<40), extension base inconsistent with specified color channel (type I) or CpG (type II) based on mapping, having a SNP in the extension base that causes a color channel switch, non-CpG sites, probes on chromosomes X, Y and M, and probes whose 5bp 3'-subsequence (including extension for type II) overlap with any of the SNPs with global MAF >1%[8]. Finally, 701,519 CpG sites (**Supplementary Table 9**) were used for further analysis for meQTL mapping (**Supplementary Fig. 5**), eQTM analysis (**Extended Data Fig. 5**) and heritability analysis (**Extended Data Fig. 6**).

*Gene expression data.* RNA was isolated using RNeasy mini kit (Qiagen No. 74106) from tubular compartment following manual microdissection. RNA quality was assessed by the Agilent Bioanalyzer 2100, and samples with a minimum 100 ng total RNA and RIN scores above 7 were used. RNA-Seq libraries were generated from total RNA with polyA+ selection of mRNA using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA). After trimming, low-quality bases using Trim-galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), RNA-seq reads were aligned to the human genome (hg19) using STAR (v2.4.1d)[9] based on GENCODE v19 annotations[10]. RSEM (v1.3.1)[11] was used to quantify gene-level read counts which were further normalized across samples using edgeR (v3.32.1)[12]. Gene expression levels were estimated as transcripts per million (TPM), and only genes with at least 0.1 TPM in at least 20% of the samples were used for further analysis. CIBERSORTx[13] was used to estimate cell fractions for each tubular sample, using single cell RNA-seq data a reference expression matrix [14]. The expression data were used for eQTM analysis (**Supplementary Fig. 11**), heritability analysis (**Extended Data Fig. 6**) and gene expression analysis (**Supplementary Fig. 10,11,14**).

*Human kidney single nucleus ATAC seq (snATAC-seq)*. Six fresh human kidneys were collected after surgical nephrectomies as described above in the “Sample procurement” section (**Supplementary Table 15**). Single nucleus ATAC-seq libraries were generated using the Chromium Single Cell ATAC Library & Gel Bead Kit according to manufacturer’s manual. After quality control, the library was sequenced on an Illumina HiSeq 2x50 paired-end kits, resulting a dataset contained 61,440 high quality cells. Reads were aligned to human genome (hg19) with SnapATAC (v2.0)[15]. After quality control and peak calling, a cell-gene activity score matrix was built by integrating all fragments overlapping with gene transcripts (in GENCODE v19 annotations)[10]. Cluster annotation was performed using a published list of cell-type marker genes[16], and 13 main clusters were identified (Proximal tubule segment 1, Proximal tubule segment 2, Proximal tubule segment 3, Loop of Henle, Distal convoluted tubule, Collecting duct principal cell, Collecting duct intercalated cell, Podocyte, Endothelial, Stroma, Immune, Lymph cell and injured proximal tubule). The injured proximal tubule cells were not included in further analysis due to its potential disease status. For each of remained 12 clusters (57,262 cells), we identified cell type-specific differentially accessible regions (DARs) by one-sided Fisher’s exact test between a given cell type and each of the other cell types for each of 410,994 peaks[17]. Peaks with FDR < 0.05 and fold change > 1 in at least half of pairwise comparisons were defined as cell type specific DARs, thus allowing inclusion of DARs shared by closely related cell types (the three segments of proximal tubules). For species conservation analysis, mouse kidney snATAC-seq data were obtained from GEO with accession number GSE157079[18]. Mouse kidney scRNA-seq data were obtained from GEO with accession number GSE107585[16] and cell type-specific expressed genes were identified as genes with cell type expression specificity weight > 0 quantified by CELLEX (v1.2.1)[19].

### GWAS independent loci comparison with previous studies

In particular, we compared independent loci identified in this study with 424 eGFRcrea GWAS loci defined by Stanzick et al. using a window-based method based on 1,201,909 cross-ancestry individuals[20]. Given the differences in locus definition, we applied both our clumping-based method and their window-based method to both datasets and then compared significant loci (**Supplementary Fig. 1**). Further, we identified novel independent signals by comparing with 634 independent signals defined by Stanzick et al. using approximate conditional analyses in 1,004,040 European individuals[20] (**Supplementary Fig. 2**). To explore the contribution of common variants and rare variants to kidney function, we compared the independent loci with creatinine-associated exome rare variants identified based on exome-sequencing data (n = 454,787 UK Biobank study participants)[21] or whole-exome imputed SNP-arrays (n = 487,409 UK Biobank study participants)[22] (**Supplementary Fig. 3**).

### Cis-eQTL meta-analysis

To obtain a comprehensive cis-eQTL map, we performed a meta-analysis based on the eQTL summary statistics obtained from four non-overlapping studies; eQTLs by Sheng et al. using imputed genotypes in 356 tubule samples[23], eQTLs by Ko et al. using imputed genotypes in 91 kidney cortex samples from The Cancer Genome Atlas (TCGA)[24], eQTLs from the Genotype-Tissue Expression (GTEx v8) study using genotypes by whole genome sequencing in 73 kidney cortex samples[25], and eQTLs from the Nephrotic Syndrome Study Network (NephQTL) using genotypes by whole genome sequencing in 166 tubulointerstitial samples[26] (see details in **Supplementary Table 3**). Four eQTL datasets were pooled by fixed effects inverse-variance meta-analysis with METAL[27], with genomic control correction for each input study (genomic control parameter 1.132 for Sheng et al.'s eQTLs, 1.110 for Ko et al.'s eQTLs, 1.050 for GTEx eQTLs and 1.064 for NephQTL eQTLs, respectively) and assessment of between-study heterogeneity with the Cochran's Q-test and $I^2$ statistic. After meta-analysis of 281,045,539 associations among 686 individual meta-analysis (72% are of European ancestry), 201,627,059 associations with a MAF > 0.05 in European population based on 1,000 Genomes phase 3 European samples (n = 503) were retained. Association summary statistics were canonicalized to make sure effect size was always reported with respect to the alternate allele as above. To define eGenes, we used the Storey approach to calculate q values[28] for all associations and q value (< 0.01) was used to identify significant eGenes. The associations from a single study and multiple studies with between-study heterogeneity (Cochran's Q-test HetISq > 50 or $I^2$ statistic HetPVal < 0.05) were selected as significant eQTLs only when they passed significance level ($q < 0.001$) in the meta-analysis and have been identified as significant eQTLs in at least one original study. In total, we identified 10,430 eGenes and 1,222,250 significant SNP-gene pairs (**Supplementary Table 4**). Novel eGenes were determined if they were not included in any of eGene lists in six reference studies[23-26,29,30].

To further define kidney-specific eGenes, meta-analysis of multiple-tissue eQTL was performed on 917,902 SNP-gene pairs which were significant in kidney and available in more than 80% of eQTL datasets mapped in 48 GTEx (v8) non-kidney tissues[25]. For each SNP-gene pair, the posterior probability that an eQTL effect exists in a given tissue (called *m* value) was calculated using a random effects model in METASOFT (v2.0.1)[31], and high-confidence eQTL was discovered by a significance cutoff of $m > 0.9$. Kidney-specific eQTLs were defined as having $m > 0.9$ in fewer than five tissues including kidney. Further, we performed enrichment analysis of kidney-specific eQTL SNPs on GWAS hits of 35 blood and urine biomarkers (including eGFRcrea) in the UK Biobank (n = 363,228 individuals)[32]. For each GWAS trait, significant variants were determined by genome-wide cutoff $p < 5\times10^{-8}$, and number of variants overlapped with kidney-specific eQTL variants and non-kidney-specific eQTL variants were counted, respectively. $\chi^2$ test was performed to calculate enrichment significance.

**Data processing for Cis-meQTL mapping**

*Sample filtering for meQTL analysis*. From 506 samples with DNA methylation, 488 samples had high quality genotype data (**Supplementary Table 8**). To exclude outliers in the methylation data, we performed Mahalanobis distance measurement using R package ClassDiscovery (v3.3.13)[33]. Briefly, methylation based PCA analysis were conducted by SamplePCA function, and then Mahalanobis distance of each sample from the center of the two-dimensional principal component space and associated chi-squared p-value were computed by function mahalanobisQC. 45 samples with chi-squared test $p < 0.05$ were identified as outliers and excluded (**Supplementary Table 9**). In total, 443 kidney samples (78.6% are of European ancestry) with both DNA methylation and genotype data remained for further analysis.

*Variant filtering for meQTL analysis*. We extracted imputed genotypes for 443 samples and performed quality control to exclude variants, including imputation confidence score INFO < 0.4 (estimated by SNPTEST v2.3.0[34]), MAF < 5%, Hardy-Weinberg equilibrium $p < 1\times10^{-6}$, missing rate < 95% for best-estimated genotypes at posterior probability > 0.9, indels with a length > 51 bp, and duplicate variants by position (**Supplementary Table 8**). Finally, 5,743,754 variants with imputed genotypes remained for the analysis. In addition, genotype based PCA analysis was conducted again using the final set of kidney samples, and then the first five PCs were used as covariates. SNP matrix with dosages for alternative allele counts was generated as an input in cis-meQTL mapping.

*PEER factor estimation for meQTL analysis*. PEER factors were estimated using PEER (v1.3)[35] based on DNA methylation with general covariates. The associations between 80 PEER factors and known clinical and demographics variables are shown in **Supplementary Fig. 5a**. We optimized the number of used PEER factors to identify the most SNP-CpG pairs on chromosome 1 (**Supplementary Fig. 5b**).

### Kidney-specific meQTLs

To identify kidney-specific meQTLs, we obtained meQTL summary results from whole blood (n = 473)[36] and skeletal muscle samples (n = 265)[37]. METASOFT (v2.0.1)[31] was applied to 5,613,318 SNP-CpG pairs that were available in all three datasets and significant in at least one dataset based on the nominal *p* thresholds above. For each SNP-CpG pair, the posterior probability that an meQTL effect exists in each study (m-value) was calculated. After excluding meQTLs in the MHC region, high-confidence tissue-specific meQTLs were identified for each study using a cutoff m-value > 0.9. To explore the function of tissue-specific mCpGs, we performed enrichment analysis for enhancers (estimated using histone modifications in human kidney, blood CD3+ cells and skeletal muscle), known transcription factor motifs and kidney cell type-specific open chromatin regions.  Further, we performed enrichment analysis of kidney-specific meQTL SNPs on GWAS hits of 35 blood and urine biomarkers (including eGFRcrea) in the UK Biobank (n = 363,228 individuals)[32]. For each GWAS trait, significant variants were determined by genome-wide cutoff

$p < 5\times10^{-8}$, and the number of variants overlapping with kidney-specific meQTL variants and non-kidney-specific meQTL variants were counted. $\chi^2$ test was performed to calculate enrichment significance.

### Cis-eQTM associations mapping and analysis

To identify associations between methylation of CpG sites and expression of genes within a ±1Mb window of the queried gene TSS, expression quantitative trait methylation (eQTM) analysis was performed using a linear regression model implemented in the MatrixeQTL R package[38]. We analyzed 414 human kidney samples used in the meQTL analysis and with available gene expression data by RNA-seq. We considered the linear model $y = \beta_0 + \beta_1 M + T\alpha + \varepsilon$ , where $y$ is the inverse normalized gene expression TPM values, $M$ the inverse normalized CpG methylation values, and $T$ the covariates. We examined and compared the following three models containing: no covariates, general covariates and PEER factors. General covariates included sample collection site, age, sex, top five genetic PCs, incomplete bisulfite conversion, sample plate, sentrix position, RNA integrity number, RNA-seq batch, and RNA-seq read types (paired-end or single-read sequencing). PEER factors were estimated using PEER R package[35] with general covariates based on CpG methylation and gene expression, respectively. The associations between PEER factors and known clinical variables were examined by Spearman's rank correlation. For each PEER factor-based model, we included general covariates and equal numbers (1-10, 15, 20, 25 and 30) of CpG methylation PEER factors and gene expression PEER factors[37]. For each model, significant CpG-gene associations were defined based on a global FDR < 0.05 to correct for multiple testing. The final eQTM model used general covariates and five PEER factors since the eQTM discovery rate changed little after correcting for more PEER factors (**Extended Data Fig.**

**5b**). To examine the robustness of the kidney eQTM association, we calculated the effect size correlation using publicly available eQTM associations identified in skeletal muscle[37], placenta[39], and primary monocytes[40].

### Functional annotation

Adult human kidney histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27ac, H3K36me3) by ChIP-seq were downloaded from GEO (GSM670025, GSM621648, GSM621651, GSM772811, GSM1112806, GSM621634 and GSM621638). Chromatin states for human adult kidneys were generated using ChromHMM (v1.17)[41], by training a 15-state model to capture all the key interactions between the chromatin marks. We also downloaded chromatin states for 127 tissues or cell types from the Roadmap epigenomics project[42]. Transcription factor enrichment of meQTL CpGs was performed using HOMER (v4.10.3)[43]. Functional enrichment analysis was performed using DAVID Bioinformatics Resources (v6.8)[44] for genes and Genomic Regions Enrichment of Annotations Tool (GREAT v4.0.4)[45] for genome regions. Protein-protein association network was generated using STRING database (v11.0)[46]. Bedtools (v2.29.2)[47] was used to process overlapping regions and deeptools (v3.5.0)[48] for quantification and profile plot of histone modifications. Drug-Gene interactions were identified using the Drug Gene Interaction Database (DGIdb v4.2.0)[49].

### GWAS summary statistics data for GWAS heritability analysis

Summary statistic data of kidney function related GWAS traits, including eGFR based on serum creatinine (eGFRcrea), eGFR based on cystatin C (eGFRcys), and blood urea nitrogen (BUN), were collected from CKDGen Consortium (https://ckdgen.imbi.uni-freiburg.de/)[50,51], the VA

Million Veteran Program (MVP)[52] and Pan-UK Biobank (https://pan.ukbb.broadinstitute.org/). Summary statistics were converted to the sumstats format using the munge_sumstats.py program included with LDSC (v1.0.1)[53]. We also obtained independent non-kidney function related traits based on UKBB GWAS summary statistic data in sumstats format from the Alkes Price lab (https://alkesgroup.broadinstitute.org/LDSCORE/independent_sumstats/)[54]. Totally, 34 GWAS traits (including six kidney function traits) with sample size larger than 200,000 were considered for further analysis (**Supplementary Table 14**).

### Validation of GWAS heritability mediated by DNA methylation and gene expression

To validate the findings based on multi-ancestry datasets, we obtained individual-level genotype, methylation, and expression data for 323 human kidney samples of European-ancestry from 414 multi-ancestry samples used above. MESC was applied to these data to estimate methylation-mediated heritability and expression-mediated heritability for each of three kidney function traits of the European-ancestry GWAS, including eGFRcrea (N=401,867 European individuals from UKBB), eGFRcys (N=402,043 European individuals from UKBB) and BUN (N=243,029 European individuals from CKDGen)[50]. LD matrix was estimated using 503 European ancestry samples from the 1000 genome Phase 3 as matching reference LD panel from the UKBB was not publicly available.

To test the effect of sample sizes, samples were randomly selected from the 414 kidneys with individual-level genotypes, methylation, and expression data to estimate methylation- and expression-mediated heritability for eGFRcrea GWAS trait (N=421,531 individuals across multiple ancestries from UKBB). We also performed down sampling analysis based on randomly selected European 323 kidney samples to estimate methylation-mediated heritability and expression-mediated heritability for eGFRcrea GWAS trait (N=401,867 individuals of European-ancestry from UKBB).

**GWAS heritability enrichment analysis**

To prioritize kidney disease relevant CpG sets, we performed methylation-mediated heritability $h^2{}_{med\sim m}$ enrichment estimates for regulatory elements (determined by chromatin states in human kidney) across 34 GWAS traits. In brief, CpG sites used for kidney meQTL mapping was categorized by adult human kidney chromatin states. For each categorized CpG set, methylation scores were estimated based on individual-level genotypes and methylation data obtained from 414 kidney samples, and then used to further estimate $h^2{}_{med\sim m}$ for the corresponding CpG set based on GWAS summary statistics. Methylation-mediated heritability $h^2{}_{med\sim m}$ enrichment was defined as the proportion of $h^2{}_{med\sim m}$ in a given CpG set divided by the proportion of CpGs in corresponding CpG set. $P$ values for $h^2{}_{med\sim m}$ enrichment was calculated by a two-tailed z test using jackknife standard errors for $h^2{}_{med\sim m}$ enrichment. To adjust $p$ values for multiple testing (374 CpG set-GWAS trait pairs =11 chromatin state CpG sets × 34 GWAS traits), q values were calculated using Storey approach[28].

To understand whether $h^2{}_{med\sim m}$ enrichment is restricted to kidney enhancers, we combined adult kidney enhancers and enhancers from 127 additional samples from the Roadmap epigenomics project[42]. For each of 128 tissue/cell type, enhancer categorized CpG set was used to estimate kidney methylation-mediated heritability $h^2{}_{med\sim m}$ enrichment and $p$ value for each of the 34

GWAS traits using individual-level genotypes and methylation data from 414 kidney samples. $q$ values were calculated using Storey approach from $p$ values for 4,352 CpG set-GWAS trait pairs (128 enhancer CpG sets × 34 GWAS traits). To explore the enrichment of heritability mediated by tissue-specific methylation, similar $h^2{}_{med\sim m}$ enrichment analysis was also applied to individual-level genotypes and methylation data from 473 blood samples for each of the 34 GWAS traits.

To explore the cell type-specificity of $h^2{}_{med\sim m}$ enrichment, cell type-specific differentially accessible regions identified from human kidney snATAC-seq data were used to annotate CpG sites. For each cell type CpG set, methylation-mediated heritability $h^2{}_{med\sim m}$ enrichment and $p$ value for each of the 34 GWAS traits were estimated using individual-level genotypes and methylation data from 414 kidney samples. $q$ values were calculated using Storey approach from $p$ values for 408 CpG set-GWAS trait pairs (12 cell type CpG sets × 34 GWAS traits).

To further explore kidney disease causing cells, we performed single cell GWAS trait enrichment using gchromVAR (v0.3.2)[55]. To this end, statistically fine-mapped regions for 94 complex traits (sample size up to 361,194 UK Biobank individuals) were downloaded from https://www.finucanelab.org/data[56]. For each trait, the causal SNPs included in the 95% credible sets by SusieR (https://stephenslab.github.io/susieR) and identified as kidney meQTLs were selected as the causal SNPs driving both CpG methylation and phenotype variations. To reduce bias, only 63 traits with more than 2,000 causal SNPs were included into the analysis. The bias-corrected enrichment statistic for 63 traits and a set of 57,262 snATAC-seq cells with 410,994 peaks was calculated by gchromVAR with input of per-variant posterior-probabilities and the peak by cell count matrix of open chromatin. Briefly, the expected number of fragments per peak per cell is computed as the proportion of all fragments across all samples mapping to the specific peaks multiplied by the total number of fragments in peaks for that cell. Similarly, the expected number of fragments weighted by the fine-mapped variant posterior probabilities was calculated for per trait per cell. Then the raw weighted accessibility deviation was calculated for each cell and trait and further 50 sets of background-weighted accessibility deviations matrix to correct for technical confounders (differential PCR amplification or variable Tn5 tagmentation conditions). For each cell and trait, the bias-corrected z score was calculated, and natural logarithm of z score (ln) was used to represent the enrichment statistic. For each trait, significance of enrichment z score (ln) variance among all cell types was determined by the Kruskal-Wallis test, and mean z score in each cell type was calculated for heatmap visualization.

### Bayesian colocalization analysis

We performed Bayesian colocalization analysis to identify the variants where the genotype effect on kidney function, methylation and gene expression were shared. In brief, variants in the MHC region were excluded first. Significant eGFRcrea GWAS variants identified above were defined as leading variants. To estimate posterior probability that a leading variant is associated with two traits (GWAS and meQTL, GWAS and eQTL, meQTL and eQTL), we extracted available variants within 100kb search window for each leading variant. In particular, the search window was narrowed (100kb / number of independent signals) for 88 GWAS loci with multiple independent signals (fine-mapped in 1 million European ancestry individuals[20]), to avoid violation to the assumption of that there is one causal variant per signal. Bayesian colocalization analysis was implemented using R package coloc (v5.1.0)[57] with default parameters ($p_1$=1×$10^{-4}$, $p_2$=1×$10^{-4}$ and $p_{12}$ =1×$10^{-5}$) and input of summary statistics for eGFRcrea GWAS (p value, MAF, sample size), meQTL (effect size and squared standard error) and eQTL (effect size and squared standard error). In the coloc results, H4 represents the posterior probability that both traits are associated and shared the same causal variants. H4 > 0.8 was used as the threshold to determine colocalization. To further refine the variants associated with all three traits (GWAS and meQTL and eQTL), we performed Bayesian multiple-trait-colocalization (moloc) analysis using R package moloc (v0.1.0)[58] with default parameters prior_var = c(0.01, 0.1, 0.5) and priors = c($1\times10^{-4}$, $1\times10^{-6}$, $1\times10^{-7}$). In moloc results, PPA.abc represents the posterior probability that three traits are associated with each other and share the same variant. PPA.abc > 0.8 was considered evidence of colocalization among all three traits.

### Summary-data-based Mendelian Randomization

We performed summary-data-based mendelian randomization (SMR) analysis in three configurations, eGFRcrea GWAS and kidney meQTL, eGFRcrea GWAS and kidney eQTL, kidney meQTL and kidney eQTL, using package SMR (v1.03)[59,60], and used heterogeneity in dependent instruments (HEIDI) to distinguish pleiotropy from linkage. To prepare the input data, GWAS effect sizes and standard errors were estimated from z statistics of the meta-analysis following a method proposed by Zhu et al.[59], and meQTL and eQTL summary data in binary format (BESD) was converted from original summary statistics following SMR data management[59].

First, we applied SMR&HEIDI approach to summary statistics data of the eGFRcrea GWAS and meQTL using "--extract-target-snp-probe" to specify SNP-CpG colocalization pairs (H4>0.8) identified above. To address issues around multiple testing, Bonferroni threshold ($1.52\times10^{-5}$, i.e.

0.05/3,286) was defined based on the number (3,286) of tested CpGs, and used to identify the CpGs whose methylation levels are associated with eGFRcrea GWAS trait. To distinguish pleiotropy from linkage, we used a p value threshold of 0.01 for the HEIDI test, without correcting for multiple tests[60]. Similarly, we applied SMR&HEIDI approach to test the colocalizations (H4>0.8) between GWAS and eQTL using "--extract-target-snp-probe" to specify SNP-gene colocalization pairs. Bonferroni threshold ($1.52\times10^{-4}$, i.e. 0.05/330) was determined based on the number (330) of tested genes, and used to identify the genes whose expression levels are associated with eGFRcrea GWAS trait. The SMR&HEIDI approach was also applied to test the colocalizations (H4>0.8) between meQTL and eQTL, using meQTL summary data as the exposure and eQTL summary data as the outcome. Bonferroni threshold ($9.98\times10^{-6}$, i.e. 0.05/5,008) was determined based on the number (5,008) of CpG~gene pairs was used to identify the significant CpG~gene pairs in which CpG methylation levels are associated with gene expression levels.

### Validation of phenome-wide association study of SLC47A1

As an independent validation, we performed PheWAS analysis based on whole exome sequencing dataset of 24,016 individuals in the BioMe Biobank (https://icahn.mssm.edu/research/ipm/programs/biome-biobank), with a loss-of-function variant annotation using Loss-Of-Function Transcript Effect Estimator (LOFTEE). Phenotypes for each individual were determined by mapping ICD-10 codes to Phecodes[61], and then phenotypic cases and controls were defined for each disease phenotype using the same method for UKBB dataset described above. 262 phenotypes with at least 300 cases, including renal dialysis (514 cases and 23,502), were included for the PheWAS analysis. Association analysis between each disease phenotype and gene burden of *SLC47A1* was implemented using SAIGE (version 0.35)[62] with covariates including sex, age, and the first 10 principal components of genetic ancestry. Finally, we also conducted PheWAS analysis for a single variant (rs111653425), which is missense variant in *SLC47A1*, in the UKBB and BioMe datasets.

**Mouse studies**

*Slc47a1* knock out mice was generated by the Yan Shu lab at the University of Maryland Baltimore[63]. All mice used in this study were housed under controlled conditions (12 h/12 h dark/light cycle, 21 ± 2 °C, humidity 60 ± 10%) and had free access to food and water. Male 8- to 10-week-old mice and littermates were used for the experiments. Freshly prepared cisplatin (Cayman Chemical; 15663-27-1) protected from light was dissolved in PBS at 1mg/kg. Mice were injected with 7mg/kg (ip) cisplatin weekly for 4 weeks. Kidneys were harvested and preserved at -80 degrees for RNA and protein and in 10% formalin for histology. All experimental protocols were approved by the Institutional Animal Care and Use Committee (IACUC) of the School of Pharmacy, University of Maryland Baltimore. All procedures were carried out in accordance with NIH guidelines for animal experimentation.

**Real time Quantitative PCR**

Mouse kidneys were homogenized, and total RNA was extracted using the Trizol method, according to manufacturer's protocol (Thermo fisher). Quality and concentration of extracted RNA was examined using nanodrop. cDNA was generated using Reverse Transcription Kit (Applied Biosystems) according to the manufacturer's protocol. Real time quantitative PCR was performed using Cyber Green Master Mix Reagents (Thermo Fisher) with ViiA 7 System (Life Technologies) instrument. Primer sequences are listed in **Supplementary Table 28**.

## Histological analysis

Kidneys were fixed in 10% neutral formalin and samples were embedded in paraffin. Kidney sections were stained with H&E and Sirius Red according to manufacturer's protocol (Polysciences, Inc 24901). Tubule injury (hydropic degeneration, hyaline casts, cytoplasmic vacuolization, loss of the brush border, tubular lumen dilation, and necrosis of tubular cells) were scored semi-quantitatively in H&E-stained images. We used the following scoring system, Score 0: no tubular injury; Score 1: <10% of tubules injured; Score 2: 10–25% of tubules injured; Score 3: 25–50% of tubules injured; Score 4: 50–74% of tubules injured; Score 5: >75% of tubules injured. Renal fibrosis was evaluated by Sirius red. Five images were randomly taken from each mouse kidney and quantified by image J software (v1.53)[64].

## Immunoblot

Whole kidney lysates were made in SDS buffer (187.5 mM Tris-HCl pH 6.8, 6% SDS, 30% glycerol and 0.03%bromophenol blue adding DTT (Dithiothreitol). Samples were sonicated using Bioruptor UCD-300 for 1 minute with high power and boiled at 95 degrees for 5 minutes. Proteins were separated by SDS-PAGE transferred onto 0.2 µm pore size PVDF membrane. After blocking with 5% non-fat dry milk, membranes were incubated with anti-RIPK3 (Millipore sigma; PRS2283;1:1000), anti-NLRP3 (cell signaling; 15101; 1:1000), anti-aSMA (Sigma; A2547; 1:1000) anti-GAPDH antibody (CST, 1:1000) at 4 degrees overnight. Membranes were incubated with appropriate secondary antibodies conjugated with HRP, and signals were detected using ECL Western Blotting Substrate (Thermo Fisher). For the quantification, Image J software was used[64].

## Reference for Supplementary Note

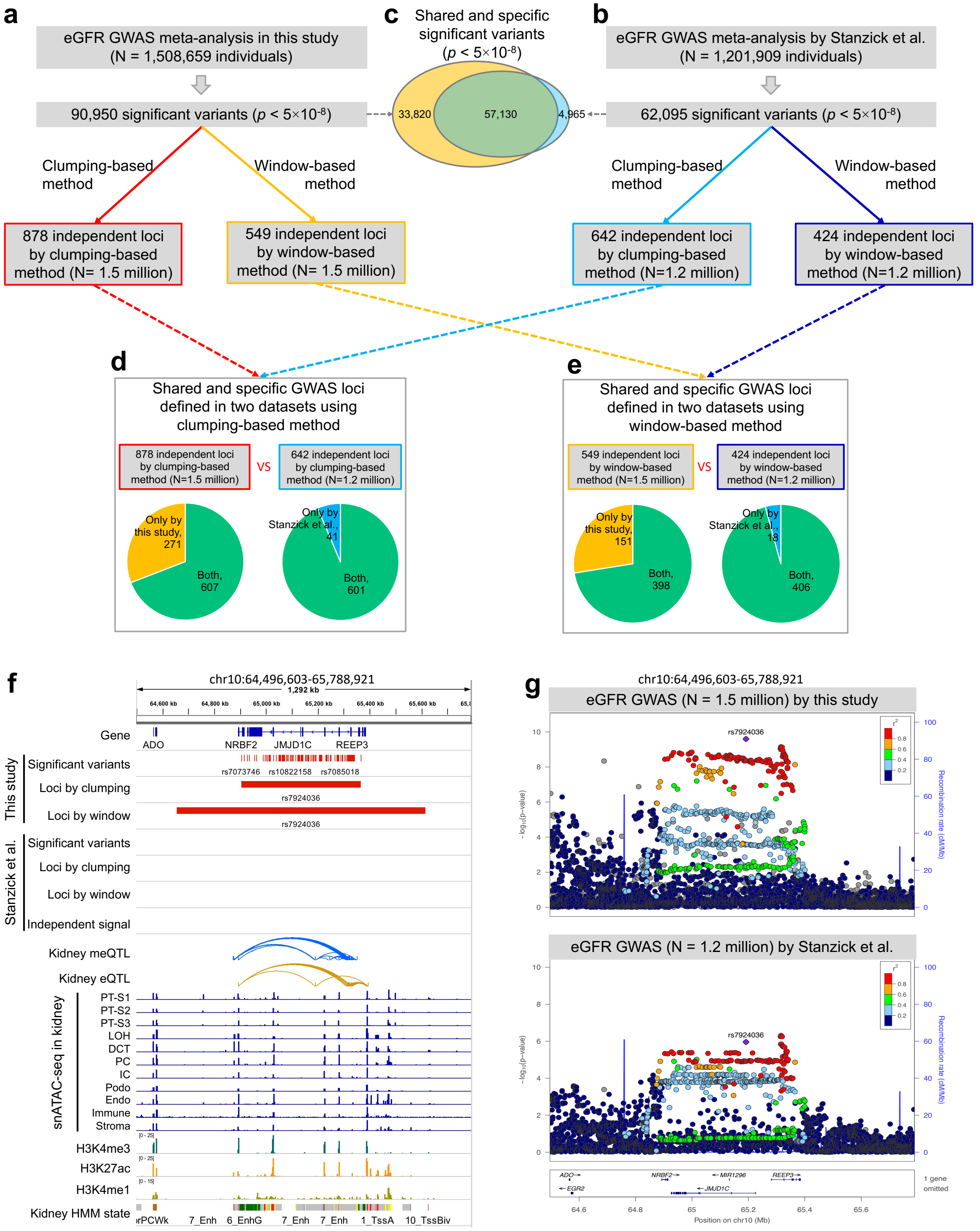
1. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
2. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5 (2013).
3. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529 (2009).
4. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-70 (2011).
5. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904 (2006).
6. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
7. Zhou, W., Triche, T.J., Jr., Laird, P.W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* **46**, e123 (2018).
8. Zhou, W., Laird, P.W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45**, e22 (2017).
9. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
10. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
11. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
12. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
13. Newman, A.M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology* **37**, 773-782 (2019).
14. Dhillon, P. *et al.* The Nuclear Receptor ESRRA Protects from Kidney Disease by Coupling Metabolism and Differentiation. *Cell Metab* **33**, 379-394 e8 (2021).
15. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* **12**, 1337 (2021).
16. Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758-763 (2018).
17. Pijuan-Sala, B. *et al.* Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat Cell Biol* **22**, 487-497 (2020).
18. Miao, Z. *et al.* Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nat Commun* **12**, 2277 (2021).
19. Timshel, P.N., Thompson, J.J. & Pers, T.H. Genetic mapping of etiologic brain cell types for obesity. *Elife* **9**(2020).
20. Stanzick, K.J. *et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat Commun* **12**, 4350 (2021).

21. Backman, J.D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628-634 (2021).
22. Barton, A.R., Sherman, M.A., Mukamel, R.E. & Loh, P.R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* **53**, 1260-1269 (2021).
23. Sheng, X. *et al.* Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nat Genet* **53**, 1322-1333 (2021).
24. Ko, Y.A. *et al.* Genetic-Variation-Driven Gene-Expression Changes Highlight Genes with Important Functions for Kidney Disease. *Am J Hum Genet* **100**, 940-953 (2017).
25. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
26. Gillies, C.E. *et al.* An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. *Am J Hum Genet* **103**, 232-244 (2018).
27. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
28. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
29. Qiu, C. *et al.* Renal compartment-specific genetic variation analyses identify new pathways in chronic kidney disease. *Nat Med* **24**, 1721-1731 (2018).
30. Eales, J.M. *et al.* Uncovering genetic mechanisms of hypertension through multi-omic analysis of the kidney. *Nat Genet* **53**, 630-637 (2021).
31. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet* **8**, e1002555 (2012).
32. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* (2021).
33. Coombes, K.R. *et al.* Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* **49**, 1615-23 (2003).
34. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**, 906 (2007).
35. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770 (2010).
36. Sheng, X. *et al.* Systematic integrated analysis of genetic and epigenetic variation in diabetic kidney disease. *Proc Natl Acad Sci U S A* **117**, 29013-29024 (2020).
37. Taylor, D.L. *et al.* Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc Natl Acad Sci U S A* **116**, 10883-10888 (2019).
38. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
39. Delahaye, F. *et al.* Genetic variants influence on the placenta regulatory landscape. *PLoS Genet* **14**, e1007785 (2018).

40. Husquin, L.T. *et al.* Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biol* **19**, 222 (2018).
41. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
42. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
43. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
44. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
45. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
46. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).
47. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
48. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-5 (2016).
49. Freshour, S.L. *et al.* Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res* **49**, D1144-D1151 (2021).
50. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
51. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat Genet* **51**, 1459-1474 (2019).
52. Hellwege, J.N. *et al.* Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nat Commun* **10**, 3842 (2019).
53. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
54. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed-model association for biobank-scale datasets. *Nat Genet* **50**, 906-908 (2018).
55. Ulirsch, J.C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet* **51**, 683-693 (2019).
56. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* (2021).
57. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
58. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538-2545 (2018).
59. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
60. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).

61. Carroll, R.J., Bastarache, L. & Denny, J.C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-6 (2014).
62. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018).
63. Li, Q., Peng, X., Yang, H., Wang, H. & Shu, Y. Deficiency of multidrug and toxin extrusion 1 enhances renal accumulation of paraquat and deteriorates kidney injury in mice. *Mol Pharm* **8**, 2476-83 (2011).
64. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-82 (2012).

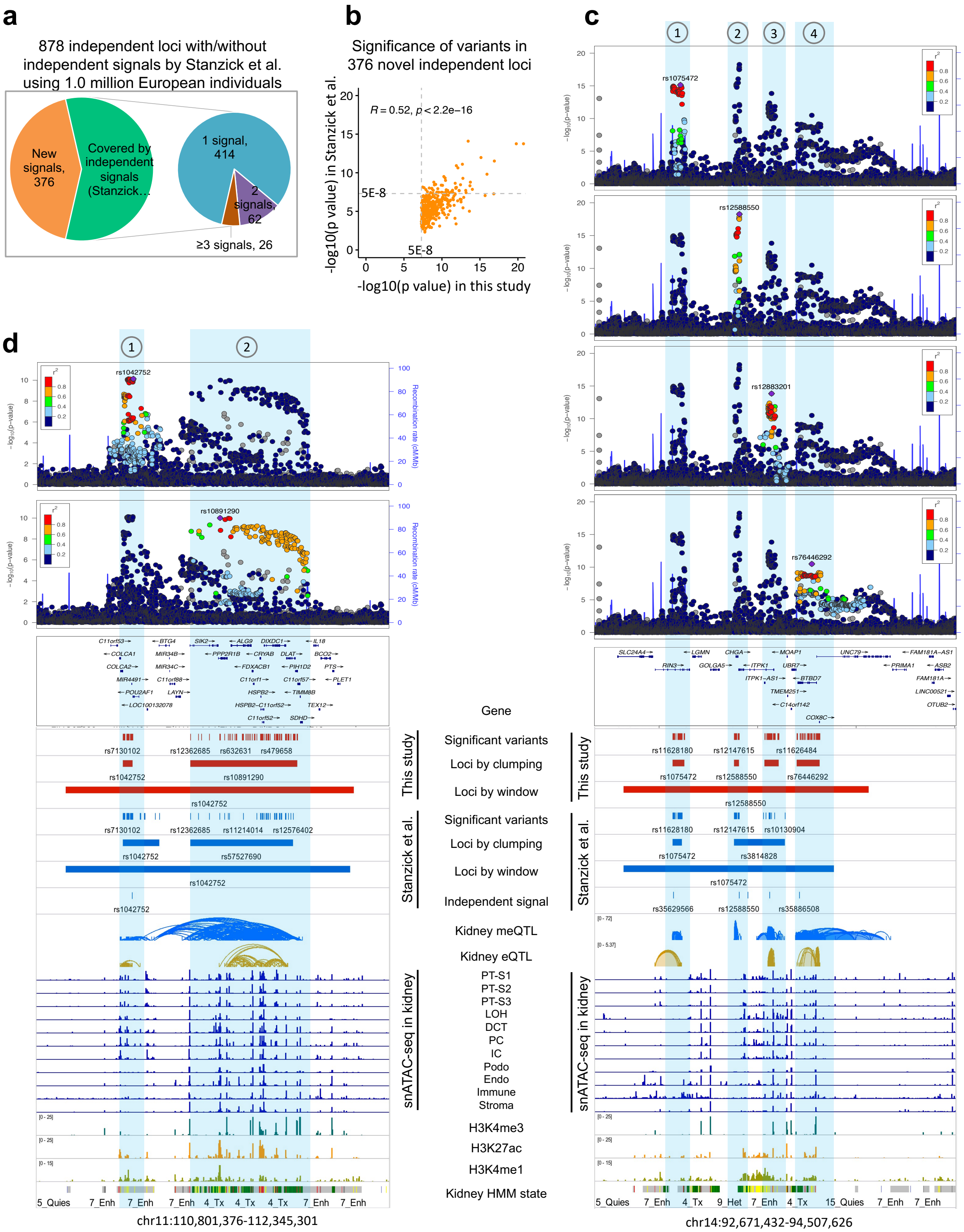
# Supplementary Figures



**Supplementary Fig. 1. Comparison of independent loci defined by meta-analysis eGFRcrea GWAS of this study and prior CKDGen study.**

**a.** Identification of independent GWAS loci using clumping-based and window-based methods, respectively, based on summary statistics of eGFRcrea meta-analysis GWAS (N = 1,508,659 individuals of trans-ancestry) mapped in this study. Window-based method was used by Stanzick et al.

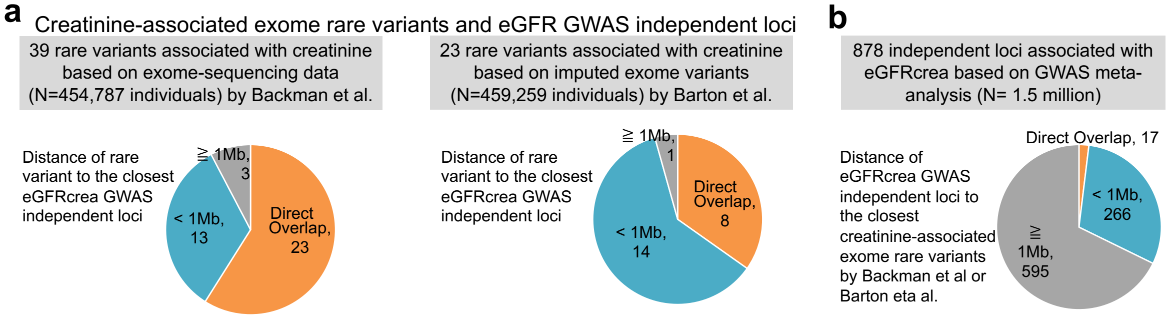
**b.** Identification of independent loci using clumping-based and window-based methods, respectively, based on summary statistics of eGFRcrea meta-analysis GWAS (N = 1,201,909 individuals of trans-ancestry) obtained from Stanzick et al.
**c.** Venn plot for the number of significant variants (GWAS two-sided $p < 5\times10^{-8}$) associated with eGFRcrea identified by this study and Stanzick et al.
**d.** Shared and specific GWAS loci defined in two datasets using clumping-based method. In each dataset, the number of independent loci overlapping and non-overlapping from the other dataset was counted.
**e.** Shared and specific GWAS loci defined in two datasets using window-based method. In each dataset, the number of independent loci overlapping and non-overlapping the other dataset was counted.
**f.** An example of an independent locus (chr10:64,496,603-65,788,921) only identified in the current GWAS. The top track shows nearby genes including *NRBF2* which was prioritized as target gene for this locus by priority score of seven in this study. The GWAS tracks shows the significant variants ($p < 5\times10^{-8}$) at this locus defined by clumping-based and window-based methods, in this study and Stanzick et al., followed by tracks showing multiple kidney omics including meQTLs, eQTLs, cell type-specific chromatin accessibility in human kidneys by single-nucleus ATAC-seq (snATAC-seq), human kidney histone modifications (H3K4me3, H3k27ac, H3K4me1) by ChIP-seq and chromatin states based on histone modifications. PT-S1-3; proximal tubule S1-3 segment, LOH, loop of Henle, DCT, distal convoluted tubule, PC, principal cell of collecting duct, IC, intercalated cell of collecting duct, endo; endothelial cells, immune: immune cell and stroma.
**g.** LocusZoom view of an independent locus (chr10:64,496,603-65,788,921) using summary statistics in this study or Stanzick et al. Y-axis is strength of association -log10(two-sided *p* value from GWAS studies).



**Supplementary Fig. 2. Comparison of 878 independent loci defined by meta-analysis eGFRcrea GWAS of this study and 634 independent signals defined by prior CKDGen study.**

**a.** Number of independent loci with/without independent signals defined by approximate conditional analyses for European 1,004,040 individuals in the latest CKDGen study by Stanzick et al.. For each of the 878 independent loci, the number of overlapping independent signals by Stanzick et al. was counted.

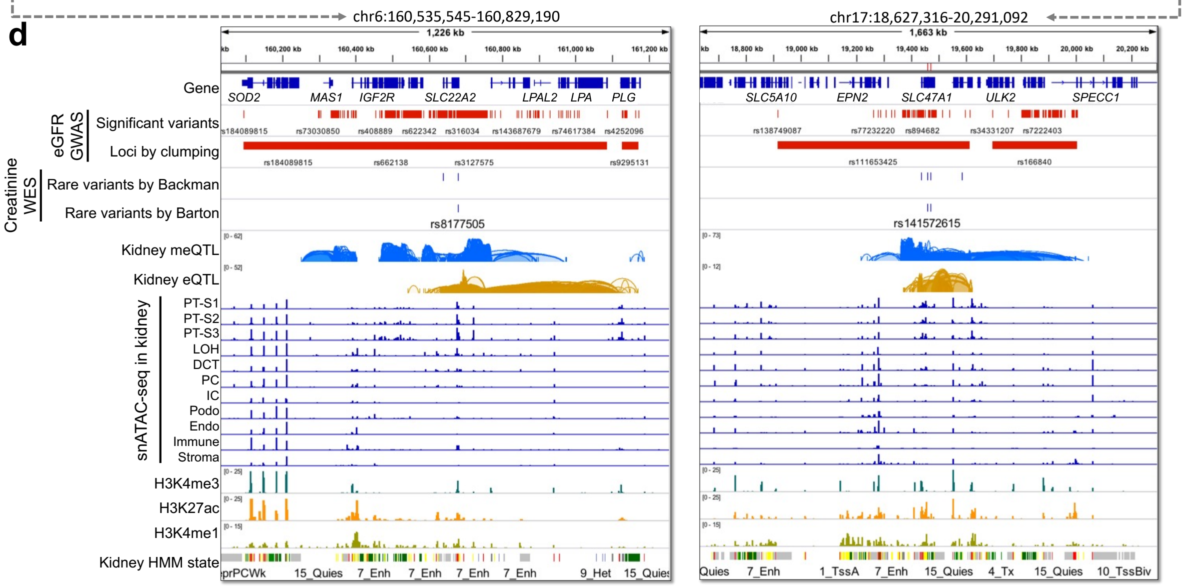
**b.** The summary statistics for variants in the 376 novel independent loci were extracted from the two datasets and common variants were used for the scatter plot. X-axis is the significance of common variants in this study, and y-axis is significance of common variants in the study by Stanzick et al. Correlation coefficient was calculated using Spearman's *rho (R)* statistic and p value was calculated using asymptotic *t* approximation.
**c.** LocusZoom view and functional annotation of shared independent loci/signals (chr14:92,671,432-94,507,626). For each independent locus, LocusZoom was plotted using summary statistics of meta-analysis eGFRcrea by this study, and the top variant was highlighted. Y-axis is strength of association -log10(two-sided *p* value from GWAS meta-analysis z-statistic). Three of the four independent loci were prioritized to different target genes (rs1075472 targeting *RIN3*, rs12588550 targeting *MOAP1*, rs12883201 targeting *ITPK1*, and rs76446292 targeting *UNC79*) in later analyses of this study.
**d.** LocusZoom view and functional annotation of multiple independent loci only separated by clumping-based method (chr14:92,671,432-94,507,626). For each independent locus, LocusZoom was plotted using summary statistics of meta-analysis eGFRcrea by this study, and top variant was highlighted. Y-axis is strength of association -log10(two-sided *p* value from GWAS meta-analysis z-statistic). Two different target genes (rs1042752 targeting *COLCA2*, and rs10891290 targeting *DIXDC1*) were prioritized in later analysis of this study (priority score ≥ 3).



c 9 rare variants associated with creatinine identified by both Backman et al. and Barton et al.

| Creatinine-associated rare variant identified by both WES studies | | | | | | | | | | The closest eGFR GWAS locus | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RSID | Chr | Position | Ref | Alt | MAF | P.value Backman | P.value Barton | Gene | Effect type | Chr | Start | End | Lead variant | MAF | P.value | Distance (bp) | LD ($R^2$) between WES variant and GWS variant * |
| rs8177505 | chr6 | 160679656 | A | AT | 0.0013 | 2.10E-63 | 2.60E-73 | *SLC22A2* | frameshift | chr6 | 160363006 | 161086716 | rs3127575 | 0.1155 | 9.00E-119 | 0 | 0.0151 |
| rs80308492 | chr6 | 43267663 | G | A | 0.0009 | 4.82E-33 | 6.80E-35 | *SLC22A7* | missense | chr6 | 43148051 | 43680896 | rs113990079 | 0.0782 | 2.10E-30 | 0 | 0.0001 |
| rs141572615 | chr17 | 19459205 | T | A | 0.0019 | 2.04E-19 | 1.00E-22 | *SLC47A1* | missense | chr17 | 18915262 | 19612945 | rs111653425 | 0.0113 | 1.04E-79 | 0 | 0 |
| rs147768037 | chr17 | 19470502 | G | A | 0.0007 | 6.80E-16 | 7.20E-11 | *SLC47A1* | missense | chr17 | 18915262 | 19612945 | rs111653425 | 0.0113 | 1.04E-79 | 0 | NA |
| rs149617956 | chr3 | 70014091 | G | A | 0.0041 | 1.74E-11 | 1.90E-11 | *MITF* | missense | chr3 | 69774170 | 70200450 | rs34297927 | 0.3975 | 4.62E-12 | 0 | 0.0014 |
| rs201194276 | chr19 | 36332622 | C | T | 0.0004 | 9.60E-12 | 3.60E-12 | *NPHS1* | missense | chr19 | 36342211 | 37024018 | rs3814995 | 0.3401 | 8.84E-15 | 9,590 | NA |
| rs150841256 | chr9 | 140127380 | G | A | 0.0005 | 2.51E-12 | 3.70E-15 | *SLC34A3* | splice_donor | chr9 | 140085029 | 140105550 | rs6606564 | 0.1548 | 1.40E-08 | 21,830 | NA |
| rs36095412 | chr1 | 20141060 | G | A | 0.0010 | 1.59E-23 | 1.30E-20 | *RNF186* | stop_gained | chr1 | 19786018 | 19792553 | rs7515104 | 0.0368 | 9.01E-09 | 348,507 | NA |
| rs1800546 | chr9 | 104189856 | C | G | 0.0055 | 1.61E-13 | 1.80E-10 | *ALDOB* | missense | chr9 | 103331404 | 103356112 | rs1226591 | 0.3991 | 2.09E-11 | 833,744 | 0.0006 |

*: LD between WES variant and GWAS variant was calculated using Ldpair (https://ldlink.nci.nih.gov/?tab=ldpair) based on European individuals of 1000G reference panel.
NA: Rare variant is not available in 1000G reference panel.



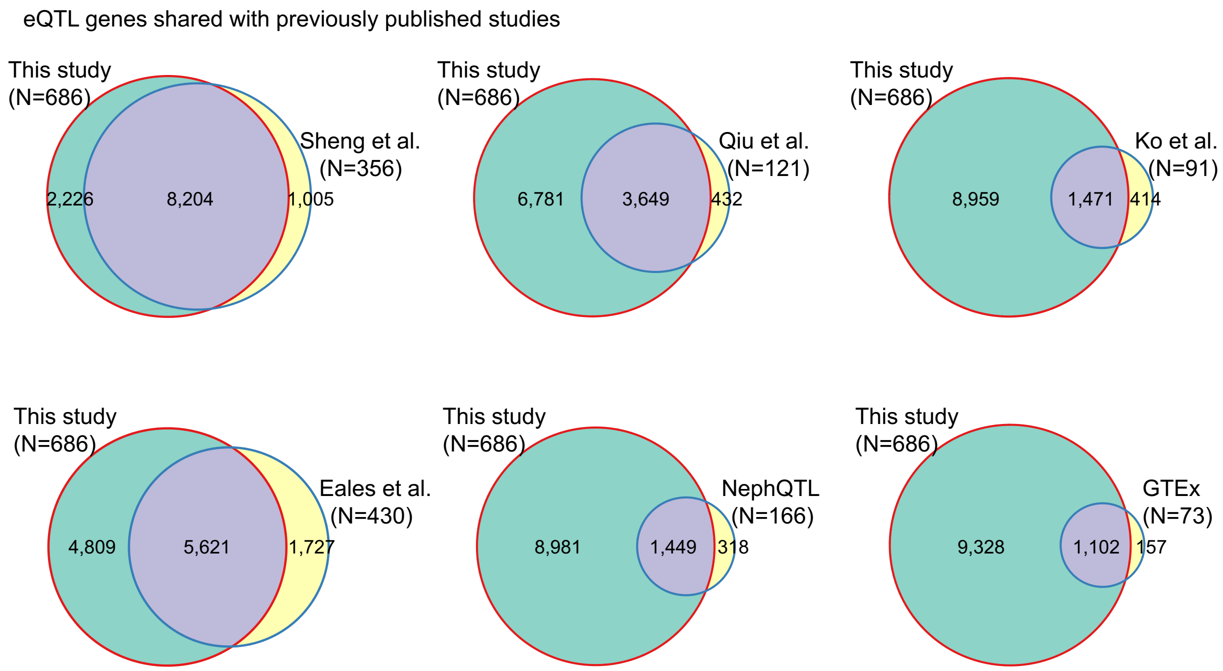
**Supplementary Fig. 3. The overlap of exome rare variants associated with creatinine-levels and meta-analysis eGFRcrea GWAS loci.**

**a.** Number of creatinine-associated exome rare variants overlapping with independent loci defined by meta-analysis eGFRcrea GWAS. Creatinine-associated exome rare variants were obtained from two whole-exome association studies by Backman et al. and Barton et al. For each creatinine-associated exome rare variant, the overlapping or the closest GWAS loci were identified.

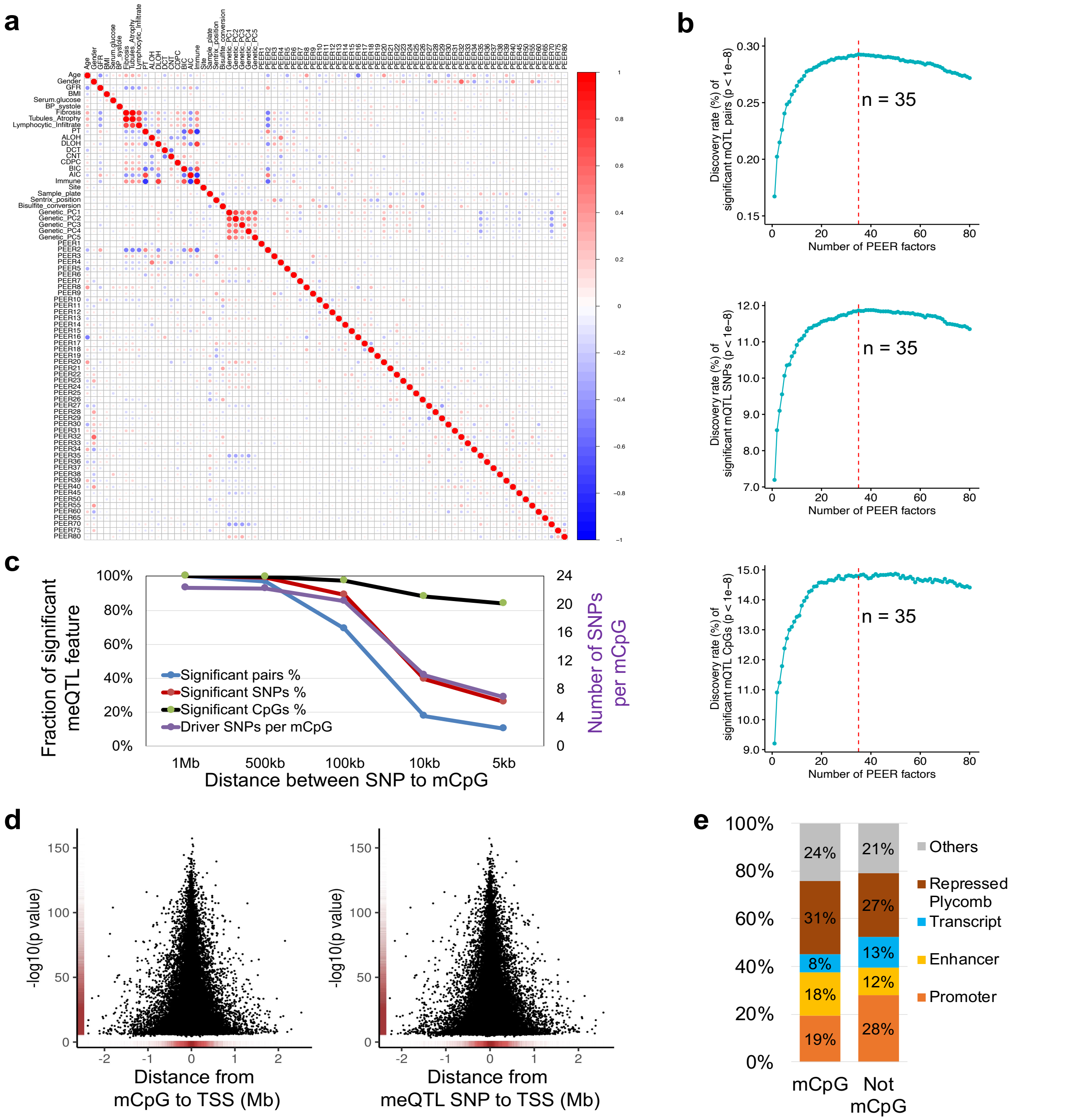
**b.** Number of eGFRcrea GWAS independent loci overlapping with creatinine-associated exome rare variants identified by Backman et al. or Barton et al.

**c.** Nine rare variants associated with creatinine identified by both Backman et al. and Barton et al.

**d.** Two examples of genomic loci where both rare variants and common variants were associated with kidney function markers.



**Supplementary Fig. 4. Comparison of eGenes identified by meta-analysis kidney eQTL of this study and previous studies.** Reported eGenes were obtained from each previous study and compared with eGenes identified by meta-analysis kidney eQTL of this study. N represents the number of individuals used for eQTL mapping in each study.



**Supplementary Fig. 5. PEER factors estimation and meQTL mapping.**

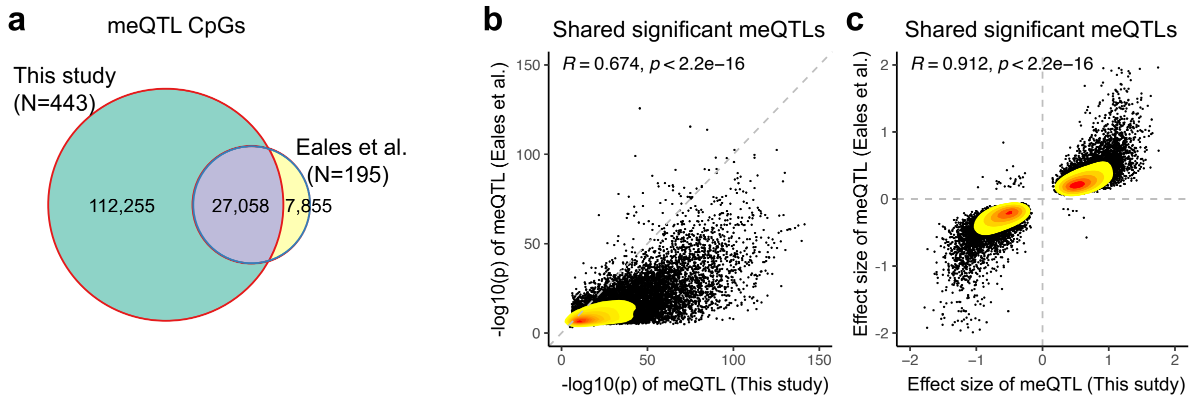
**a.** Correlation of PEER factors and clinal, technical variables, estimated cell fraction, and genotyping PCA. Heatmap of Spearman's rank correlation coefficient (rho) was calculated and shown in blue (negative correlation) and red (positive correlation).

**b.** The effect of the number of included PEER factor on the meQTL discovery rate. The number of identified features (CpG~SNP pairs, CpGs, SNPs on chromosome 1) (y-axis) vs. the number of PEER factors (x-axis) included in the linear regression model.

**c.** The relationship between significant meQTLs and the distance between SNPs and mCpGs.

**d.** The strength of association (y-axis) (-log10(p value calculated using linear regression meQTL model)) of the best mSNPs (the lead meQTL) decreases with the increasing distance (x-axis) from the CpG site to transcription start site (TSS) and from the SNP to TSS.

**e.** Chromatin state (human kidney ChromHMM) based functional annotation of meQTL significant CpGs.

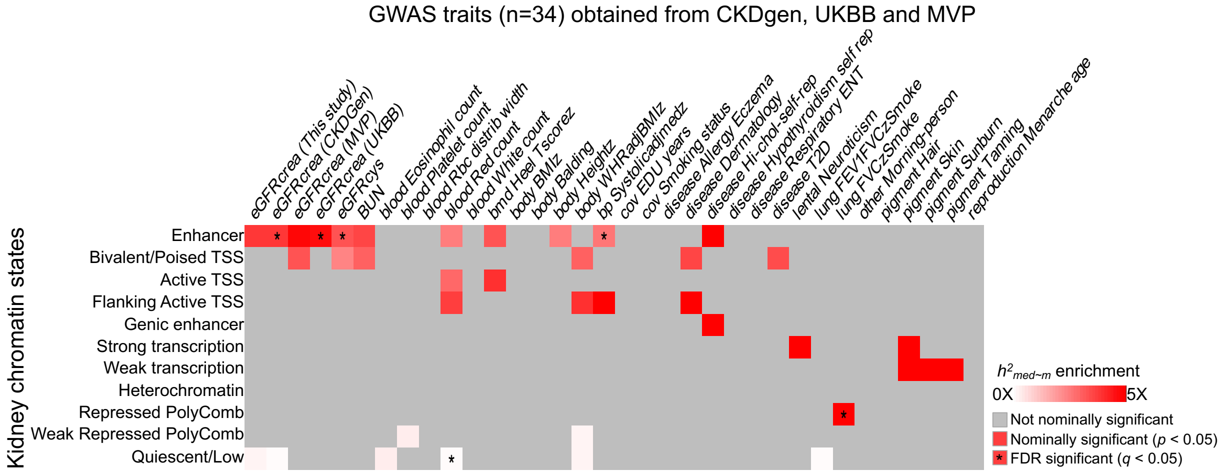


**Supplementary Fig. 6. Comparison of kidney meQTLs identified in this study and prior study.**

**a.** Venn plot of kidney meQTL CpGs identified in this study and a recently published study by Eales et al. N represents the number of individuals used for eQTL mapping in each study.

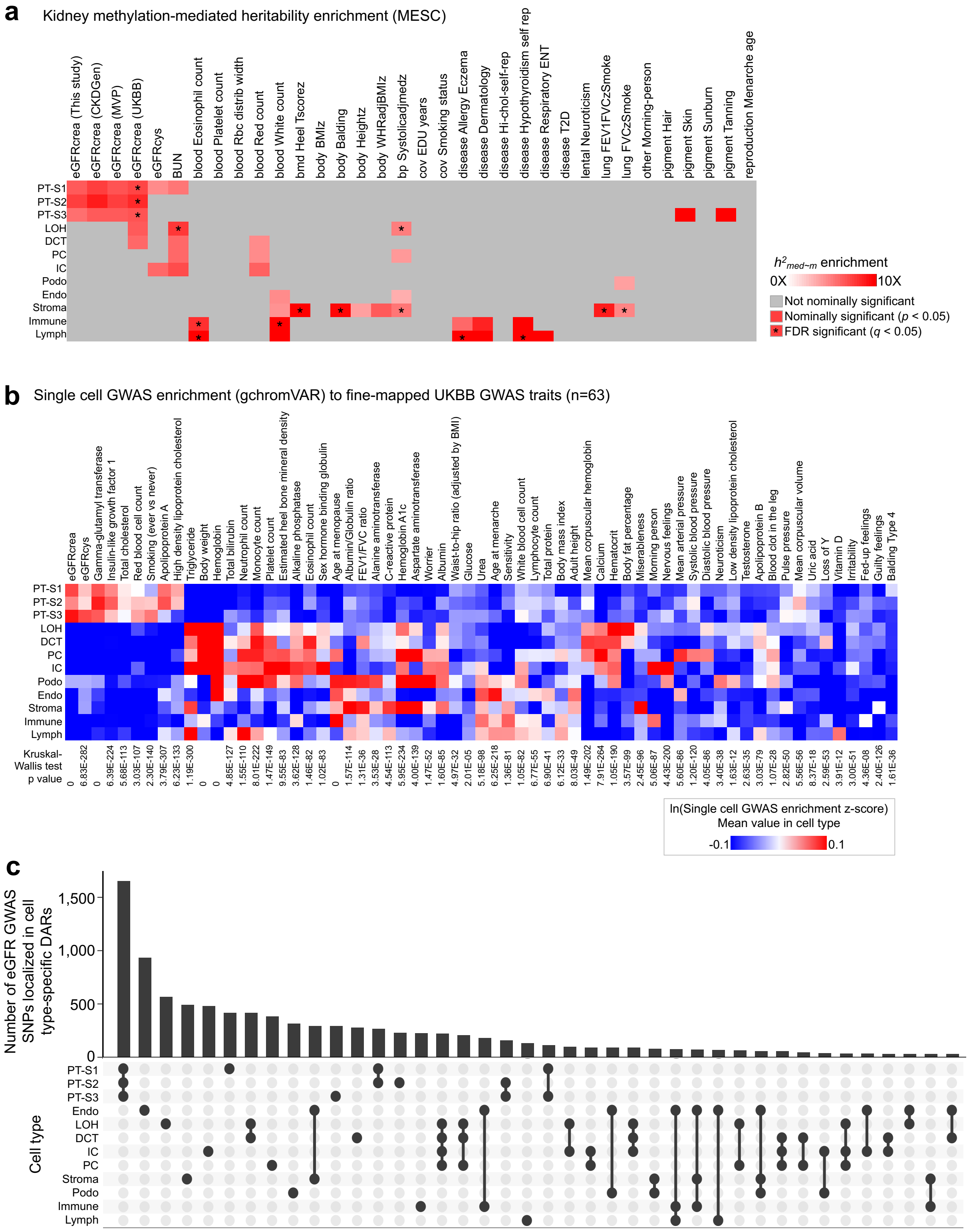
**b.** Correlation of meQTL significance between the two studies. Publicly accessible top variants and their association with meQTL CpGs were obtained from supplementary table of study by Eales et al., and then overlapped with meQTLs identified in this study. The shared meQTLs were used for the scatter plot of meQTL signification (-log10(p)) in this study (x-axis) and Eales et al. (y-axis). Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation.

**c.** Correlation of meQTL effect sizes between the two studies. Similarly, the shared meQTLs were used for the scatter plot of effect sizes in this study (x-axis) and Eales et al. Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation.



**Supplementary Fig. 7. Kidney methylation-mediated heritability enrichment of kidney chromatin states for GWAS traits.**

The x-axis shows the GWAS traits, while the y-axis shows the kidney chromatin states estimated by ChromHMM. Gray; non-significant. White to red indicates $h^2_{med}$ enrichment (nominal two-sided $p < 0.05$ calculated by MESC). Asterisk indicates $h^2_{med}$ enrichment passing FDR q < 0.05 (accounting for 374 tests for 11 chromatin state CpG sets and 34 GWAS traits).

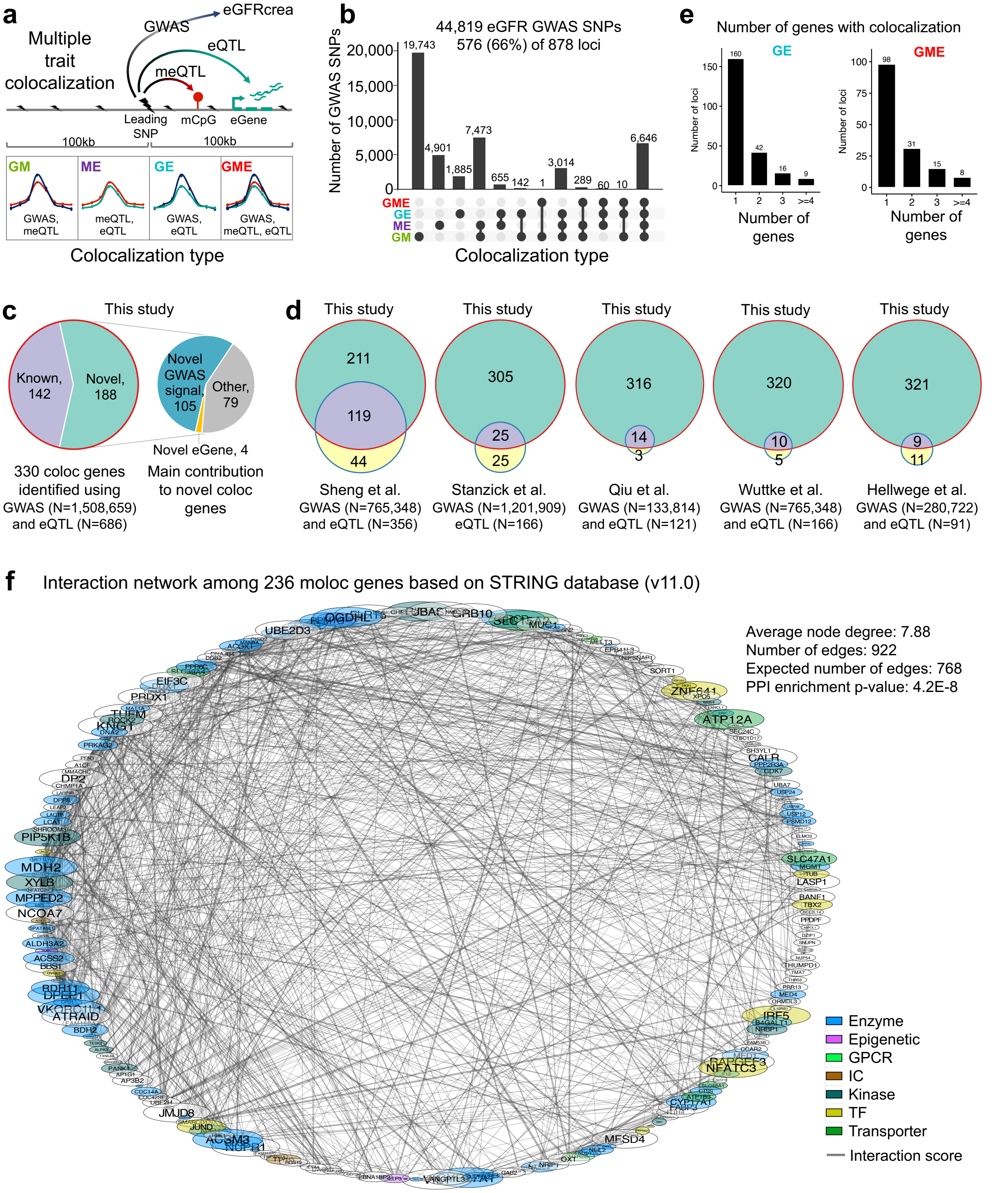


**Supplementary Fig. 8. Renal trait GWAS heritability mediated by kidney cell type open chromatin regions.**

**a.** Enrichment of methylation-mediated GWAS heritability in human kidney cell type-specific accessible regions. The x-axis shows the GWAS traits, while the y-axis shows cell types identified by snATAC-seq. Gray, non-significant, while white to red indicates $h^2{}_{med}$ enrichment (nominal two-sided $p < 0.05$ calculated by MESC). Asterisk indicates $h^2_{med}$ enrichment passing FDR q < 0.05 (accounting for 408 tests for 12 cell type CpG sets and 34 GWAS traits).

**b.** Single cell GWAS enrichment in human kidney cell type-specific accessible regions by gchromVAR. The x-axis shows the 63 fine-mapped GWAS traits, the y-axis the cell types clustered in the snATAC-seq. For each trait, gchromVAR was used to estimate the enrichment (z-score) of fine-mapped loci to the open chromatin peaks in each cell from the snATAC-seq. Then ln transformed z-scores of cells in the same cluster of given cell type was averaged for the heatmap plot. The single cell GWAS enrichment z-score mean values in each cell type for each trait is represented in color blue (low) to red (high). The two-sided *p* value was calculated by Kruskal-Wallis test.

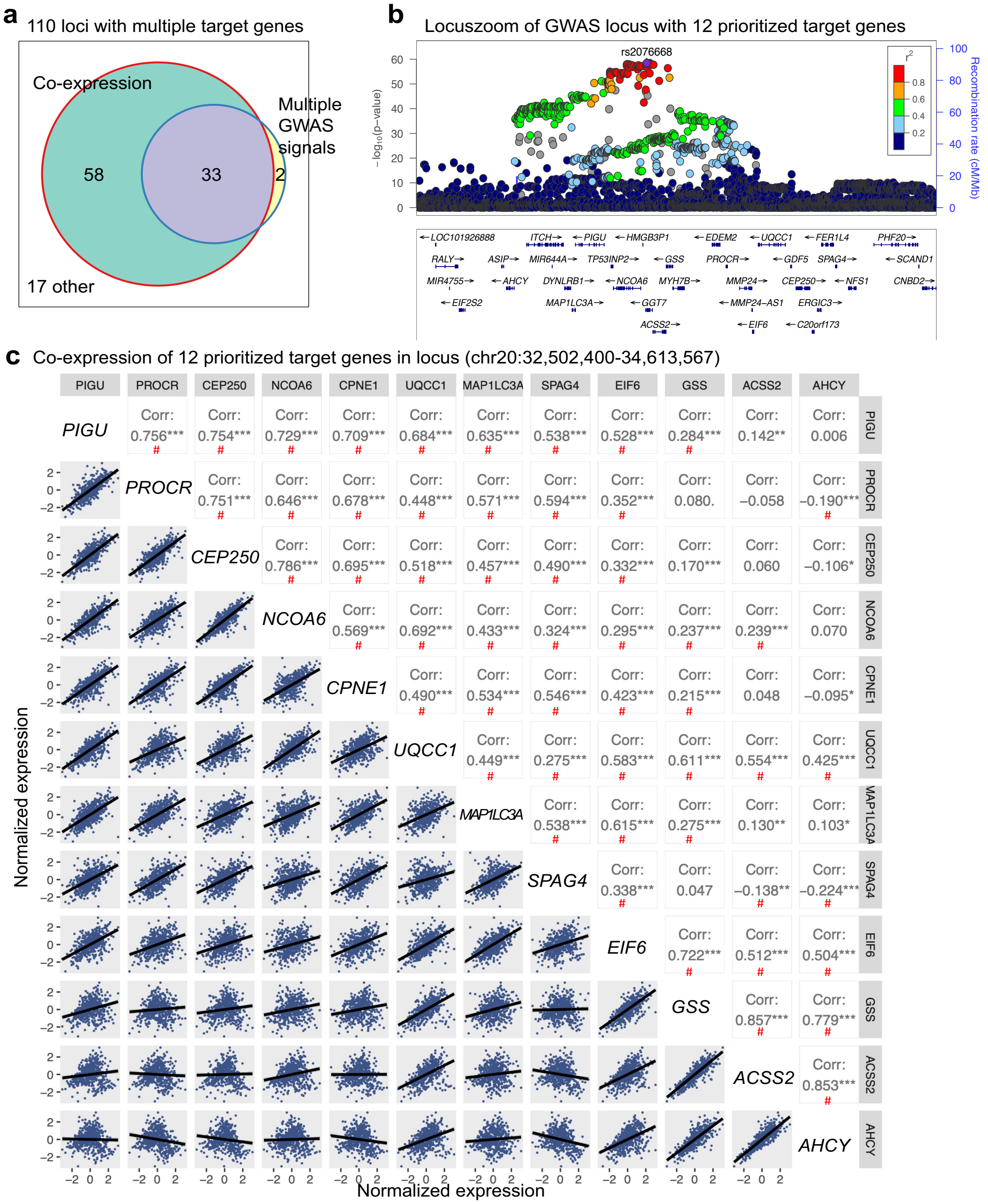
**c.** Number of genome-wide significant eGFRcrea GWAS SNPs localized in cell type-specific accessible regions. This figure was plotted using R package UpSetR (https://github.com/hms-dbmi/UpSetR/) to show the number of genome-wide significant eGFRcrea GWAS SNPs overlapping with cell type-specific differentially accessible regions (DARs). As some cell type specific DARs are shared between cells, therefore a SNP may be overlapped with DARs identified in different cells. Y-axis is the number of SNPs overlapped with DARs in different cell types (dots connected by black line), while the x-axis lists all available combinations of cell types.



**Supplementary Fig. 9. Colocalization analysis among eGFRcrea GWAS, kidney meQTL and eQTL.**
**a.** Schematic representation of multiple trait colocalization analysis of eGFRcrea GWAS, meQTL and eQTL.
**b.** Number of significant eGFRcrea GWAS variants showing colocalization across eGFRcrea GWAS, meQTL and eQTL.
**c.** Fraction of known and novel of colocalization of eGenes for eGFRcrea GWAS and kidney eQTL.
**d.** Venn plot of colocalization eGenes identified in this study and previous studies.

**e.** Number of independent loci and genes showing colocalization between GWAS and eQTL (left panel), and colocalization among GWAS, meQTL and eQTL (right panel).

**f.** Protein-protein associations of 236 moloc prioritized genes using the STRING database. The line between two dots represents the association between two proteins. The thickness indicates interaction score of associations quantified based on experiments, text mining and functional databases. The size of dot indicates the expression level of the gene in kidney, the color of dot represents gene family. The protein-protein interaction (PPI) enrichment p-value was calculated by hypergeometric test. Cytoscape file for this plot is available at github (https://github.com/hbliu/Kidney_Epi_Pri/blob/main/Gene_prioritization/Interaction_network_of_moloc_Genes.cys)

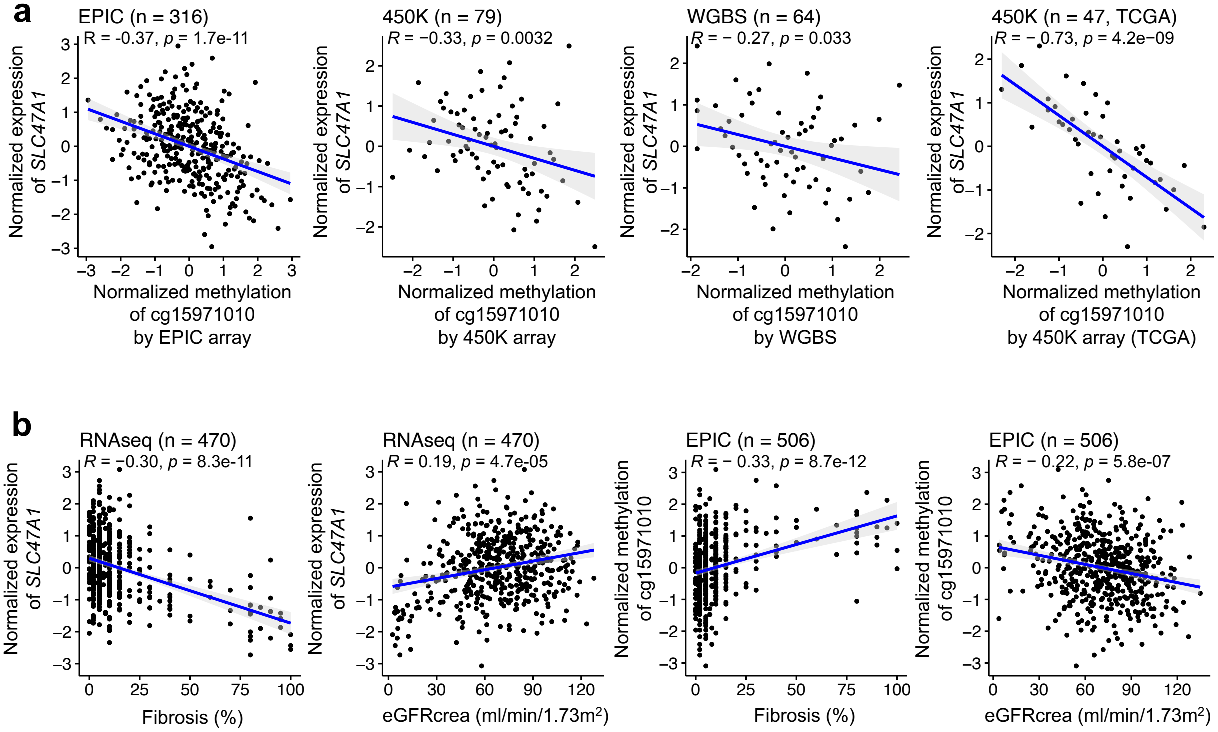


**Supplementary Fig. 10. Contribution of co-expression and multiple GWAS signals to 110 loci with multiple targe genes.**

**a**. Venn plot of loci prioritized to multiple target genes with co-expression and multiple GWAS signals.

**b**. Locuszoom of GWAS locus (chr20:32,502,400-34,613,567) with 12 prioritized target genes. Top variant rs2076668 was highlighted. Y-axis is strength of association -log10(two-sided *p* value from GWAS meta-analysis z-statistic).

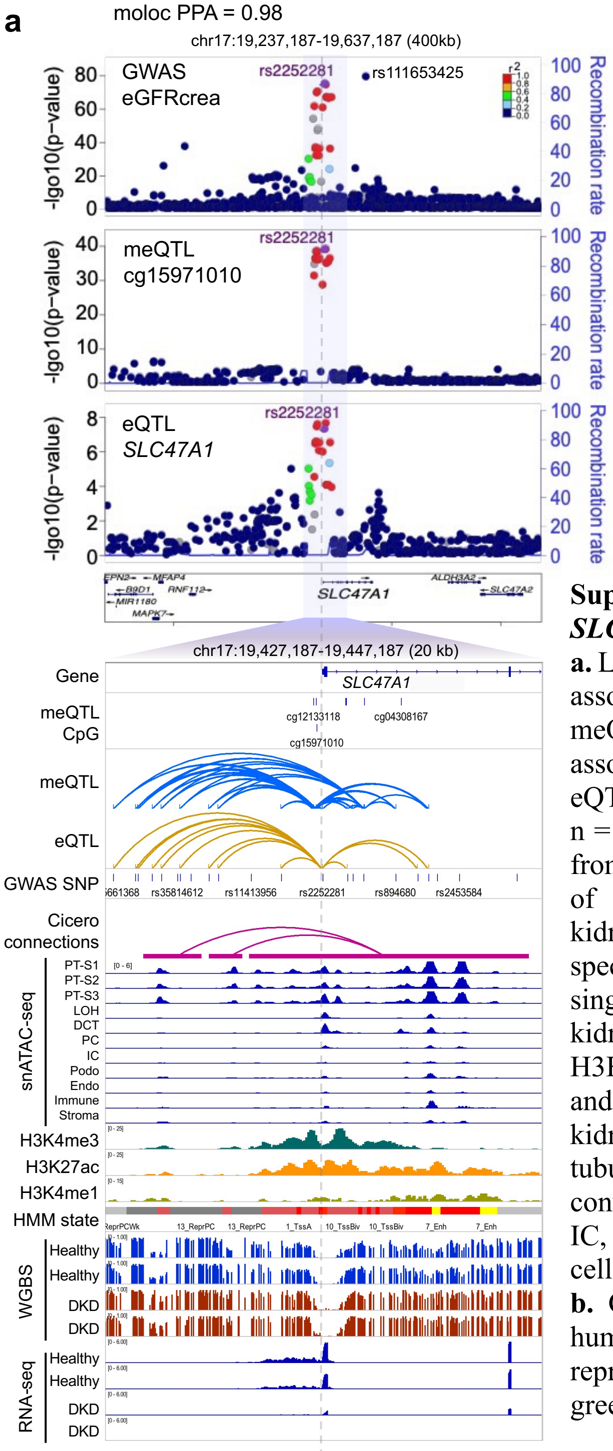
**c.** Co-expression of 12 prioritized target genes at this locus (chr20:32,502,400-34,613,567). For each gene, the normalized expression (INT transformed TPM) values in 470 kidney samples were used for the scatterplot (bottom panel). Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation. Significance of correlation was showed using “***” if the p-value is < 0.001, “**” if the p-value is < 0.01, “*” if the p-value is < 0.05 and # is FDR < 0.05.

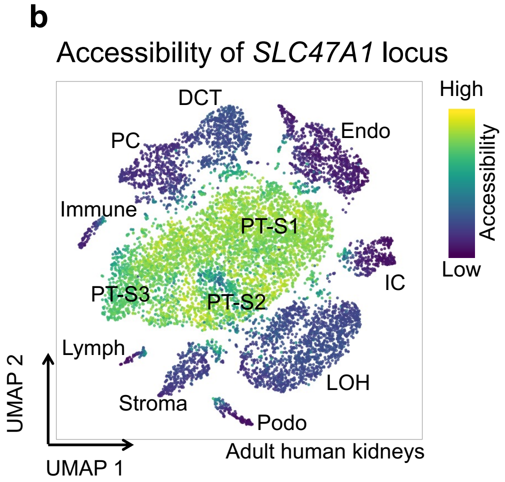


**Supplementary Fig. 11. Regional methylation data and gene expression at the *SLC47A1* locus**

**a.** Correlation between local CpG methylation and *SLC47A1* expression in human kidneys. X-axis shows normalized methylation, y-axis normalized expression data. Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation.

**b.** Correlation between local CpG methylation and *SLC47A1* expression and kidney functions in human kidneys. X-axis eGFRcrea (ml/min/1.73m$^2$) or fibrosis (%). Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation.

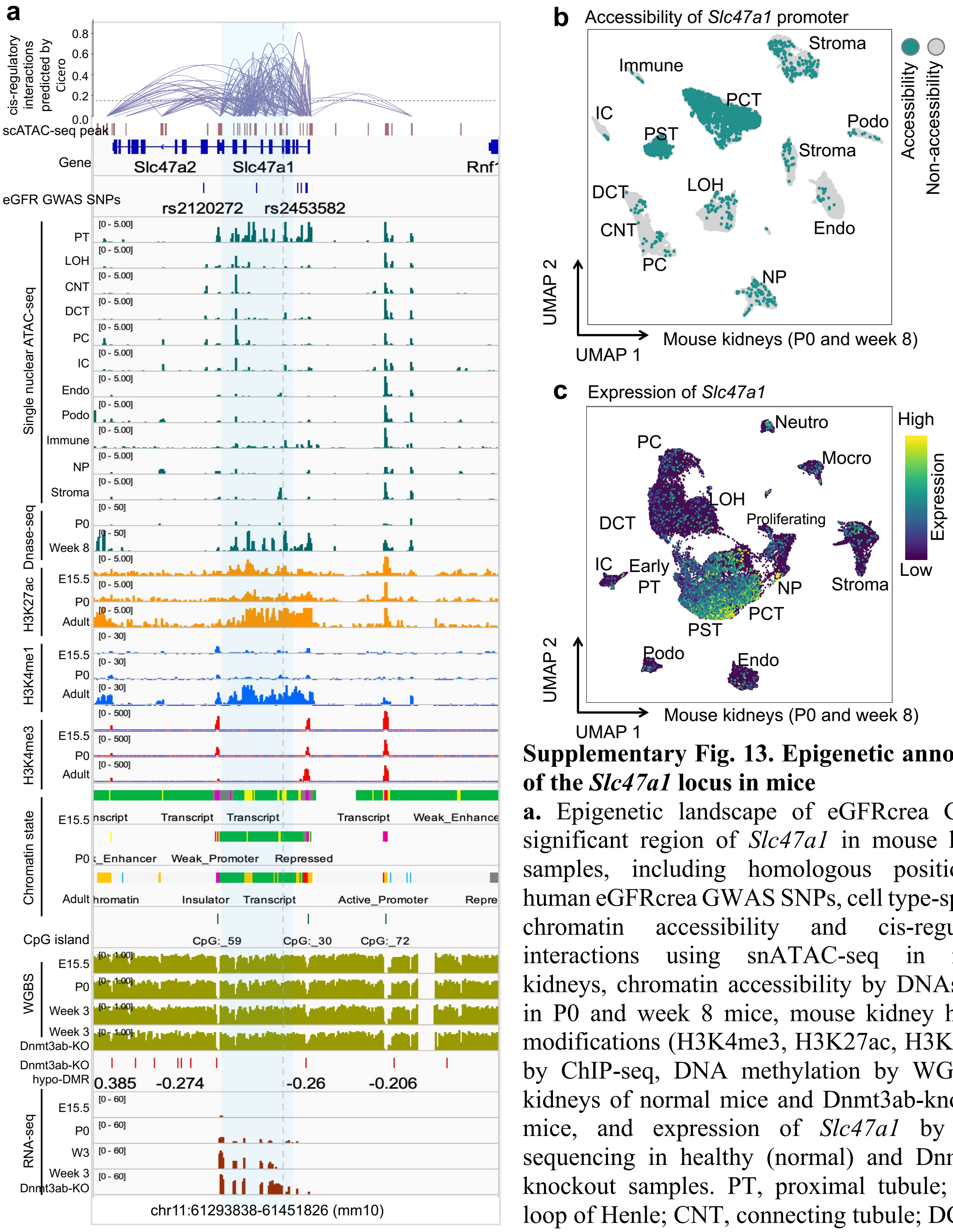




**Supplementary Fig. 12. Epigenetic annotation of in *SLC47A1* locus**

**a.** LocusZoom plots of GWAS (genotype and eGFRcrea association, N = 1,508,659), kidney CpG cg15971010 meQTL (genotype and cg15971010 methylation association n = 443) and kidney tubule SLC47A1 eQTLs (genotype and *SLC47A1* expression association n = 356). The y axis shows -log10($p$ value) calculated from GWAS, meQTL, and eQTL. Epigenetic landscape of eGFRcrea GWAS significant region in human kidney samples, including meQTLs, eQTLs, cell type-specific chromatin accessibility in human kidneys by single-nucleus ATAC-seq (snATAC-seq), human kidney histone modifications (H3K4me3, H3k27ac, H3K4me1) by ChIP-seq, DNA methylation by WGBS, and RNA sequencing in healthy (normal) and diabetic kidney disease (DKD) samples. PT-S1-3; proximal tubule S1-3 segment, LOH, loop of Henle, DCT, distal convoluted tubule, PC, principal cell of collecting duct, IC, intercalated cell of collecting duct, endo; endothelial cells, immune: immune cell and stroma.

**b.** Chromatin accessibility at the *SLC47A1* locus in human kidney analyzed by snATAC-seq. Each dot represents a cell, dark blue indicates lower and bright green higher expression.

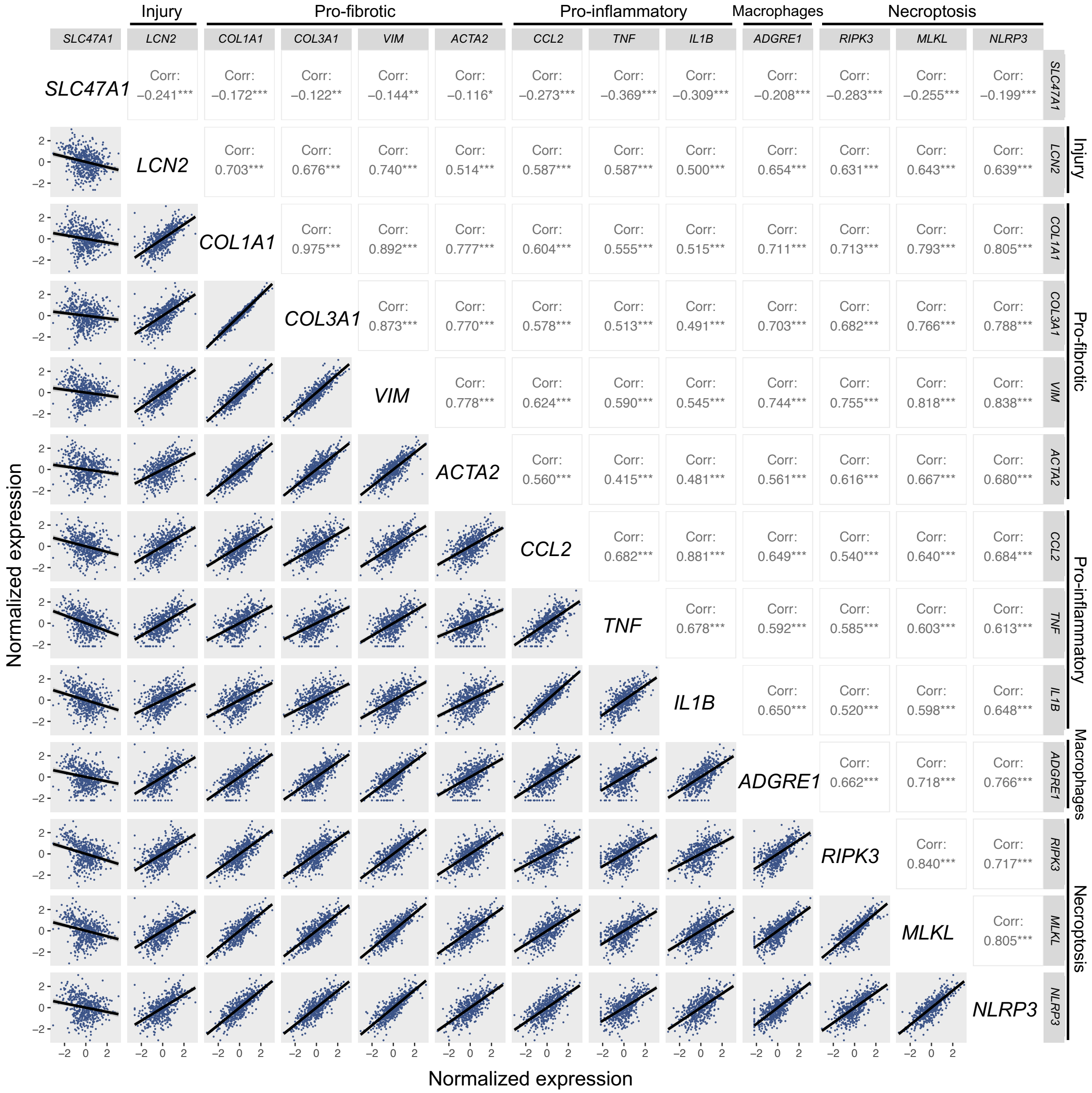


**Supplementary Fig. 13. Epigenetic annotation of the *Slc47a1* locus in mice**

**a.** Epigenetic landscape of eGFRcrea GWAS significant region of *Slc47a1* in mouse kidney samples, including homologous position of human eGFRcrea GWAS SNPs, cell type-specific chromatin accessibility and cis-regulatory interactions using snATAC-seq in mouse kidneys, chromatin accessibility by DNAse-Seq in P0 and week 8 mice, mouse kidney histone modifications (H3K4me3, H3K27ac, H3K4me1) by ChIP-seq, DNA methylation by WGBS in kidneys of normal mice and Dnmt3ab-knockout mice, and expression of *Slc47a1* by RNA sequencing in healthy (normal) and Dnmt3ab-knockout samples. PT, proximal tubule; LOH, loop of Henle; CNT, connecting tubule; DCT, distal convoluted tubule; PC, principal cell of collecting duct; IC, intercalated cell of collecting duct, Endo, endothelial cells; Podo, podocytes; Immune, immune cell; NP, nephron progenitors; and stroma.

**b.** Cell type-specific chromatin accessibility at the *Slc47a1* gene promoter in mouse kidneys (P0 and 8 weeks). Each dot represents a cell, with green color indicates accessible and grey not-accessible region in denoted cells. Proximal tubule cells represented by two clusters: proximal straight tubule (PST), proximal convoluted tubule (PCT).

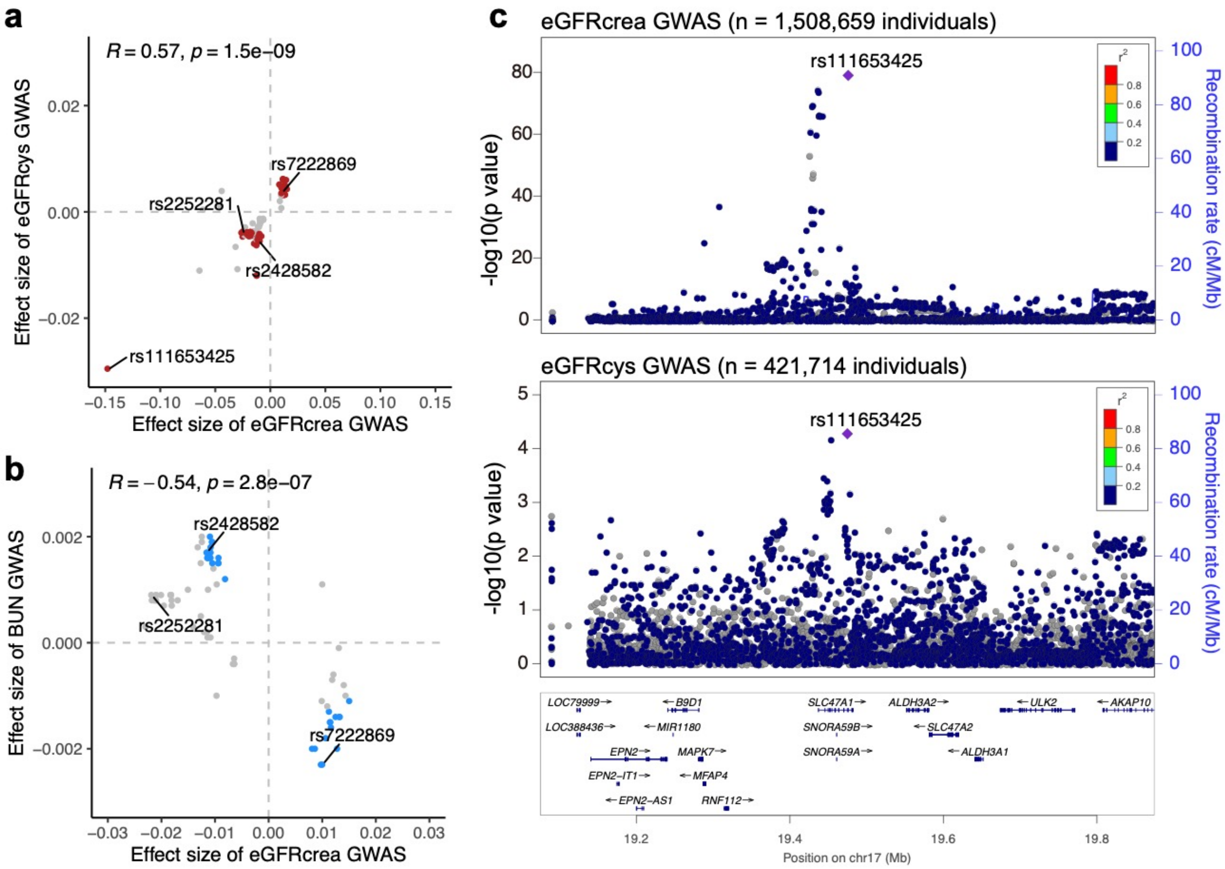
**c.** Cell type specific expression of *Slc47a1* in mouse kidneys (P0 and 8 weeks GSE157079). Each dot represents a cell, dark blue indicates lower and bright green higher expression.



**Supplementary Fig. 14. Expression of *SLC47A1* and markers of kidney injury, fibrosis, and inflammation in human kidneys.**

From top to bottom (and left to right), the genes are solute carrier family 47 member 1 (*SLC47A1*)**,** Lipocalin 2 **(***LCN2*), Collagen 1 (*COL1A1*), Collagen3 (*COL3A1*), Vimentin (*VIM*), Actin alpha 2 (*ACTA2*), Chemokine ligand 2 (*CCL2*), Tumor necrosis factor (*TNF*), Interleukin 1beta (*IL1B*), Adhesion G protein-coupled receptor E1 (*ADGRE1*), Receptor interacting serine/threonine kinase 3 (*RIPK3*), Mixed Lineage Kinase Domain Like Pseudokinase (*MLKL*), and NLR family pyrin domain containing 3 (*NLRP3*).

For each gene, the normalized expression (INT transformed TPM) values were used for scatterplot (bottom panel). Correlation coefficient was calculated using Spearman's *rho (R)* statistic and two-sided *p* value was calculated using asymptotic *t* approximation. Significance of correlation was showed using "***" if the p-value is < 0.001, "**" if the p-value is < 0.01 and "*" if the p-value is < 0.05.



**Supplementary Fig. 15. Validation of eGFRcrea GWAS in *SLC47A1* locus using eGFRcys GWAS and BUN GWAS.**

**a**. Scatter plot of effect sizes between eGFRcrea GWAS (N=1,508,659 individuals) and eGFRcys GWAS (N=421,714 individuals) in *SLC47A1* locus. Significant eGFRcrea GWAS variants passing two-sided $p < 5\times10^{-8}$ in *SLC47A1* locus were used for the plot. Red dots represent validated variants showing nominally significant (two-sided GWAS $p < 0.05$) association with eGFRcys in the same effect direction. Correlation coefficient was calculated using Spearman's rho (R) statistic and two-sided p value was calculated using asymptotic t approximation.

**b**. Scatter plot of effect sizes between eGFRcrea GWAS (N=1,508,659 individuals) and BUN GWAS (N=852,678 individuals) in *SLC47A1* locus. Significant eGFRcrea GWAS variants passing two-sided $p < 5\times10^{-8}$ in *SLC47A1* locus were used for plot. Blue dots represent validated variants showing nominally significant (two-sided GWAS $p < 0.05$) association with BUN in the opposite effect direction. Correlation coefficient was calculated using Spearman's rho (R) statistic and two-sided p value was calculated using asymptotic t approximation.

**c**. Locuszoom of eGFRcrea GWAS (n = 1,508,659 individuals) and eGFRcys GWAS (n = 421,714 individuals) GWAS in *SLC47A1* locus. The top SNP for eGFRcrea GWAS in *SLC47A1* locus was highlighted in two Locuszoom plots.