PNAS
www.pnas.org

Supplementary Information for

Systematic integrated analysis of genetic and epigenetic variation in diabetic kidney disease

Xin Sheng[1,2], Chengxiang Qiu[1,2], Hongbo Liu[1,2], Caroline Gluck[1,2], Jesse Y. Hsu[3,4], He Jiang[5], Chi-yuan Hsu[6], Daohang Sha[4], Matthew R. Weir[7], Tamara Isakova[8,9], Dominic Raj[10], Hernan Rincon-Choles[11], Harold I. Feldman[1,3,4], Raymond Townsend[1], Hongzhe Li[3,4], Katalin Susztak[1,2]

Katalin Susztak

Email: ksusztak@pennmedicine.upenn.edu

This PDF file includes:

Supplementary text

Figs. S1 to S24

Legends for Datasets S1 to S31

SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S31

Supplemental Information Text

Supplemental Methods

DNA methylation quantification and quality control

Cytosine methylation levels at 866,836 sites were quantified using the Illumina Infinium MethylationEPIC Beadchip (Illumina). The *minfi* package (1) was used to extract the intensity methylation value for each site and for pre-processing and quality control. Background correction and control normalization for raw intensity signals were performed using *preprocessIllumina* function (2). We excluded probe sites with P>0.01 in >10% of the sample and samples (N=0) having >10% of the probe sites with poor detection p value (p>0.01). A Beta-mixture quantile (BMIQ) normalization method was adopted to adjust beta values of the type 2 probes to the statistical distribution characteristic of type 1 probes on the array (3). Probes located on sex chromosomes, cross-reactive probes and probes located within 10 bp of 1,000 Genomes Project (4) SNPs (MAF>0.01) were removed (5, 6). But probes located within 10 bp of SNPs were included in mQTL analysis. One outlier sample was excluded after principal component analysis (PCA) (7) (**Fig. S22**). We used the *impute* package for probes with missing values (8) to estimate the missing methylation values of k-nearest neighboring probes (9) with the default value k=10. Methylation value at each site was represented as a beta (**β**) value, the ratio of the methylated signal (M) to the sum of the methylated and unmethylated signals (U), $\boldsymbol{\beta}$ = M/(M+U), ranging from 0 (non-methylated) to 1 (completely methylated). Due to the fundamental heteroscedasticity of beta value(10), this was transformed to an M value ($M=\log_2(\boldsymbol{\beta}/(1-\boldsymbol{\beta}))$) in all the linear regression models (11).

Genome-wide SNP genotyping and quality control

A total of 3,598 CRIC Study participants had available genotype information using the Illumina HumanOmni 1-Quad v1.0 microarrays. Raw genotype data were downloaded from dbGaP in PLINK format (12). Of the 500 matched participants, 481 participants had available genotype data. PLINK (13) was used to remove samples with >5% missing values, monomorphic SNPs and SNPs with <90% call rate, Hardy-Weinberg equilibrium $P<1\times10^{-6}$, and a minor allele frequency of <1%. Association with chemistry plate ID was estimated using the PLINK function *–loop-assoc*. To identify poor DNA quality or sample contamination, we applied a heterozygosity test to variants locates in autosomes using the inbreeding coefficient. Five samples were excluded because of reduced genotypic heterozygosity (inbreeding coefficient cutoff: heterozygosity rate ± 3-fold standard deviations from the mean). Identity-by-descent (IBD) was computed for all pairwise sample combinations to further identify potential sample contamination (3 samples were excluded due to their high PIHAT value (>0.2)). Sex was verified by using X and Y SNP genotypes. SNPs located on sex chromosome were excluded (14). After quality control, 473 participants and 809,455 SNPs were used for the mQTL analysis (**Table S31**).

Population Structure and Imputation

We performed Principal Component Analysis (PCA) implemented in EIGENSTRAT (15) on the genotype data of the 473 samples with the additional 2,504 genotype data from the 1000 Genomes Project Phase 3 (661 AFR, 347 AMR, 504 EAS, 503 EUR, 489 SAS) (4). Samples from our study and samples from 1,000 Genomes Project clustered with respect to their ancestral background (**Fig. S23**). For all further linear regression analysis (MWAS and mQTL), the first 10 principal components, generated from the 1,000 Genomes Project Phase 3 as the reference panel with the 473 CRIC samples, were used as covariates. To increase the power and coverage of mQTL discovery in our study, we pre-phased genotype data of 473 samples by SHAPEIT2 (16), and imputed by IMPUTE2 (17, 18), using 1,000 Genomes Phase 3 (NCBI build 37, release date October 2014) as the reference panel. We also applied additional quality control steps, including removing indels, SNPs not presented in the 1,000 Genomes dataset and alleles incompatible with reference panel before imputation. The imputed SNPs with missing rate cutoff <95% for best-estimated genotypes at posterior probability >0.9, Hardy-Weinberg Equilibrium $p<1\times10^{-6}$, imputation confidence score, INFO (a measure of r^2) <0.4 (calculated by SNPTEST (19)), and MAF<5% were excluded. We used 6,177,888 high quality autosomal SNPs for the mQTL analysis.

Model selection

In order to improve model accuracy and reduce model over-fitting in the linear regression analysis, we adopted backward stepwise procedure to select the corresponding confounders for each response variable. For each response variable, first we included all available variables in its full model, then built a null model by eliminating one variable iteratively each time (20). We calculated the Akaike Information Criterion (AIC) values for both full and null models, and used ANOVA test to estimate whether the eliminated variable has significant influence on the response variable. Both full and null linear mixed effect models were fitted by maximizing the log-likelihood criterion. Only if the difference significance P-value (ANOVA test) between both models was below 0.05, and the AIC value of the full model reduced comparing to the null model, the eliminated variable was retained as an independent variable for the response variable. First, we selected the most significant probes associated with eGFR by only including M-value as independent variable in the linear regression model. Then, we adopted the M-values of these probes as dependent variable to select independent variables that significantly affect M-values. At last, age, batch effect, top 10 PCs of genetic background, hypertension and whole blood cell subtype composition were included in the model to adjust M-value. For baseline eGFR, age, sex, hypertension, and hemoglobin A1c were selected as confounders. The top 10 PCs of genetic background and blood cell proportions were also added to improve the cross compatibility of the data obtained from different populations. Baseline eGFR was adjusted (we used residualized eGFR: baseline eGFR adjusted for age, sex, hypertension, hemoglobin A1c, top 10 genetic PCs, and whole blood cell proportions) before we conducted linear regression. The distribution of residualized eGFR is shown in **Fig. S2D**. Similarly, the residualized eGFR slope was used in the linear regression model to calculate the association between kidney function decline and DNA methylation changes. It was defined as eGFR slope by adjusting for baseline eGFR, sex, age, top 10 genetic PCs, hemoglobin A1c, urine albumin creatinine ratio, and whole blood cell proportions. The distribution of residualized eGFR slope are shown in **Fig. S24**. We used baseline eGFR to represent kidney function and eGFR slope to represent kidney function decline in our MWAS calculations.

Robustness of the MWAS analysis for baseline kidney function

As multiple environmental factors can influence methylation levels, we performed sensitivity analysis to test the robustness of our results. First, we analyzed the relationship of the top 10 principal components (PCs) derived from the DNA methylation data and measured phenotypes. We observed a weak correlation (r<0.2) between smoking and PC2 (r=-0.108), and BMI and PC10 (r=-0.190). Next, we iteratively compared the regression coefficients from our initial model (adjusted for age, batch effect, top 10 genetic PCs, hypertension, imputed whole blood cell proportions, and hemoglobin A1c) to models including smoking history, smoking status and BMI. The associations of DMPs with kidney function remained robust, even after adjusting for smoking history (**Fig. S7A and B**), smoking status (**Fig. S7C and D**), and BMI (**Fig. S7E and F**) (Pearson's correlation > 0.99). Adding genotype information improved the stability of the model (Pearson's correlation ~ 0.984 to 0.998) (**Fig. S7I-L**). We also performed *post hoc* analyses to investigate the contribution of age and smoking. We confirmed the association between published smoking-associated methylation probes in our dataset (Pearson's correlation=0.987, two-sided P=1.003E-19, **Fig. S7M**) (21). The 'epigenetic age', calculated by a modified version of the method described by Horvath (22), highly correlated with the actual age of the subjects (Pearson's correlation=0.830, two-sided P=2.25E-121, **Fig. S7N**). To conclude, Smoking- (21), age- (22), and BMI- (23) associated methylation changes did not show significant association (two-sided P <5E-05) with kidney function in our study.

Methylome wide association study for kidney function decline (conditional logistic regression)

The distribution of the significance across the genome is shown in **Fig. S8A**. As the methylation difference between the two groups increased, the significance of MWAS association increased (**Fig.**

S8C). The methylation levels of cg11244695 showed the strongest significance between the two groups (t-test two-sided P=3.4E-06) (**Fig. S8D**). We further interrogated the robustness of the kidney function decline associated DMPs by sensitivity analyses. We compared the regression coefficients from our initial model to models that included smoking history, smoking status and BMI. The Pearson's correlation coefficients for these models were above 0.99, indicating that DMPs were related to kidney function decline and in line with models adjusting for smoking history (**Fig. S10A and B**), smoking status (**Fig. S10C and D**), and BMI (**Fig. S10E and F**).

Sensitivity analysis

To support the associations between DNA methylation and kidney function or functional decline, we conducted additional sensitivity analyses by iteratively including race, top 3 or 5 PCs of genetic background, current smoking status, smoking history in the past 100 days, and BMI to the primary association analysis model of baseline eGFR, kidney function decline rate group and eGFR slope, respectively. Results of the statistical analyses (effect size and P value) of the different models were compared.

RNA-seq and DNA methylation data of human kidney samples

Human kidney samples were collected from routine surgical nephrectomies with the approval of the institutional review board (IRB) of the University of Pennsylvania. Samples were de-identified and clinical information was obtained via an honest broker. Pathology examination was conducted by local nephropathologist. The kidney tissue was immediately placed and stored in RNAlater (Ambion) based on manufacturer's instruction. Cytosine methylation levels at 866,836 sites were quantified using the Illumina Infinium MethylationEPIC Beadchip (Illumina) of 227 whole kidney samples (**Table S13**). The tissue was further manually microdissected under a microscope in RNAlater for tubular compartment. Dissected tubule was homogenized, and RNA was prepared using RNAeasy mini columns (Qiagen, Valencia, CA) according to the manufacturer's instructions. RNA quality was assessed with the Agilent Bioanalyzer 2100 and samples with RIN scores above 7 were adopted. Libraries were prepared by the Illumina TruSeq RNA Preparation Kit. 433 Samples were sequenced by Illumina HiSeq for single-end 100 bp (**Table S16**). Low quality reads and adapters were trimed by Trim-galore (24), and trimmed reads were further aligned to the human genome (hg19/GRCh37) with STAR (2.4.1d) based on GENCODE v19 annotations (25). Gene-level expression was estimated with uniquely mapped reads as reads per kilobase of transcript per million mapped reads (RPKM) using HTSeq (0.9.1) (26).

Human Kidney single cell RNA-seq data (27)

We downloaded the Supplementary Data S1 of Young et al. (27), and used the data of 4,796 normal immune cells, which originated from five healthy adult kidneys. We adapted the author's R code (https://github.com/constantAmateur/scKidneyTumors) to process and normalize the data using R package Seurat v2.3.4 (28). The individual 10X sequencer channel information was used to adjust batch effect by ComBat (29), while preserving zero expression values. We further normalized and log transformed expression data by the NormalizeData function in Seurat with scale factor =1E04. Cluster assignment of each cell was based on the information in Supplementary Data S11 and cell labels were given to the 11 normal immune cell clusters according to the information in the Supplementary Data S2 of Young et al. (27).

Enrichment test of functional categories

We used the chromatin state annotation from human PBMCs, as our methylation data was measured from PBMCs. The chromatin states annotation data of human PBMCs, heart and liver were downloaded from ROADMAP Epigenomics Data (https://egg2.wustl.edu/roadmap/web_portal/) (30). The core 15-state model were merged into 4 main functional categories (enhancer, promoter, transcribed, inactive). Human kidney ChIP-seq data for histone H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9ac, and H3K9me3 (GSM621634, GSM621648, GSM621651, GSM670025, GSM772811, GSM1112806) were used to annotate chromatin states using ChromHMM (31). To estimate the distribution under the null hypothesis that there is no enrichment, we randomly selected 99 and 111 probes with similar methylation variance (difference below 0.1) of the 99 and 111 DMPs as background for eGFR and eGFR slope, respectively, and repeated the analysis 500 times. The fold change was calculated by the frequency ratio of DMPs and background probes that fell into each chromatin state.

Pathway analysis of candidate causal genes.

Web-based programs, DAVID (https://david.ncifcrf.gov) and GIANT (http://giant.princeton.edu) (32) were used to define pathway enrichment for MWAS results and high fidelity CKD risk genes.

Supplementary Figures

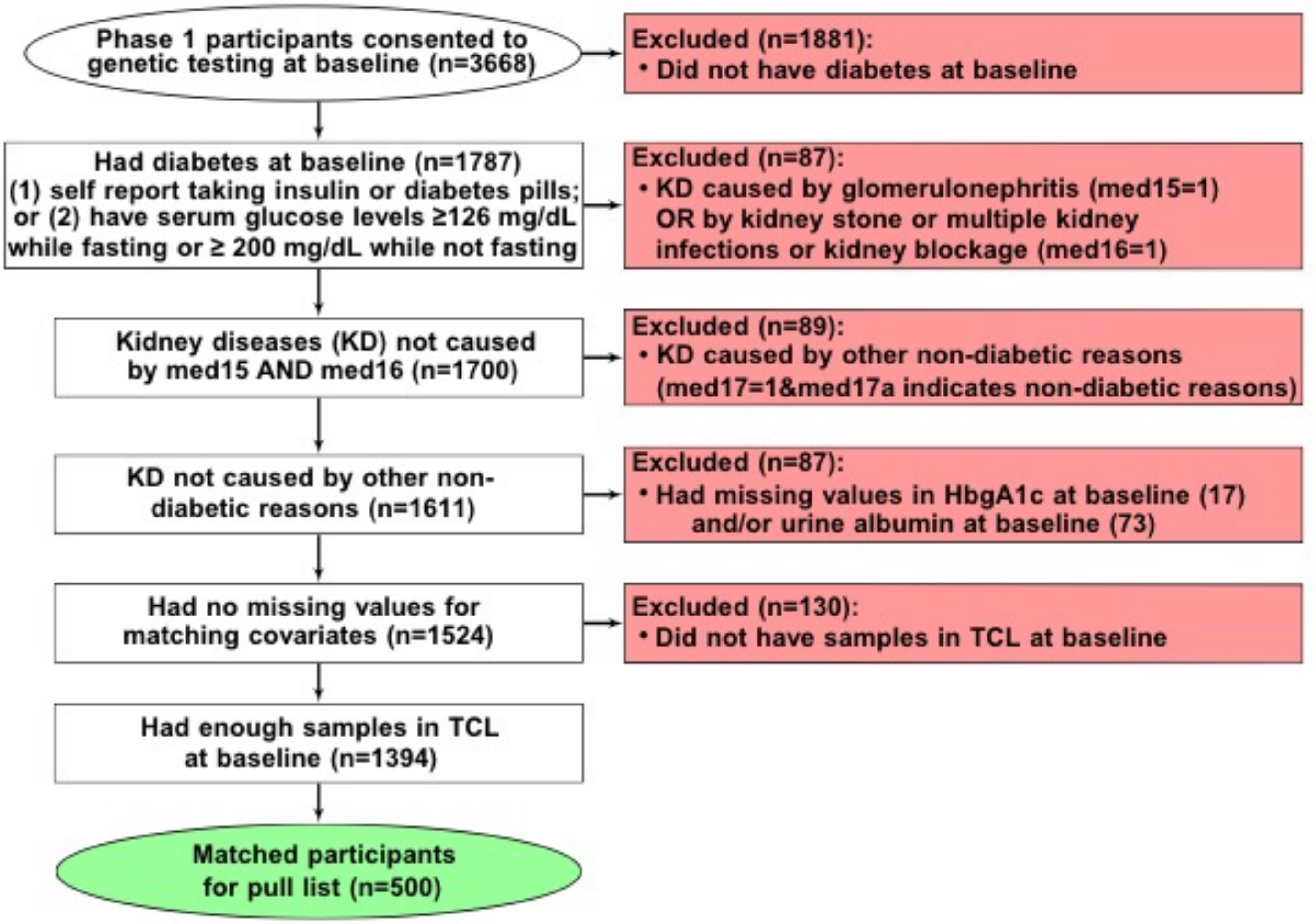
Phase 1 participants consented to genetic testing at baseline (n=3668)
Excluded (n=1881):
• Did not have diabetes at baseline
Had diabetes at baseline (n=1787)
(1) self report taking insulin or diabetes pills; or (2) have serum glucose levels ≥126 mg/dL while fasting or ≥ 200 mg/dL while not fasting
Excluded (n=87):
• KD caused by glomerulonephritis (med15=1) OR by kidney stone or multiple kidney infections or kidney blockage (med16=1)
Kidney diseases (KD) not caused by med15 AND med16 (n=1700)
Excluded (n=89):
• KD caused by other non-diabetic reasons (med17=1&med17a indicates non-diabetic reasons)
KD not caused by other non-diabetic reasons (n=1611)
Excluded (n=87):
• Had missing values in HbgA1c at baseline (17) and/or urine albumin at baseline (73)
Had no missing values for matching covariates (n=1524)
Excluded (n=130):
• Did not have samples in TCL at baseline
Had enough samples in TCL at baseline (n=1394)
Matched participants for pull list (n=500)

Fig. S1. The CRIC study subsampling flowchart (n=500, 250+250 progressive DKD, 1:1 stratified design, 250 strata). 473 samples retained after combining with good quality genetic data. TCL: Translational Core Laboratory, which is a central lab storing all CRIC samples.

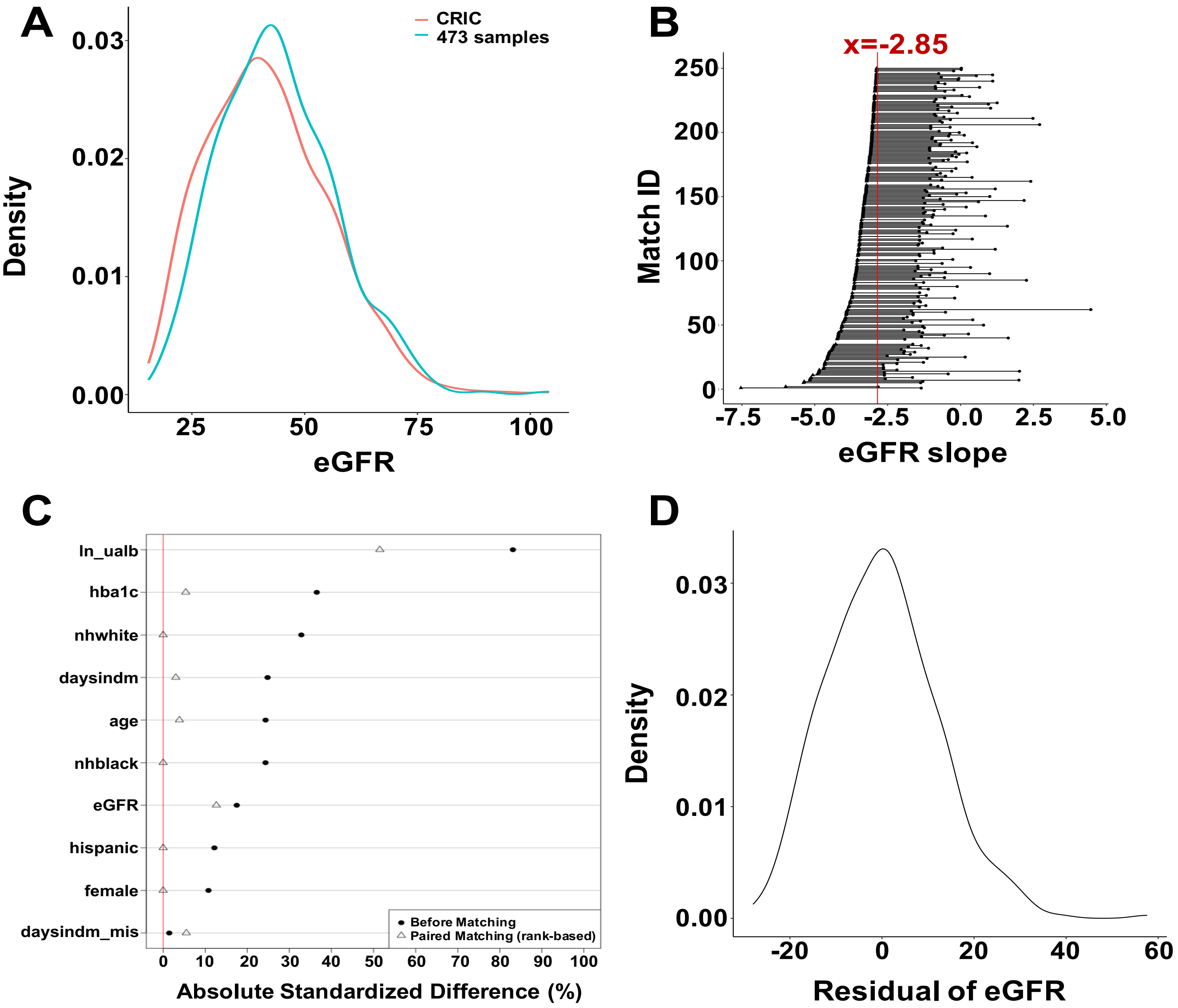

Fig. S2. Study population

A). The density plot of baseline eGFR in our dataset (n=473 samples with high quality genotype data). Red line: entire CRIC population, blue line: current study (473 samples)

B). Subjects with the BLUP adjusted eGFR slope < -2.85 were assigned to the ‘fast-progress group’, and subjects with the BLUP adjusted eGFR slope > -2.85 were assigned to the ‘slow-progress group’. 1:1 stratified design for kidney function decline (eGFR slope) (205 strata, n=410 [205×2] samples with high quality genotype data).

C). The dataset for eGFR slope was designed by 1:1 stratified design to control covariates, including age, hemoglobin A1c, baseline eGFR, logarithm of urine albumin, gender, race, and days with diabetes (self-reported).

D). The density plot of the residualized eGFR (n=473) (by adjusting for age, genetic background, and gender etc. as described in the methods).

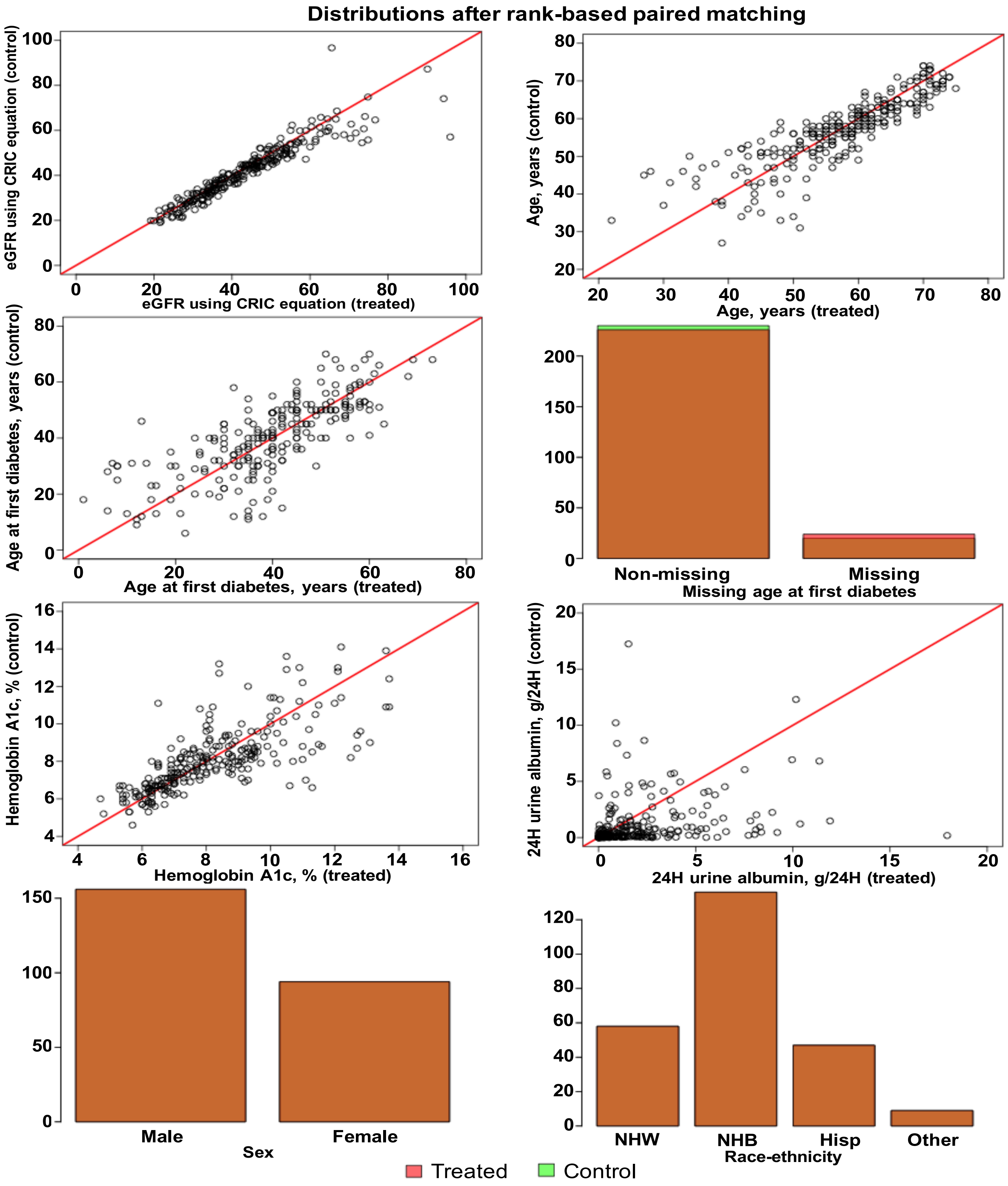

Fig. S3. The distribution of variables after stratification Comparisons of covariates between stratified pairs. The 1:1 stratified case-control design controlling for eGFR, age, sex, race, the age of diabetes diagnosis hemoglobin A1c, but not for 24 hours urine albuminuria.

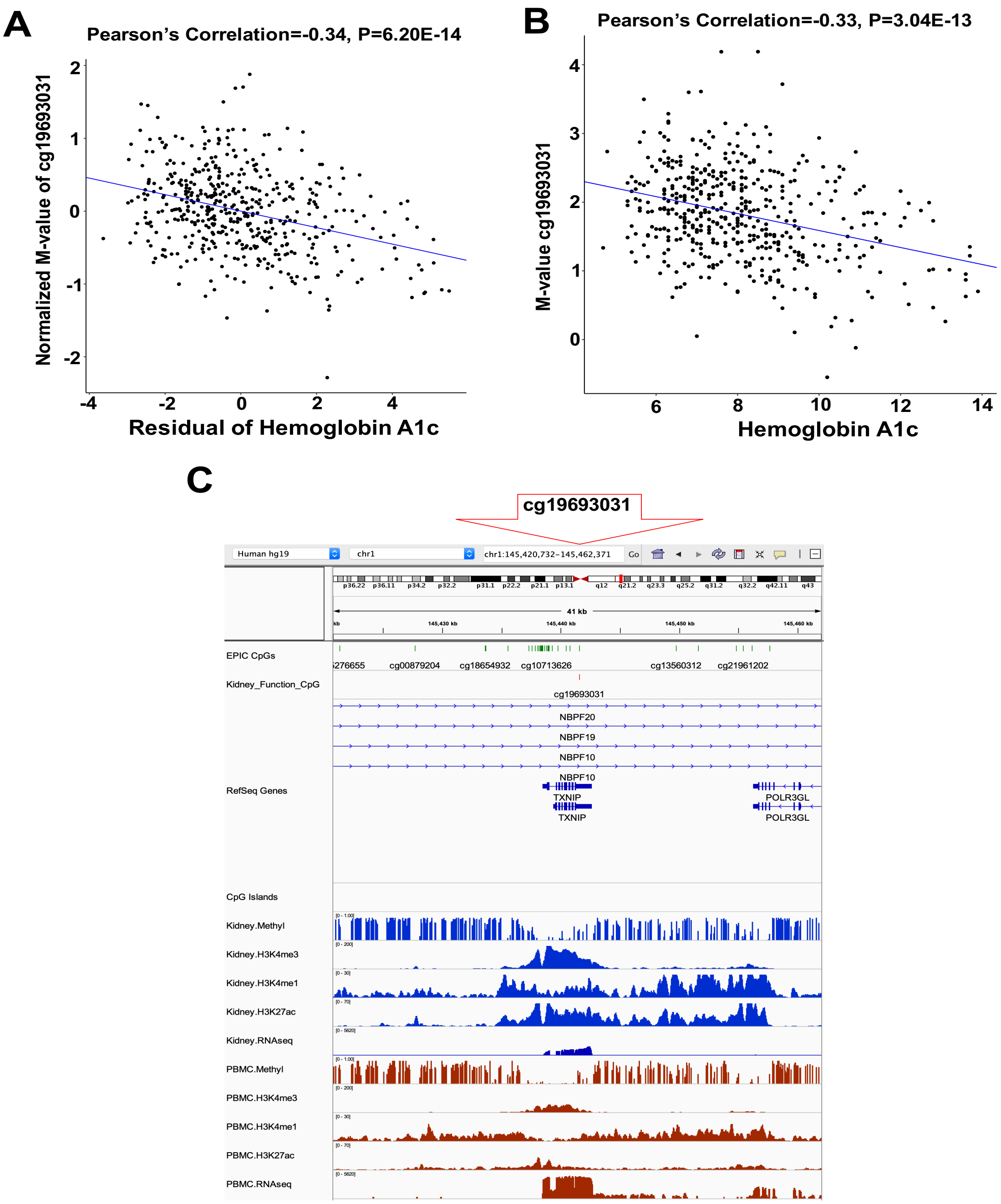

Fig. S4. Functional annotation of cg19693031

A). The association between methylation levels of cg19693031 (normalized M-value) and residualized hemoglobin A1c.

B). The association between methylation levels of cg19693031 (normalized M-value) and hemoglobin A1c.

C). Genomic annotation around the index probe cg19693031 in region chr1:228,688,040-228,689,772. Whole genome bisulfate sequencing (WGBS), histone modification marks and RNA-seq from healthy human kidneys and PBMC are shown. H3K4me1 marks poised enhancers, H3K27ac marks active enhancers, and H3K4me3 marks active promoters.

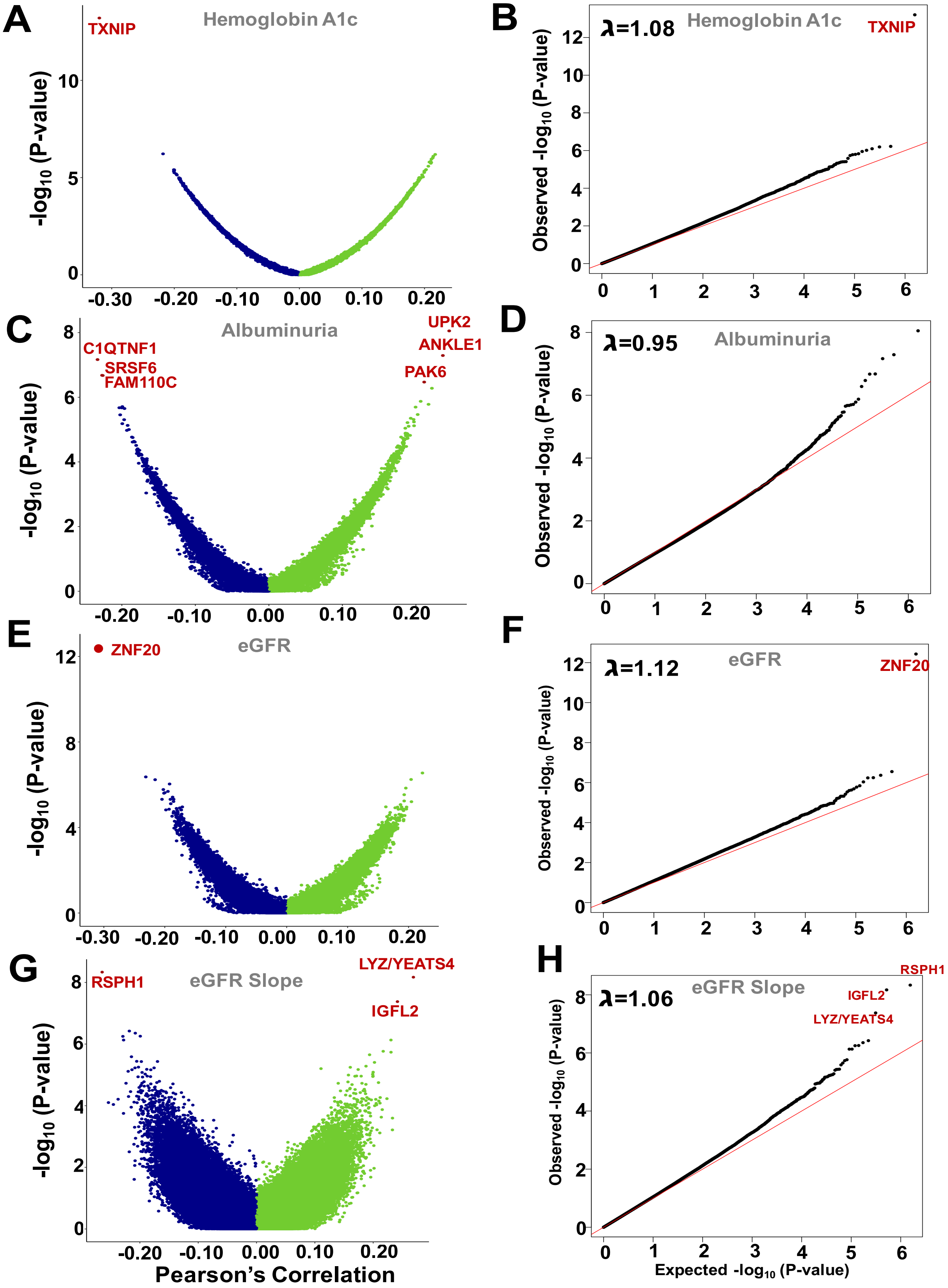

Fig. S5. Methylation changes associated with hemoglobin A1c, albuminuria, eGFR,and eGFR slope in whole blood samples of 473 CRIC study participants.

A). Volcano plot. The x-axis represents the Pearson's correlation between DNA methylation (M-value) and hemoglobin A1c. The y-axis is the negative base 10 log of the association P-value.

B). Quantile-quantile (QQ) plot. The observed P-value distribution versus the expected P-value distribution (hemoglobin A1c).

C). Volcano plot. The x-axis represents the Pearson’s correlation between DNA methylation (M-value) and albuminuria. The y-axis is the negative base 10 log of the association P-value.

D). Quantile-quantile (QQ) plot. The observed P-value distribution versus the expected P-value distribution (albuminuria).

E). Volcano plot. The x-axis is the Pearson’s correlation between DNA methylation of each CpG site and baseline eGFR. The y-axis is the negative base 10 log of the association P-value.

F). Quantile-quantile (QQ) plot. The observed P-value distribution versus the expected P-value distribution (baseline eGFR).

G). Volcano plot. The x-axis represents the Pearson’s correlation between DNA methylation (M-value) and kidney function decline (eGFR slope). The y-axis is the negative base 10 log of the association P-value.

H). Quantile-quantile (QQ) plot. The observed P-value distribution versus the expected P-value distribution (eGFR slope).

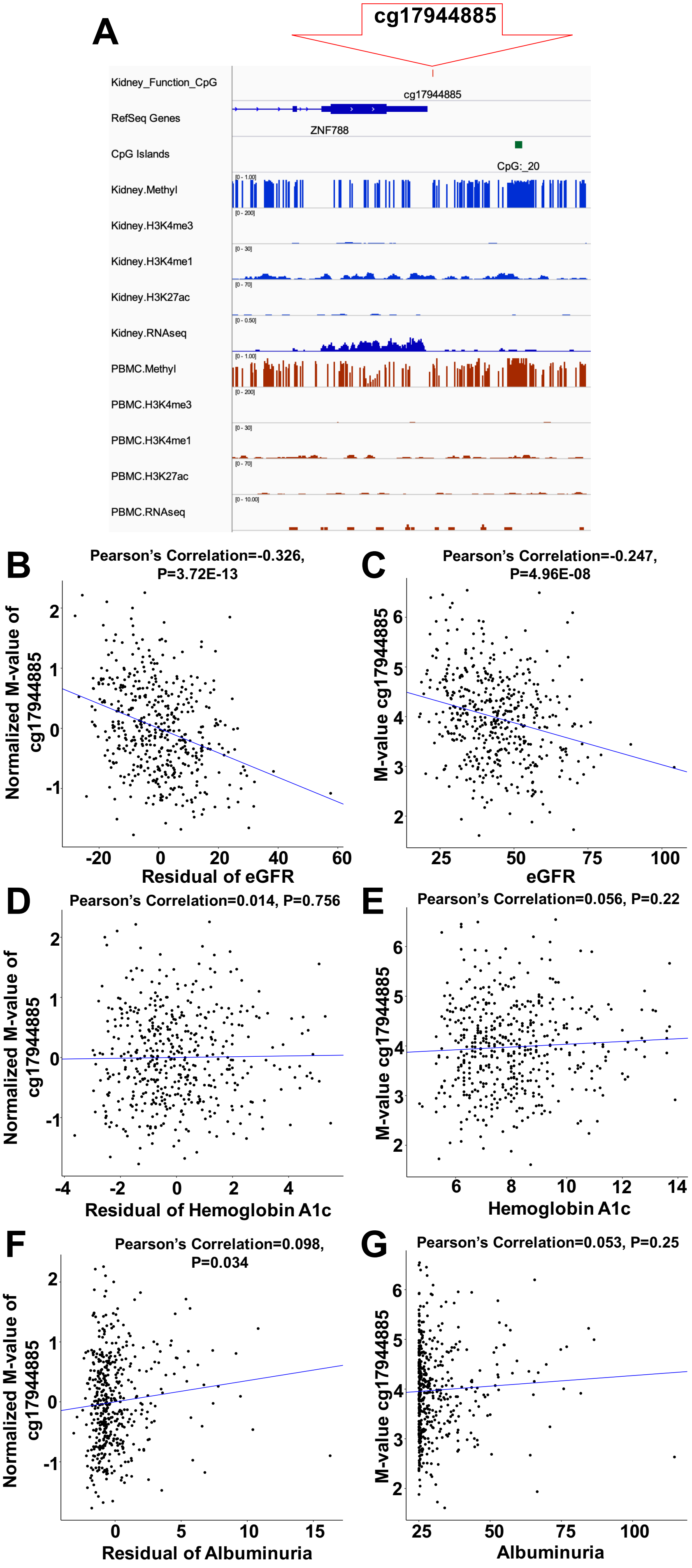
A
cg17944885
Kidney_Function_CpG
cg17944885
RefSeq Genes
ZNF788
CpG Islands
CpG:_20
Kidney.Methyl
[0 - 1.00]
Kidney.H3K4me3
[0 - 200]
Kidney.H3K4me1
[0 - 30]
Kidney.H3K27ac
[0 - 70]
Kidney.RNAseq
[0 - 0.50]
PBMC.Methyl
[0 - 1.00]
PBMC.H3K4me3
[0 - 200]
PBMC.H3K4me1
[0 - 30]
PBMC.H3K27ac
[0 - 70]
PBMC.RNAseq
[0 - 10.00]
B
Pearson's Correlation=-0.326, P=3.72E-13
Normalized M-value of cg17944885
Residual of eGFR
-20
0
20
40
60
-1
0
1
2
C
Pearson's Correlation=-0.247, P=4.96E-08
M-value cg17944885
eGFR
25
50
75
100
2
3
4
5
6
D
Pearson's Correlation=0.014, P=0.756
Normalized M-value of cg17944885
Residual of Hemoglobin A1c
-4
-2
0
2
4
-1
0
1
2
E
Pearson's Correlation=0.056, P=0.22
M-value cg17944885
Hemoglobin A1c
6
8
10
12
14
2
3
4
5
6
F
Pearson's Correlation=0.098, P=0.034
Normalized M-value of cg17944885
Residual of Albuminuria
0
5
10
15
-1
0
1
2
G
Pearson's Correlation=0.053, P=0.25
M-value cg17944885
Albuminuria
25
50
75
100
2
3
4
5
6

Fig. S6. Functional annotation of cg17944885

A). Genomic annotation around the index probe cg17944885 in region chr19:12,219,104-12,231,129. Whole genome bisulfate sequencing (WGBS), histone modification marks and RNAseq from healthy human kidneys and PBMC are shown. H3K4me1 marks poised enhancers, H3K27ac marks active enhancers, and H3K4me3 marks active promoters. Zoom-in version of Figure1D.

B). The association between methylation levels of cg17944885 (normalized M-value) and residualized eGFR.

C). The association between methylation levels of cg17944885 (normalized M-value) and eGFR.

D). The association between methylation levels of cg17944885 (normalized M-value) and residualized hemoglobin A1c.

E). The association between methylation levels of cg17944885 (normalized M-value) and hemoglobin A1c.

F). The association between methylation levels of cg17944885 (normalized M-value) and residualized albuminuria.

G). The association between methylation levels of cg17944885 (normalized M-value) and 24 hours urine albuminuria.

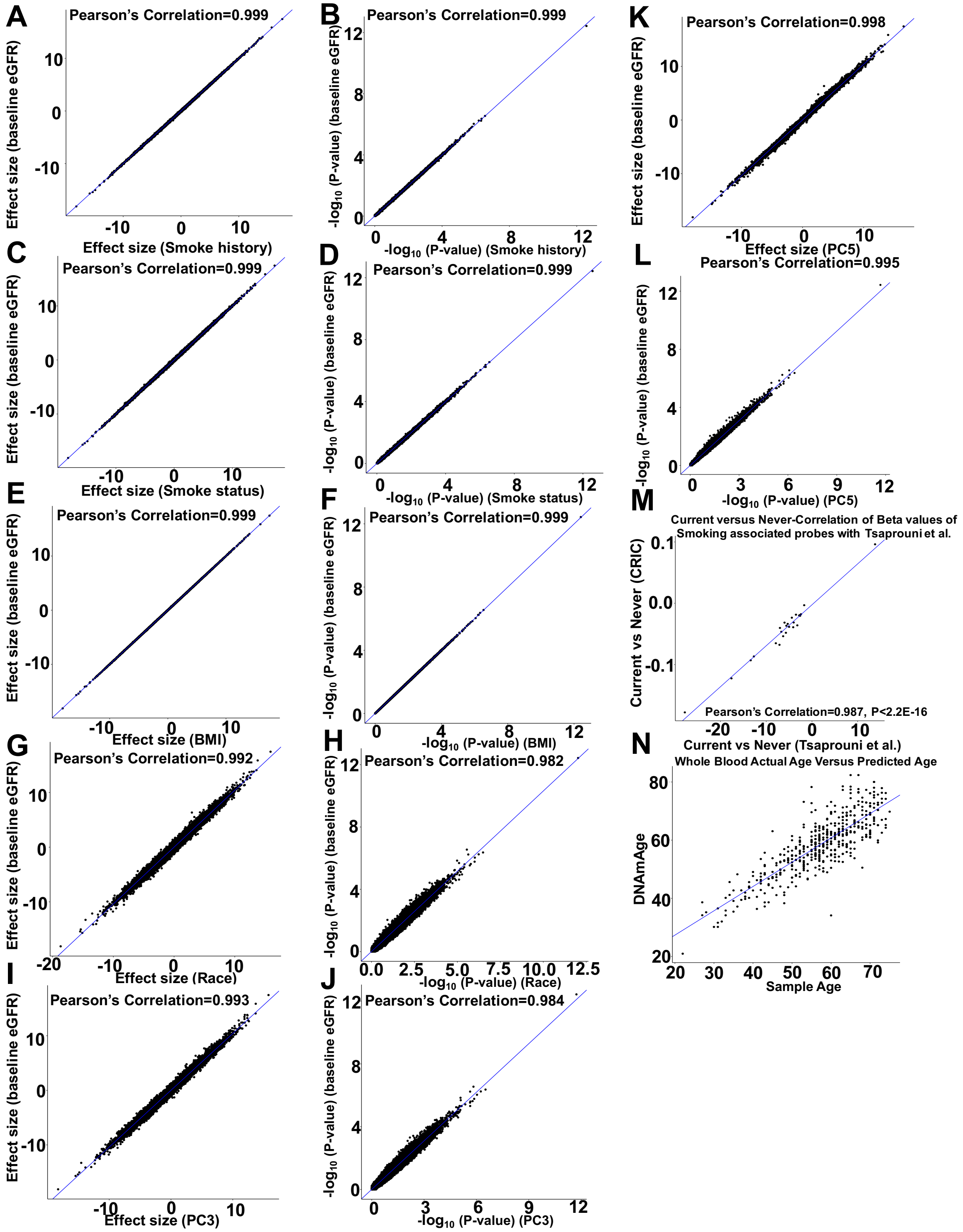

Fig. S7. Sensitivity analysis of eGFR MWAS

A). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included smoking history.

B). The Pearson's correlation coefficient of -log$_{10}$(P-value) of the initial eGFR linear regression model or model that included smoking history.

C). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included smoking status.

D). The Pearson's correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR linear regression model or model that included smoking status.

E). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included BMI.

F). The Pearson's correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR linear regression model or model that included BMI.

G). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included race.

H). The Pearson's correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR linear regression model or model that included race.

I). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included top 3 genetic PCs.

J). The Pearson's correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR linear regression model or model that included top 3 genetic PCs.

K). The Pearson's correlation coefficient of effect sizes of the initial eGFR linear regression model and the model that included top 5 genetic PCs.

L). The Pearson's correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR linear regression model or model that included top 5 genetic PCs.

M). Correlation between smoking related DNA methylation CpG sites in our dataset (Smokers versus non-smokers) and the previously published dataset by Tsapouroni et al.(21)

N). Correlation between actual baseline age of whole blood subjects in this study and estimated epigenetic age using a modification of the Horvath method(22).

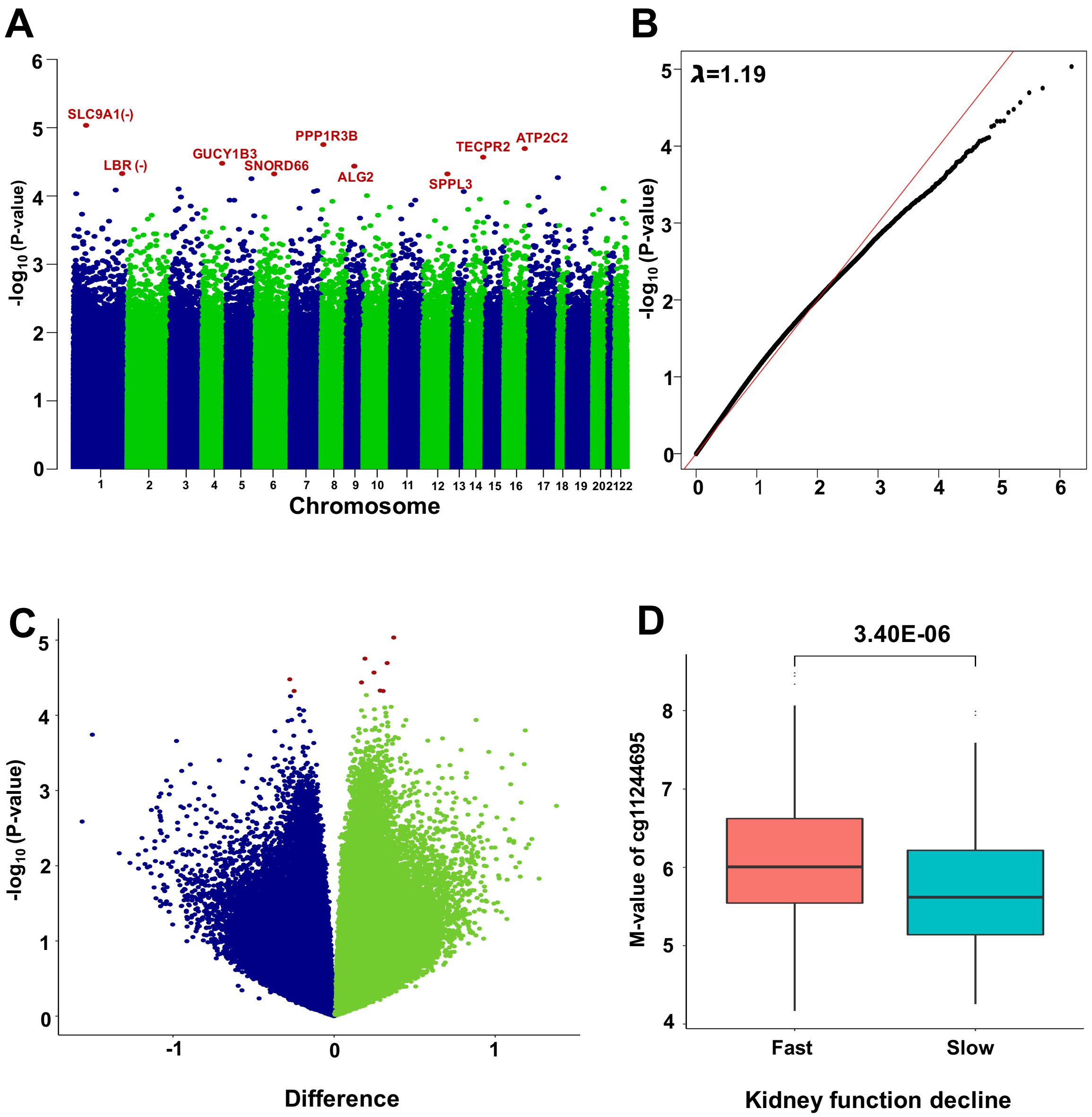

Fig. S8. Methylation changes associated with kidney function decline (eGFR slope) by conditional logistic regression model in whole blood among 410 participants (205 strata) of the CRIC study.

A). Manhattan plot. The y-axis is the negative base 10 log of the association P-value. The x-axis represents the genomic location of probes. 9 CpG sites, which passed the pre-specified discover P-value threshold of 5E-05 were highlighted in dark red color.

B). Quantile-quantile (QQ) plot. The observed P-value distribution versus the expected P-value distribution.

C). Volcano plot. The x-axis is the Pearson's correlation between DNA methylation of each CpG site and baseline eGFR. The y-axis is the negative base 10 log of the association P-value.

D). Mean methylation levels of cg11244695 was significantly different between fast and slow progressors (two-tailed t-test P-value=3.40E-06). Center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend to the 5th and 95th percentiles; outliers are represented by dots.

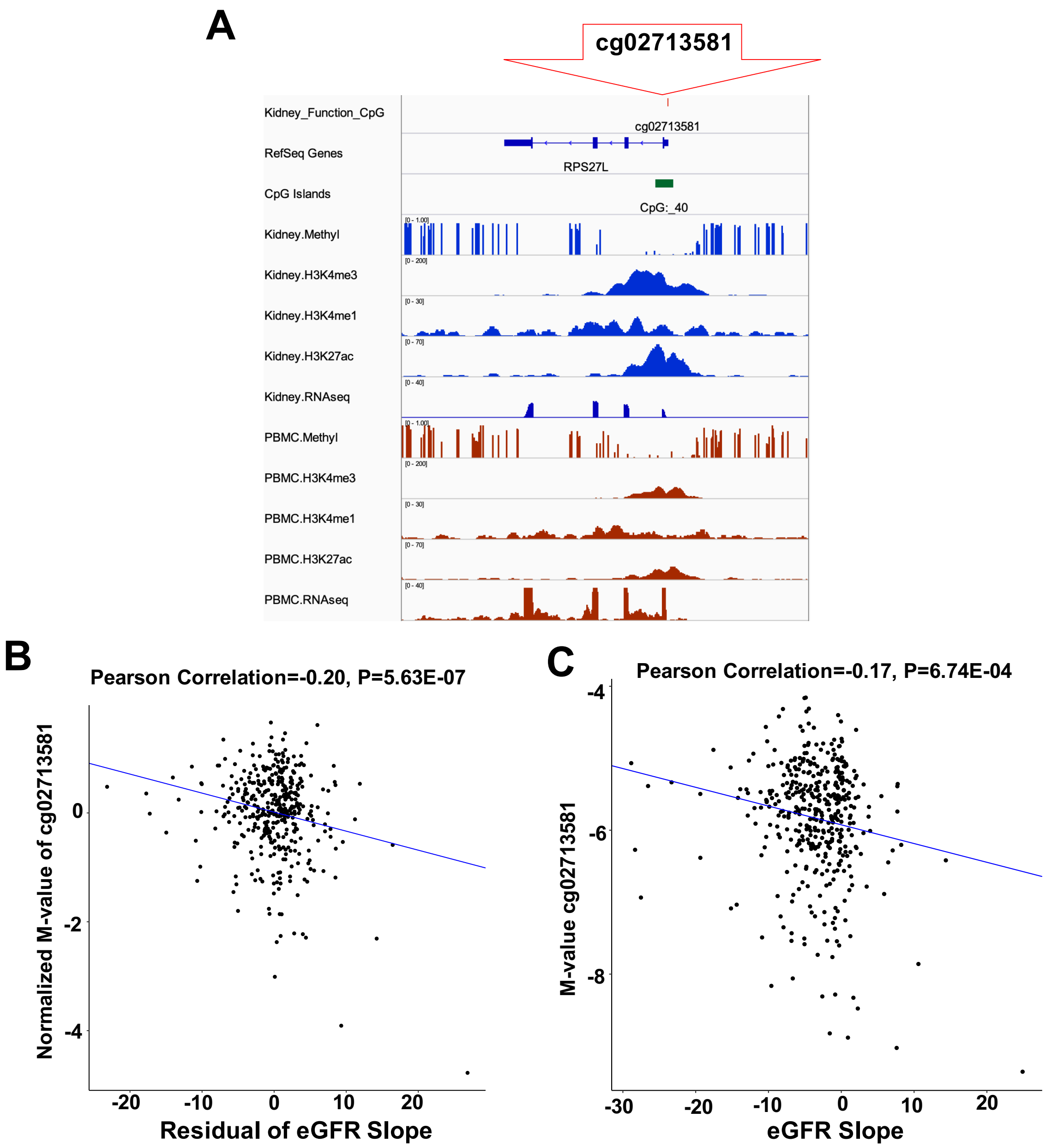

Fig. S9. Functional annotation of the cg02713581

A). Genomic annotation around the index probe cg02713581 in region chr15:63,442,898-63,453,470. Whole genome bisulfate sequencing (WGBS), histone modification marks and RNA-seq from healthy human kidneys and PBMC are shown. H3K4me1 marks poised enhancers, H3K27ac marks active enhancers, and H3K4me3 marks active promoters.

B). The association between methylation levels of cg02713581 (normalized M-value) and residualized eGFR slope.

C). The association between methylation levels of cg02713581 (normalized M-value) and eGFR slope.

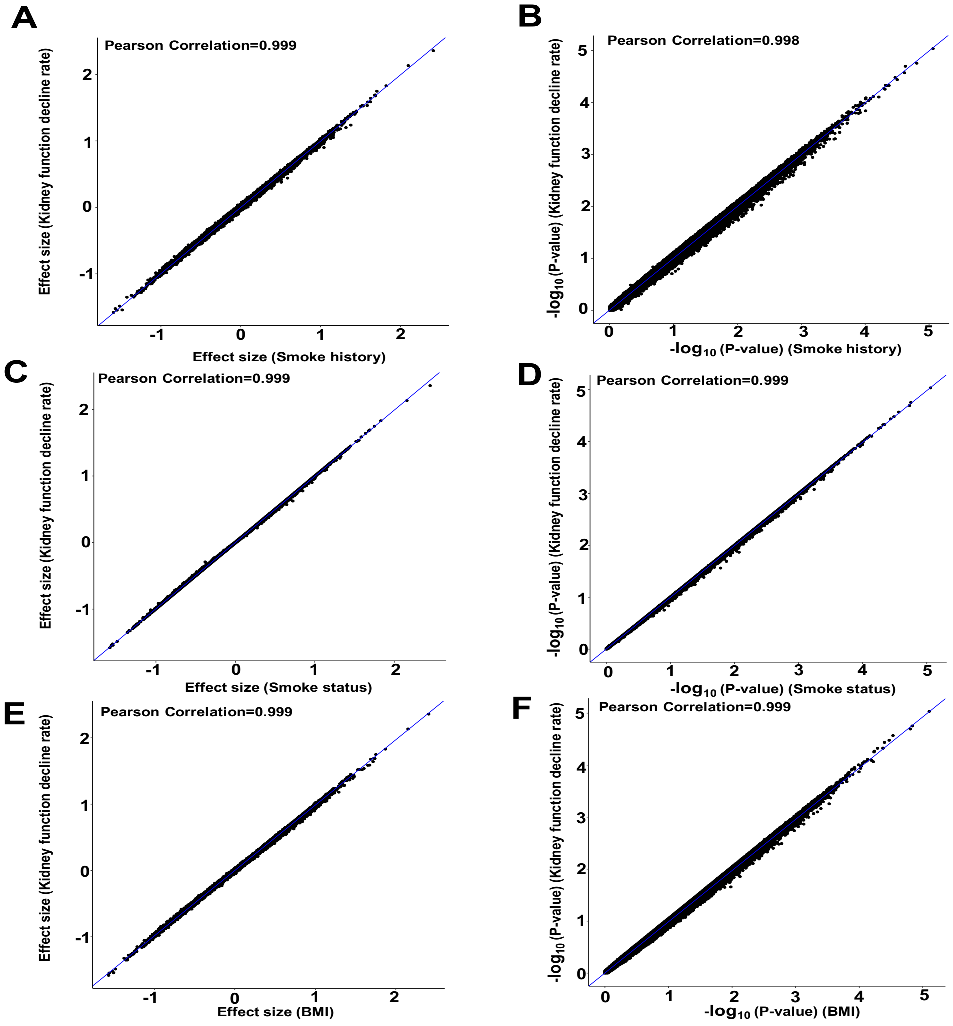

Fig. S10. Sensitivity analysis of kidney function decline MWAS (conditional logisitic regression)

A). The Pearson's correlation coefficient of effect sizes of the initial conditional logistic regression model and the model that included smoking history.

B). The Pearson's correlation coefficient of -log$_{10}$(P-value) of the initial conditional logistic regression model and the model that included smoking history.

C). The Pearson's correlation coefficient of effect sizes of the initial conditional logistic regression model and the model that included smoking status.

D). The Pearson's correlation coefficient of -log$_{10}$(P-value) of the initial conditional logistic regression model and the model that included smoking status.

E). The Pearson's correlation coefficient of effect sizes of the initial conditional logistic regression model and the model that included BMI.

F). The Pearson's correlation coefficient of -log$_{10}$(P-value) value of the initial conditional logistic regression model and the model that included BMI.

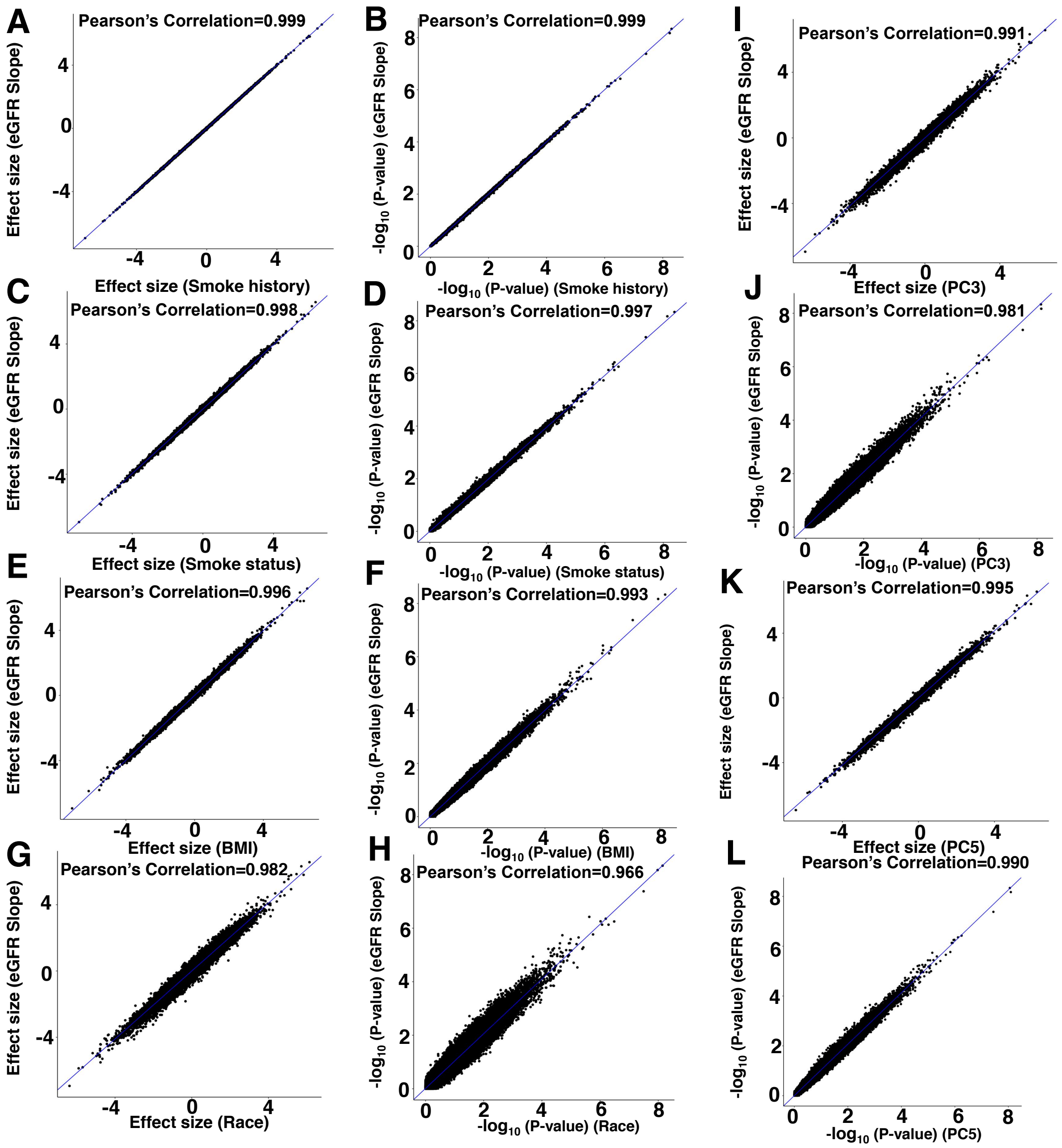

Fig. S11. Sensitivity analysis of eGFR slope MWAS (linear regresion)

A). The Pearson's correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included smoking history.

B). The Pearson's correlation coefficient of $-\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included smoking history.

C). The Pearson's correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included smoking status.

D). The Pearson's correlation coefficient of $-\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included smoking status.

E). The Pearson's correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included BMI.

F). The Pearson’s correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included BMI.

G). The Pearson’s correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included race.

H). The Pearson’s correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included race.

I). The Pearson’s correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included top 3 genetic PCs.

J). The Pearson’s correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included top 3 genetic PCs.

K). The Pearson’s correlation coefficient of effect sizes of the initial eGFR slope linear regression model and the model that included top 5 genetic PCs.

L). The Pearson’s correlation coefficient of -$\log_{10}$(P-value) of the initial eGFR slope linear regression model or model that included top 5 genetic PCs.

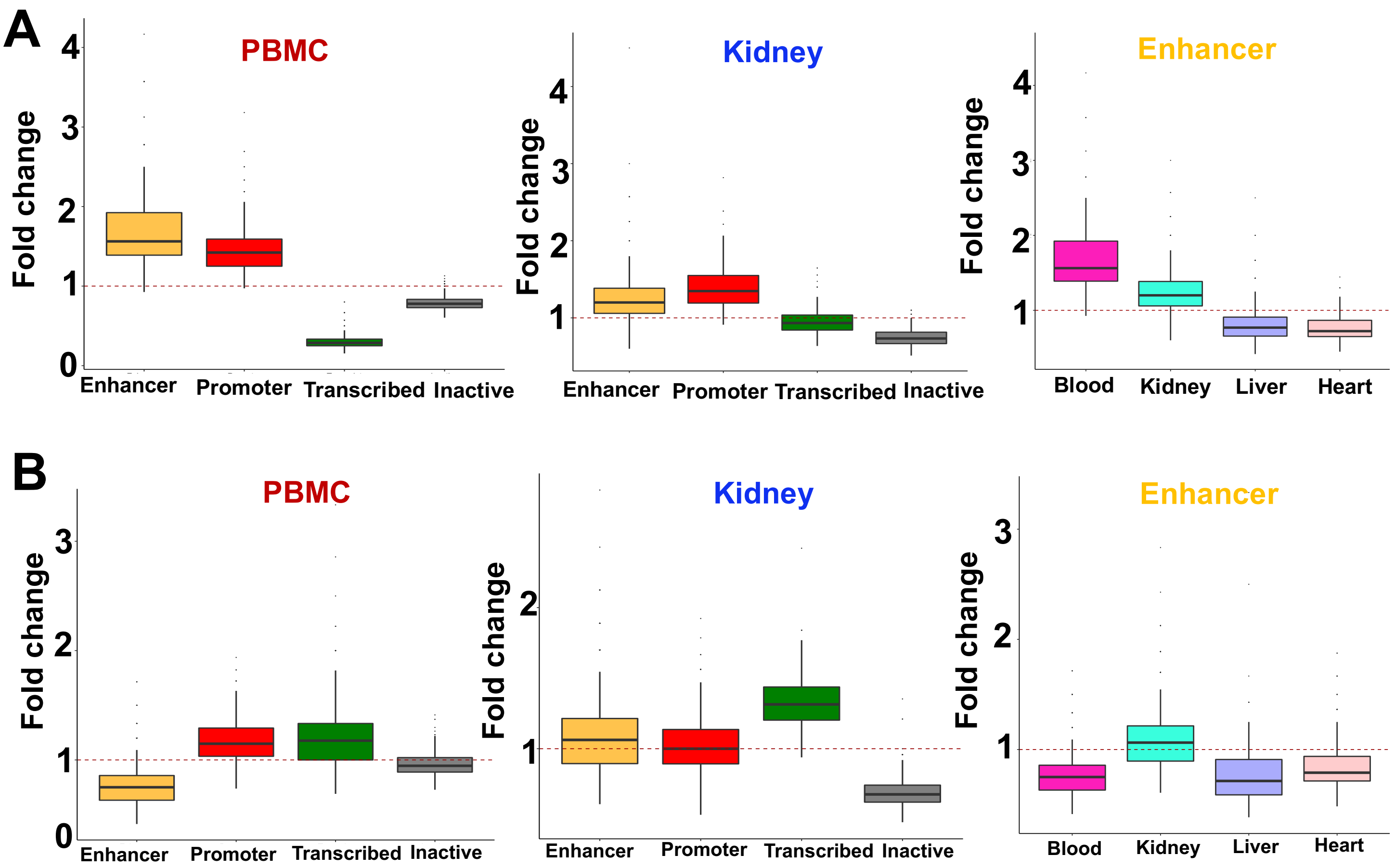

Fig. S12. Functional annotation of DKD phenotypes associated loci

A). eGFR-associated DMPs were enriched in enhancer and promoter regions in PBMC and kidney (the left and middle panels, respectively). The y-axis represents the fold change, which was calculated by the frequency ratio of DMPs and background probes that fell into each chromatin states. While comparing across different organs, these DMPs were enriched in enhancer regions of blood and kidney (the right panel). (Center line, median fold change; box limits, upper and lower quartiles; whiskers, 1.5×interquartile range)

B). Functional enrichment analysis of the 111 eGFR slope-associated DMPs. The y-axis represents the fold change, which was calculated by the frequency ratio of DMPs and background probes (with similar variance in methylation levels of DMPs) that fell into each chromatin states. While comparing across different organs, these DMPs were enriched in promoter and transcribed regions in PBMCs and enhancers and transcribed regions in kidney (the left and middle panels, respectively). These DMPs were enriched in kidney specific enhancers (the right panel) (Center line, median fold change; box limits, upper and lower quartiles; whiskers, 1.5×interquartile range).

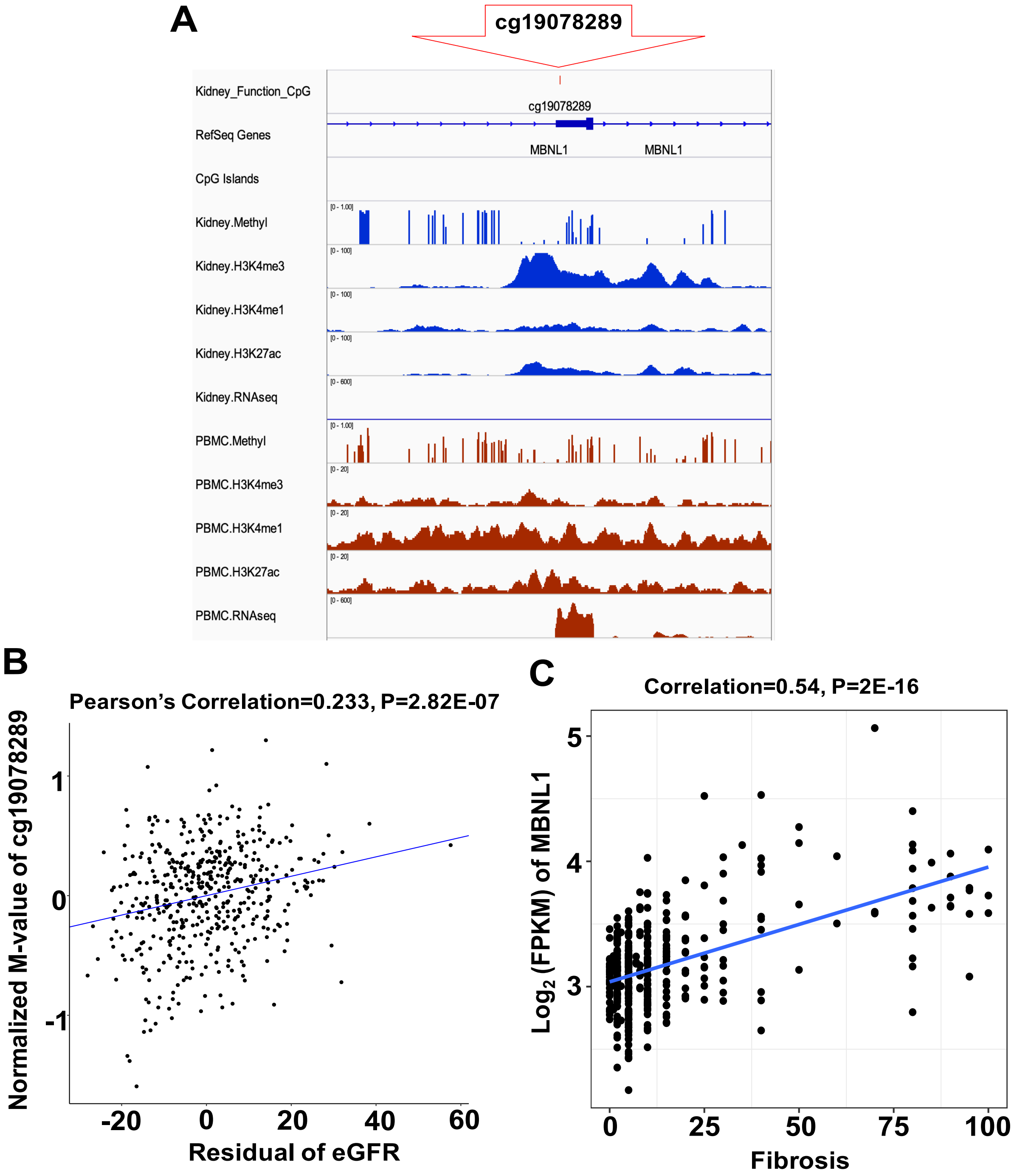

Fig. S13. Functional annotation of cg19078289 and its nearby gene MBNL1

A). Genomic annotation around the index probe cg19078289 in region chr3:152,011,294-152,022,869. Whole genome bisulfate sequencing (WGBS), histone modification marks and RNA-seq from healthy human kidney and PBMC were illustrated here. H3K4me1 marks poised enhancers, H3K27ac marks active enhancers, and H3K4me3 marks active promoters.

B). The association between methylation levels of cg19078289 (normalized M-value) and residualized eGFR (MWAS).

C). Correlation between MBNL1 expression in 433 microdissected human kidney samples and fibrosis

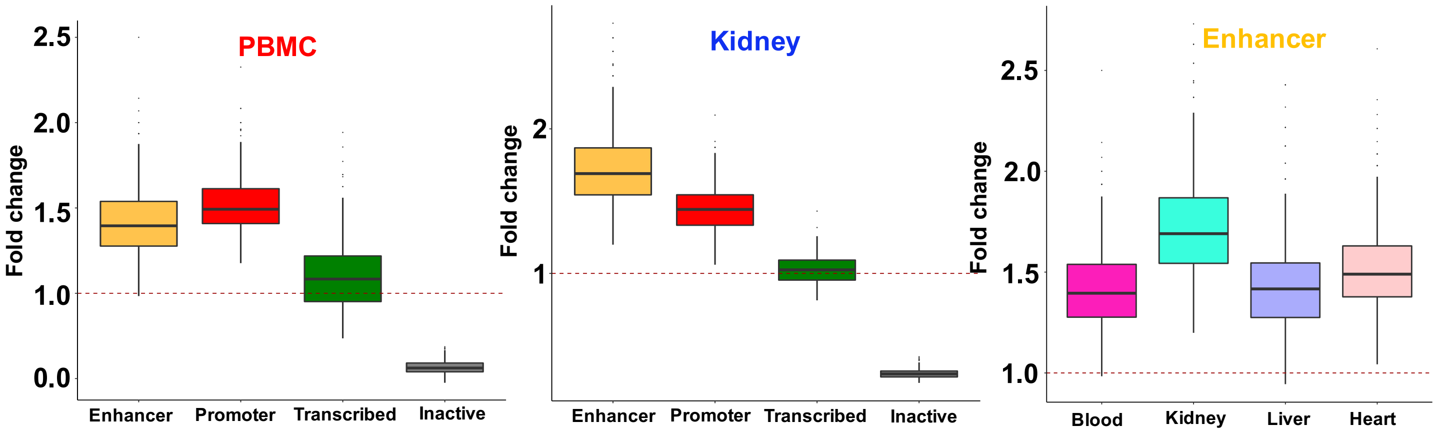

Fig. S14. Functional enrichment analysis of the 267 CpG sites identified by the moloc analysis The y-axis represents the fold change, which was calculated by the frequency ratio of the 267 identified CpGs and background probes (with similar variance in methylation levels of the 267 CpGs) that fell into each chromatin states. These DMPs were enriched in promoter and transcribed regions in PBMCs and enhancers and transcribed regions in kidney (the left and middle panels, respectively). While comparing across different organs, these DMPs were enriched in kidney specific enhancers (the right panel) (Center line, median fold change; box limits, upper and lower quartiles; whiskers, 1.5×interquartile range).

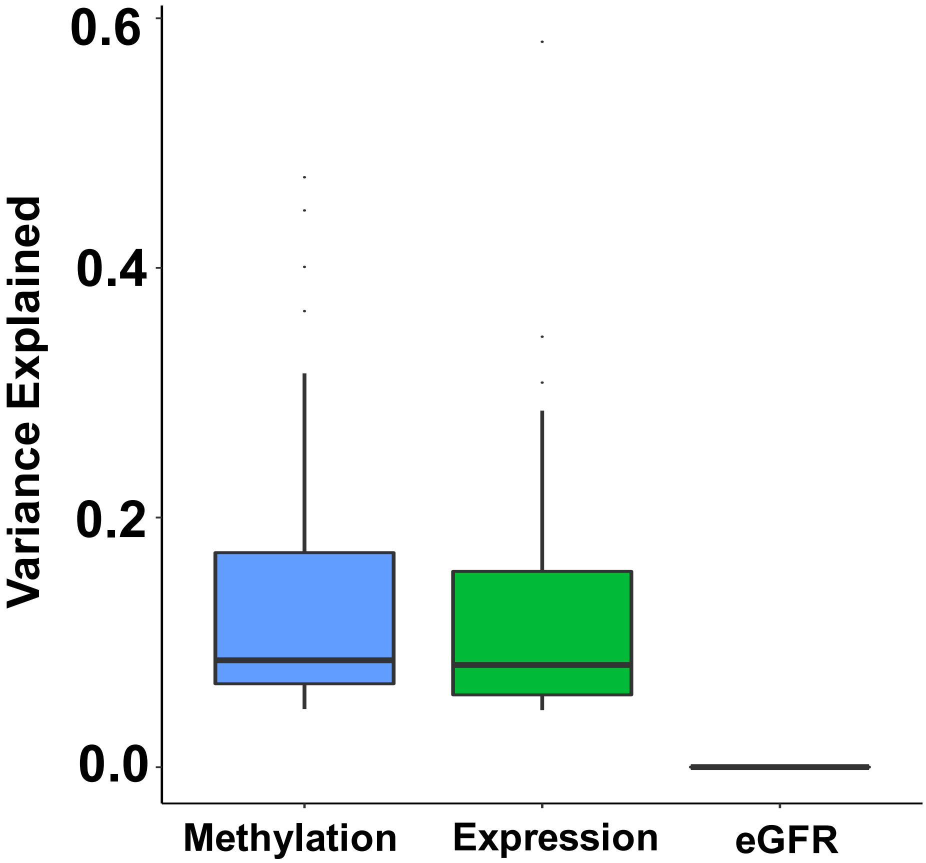

Fig. S15. The attenuation of effect sizes of genetic variants on methylation and gene expression towards kidney function (eGFR) Distribution of the variance explained in methylation, gene expression and kidney function by the top mQTLs for the 102 mCpGs that are significantly associated with the expression of 40 high confidence kidney risk causal genes and kidney function (eGFR).

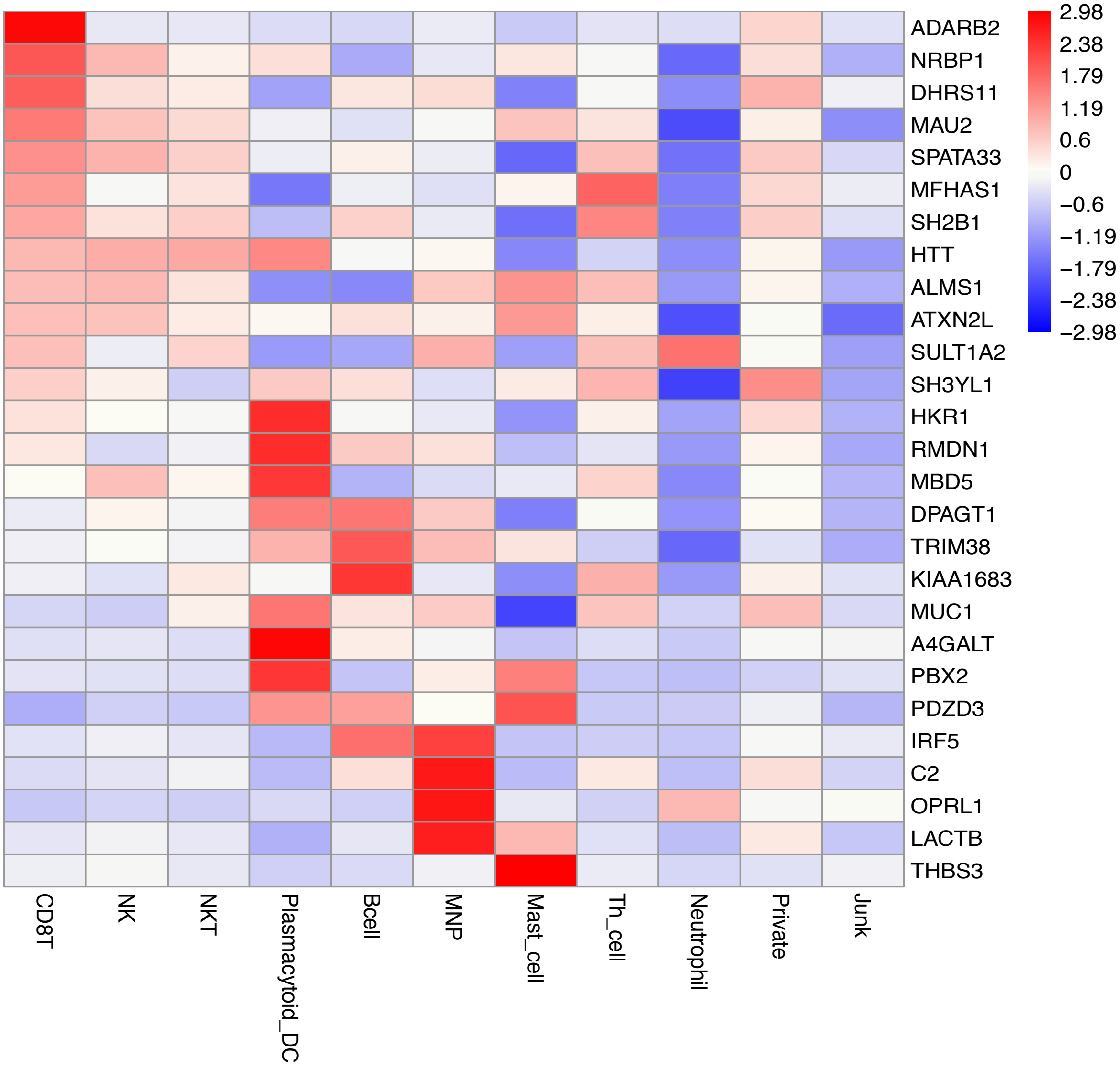

Fig. S16. 27 high fidelity protein-coding genes expression in the immune cells of healthy human kidney single cell dataset (27)**.** Mean expression values of the genes were calculated in each cluster. The color scheme is based on Z-score (calculated from the normalized gene expression levels) of 27 high fidelity protein-coding genes (excluding the genes located in MHC region) in the immune cells from human kidney single cell RNA-seq dataset. Z-score show the relatively expression levels compare to other cell types.

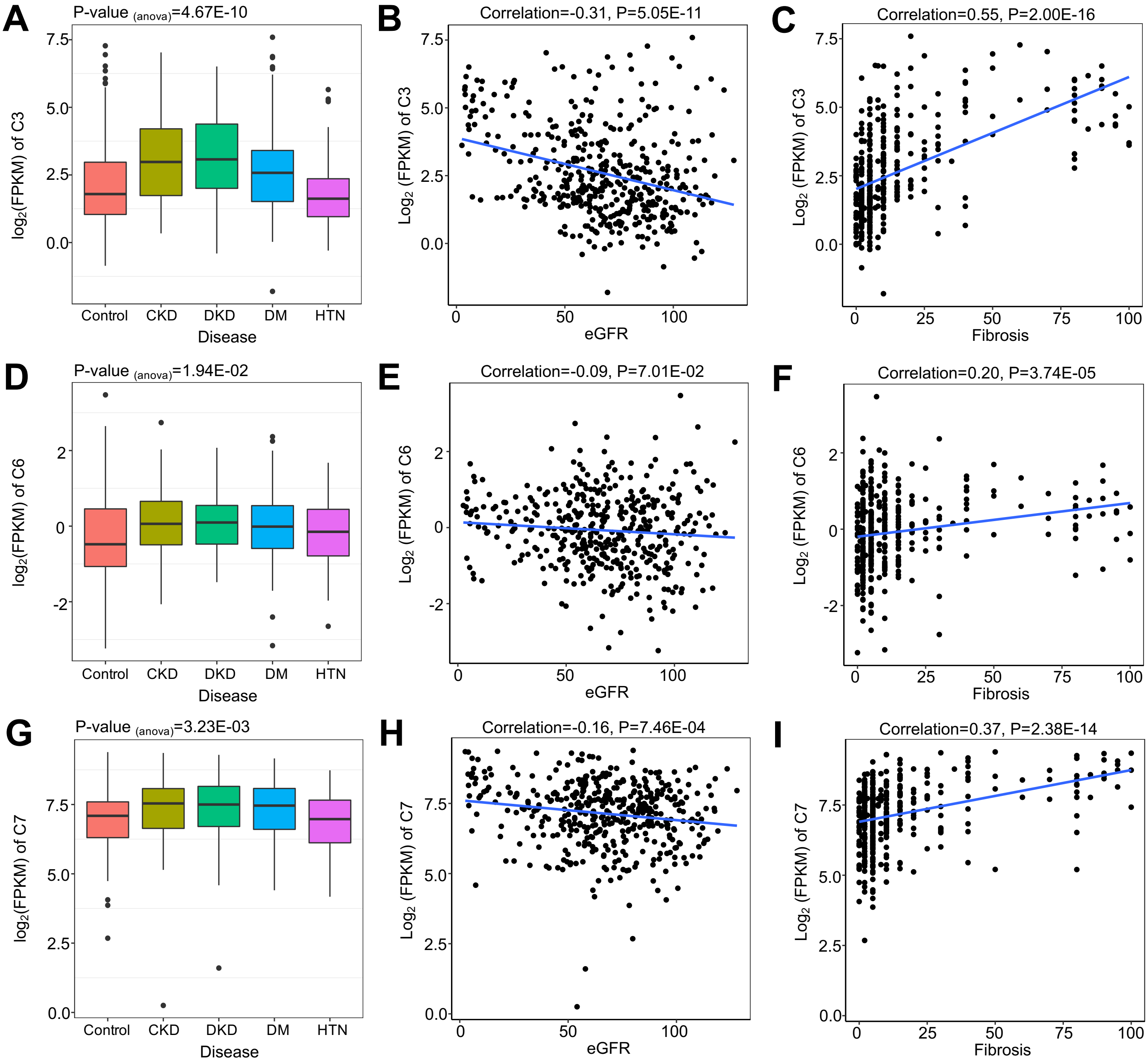

Fig. S17. The role of complement pathway in affecting kidney function.

A). C3 expression across different disease groups of samples

B). Correlation between C3 expression in 433 microdissected human kidney samples and eGFR

C). Correlation between C3 expression in 433 microdissected human kidney samples and fibrosis

D). C6 expression across different disease groups of samples

E). Correlation between C6 expression in 433 microdissected human kidney samples and eGFR

F). Correlation between C6 expression in 433 microdissected human kidney samples and fibrosis

G). C7 expression across different disease groups of samples

H). Correlation between C7 expression in 433 microdissected human kidney samples and eGFR

I). Correlation between C7 expression in 433 microdissected human kidney samples and fibrosis

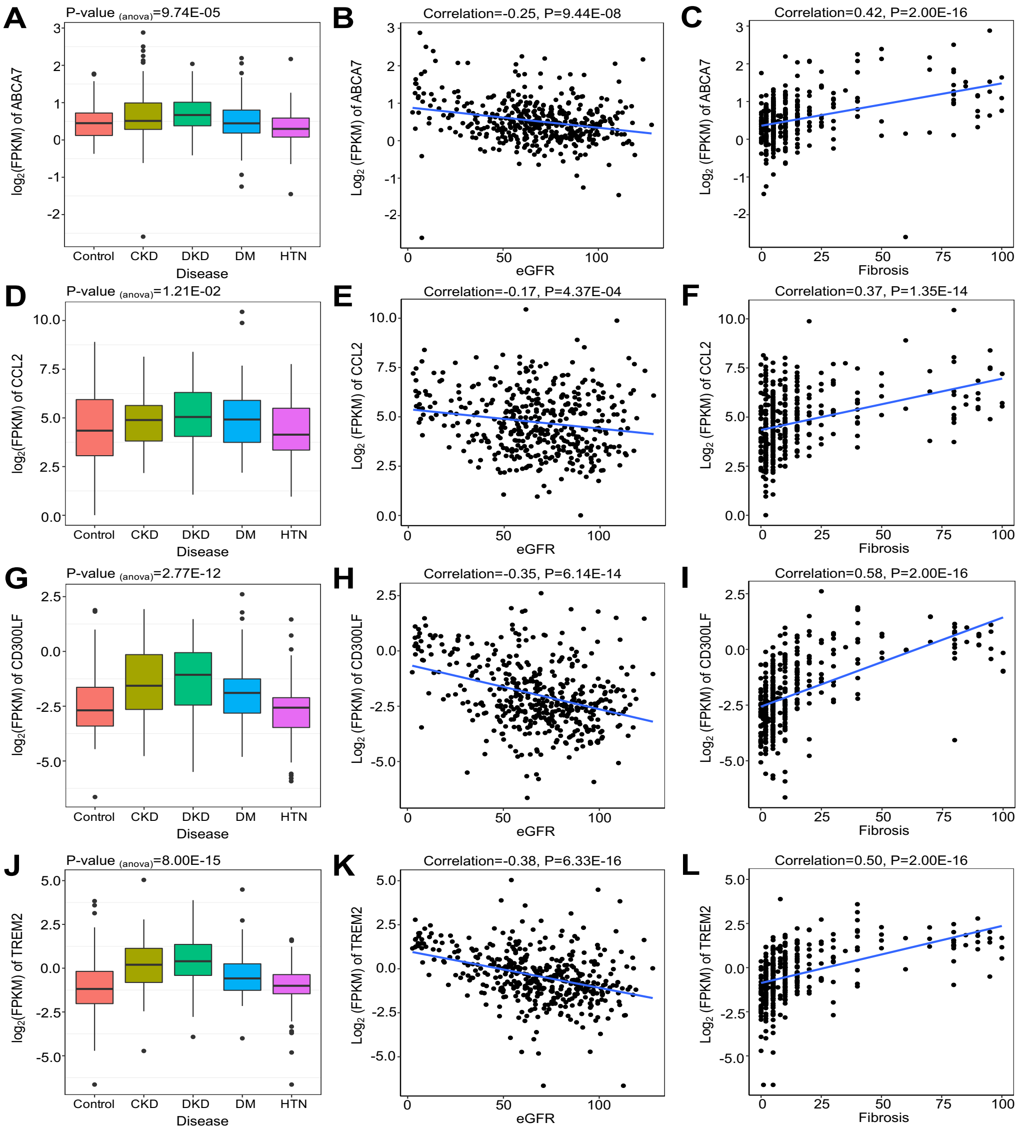

Fig. S18. The role of the positive regulation of apoptotic cell clearance pathway in affecting kidney function.

A). ABCA7 expression across different disease groups of samples

B). Correlation between ABCA7 expression in 433 microdissected human kidney samples and eGFR

C). Correlation between ABCA7 expression in 433 microdissected human kidney samples and fibrosis

D). CCL2 expression across different disease groups of samples

E). Correlation between CCL2 expression in 433 microdissected human kidney samples and eGFR

F). Correlation between CCL2 expression in 433 microdissected human kidney samples and fibrosis

G). CD300LF expression across different disease groups of samples

H). Correlation between CD300LF expression in 433 microdissected human kidney samples and eGFR

I). Correlation between CD300LF expression in 433 microdissected human kidney samples and fibrosis

J). TREM2 expression across different disease groups of samples

K). Correlation between TREM2 expression in 433 microdissected human kidney samples and eGFR

L). Correlation between TREM2 expression in 433 microdissected human kidney samples and fibrosis

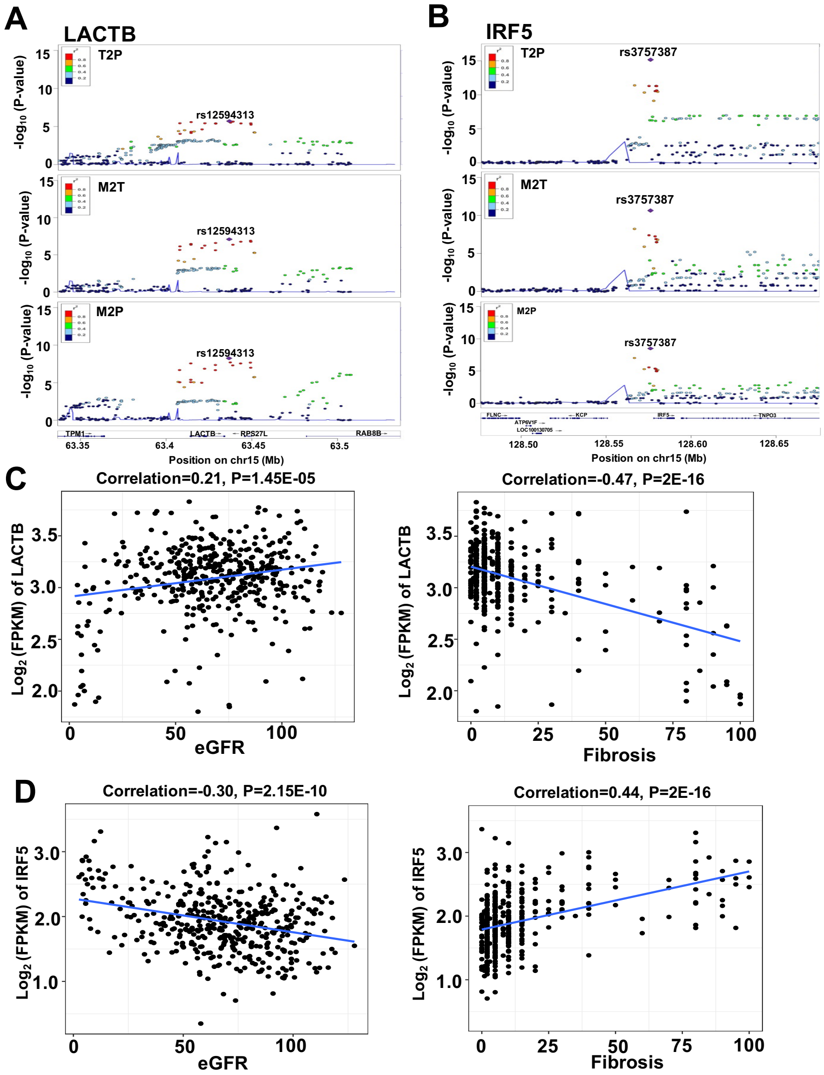

Fig. S19. The pleiotropic associations of LACTB and IRF5

A). The pleiotropic associations of gene expression of LACTB and kidney function (eGFR) (T2P; top panel), methylation of cg02713581 and gene expression of LACTB (M2T; middle panel), and methylation of cg02713581 and kidney function (eGFR) (M2P; bottom panel) by using variants locate within ±100 kb of rs12594313 as instrumental variables individually.

B). The pleiotropic associations of gene expression of IRF5 and kidney function (eGFR) (T2P; top panel), methylation of cg04864179 and gene expression of IRF5 (M2T; middle panel), and methylation of cg04864179 and kidney function (eGFR) (M2P; bottom panel) by using variants locate within ±100 kb of rs3757387 as instrumental variables individually.

C). Correlation between LACTB expression in 433 microdissected human kidney samples and eGFR (left) and fibrosis (right)

D). Correlation between IRF5 expression in 433 microdissected human kidney samples and eGFR (left) and fibrosis (right)

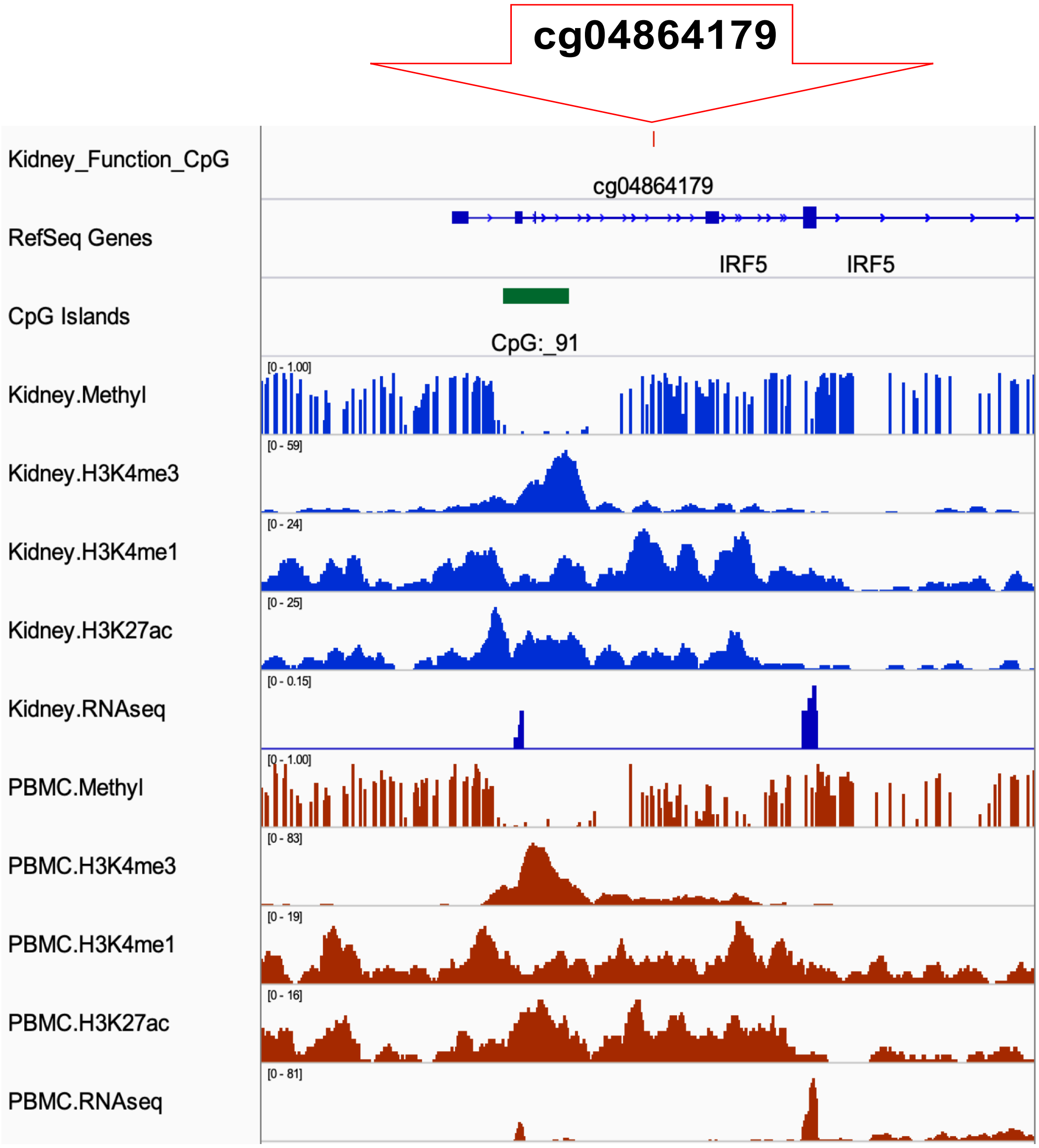

Fig. S20. **Functional annotation of cg04864179** Genomic annotation around the index probe cg04864179 in region chr7:128,574,317-128,585,610. Whole genome bisulfate sequencing (WGBS), histone modification marks and RNAseq from healthy human kidneys and PBMC are shown. H3K4me1 marks poised enhancers, H3K27ac marks active enhancers, and H3K4me3 marks active promoters.

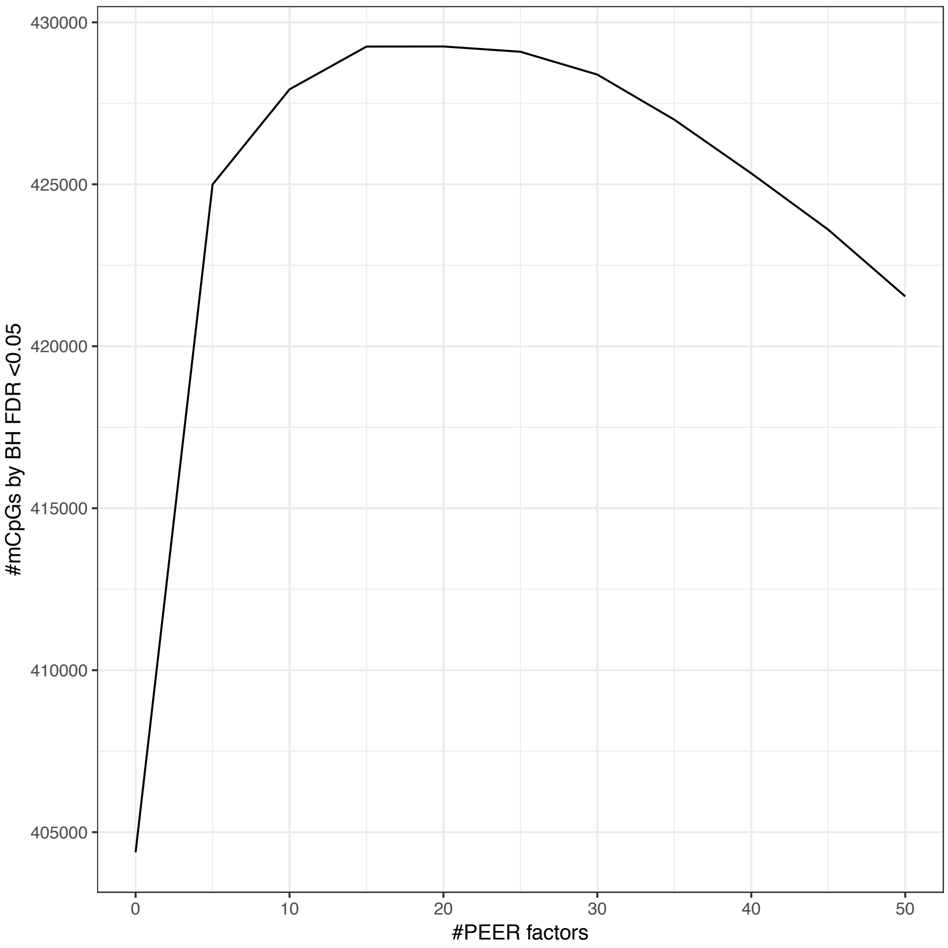

Fig. S21. PEER factor optimization The number of significant mQTLs (mCpGs) depends on the number of included PEER factors**.** The mCpGs identified were identified by Benjamini-Hochberg FDR<0.05 to ensure the rapid optimization.

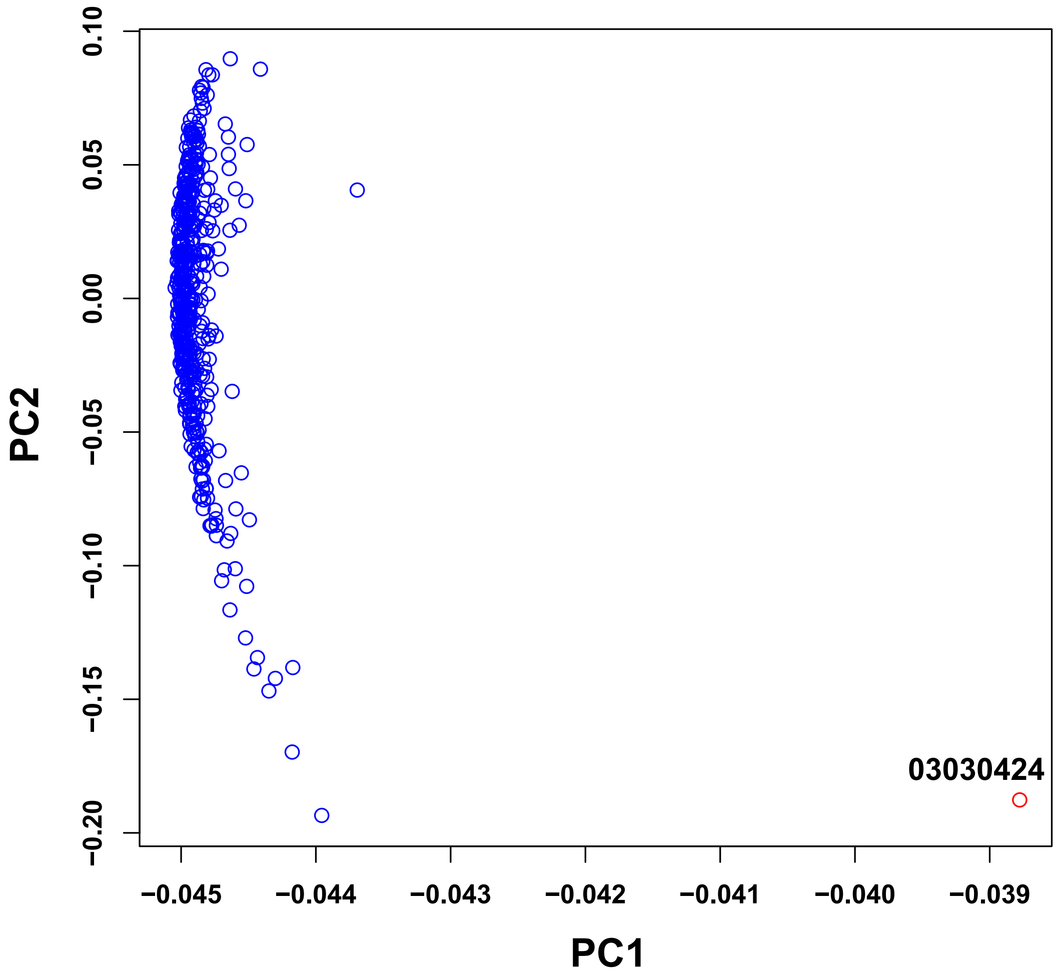

Fig. S22. Methylation data distribution Principal Component Analysis of the methylation data (beta values) of the cohort (n=473) after combining with genotype data.

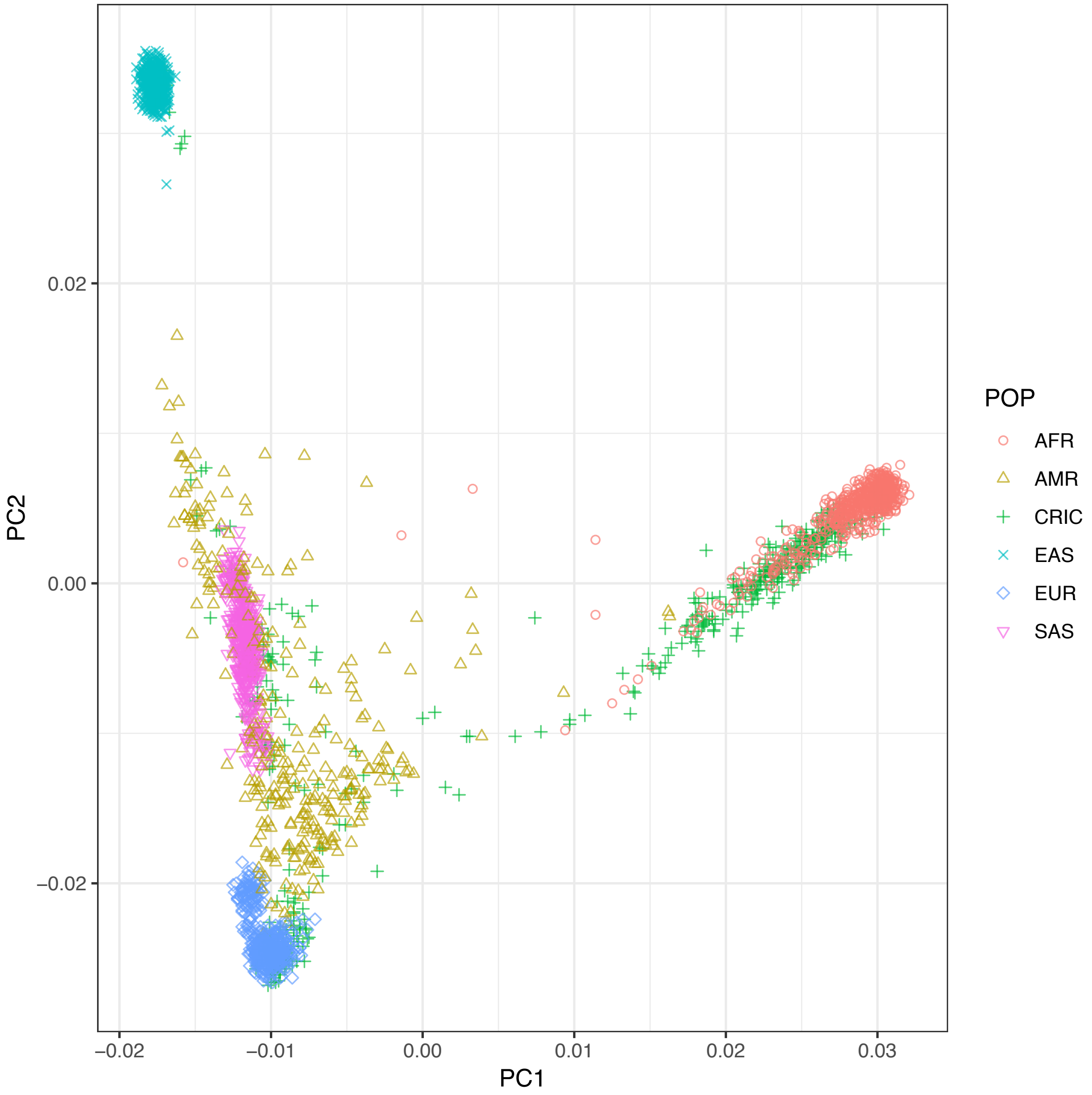

Fig. S23. Genotype data distribution Principal component analysis of genotype data of 473 individuals (samples with high quality genotype data available) and 1000 Genomes Project Phase 3 reference genome data. Green cross: CRIC subjects.

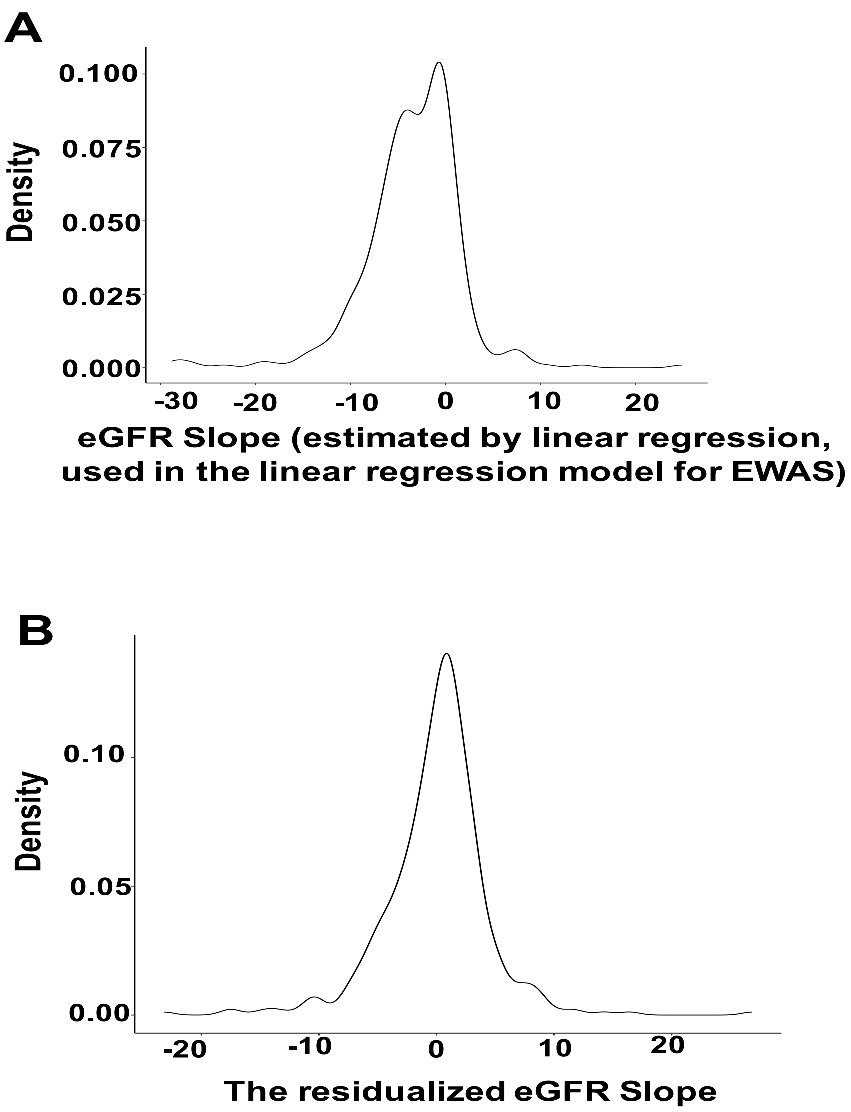

Fig. S24. The eGFR slope data distribution

A). The density plot of eGFR slope which was estimated from multiple data points using linear regression, and was used for the eGFR slope MWAS analysis (n=410).

B). The density plot of the residualized eGFR slope (adjusted for age, genetic background, eGFR and gender etc.).

Legends for Datasets S1 to S31

Dataset S1. Baseline demographic and clinical characteristics of the participants

Dataset S2. DNA Methylation sites associated with hemoglobin A1c at p<5e-5 in the CRIC study

Dataset S3. DNA Methylation sites associated with albuminuria at p<5e-5 in the CRIC study

Dataset S4. DNA Methylation sites associated with eGFR at p<5e-5 in the CRIC study

Dataset S5. DNA Methylation sites associated with kidney function decline rate at p<5e-5 in the CRIC study (results of conditional logistic regression)

Dataset S6. DNA Methylation sites associated with eGFR slope at p<5e-5 in the CRIC study (result of linear regression)

Dataset S7. Demographic and Clinical Characteristics of kidney tubule data

Dataset S8. The replicated probe associated with baseline eGFR

Dataset S9. Functional annotation of 110 CpGs associated with hemoglobin A1c at p<5e-5 in the CRIC Study

Dataset S10. Functional annotation of 73 CpGs associated with albuminuria at p<5e-5 in the CRIC Study

Dataset S11. Functional annotation of 99 CpGs associated with eGFR at p<5e-5 in the CRIC Study

Dataset S12. Functional annotation of 111 CpGs associated with kidney decline rate at p<5e-5 in the CRIC Study (results of linear regression)

Dataset S13. Demographic and Clinical Information of kidney tubule samples (methylation and gene expression data of human kidney)

Dataset S14. The overlap in Gaunt et al.'s (33) mQTL results (π1)

Dataset S15. Kidney function associated CpGs that were driven by genetic variations

Dataset S16. Demographic and Clinical Information of kidney tubule participants (RNA-seq data of human kidney)

Dataset S17. Moloc regions where causal variants were shared by gene expression, DNA methylation and CKD with posterior probability abc_PP≥0.8 (organized by each signifcantly moloc triplets, based on the study of Wuttke et al. (34))

Dataset S18. Moloc regions where causal variants were shared by gene expression, DNA methylation and CKD with posterior probability abc_PP≥0.8 (organized by each signifcantly moloc triplets, based on the MVP study)

Dataset S19. Moloc regions where causal variants were shared by gene expression, DNA methylation and CKD with posterior probability abc_PP≥0.8 (organized by each signifcantly moloc triplets, based on the study of Morris et al.)

Dataset S20. Moloc genes locates in MHC regions

Dataset S21. Top enriched biological process (BP) of 71 'moloc' protein-coding genes

Dataset S22. eGene-mCpGs-GWAS triplets that showed pleiotropic associations of methylation, transcription and phenotype (GWAS data were from the Study of Wuttke et al. (34))

Dataset S23. eGene-mCpGs-GWAS triplets that showed pleiotropic associations of methylation, transcription and phenotype (GWAS data were from the Study of Hellwege et. al. (35))

Dataset S24. eGene-mCpGs-GWAS triplets that showed pleiotropic associations of methylation, transcription and phenotype (GWAS data were from the Study of Morris et. al. (36))

Dataset S25. Coefficient of Variation of methylation levels (across 473 samples) of 102 CpG sites identified by pleiotropic association test

Dataset S26. Functional annotation results of all genes that show pleiotropic associations of methylation, transcription and phenotype

Dataset S27. Functional enrichment of the highly connected function group found in NK cells

Dataset S28. Functional enrichment of the highly connected function group found in CD8+ T cells

Dataset S29. 53 moloc regions of LACTB transcription, cg02713581 methylation and eGFR variations

Dataset S30. 48 moloc regions of IRF5 transcription, cg04864179 methylation and eGFR variations

Dataset S31. Quality Control of Genotype data

SI References

1. Aryee MJ*, et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10):1363-1369.
2. Fortin J-P, Triche Jr TJ, & Hansen KD (2016) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* 33(4):558-560.
3. Teschendorff AE*, et al.* (2012) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2):189-196.
4. Siva N (2008) 1000 Genomes project. (Nature Publishing Group).
5. Barfield RT*, et al.* (2014) Accounting for population stratification in DNA methylation studies. *Genetic epidemiology* 38(3):231-241.
6. De Jager PL*, et al.* (2014) Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience* 17(9):1156.
7. Wold S, Esbensen K, & Geladi P (1987) Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1-3):37-52.
8. Hastie T, Tibshirani R, Narasimhan B, & Chu G (2001) impute: Imputation for microarray data. *Bioinformatics* 17(6):520-525.
9. Peterson LE (2009) K-nearest neighbor. *Scholarpedia* 4(2):1883.
10. Teschendorff AE & Relton CL (2018) Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics* 19(3):129.
11. Du P*, et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* 11(1):587.
12. Mailman MD*, et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics* 39(10):1181.
13. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81(3):559-575.
14. Consortium G (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348(6235):648-660.
15. Price AL*, et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 38(8):904.
16. Delaneau O, Marchini J, & Zagury J-F (2012) A linear complexity phasing method for thousands of genomes. *Nature methods* 9(2):179.
17. Delaneau O, Zagury J-F, & Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* 10(1):5.
18. Howie BN, Donnelly P, & Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* 5(6):e1000529.
19. Marchini J, Howie B, Myers S, McVean G, & Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* 39(7):906.
20. Yuan M & Lin Y (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49-67.
21. Tsaprouni LG*, et al.* (2014) Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 9(10):1382-1396.
22. Horvath S (2013) DNA methylation age of human tissues and cell types. *Genome biology* 14(10):3156.
23. Wahl S*, et al.* (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 541(7635):81.
24. Krueger F (2012) Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. *URL* http://www/*.*

bioinformatics. babraham. ac. uk/projects/trim_galore/.(Date of access: 28/04/2016).

25. Harrow J*, et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 22(9):1760-1774.
26. Anders S, Pyl PT, & Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166-169.
27. Young MD*, et al.* (2018) Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* 361(6402):594-599.
28. Butler A, Hoffman P, Smibert P, Papalexi E, & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* 36(5):411-420.
29. Leek JT, Johnson WE, Parker HS, Jaffe AE, & Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6):882-883.
30. Kundaje A*, et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317.
31. Ernst J & Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* 9(3):215.
32. Greene CS*, et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 47(6):569.
33. Gaunt TR*, et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome biology* 17(1):61.
34. Wuttke M*, et al.* (2019) A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics* 51(6):957.
35. Hellwege JN*, et al.* (2019) Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nat Commun* 10(1):3842.
36. Morris AP*, et al.* (2019) Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nature communications* 10.