Understanding Chronic Kidney Disease: Genetic And Epigenetic Approaches

Yi-An Ko Ko
University of Pennsylvania, yian.kia.ko@gmail.com

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Bioinformatics Commons, Genetics Commons, and the Systems Biology Commons

Recommended Citation
https://repository.upenn.edu/edissertations/2404

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/2404
For more information, please contact repository@pobox.upenn.edu.
Understanding Chronic Kidney Disease: Genetic And Epigenetic Approaches

Abstract
The work described in this dissertation aimed to better understand the genetic and epigenetic factors influencing chronic kidney disease (CKD) development. Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) significantly associated with chronic kidney disease. However, these studies have not effectively identified target genes for the CKD variants. Most of the identified variants are localized to non-coding genomic regions, and how they associate with CKD development is not well-understood. As GWAS studies only explain a small fraction of heritability, we hypothesized that epigenetic changes could explain part of this missing heritability.

To identify potential gene targets of the genetic variants, we performed expression quantitative loci (eQTL) analysis, using genotyping arrays and RNA sequencing from human kidney samples. To identify the target genes of CKD-associated SNPs, we integrated the GWAS-identified SNPs with the eQTL results using a Bayesian colocalization method, coloc. This resulted in a short list of target genes, including PGAP3 and CASP9, two genes that have been shown to present with kidney phenotypes in knockout mice. To examine the functional role of a newly identified gene from the integrative analysis, MANBA, we knocked down this gene in zebrafish. This resulted in pericardial edema, a phenotype seen with kidney developmental defect.

While epigenetic dysregulation has been suggested as a mechanism for the development of many diseases, little was known about the epigenome of normal and diseased human kidneys. We performed cytosine methylation analysis in microdissected human kidney tubules in both CKD samples and controls, and found differentially methylated regions (DMRs) in CKD. These DMRs are further validated in the second cohort. The DMRs were mostly localized outside of promoter areas and enriched in intronic enhancer regions, and we found that the DMRs contain consensus-binding motifs for key renal transcription factors (HNF, TCFAP, SIX2). Furthermore, we found these DMRs correlated with the transcript changes. A network analysis of these correlated DMR-transcript pairs revealed an enrichment of a fibrosis network highlighted the TGF-β pathway. In vitro validation experiments established a likely causal relationship between epigenetic changes and CKD development.

In summary, integration of eQTL analysis with GWAS-identified CKD variants has facilitated the identification of likely CKD target genes. We discovered that cytosine methylation changes in human kidney tissue samples in the major fibrotic pathways. Overall, genetic variations as well as cytosine methylation levels have significant impact on gene expression regulation, possibly downstream organ function, and CKD development.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Cell & Molecular Biology

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/2404
UNDERSTANDING CHRONIC KIDNEY DISEASE:
GENETIC AND EPIGENETIC APPROACHES

Yi-An Ko
A DISSERTATION
in
Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy
2017

__________________________________________
Supervisor of Dissertation
Katalin Susztak, M.D. Ph.D., Professor of Medicine Department

__________________________________________
Graduate Group Chairperson
Daniel S. Kessler, Ph.D., Associate Professor of Cell and Developmental Biology

Dissertation Committee:
Christopher D. Brown, Ph.D., Assistant Professor of Genetics
Russ P. Carstens, M.D., Associate Professor of Medicine
Michael A. Pack, M.D., Professor of Medicine
Daniel J. Rader, M.D., Seymour Gray Professor of Molecular Medicine
DEDICATION

To the contributors of my two alleles
Jung-Shiang Tu and Der-Shyang Ko
For their love and support
ACKNOWLEDGMENT

During the course of my Ph.D. studies at University of Pennsylvania. I have had the pleasure to work with many inspiring and knowledgeable people. I am grateful to the Cell and Molecular Biology Program in the University of Pennsylvania Perelman School of Medicine for creating a very stimulating and inspiring environment. I thank to my thesis committee members, Dr. Katalin Susztak, Dr. Christopher D Brown, Dr. Russ P. Carstens, Dr. Michael A. Pack, and Dr. Daniel J. Rader, for all of their encouragement and hard work in directing my thesis.

My development in science has been shaped by a number of people, but none more than Dr. Katalin Susztak. Working with Dr. Susztak has been a great journey of research, work, discovery, mistakes, laughter, celebration, and understanding. I have learned far more from her than only science, and look forward to a what I hope will be a long and productive collaboration. I am fortunate to have met Dr. Susztak soon after arriving in the U.S., and I hope to make her proud.

Thanks to the fourth and fifth floor Clinical Research Building members for being my GPS and friends in life and work, including members from Holtzman Lab, Epstein Lab, Tishkoff Lab, Zhou Lab, Brown, Wu Lab, Murray Lab and Sundaram Lab. I especially want to thank Juni, Ivy, Cody, Ian, Arcy, Chuck, Jay for being there since my first day on the floor, all the chats, help, encouragements and our Friday night movie club. I have extra-special affections for Elicia and Gleb for their day-to-day presence in my life, the sense and nonsense moments, scientific and non-scientific conversations. These will always be among my best memories as a graduate student.

Thanks too to my musical friends, Pasquale, Toby, Neil, Diane, Gary, Eleanor, Sara, Ralph, Eva, Helga, Bob, Sue, and Joyce. You are my extended cultural family in the U.S., and we will continue the celebration of art and life together. Thanks to my biking team leader Rick -- we will not be defeated by the cold weather! Thanks for the volunteering coordinators at Action Wellness, Jay, Ron, Harley, Jay-Bird; I am so happy to be part of you, and I will continue to be there. Finally, I would like to give special thanks to Dr. Paul Li and Dr. Akon Higuchi for directing my career since I was fresh out of undergrad; your influence on me has been much greater than you can imagine.

Last but not least, I wish to thank Peter for his everlasting love, encouragement and support.

ABSTRACT

UNDERSTANDING CHRONIC KIDNEY DISEASE: GENETIC AND EPIGENETIC APPROACHES

Yi-An Ko
Katalin Susztak

The work described in this dissertation aimed to better understand the genetic and epigenetic factors influencing chronic kidney disease (CKD) development. Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) significantly associated with chronic kidney disease. However, these studies have not effectively identified target genes for the CKD variants. Most of the identified variants are localized to non-coding genomic regions, and how they are related to CKD development is not well-understood. As GWAS studies only explain a small fraction of heritability, we hypothesized that epigenetic changes could explain part of this missing heritability.

To identify potential gene targets of the genetic variants, we performed expression quantitative loci (eQTL) analysis, using genotyping arrays and RNA sequencing from human kidney samples. To identify the target genes of CKD-associated SNPs, we integrated the GWAS-identified SNPs with the eQTL results using a Bayesian colocalization method, coloc. This resulted in a short list of target genes, including PGAP3 and CASP9, two genes that have been shown to present with kidney phenotypes in knockout mice. To examine the functional role of a newly identified gene from the integrative analysis, MANBA, we knocked down this gene in zebrafish. This resulted in pericardial edema, a phenotype seen with kidney developmental defect.

While epigenetic dysregulation has been suggested as a mechanism for the development of many diseases, little was known about the epigenome of normal and diseased human kidneys. We performed cytosine methylation analysis in microdissected human kidney tubules in both
CKD samples and controls, and found differentially methylated regions (DMRs) in CKD. These DMRs are further validated in the second cohort. The DMRs were mostly localized outside of promoter areas and enriched in intronic enhancer regions, and we found that the DMRs contain consensus-binding motifs for key renal transcription factors (HNF, TCFAP, SIX2). Furthermore, we found these DMRs correlated with the transcript changes. A network analysis of these correlated DMR-transcript pairs revealed an enrichment of a fibrosis network highlighted the TGF-β pathway. *In vitro* validation experiments established a likely causal relationship between epigenetic changes and CKD development.

In summary, integration of eQTL analysis with GWAS-identified CKD variants has facilitated the identification of likely CKD target genes. We discovered that cytosine methylation changes in human kidney tissue samples in the major fibrotic pathways. Overall, genetic variations as well as cytosine methylation levels have significant impact on gene expression regulation, possibly downstream organ function, and CKD development.
# TABLE OF CONTENTS

ACKNOWLEDGMENT .................................................................................................................. III
ABSTRACT .................................................................................................................................... IV
TABLE OF CONTENTS .................................................................................................................. VI
LIST OF TABLES ........................................................................................................................... VII
LIST OF FIGURES ........................................................................................................................ VIII
LIST OF APPENDIX .................................................................................................................... XI

## CHAPTER 1: Introduction
1.1 Chronic kidney disease: approaches for understanding the disease ........................................ 1
   1.1.1. Overview of chronic kidney disease ........................................................................ 1
   1.1.2. Genetics and CKD .................................................................................................... 1
   1.1.3. Epigenetics and CKD ............................................................................................. 4
1.2 Genome-wide association studies ....................................................................................... 5
   1.2.1. The history of genetic mapping .............................................................................. 5
   1.2.2. GWAS methods ..................................................................................................... 7
   1.2.3. GWAS and chronic kidney diseases ..................................................................... 9
1.3 Expression quantitative locus analysis (eQTL) .................................................................. 12
   1.3.1. The development of eQTL analysis ..................................................................... 12
   1.3.2. eQTL analysis ....................................................................................................... 13
   1.3.3. Finding colocalization of GWAS and eQTL loci .................................................. 14
1.4 Epigenetics ......................................................................................................................... 15
   1.4.1. Epigenetics in chronic kidney disease .................................................................. 15
   1.4.2. Epigenetics and CKD development in mouse models ......................................... 17
   1.4.3. Epigenetics and CKD development human studies ............................................. 19
   1.4.4. Epigenetics Overview .......................................................................................... 20
   1.4.5. Methods to study the epigenome .......................................................................... 26

## CHAPTER 2: Genetic-variation-driven gene-expression changes highlight genes with important
functions for kidney disease ........................................................................................................ 33
2.1. Abstract ............................................................................................................................... 33
2.2. Introduction ........................................................................................................................ 34
2.3. Results ................................................................................................................................. 35
   2.3.1. eQTL analysis of human kidney tissues ................................................................. 35
   2.3.2. Validation and replication of kidney cis-eQTL signals ........................................... 37
   2.3.3. eGenes are enriched for metabolism-associated genes ....................................... 41
   2.3.4. Kidney eQTL highlights the genetics of disease traits .......................................... 43
   2.3.5. Defining the genetic network underlying CKD ..................................................... 44
   2.3.6. Integrative analysis identifies MANBA as a potential gene for CKD ................. 47
   2.3.7. Analysis of MANBA gene expression ................................................................. 51
   2.3.8. Knockdown of Manba results in kidney defect in zebrafish .................................. 52
2.4. Discussion and Conclusion .............................................................................................. 55
2.5. Materials and Methods .................................................................................................... 57

## CHAPTER 3: Cytosine methylation changes in enhancer regions of core pro-fibrotic genes
characterize kidney fibrosis development ................................................................................. 62
3.1. Abstract ............................................................................................................................... 62
3.2. Introduction ........................................................................................................................ 62
3.3. Results ................................................................................................................................. 64
LIST OF TABLES

Table 1-1 | Gene regulatory function of histone marks
Table 2-1 | SNPs identified in human kidney eQTL that are in tight LD with CKD GWAS SNPs
Table 2-2 | Significant genes identified through Coloc that are associated to both CKD GWAS- and human kidney-identified SNPs
Table 3-1 | Demographic, clinical and histological characteristics of the samples
LIST OF FIGURES

Figure 1-1 | Illustration of cis-eQTL.
Figure 1-2 | The eukaryotic transcription unit.
Figure 1-3 | Schematic diagram of the restriction enzyme based method to identify differentially methylated regions.
Figure 2-1 | PCA plot of kidney samples against 1000 genomes reference genome.
Figure 2-2 | Cis-eQTL analysis of human kidney genome.
Figure 2-3 | Positive correlation between sample number and identified eGenes.
Figure 2-4 | Positive correlation between sample number and replicated significant SNP-eGene pairs.
Figure 2-5 | Significant SNPs in the eQTL are enriched at regulatory regions.
Figure 2-6 | An example of significant SNPs in the eQTL that is enriched at regulatory regions.
Figure 2-7 | Network of eGenes.
Figure 2-8 | Density distributions of gene expression of eGenes and all genes in kidney samples.
Figure 2-9 | Disease trait GWAS Enrichment analysis.
Figure 2-10 | Integrative analysis of GWAS and eQTL SNPs.
Figure 2-11 | Integrative analysis of GWAS and eQTL SNPs.
Figure 2-12 | Genetic and epigenetic annotation of MANBA-locus SNPs.
Figure 2-13 | Locuszoom plots in MANBA region at chr4:103.4-103.8Mb.
Figure 2-14 | Genetic and epigenetic annotation of MANBA-locus SNPs.
Figure 2-15 | Gene expression analyses of MANBA.
Figure 2-16 | Kidney defect phenotype observed in zebrafish after knockdown MANBA.
Figure 2-17 | Kidney defect phenotype observed in zebrafish after knockdown MANBA.
Figure 3-1 | Statistically significant cytosine methylation differences in chronic kidney disease.
Figure 3-2 | External and internal validation of the observed changes.
Figure 3-3 | Chronic kidney disease differentially methylated regions are localized to kidney-specific enhancer regions.

Figure 3-4 | Chronic kidney disease differentially methylated regions are enriched for kidney-specific transcription factor binding sites.

Figure 3-5 | Differentially methylated regions correlate with transcript changes.

Figure 3-6 | Regulation of transcripts by a DNA methyltransferase inhibitor in in vitro cultured human tubular epithelial cells.

Figure 3-7 | Gene body cytosine methylation changes drive gene expression differences.
LIST OF APPENDIXES

Appendix 1 | List of CKD associated SNPs. List of SNPs curated manually from all kidney-function-associated GWAS.
Appendix 2 | All significant eGenes
Appendix 3 | Differentially methylated loci and the corresponding differentially expressed transcripts
Appendix 4 | Infinium DKD dataset and HELP CKD dataset differentially methylated loci
Appendix 5 | Technical validation of methylation level using MassArray Epityper
Appendix 6 | RefSeq annotated genomic distribution of the hypo- and hypermethylated regions
Appendix 7 | List of differentially methylated regions with p-value methylation level, genomic locus and nearest annotated transcript

xi
CHAPTER 1: Introduction

1.1 Chronic kidney disease: approaches for understanding the disease

1.1.1. Overview of chronic kidney disease

The major function of the human kidney is to filter blood and eliminate metabolic waste through urine. The healthy human kidney filters 180 liters of blood each day. Most electrolytes and small molecules are reabsorbed from the primary filtrate while some of them are secreted in the urine. This process requires a large quantity of active molecule transport and associated energy expenditure in the kidney.

Kidney function is mostly measured by filtration capacity (glomerular filtration rate or GFR), based on plasma clearance of endogenous creatinine. Chronic kidney disease (CKD) is a common condition, estimated to affect 13.6% of the US population [2] and 11 to 13% of the global population [3]. CKD is characterized by a 40% decline in eGFR or leakiness of the filter to plasma proteins, specifically albumin. More than 75% of persons with CKD have hypertension and diabetes in the U.S. [2].

CKD is associated with a 3- to 5-fold increase in risk of mortality and is an irreversible disease process that culminates in end-stage renal disease (ESRD) [4, 5]. Current treatments for ESRD are limited to dialysis and kidney transplantation. Understanding of the mechanisms of CKD development is relatively limited with respect to the roles of genetic and environmental factors (hypertension, diabetes). No new drugs have been registered for CKD since 2001.

1.1.2. Genetics and CKD

A small proportion of CKD cases has monogenetic causes, including cystic kidney diseases and glomerular diseases. However, in the majority of cases, CKD is a complex-trait disease with both genetic and environmental factors, including diabetes and hypertension [6, 7]. One of the key questions in the study of CKD is whether there is evidence suggesting a causal role for genetics, specifically whether CKD is heritable. Heritability is the proportion of phenotype
variation that is the result of the additive genetic effect of a disease or other condition, e.g. a positive or negative correlation of a given trait between parents and offspring. Heritability can be assessed through different methods including familial aggregation, adopted children (to distinguish environmental factors from heritability), and twin studies.

Family aggregation studies in the general population have addressed the heritability of kidney function, and are also useful for assessing rare variants traveling through pedigrees [8-12]. In the Framingham Heart Study offspring cohort with more than 300 families [8], the multivariable-adjusted heritability estimates were 0.29 for creatinine and 0.33 for eGFR. The heritability of eGFR estimated in a normal kidney function Utah pedigree cohort was 0.3 to 0.5 [13]. (Kidney traits are further discussed in Chapter 1.2.3.)

Studies have also revealed that association between kidney disease and family history is stronger in African Americans [14, 15] than European-descent Americans [14], suggesting potential differences in genetic variation between different populations. In short, kidney function is heritable and shows phenotype variations that are specific to different populations.

Familial linkage analyses have helped identify the causal genes for strongly heritable kidney diseases that follow simple Mendelian inheritance patterns. For example, the autosomal dominant polycystic kidney disease (ADPKD) gene on chromosome 16 [16], and the Alport syndrome gene on the X chromosome [17], were both discovered through genomic linkage analysis. Mutations in these genes have low frequency but the effect size for disease risk is large.

Linkage analysis, a statistical method that associates the function of genes or variants with their chromosomal location, has also identified a few genomic loci associated with CKD. These include the identification of a chromosome 8 locus link to urinary microalbumin (Fox et al., 2005), a link between chromosome 19 and 12 with urine albumin to creatinine ratio (UACR; see Chapter 1.2.3 for additional detail) (Freedman et al., 2003), a link between the q arm of chromosome 7 with diabetic nephropathy [18], a link between chromosome 2q36 and 6q22-23 with IgA nephropathy [19, 20], and linkage between genes on chromosomes 22q, 5q and 7q with
Although linkage analysis was a successful approach to detect rare genetic variants with large effects on disease risk, diseases with different penetrance in affected individuals make disease classification difficult (e.g. not all offspring that carry the same genomic mutation will develop the same disease) and decrease study power. Pedigree analyses have generated results that map genetic loci to kidney diseases using multiple families. However, it is very unlikely that each family has the same sets of genetic variants, so the linkage results can conflict and the significance decreases. Linkage analysis often identifies rare variants, with large effect size, that are segregated within tested families. This provides useful insight for monogenic traits, but for complex traits, this approach has generated few results and often cannot be reproduced in the general population.

Since complex traits often have multiple genetic factors, these genetic factors should be surveyed in general populations that carry the disease, rather than by linkage analysis on family pedigrees. To study genetic variants that are associated with specific diseases in the general population, the “common disease, common variant” (CD-CV) hypothesis was developed in the 1990s. The CD-CV hypothesis proposes that genetic variants in normal individuals contribute to overall risk of complex-trait diseases. It was first described in a few publications [24-26], then supported by comprehensive empirical data [27]. Later the CD-CV hypothesis was advanced through the completion of the human genomic map [28, 29], generation of public SNP databases [30], and the International HapMap project [31, 32].

Based on the CD-CV hypothesis, GWAS were devised to study the association between common variants and complex-trait phenotypes. GWAS with genotyping arrays have facilitated the identification of multiple loci associated with kidney function. Genotyping arrays are designed so that through statistical inference, they can recover most of the HapMap and 1000 Genomes cataloged variants. The array is an economical substitute for whole-genome sequencing [33]. More than 100 loci have been identified, and these variants show strong and reproducible
association with CKD. The majority of the variants are localized to non-coding regions of the genome [34]. More detailed development of GWAS can be found in Chapter 1.2.

The detection of genetic variants that exert small effects or appear with very low frequency in complex-trait disease development requires very large study cohorts for sufficient statistical power. These variants may therefore be missed by GWAS with current sample size. To date, the identified genetic variants explain only a small fraction of the heritability of CKD. Single nucleotide polymorphisms (SNPs) associated with kidney function explain only approximately 3% for eGFR thus far [35-38].

Since the majority of the GWAS-identified variants are localized to the non-coding regions of the genome, it is unclear which genes are affected by the genotypic changes. In order to develop a more complete understanding of the genetic basis for CKD, CKD-associated SNPs need to be tied to their target genes. With the advance of microarrays and RNA sequencing, it is feasible to assess genome-wide gene-expression changes. A few studies have shown the heritability of gene-expression traits [39-42], and thus we can use gene expression as a quantitative trait for performing expression quantitative trait loci analysis (eQTL), a method that will be discussed further in Chapter 1.3. CKD-associated SNPs can be more accurately understood by using eQTL to link to potential target genes, and could be studied for their relevant biological functions.

In Chapter 2, I focus on understanding the genetics of CKD, with an aim of identifying target genes in GWAS studies. I use eQTL analysis to identify target genes for the CKD GWAS hits. Methods used in this study have identified potential target genes for mechanistic and causal studies of kidney disease development.

1.1.3. Epigenetics and CKD

A plausible mechanism that could explain the missing heritability for phenotypes is epigenetic change [43]. In general, epigenetics studies chemical modification changes in the genome that do not alter DNA sequences. Epigenetic traits are DNA-sequence-independent
transgenerational changes, brought on by changes to the environment of the affected individual that propagate environmental effects on phenotypic development. Transient metabolic changes, for example, can result in long-term phenotypic alterations by epigenetic reprogramming.

One popular hypothesis for the missing heritability is the “developmental origins of adult diseases” hypothesis, also referred to as the Brenner-Barker hypothesis or fetal programming. For example, an adverse nutritional environment during development has been shown in animal models to be associated with worse lifetime outcomes including blood pressure, metabolic and renal traits, even after the subject animals have been separated from the adverse conditions.

In addition to fetal programming, the effect of poor glycemic control can be detected even a quarter-century after cardiovascular and kidney disease development, a phenomenon called “hyperglycemic memory.” More detail is provided in Chapter 1-4.

To understand whether epigenetic changes potentially contribute to CKD development, I performed genome-wide cytosine methylation analysis of microdissected human kidney samples. I discuss this work and its results in Chapter 3.

1.2 Genome-wide association studies

1.2.1. The history of genetic mapping

Genetic mapping of inheritance dates to the late 19th century with Gregor Mendel’s publication of the laws underlying inheritance patterns he observed in *Pisum sativum* [44]. However, in the beginning of the 20th century, cases involving more than two traits were found not to follow Mendelian random segregation principles consistently. For example, while eye and body color in *Drosophila melanogaster* tend to pass down together, this is not always true. These inconsistencies motivated researchers to look into the genetic mechanisms underlying multiple-trait inheritance.

Around the same time, cell biology was progressing quickly along with increased microscope resolution. Frans Alfons Janssens discovered the crossing-over of genes during
meiosis and termed the process “chiasma.” Thomas Hunt Morgan found that this process explained his own observations in *D. melanogaster* cross-breeding experiments. He adopted the evidence collected by Janssens and developed the theory of genetic linkage. Morgan stated that as a result of meiotic recombination, any markers that are correlated in segregation are located close to one another in the genome [45]:

If the materials that represent these factors are contained in the chromosomes, and if those factors that “couple” be near together in a linear series, then when the parental pairs (in the heterozygote) conjugate like regions will stand opposed. There is good evidence to support the view that during the strepsinema stage homologous chromosomes twist around each other, but when the chromosomes separate (split) the split is in a single plane, as maintained by Janssens. In consequence, the original materials will, for short distances, be more likely to fall on the same side of the split, while remoter regions will be as likely to fall on the same side as the last, as on the opposite side.

Morgan highlighted the idea that traits are linked and his student Alfred Sturtevant took the analysis forward. He realized that all *D. melanogaster* have the same chromosome length. This facilitated construction of a genetic map, because the frequency of specific phenotypic pairs among all offspring could be used to calculate the relative distance of the relevant genetic markers. He proposed that the distance between any two genetic markers on a *D. melanogaster* chromosome was inversely related to the ease of separating the traits. Sturtevant successfully tested the hypothesis and produced the first chromosome map, which included the sex-linked genes in the correct order and relative spacing [46]. The non-random assortment of genetic markers in the chromosomes was later termed linkage disequilibrium (LD) [47].

Since the first genetic mapping of *Drosophila* traits, geneticists have mapped numerous other model organisms by intensive interbreeding for the mutants of detectable phenotypes. However, this cannot be carried out in humans for obvious ethical reasons. There were some successes in identifying linkage, including the first autosomal linkage in human, the Lutheran blood group antigen system [48]. Then, in 1980, Botstein suggested the first systematic method to map genes, using an existing method involving restriction fragment length polymorphisms (RFLPs). RFLP detects DNA polymorphisms through restriction endonucleases. The restriction
endonucleases are sequence-specific restriction enzymes that cannot recognize and digest the DNA sequences when there is polymorphism in the DNA sequence. Applying this method to human genomic DNA would result in different lengths of DNA sequence, indicating the presence of polymorphisms [49, 50]. Combining the polymorphisms as genetic markers with pedigree information, linkage of human genetic diseases was enabled, for example Huntington’s disease [51], adult polycystic kidney disease [16], cystic fibrosis [52-54], and retinoblastoma [55, 56].

Since the 1980s, genetic mapping has identified mutations in thousands of genes that are associated with Mendelian diseases through relatively low-density panels of markers. However, common diseases exhibit more complex inheritance patterns in the general population, and traditional gene-mapping methods cannot identify genetic causes effectively for such common diseases as hypertension, cardiovascular diseases, obesity, and CKD.

1.2.2. GWAS methods

GWAS examines genetic variants across the human genome to identify associations between variants and phenotypes. Study power is one of the main considerations in designing a GWAS. Study power is the probability of the null hypothesis being correctly rejected. Large sample size is necessary to provide enough study power, especially when the association between a single variant and the disease is weak.

Association studies are carried out in two stages. In the first stage (the discovery phase), the association of variants and phenotypes is determined. Variants that are identified if they have statistically significant allele frequency differences associated with disease phenotypes and quantitative traits. In the second stage (the replication phase), significantly associated markers from the discovery phase are evaluated for association in additional independent study samples that resemble the initial discovery sample. Replication serves to confirm association and to detect potential bias. The replication analyses reduce the incidence of technical mistakes and spuriousness in the identified associations.

To date, more than 88 million genomic variants have been cataloged in the 1000
Genomes Project [57]. Directly assaying all known common polymorphisms is inefficient and impractical, because many variants are highly correlated due to LD. Genotyping a reduced set of variants in one strongly linked block can be used to infer the remaining variants through genotype imputation (discussed below). It is therefore important to select variants that can achieve genome-wide coverage within a feasible budget. Generating genetic maps based on direct genomic distance is rarely used for marker selection (e.g. using evenly spaced markers across the autosomal chromosome). On the other hand, cataloged common human sequence variations are more commonly used to calculate LD and select genetic markers for the genotyping arrays [58, 59].

In the 1990s, array-based genotyping platforms assayed fewer than 10,000 SNPs, but this has subsequently increased to between 300,000 and 5,000,000 SNPs per array in recent years. GWAS efficiency and statistical power can be improved through genotype phasing and imputation. Phasing is a method to estimate haplotype. When genomic variants are genotyped, we observe pairs of alleles, but the strand and haplotype (a series of genomic variants that are on the same chromosome) to which they belong are unknown; in other words, we cannot identify the parent from which the individual acquired the haplotype.

Once the haplotype is phased, imputation is performed. Imputation is the statistical inference of unobserved genotypes using a set of observed-genotype-phased haplotypes as a reference [31, 33, 57, 60, 61]. This technique allows us to estimate the genetic markers that are not directly genotyped. The HapMap, 1000 Genomes Project and Haplotype Reference Consortium have produced extensive LD maps, and there are several methods to impute genotypes based on these LD relationships [62-65]. When combining genotype data from different studies, often the genotype data are generated on different platforms. By imputing genotypes, meta-analysis can infer genotypes that are unknown in some of the studies, allowing GWAS to increase sample size and power.

Interpretation of the results of large association studies faces the statistical difficulty of
type I errors, where incorrect rejection of the null hypothesis yields false positive results. To cope with false positives, a multiple-testing corrected p-value is typically used. One of the most frequently used is to take the Bonferroni correction for multiple tests, where the cutoff p-value of 0.05 is corrected by the approximate one million independent tests (the number of common variants tested in the genomes of individuals of European ancestry) to generate the threshold [66, 67].

Hundreds of traits have been examined by GWAS, and variants that have reached genome-wide significance can be accessed at the GWAS Catalog supported by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) [34].

The majority of GWAS-identified loci are non-coding variants and this limits the ability of GWAS to identify the causal variants and to link these genetic variants to the disease mechanisms.

1.2.3. GWAS and chronic kidney diseases

The National Kidney Foundation defines chronic kidney disease according to any of the following criteria: pathologic kidney abnormalities, persistent proteinuria, other urine abnormalities (e.g., renal hematuria), imaging abnormalities, and eGFR <60 mL/min/1.73 m² on two occasions separated by more than 90 days [68]. CKD is classified in five stages, where stage 1 is the mildest form and stage 5 is the most severe with eGFR <15 mL/min/1.73m². When stage 5 CKD is treated by dialysis or kidney transplant, it is also categorized as ESRD or renal failure. The current treatment options for ESRD are limited to dialysis or kidney transplant.

Most of these measurable traits of kidney dysfunction can be used as a quantitative trait in GWAS studies, including eGFR, proteinuria, histologic morphology, and metabolites observed in urine. eGFR is estimated using equations that incorporate endogenous filtration markers, for example serum creatinine (SCr) or cystatin C (CysC). Case-control studies can also use eGFR to define CKD as patients with eGFR below 60 mL/min/1.73m². The CKDGen Consortium published
a meta-analysis of multiple cohorts with the largest sample size to date for kidney function eGFR in 2016. In total, 175,000 individuals were included and 53 loci were identified (29 known and 24 novel loci; most of these variants are associated with eGFRcrea, one with eGFRcys and four with CKD) [69].

Another key marker for CKD is urinary albumin. Albumin is a 66 kDa protein that circulates in the blood stream and cannot pass through the glomerular filtration barrier. However, when kidney structural integrity of the glomerular filtration barrier is injured or albumin reabsorption along the proximal tubules is impaired, larger proteins like albumin will leak into the urine.

To measure proteinuria, a frequently used method is to collect samples at only one time point and measure the albumin and creatinine concentrations. Creatinine level is used for normalization for the albumin because albumin levels will vary with urine concentration. This method yields the urine albumin-to-creatinine ratio (UACR); microalbuminuria is defined as 30-300 mg/g, and macroalbuminuria as UACR >300 mg/g.

In a GWAS study examining UACR, a total of 67,000 samples was used for meta-analysis and independent replication in more than 51,000 samples, which identified 1 locus with a missense variant in the cubilin gene (CUBN) [70, 71]. As mentioned earlier, CKD is defined by reduced eGFR or by microalbuminuria. Several attempts were made to validate eGFR-associated SNPs in UACR; however, none of the eGFR-associated SNPs were associated with UACR or albuminuria, suggesting different genetic components for the two traits [72, 73].

Several specific etiologies of CKD were tested in case-control studies, including diabetic kidney disease (DKD), ESRD, IgA nephropathy, steroid sensitive nephrotic syndrome (SSNS), and membranous nephropathy (MN).

It is estimated that approximately 45% of CKD patients have diabetic kidney disease (DKD), which is diabetes accompanied by CKD markers including increased UACR and decreased eGFR. The Diabetes Control and Complications Trial (DCCT)/Epidemiology of
Diabetes Interventions and Complications (EDIC) performed GWAS in 1,304 participants and found 13 independent loci [74] in type I diabetes; however, there were no SNPs that reached genome-wide significance. Similarly, no variant reached genome-wide significance in a trans-ethnic meta-analysis of type II diabetes in the Family Investigation of Nephropathy and Diabetes (FIND) cohort in 2008 [75].

In 2015, 1 locus on chromosome 6 reached genome-wide significance in a study of more than 6,000 unrelated individuals in the FIND cohort for DKD, and this result was replicated in a separate study of more than 7,000 individuals [76]. A meta-analysis showed 21 significant loci, which replicated loci that were close to genome-wide significance in both Pezzolesi et al and the FIND study [77]. And one significant locus was found with combined discovery and replication cohorts [78]. Other studies have discovered loci associated with type I or type II diabetes-associated ESRD [78-80]. However, there was no overlap between DKD- and CKD-identified loci.

The disparities among studies possibly arose from the use of different diagnostic parameters for DKD. Some studies used diabetes mellitus (DM) combined with eGFR to classify DKD, while others combined DM with albuminuria levels. The other possible reason may be due to differences in environmental effects such as glycemic control, DM duration, and blood pressure control, as they influence disease development and perhaps the association results.

Important loci on chromosome 22 have been discovered using admixture mapping associated with non-DM ESRD in African American cohorts [81, 82]. This locus was later fine-mapped to three variants close to the APOL1 gene, which showed an evolutionary advantage of resistance to trypanosome-mediated sleeping sickness in West Africa [83, 84].

Finally, IgA nephropathy (IgAN) is characterized by the deposit of IgA immune complexes in the glomerular mesangium, which causes progressive kidney injury [85]. The genetic component of IgAN has recently been shown by many linkage studies as well as GWAS results. Two major GWAS meta-analyses replicated 9 loci and identified 6 novel loci [86, 87].

A collection of CKD associated SNPs is summarized in Appendix 1, which lists of 110
independent loci of GWAS findings for different kidney disease associated traits; all neighboring
SNPs with LD \( \geq 0.2 \) were removed.

1.3 Expression quantitative locus analysis (eQTL)

1.3.1. The development of eQTL analysis

As discussed in the preceding section, the majority (around 90\%) of the variants
identified in GWAS are in non-coding regions of the genome. The non-coding variants present
special challenges for research, because predicting functional consequences of genetic variants
in non-coding regions has been difficult. Identifying quantitative phenotypes that are associated
with these variants (QTL) can facilitate the mechanistic studies for disease development, for
example gene expression levels, or expression QTL (eQTL).

The systematic measurement of gene expression levels through microarray methods
facilitated the first few major eQTL analyses, proceeding from the yeast [88], maize [89], mouse
[89], to human genome [90], and the eQTL method has been applied to several diseases.
However, most eQTL analyses of human samples were performed in immortalized cell lines or
circulating cells [91]. The transcriptome is tissue-type specific, thus surrogate cell types cannot
represent organ-specific regulation of gene expression by variants. There is a large number of
ubiquitously expressed genes that show shared regulation of gene expressions in the genome, as
shown in GTEx consortium [92] and other international projects that have generated
comprehensive catalogs of human eQTL maps.

Adopting transcription levels measured by RNA-sequencing in eQTL was advanced
through the work of the GTEx consortium [92]. The application of RNA-sequencing enabled the
quantification of transcript levels, identification of isoforms, and examination of allele-specific
expression, as well as long non-coding RNAs and novel transcripts in a much higher resolution
than the array-based method [91, 93].
1.3.2. eQTL analysis

eQTL mapping examines the genotype of a genetic locus that is significantly associated with the transcript or gene expression levels local (cis) or distal (trans) to the transcription start site of each gene. The cis-eQTL concept is illustrated in Figure 1-1, where the distance is conventionally specified as 1Mb from the transcription start site of each gene, as most regulatory control takes place in the vicinity of genes. As shown in the figure, an eQTL may contain multiple genetic variants that tag the actual causal variant or variants in close LD, which accounts for the genotype-dependent gene expression.

Stringent genotype quality control (QC) is crucial to reduce the incidence of technical problems and spuriousness in the identified associations. After QC, genotypes are subject to phasing and imputation using large reference panels to infer the untyped genotypes. It also increases the accuracy of imputation for variants with lower frequencies [94-96]. Gene expression analysis using RNA-seq instead of microarray to quantify the transcript expression level provides a few advantages. RNA-seq provides transcript isoform expression as well as information for allele-specific expression.
eQTL analysis of peripheral blood samples that the number of genes with identified local
eQTLs rose from around 5% in 100 samples to approaching 50% of all genes after meta-analysis
using several thousand samples [97]. This establishes that increased sample size can increase
resolution of variants explaining the transcript variance [97, 98].

Recent reports from GTEx with 44 tissue types showed that although increased sample
size is positively related to more shared cis-eQTL, weaker associations still remain that are
unique to the individual tissues and have higher effect size than shared eQTL [98].

1.3.3. Finding colocalization of GWAS and eQTL loci

The general model used to understand GWAS signals is that a causal variant is localized
to a regulatory region in the disease-causing cell type. The variant alters TF binding and further
affects the expression level of the genes.

One way to prioritize studying regions associated with potential causal variant is to
combine statistical association of genetic variants with gene expression, and with complex traits,
by colocalization of signals.

An important example of the integrative analysis from GWAS to eQTL to phenotype was
demonstrated with the FTO locus. This locus was discovered to be associated with obesity in
several GWAS [99, 100], with the reproducibility suggesting that it was the most significant signal
for obesity. The obesity-associated FTO region was shown to have enhancer properties identified
using histone mark modifications as well as enhancer assays [101].

The FTO region was also demonstrated to interact with the promoters of the IRX3 gene
through circular chromosome conformation capture followed by sequencing (4C-seq) [102]. In
short, the FTO region regulates the IRX3 gene through its enhancer activity. The rs1421085 T-to-
C variant in the obesity-associated FTO region was shown to disrupt a conserved ARID5B
repressor motif. Under the normal condition of the T allele at rs1421085, ARID5B binds to the
sequence and suppresses the enhancer, causing a drop of IRX3 and IRX5 expression. These
changes shift the mesenchymal progenitors from white adipocytes, browning to beige adipocytes (lipid storage), and leading to decreased mitochondrial respiration and thermogenesis with increased adipocyte size. When there is disruption of the ARID5B binding site with C allele at rs1421085, IRX3 and IRX5 expression increases and the pre-adipocyte differentiates into lipid-storage white adipocyte. [101].

Studies have shown that trait-associated GWAS SNPs are significantly more likely to be eQTLs than expected by chance [103]. To highlight the target genes of GWAS SNPs, we can colocalize the eQTL and GWAS signals. The colocalization of genetic variant signals implies that the gene with eQTL at the same locus as GWAS can influence the disease trait [104]. Several methods have been proposed, and each has a different hypothesis and approach for mapping the location and signals from the summary statistic in eQTL and GWAS.

Among many GWAS and eQTL integration studies, I reviewed two that were used in Chapter 2 as they provided superior understanding of colocalization. The Sherlock method [105] uses Bayesian statistics to match the association signals from GWAS and eQTL for individual genes, in order to distinguish whether the significance are coming from the same variant. A second method, Coloc [106], also based on Bayesian statistics, tests the hypotheses that the colocalizations of GWAS and eQTL signals represents a causal or coincidental relationship.

These methods explain the significance of using both eQTL and GWAS summary statistics to assist in identification of potential causal variant(s) at a specific locus. To further narrow down and identify the potential disease-causing variants, we need to further annotate the genome with regulatory elements in a cell-type specific manner, including histone tail modifications and cytosine methylation levels.

1.4 Epigenetics

1.4.1. Epigenetics in chronic kidney disease

As GWAS only explains 3 % of CKD heritability, there is a need to understand the
“missing” heritability. Increasing the samples size of GWAS cohorts, using denser genotyping platforms and different ethnic groups will be critical for this effort. Another potentially important mechanism for understanding missing heritability is transgenerational epigenetics.

There is a large variation in the number of nephrons in the human kidney, ranging from approximately 200,000 to 2 million. The number of nephrons at birth is inversely correlated with the risk of hypertension and CKD [107]. Low nephron count is correlated with low birth weight [108], which often results from intrauterine malnutrition [109, 110]. Low nephron number causes increased glomerular pressure on the remaining nephrons, and the increased workload on each glomerulus can induce microalbuminuria and glomerulosclerosis, resulting in nephron loss and renal injury [111].

Intrauterine malnutrition is associated with metabolic syndrome, hypertension, and renal disease development by fetal programming [111-114]. A direct correlation of low birth weight to microalbuminuria in type I diabetes has been established in the general population [115]. Rodent models of intrauterine growth retardation have indicated that lack of nutrient or oxygen availability results in the development of salt-sensitive hypertension and microalbuminuria [116]. Together, these results indicate that the epigenome could be an important mediator of long-lasting intrauterine environmental effects, particularly because the epigenome is plastic during development [117-119].

In addition, several environmental factors have been shown to have a long-lasting effect on CKD development [120]. Environmentally induced epigenetic modification could represent a plausible mediator explaining the impact of environmental factors on CKD development. Although differentiated cells have a stable epigenome, lately the cellular memory or programming effect of the environment has been described to extend beyond development as well (Fig. 1-2).

Epigenetics can explain missing heritability in cases of transgenerational changes. Transgenerational epigenetic effects have been shown in rodent models [121, 122] and also proposed in a two human studies [123, 124]. The association between intrauterine exposure to
maternal diabetes has been established as having an adverse effect on BMI and insulin secretory function in offspring [125]. The transgenerational effect was further studied in a Pima Indian (U.S.) cohort with or without intrauterine exposure to diabetes. The Pima Indian study showed that the risk can be explained by cytosine methylation differences that are enriched in type 2 diabetes, maturity onset diabetes of the young (MODY), and Notch pathways [126].

One key question is how does nutrient availability change the epigenome? Most chromatin-modifying enzymes require substrates or cofactors that are intermediates of cell metabolism [127-130]. For example, acetylation by histone acetyltransferases depends on the local subcellular acetyl-CoA concentration [131, 132]. In vitro evidence supports the conclusion that fluctuations of metabolite levels modulate activities of chromatin-modifying enzymes, thereby influencing chromatin dynamics [133-135].

Cells and tissues are sensitive to epigenetic shift while the epigenome is plastic and being established during development and differentiation. Change in the epigenome during development plays an important role in phenotype development later on, by modulating how cells respond to stimuli [136, 137]. For example, while podocytes without an essential component of H3K4 methyltransferase develop into functionally competent cells, it was shown that they have an abnormal response to environmental stressors (e.g. aging) [138]. Recent studies suggest that we can predict cellular aging based on the epigenetic signature [139-142]. This indicates that the epigenome could provide a traceable link between the genome and the environment. In short, the epigenome is emerging as a biochemical record of relevant life events.

1.4.2. Epigenetics and CKD development in mouse models

Recent research supports the role of epigenetics in CKD development. Bechtel et al. showed differences in the cytosine methylation profiles of fibroblasts isolated from control and CKD mouse kidneys [143]. Using a folic acid-induced kidney fibrosis model, the investigators showed that hypermethylation of a guanosine triphosphatase-activating protein, RASAL1, causes increased Ras activation in fibroblasts, resulting in fibroblast proliferation and fibrosis
development. Genetic deletion or chemical inhibition of DNMT1 is able to reduce fibrosis development in the folic acid-induced kidney fibrosis model. Future studies using DNMT inhibitors in other progressive kidney disease models, as well as cell-type-specific Dnmt1 deletion in mice, will aide in elucidating how changes in cytosine methylation contribute to kidney fibrosis development.

The contribution of histone deacetylase (HDACs) has also been described in different mouse models of acute and chronic kidney injury [144-150]. The HDAC inhibitor Trichostatin A (TSA) ameliorates proliferative glomerulonephritis [151], long-term glomerulosclerosis [152, 153], and proteinuria in a nephrotoxic serum nephritis model [152]. Pretreating mice with valproic acid, a class I-selective HDAC inhibitor, was highly effective in reducing proteinuria as well as sclerosis in the doxorubicin-induced focal segmental glomerulosclerosis model [154]. Two separate studies have shown that HDAC inhibitors can also attenuate diabetes-induced renal hypertrophy and renal damage in rodents [155, 156].

Together, these chemical inhibitor-based studies strongly suggest that histone deacetylases play a major role in CKD development. The proposed mechanism is that HDAC inhibitors are involved in reactivating major developmental pathways and promoting tissue repair after injury. Unfortunately, HDAC inhibitors, similar to many inhibitors, have cell-type specificity issues. Therefore, follow-up studies using genetically engineered site-specific knockout animals will be essential to determine the role and contribution of histone acetylation in kidney fibrosis.

Cell-culture-based studies from the Natarajan and El-Osta laboratories have indicated that cells that are grown in a high-glucose medium develop rapid and persistent changes in their histone tail modification patterns [157-160]. In cultured aortic endothelial cells, transient hyperglycemia caused changes in H3K4me1 on the promoter of the p65 subunit of nuclear factor-κB, and subsequently was associated with increased p65 transcript levels. The expression of p65 was increased in endothelial cells even several days after they were returned to a normal glucose medium.
A similar mechanism has also been shown in diabetic mice. The sustained proinflammatory phenotype observed in vascular smooth muscle cells cultured from type 2 diabetic db/db mice was associated with reduced levels of the repressive mark H3K9me3 on the affected inflammatory gene promoters [161]. Treatment of glomerular mesangial cells with transforming growth factor (TGF)-β1 or high glucose could lead to changes in key active and repressive histone modifications at the promoters of inflammatory and fibrotic genes [162, 163]. These observations link episodes of hyperglycemia with a persistent increase of inflammation and fibrosis, which are important characteristics of diabetic kidney disease.

1.4.3. Epigenetics and CKD development human studies

Large-scale human studies showing epigenetic differences in CKD have not been published, because the epigenome is cell-type specific and therefore human kidney tissue and kidney cell lines are necessary for experiments. There are a few human epigenome-wide association studies that used peripheral blood mononuclear cells or other surrogate cell types. Sapienza et al [164] described differences in cytosine methylation profiles of diabetic patients with and without ESRD. Infinium HumanMethylation27 BeadChip arrays were used to determine cytosine methylation changes. Although there was significant patient heterogeneity, Sapienza et al identified close to 110 loci with statistically significant methylation differences in ESRD samples compared with those without ESRD. Most loci showed lower methylation levels in the ESRD cases, and many of the differentially methylated regions were in close proximity to genes that previously had been shown to express differently in patients with kidney diseases. Although this is an important first step to understand the epigenome of CKD, future research will be required to determine whether markers found in human urine correlate to epigenetic changes in renal epithelial cells, which can further be used to develop disease markers using non-invasive sample collecting method.

Another study, by Ciavatta et al [165], examined the potential contribution of the epigenome to antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis. ANCA disease is
associated with spontaneous development of autoantigens against myeloperoxidase (MPO) and proteinase 3 (PR3) proteins. Ciavatta et al found that H3K27me3, a chromatin modification associated with silenced promoters, was decreased at the PR3 and MPO promoter areas in ANCA patients compared with healthy controls. H3K27me3 was also associated with decreased cytosine methylation of PR3 and MPO CGI promoter areas. The investigators proposed that normally higher RUNX3 levels were responsible for recruiting H3K27 methyltransferase, a subunit of enhancer of zeste homolog 2 (EZH2), to PR3 and MPO loci. Loss of RUNX3 expression is a plausible cause for the loss of H3K27 methylation and the expression of MPO and PR3 in ANCA patients. This was an elegant illustration that epigenetic modifications associated with gene silencing are perturbed at ANCA autoantigen-encoding genes, potentially contributing to autoimmunity development.

Future studies could determine whether similar mechanisms play a role in other autoimmune diseases as well. In Chapter 3, we discuss the regulation of gene expression in the pathogenesis of CKD through cytosine methylation. The alterations in both gene expression and cytosine methylation in the TGF-β pathway (TGFB3, SMAD3, SMAD6) were observed in the CKD patient samples, and these genes were known to be the key regulators of kidney fibrosis development and have important roles in CKD development [166].

1.4.4. Epigenetics Overview

The epigenetic mark is characterized by chemically modified DNA or associated proteins. Epigenetic marks include cytosine modification (mainly methylation) and histone tail modification, which are changes caused by nutrition and other environmental factors [167, 168]. The epigenetic mark is inherited during cell division to maintain cell identity, thereby mediating between the stably inherited genome and the changing environment. Characterization of the epigenetic mark has played a major role in defining cell-type-specific gene expression.

A classic transcription unit in a multicellular eukaryote contains both clusters of proximal promoter elements and several types of cis-acting regulatory sequences including insulators,
promoters, enhancers, silencers, and locus-control regions [169]. Insulators border and separate transcriptional units. Promoters are usually located at the 5' end of transcription start site, the minimal sequences required for accurate transcription initiation. The promoter region can encompass a cluster of cytosine-guanine nucleotide pairs within the linear sequence, known as CpG islands (CGI) [169, 170].

For gene transcription to occur, not only promoters but also long- and short-range regulatory regions are needed. These cis-type gene-regulatory regions are often highly cell-type-specific. Simultaneous binding of transcription factors (TFs) to each other and to the long- and short-range regulatory regions results in genomic DNA loops that join distant regulatory DNA sequences together (Figure 1-2) [171, 172].

Enhancers are also of critical importance, and can be located from between a few kilobases (short-range) to hundreds of kilobases (long-range) away from the regulated gene [169, 173, 174]. Enhancers can be found in the opposite DNA strand, downstream of the regulated gene, or at intronic regions. Enhancers usually show significant enrichment for the binding of cell-type-specific TFs [174]. TF binding, or the presence of multiple enhancers, is critical for regulating the optimal transcription strength following an external stimulus. However, sometimes the loss of even one enhancer can result in loss of expression level of a gene [175].
While the critical importance of enhancers has been increasingly appreciated, genome-wide identification of enhancers has been exceedingly difficult. Recent advances in epigenetics, sequencing and computational methods have helped to define these important gene regulatory regions and will be discussed in detail in this chapter.

**Histone modification:** The DNA in the cell nucleus is wrapped around small, positively charged proteins called histones, to form a larger organized structure called chromatin. When 147 base pairs of DNA are wrapped around histones, it forms a compact structure called a nucleosome [176, 177]. In addition to DNA, nucleosomes are comprised of a histone octamer consisting of two copies each of the core histones H2A, H2B, H3, and H4 [175].

Nucleosomes play important roles in compacting DNA into the nucleus. When the nucleosome is in condensed condition, called heterochromatin, it is inaccessible to transcription factors and transcriptionally inactive [178]. Modification of the amino acids in the histone tails appears to be central to the processes that regulate transcription, including methylation, acetylation, phosphorylation, and ubiquitination [179]. The histone code hypothesis proposes that

![Figure 1-2 | The eukaryotic transcription unit.](image)

The eukaryotic transcription complex is characterized by a combination of transcription factors binding to promoters and one or more looping enhancers. Enhancers can be upstream, downstream or intronic. Enhancers are cell-type-specific. Both enhancer looping and TF binding are essential for messenger RNA transcription. Different chromatin regions are characterized by different histone tail marks. CTCF, CCCTC-binding factor. TF, transcription factor.
chromatin-DNA interaction (i.e. the gene regulatory complex) is guided by histone tail modification [170]. There are molecular histone code “writers” that establish these modifications, and “readers”, the enzymes that use these histone tail modifications to assign gene regulatory function to DNA [180-183]. More than 60 histone modifications have been described [184, 185], although some are redundant and the overall set can be simplified into 10-15 different patterns, each with specific gene regulatory functions (Table 1-1).

Functional genomic elements are identified through sequence-association with histone modification, TF binding and DNaseI hypersensitivity [186]. Cell-type-specific gene regulation can be understood by defining the cellular histone code (the epigenome). The pioneering ENCODE and Roadmap Epigenome projects have resulted in the ongoing characterization of cell-type-specific annotations of gene regulatory regions through the use of cultured human cell lines and human tissues [187].

Transcription factor binding is a key characteristic of a regulatory region; gene regulatory regions are mostly open and nucleosome-free. Nucleosome-free DNA is sensitive to DNaseI digestion, allowing the identification of gene regulatory regions by DNaseI hypersensitivity methods [188-190]. Nucleosomes directly adjacent to DNaseI-hypersensitive sites are characterized by specific histone tail modifications.

For example, H3K4me3 in combination with H3K4me2 and H3K4me1 H3K27ac usually marks active promoter areas. On the other hand, H3K27me3 and H3K9me3 usually mark inactive promoters. H3K4me1, H3K27ac and H3K9ac marks together are associated with enhancer activity, and H3K36me3 and H3K20me1 mark transcribed regions both for coding and non-coding transcripts [170, 191]. The binding of the transcriptional repressor, CTCF, defines insulator areas that separate different transcriptional units.

**Table 1-1** Gene regulatory function of histone marks

<table>
<thead>
<tr>
<th>Chromatin Mark</th>
<th>Chromatin State</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCF&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Insulator/repetitive/CNV&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Poised promoter/repressed promoter</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Transcription transition and elongation</td>
</tr>
<tr>
<td>Histone Modification</td>
<td>Function</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>Transcription transition</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>Enhancer</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Promoter</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Promoter/enhancer</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Active enhancer</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Active Promoter</td>
</tr>
</tbody>
</table>

§ CTCF, CCCTC-binding factor; ¶ CNV, copy number variation

Combinations of histone marks further tune gene regulatory regions within the genome. In particular, H3K4me1 and H3K27ac together have been shown to mark the location of cis-regulatory active enhancers. H3K9me3 and H3K20me3 histone tail modifications are used to locate satellite, telomeric, and repeat regions. In embryonic stem cells, large genomic regions have histone modifications with characteristics opposite to those seen in differentiated cell types. For example, many developmentally important transcription factor-encoding regions in embryonic stem cells contain both the active mark of H3K4me3 and the transcriptional repressor mark of H3K27me3. These regions are called bivalent domains [192-194]. Promoters of key developmental transcription factors that control differentiation typically have bivalent domains [195-197].

During differentiation, cells will take on either the active or the repressive mark depending on which pathway they follow. Adding methyl groups to histones is attributed to three families of enzymes that catalyze the addition of methyl groups donated from S-adenosylmethionine to histones. The SET domain-containing proteins [198], DOT1-like proteins, have been shown to methylate lysines (K) [199, 200], and members of the protein arginine (R) N-methyltransferase family [201] have been shown to methylate arginine. H3K27me3 is catalyzed by the EZH2 subunit of the Polycomb repressive complex 2 (PRC2) [202]. Histone acetylation is catalyzed by histone acetyltransferases (HATs) [203-205], and acetyl groups are removed by histone deacetylases (HDACs) [206-208]. These enzymes can be targeted HDAC inhibitors such as valproic acid and hydroxamic acids derivatives [209].

24
**Cytosine methylation:** Methylation of cytosines in the fifth position (5-methylcytosine, 5-mC) is an epigenetic modification. The majority of the genome has a low cytosine/guanine content, and these cytosines usually are methylated, including transposons and other repeated elements. The cytosine nucleotide followed by the guanine nucleotide is expressed as “CpG”. CpGs often spontaneously undergo deamination and become thymine (CpT), resulting in underrepresentation of CpG [29]. Cytosine-rich regions are organized into CpG islands (CGIs) in the genome [210, 211]; these are short regions, about 300-3,000 bp, characterized by high cytosine content ( > 60%) [212, 213].

These CGIs are enriched on gene promoter regions, and cytosines in CGIs usually are unmethylated and the methylation of promoters negatively correlates with transcript levels. There are two basic models for promoter hypermethylation-induced transcriptional silencing: DNA methylation can repress transcription directly by blocking transcriptional activators from binding to cognate DNA sequences, and methyl binding proteins recognize methylated DNA and recruit corepressors to silence gene expression directly. Genome-wide studies have also indicated that methylation of regions adjacent to CGIs, known as CpG shore regions, play a key role in diverse biological processes [214]. While the functional effect of gene body methylation has been debated, it has been shown that regulation of splicing through binding of CTCF (which is sensitive to cytosine methylation levels) results in pausing of RNA polymerase II [215-218]. It may also enhance transcription through inhibition of initiation of cryptic intragenic transcription [219-221].

Cytosine methylation levels are usually the highest in fully differentiated cells; however, cytosine methylation is erased after fertilization, with only a handful of imprinted regions remaining methylated in zygotes [222-224]. During cell-type-specific differentiation, cytosine methylation is established by DNA methyltransferases (DNMTs) [225-229]. Dnmt1 is responsible for maintaining the established DNA methylation levels during cell division. Homozygous Dnmt1 embryos are stunted, delayed in development, and do not survive past midgestation [225]. Dnmt3a and Dnmt3b participate in de novo DNA methylation of unmethylated cytosines and
appear to methylate different genomic DNA regions \textit{in vivo}. Dnmt3a actively methylates naked DNA and open nucleosomal DNA [230], while Dnmt3b is specifically required for methylation of centromeric minor satellite repeats. It has been shown Dnmt3a and Dnmt3b double-mutant mice have severe growth defects and do not survive over 4 weeks after birth. While Dnmt3a mutant mice develop to term, there is no viable Dnmt3b knockout and severe developmental defect at later stages of development is found [225, 226, 231-234]. Mutations of human DNMT3B are found in immunodeficiency, centromeric instability, and facial anomalies syndrome, a developmental defect characterized by hypomethylation of pericentromeric repeats [235, 236].

In addition to cytosine methylation, the TET (Ten-Eleven-Translocation) enzymes are known to oxidize 5-mC to 5-hydroxymethylcytosine (5-hmC), adding a hydroxyl group to the methylation of cytosines at C5 position. 5-hmC was identified as having the ability to regulate gene expression in coordination with transcription factors [237]. When hydroxymethylation is disrupted, it can result in disordered cell function, for instance, myeloid cancers [238-240].

1.4.5. Methods to study the epigenome

Analyzing condition-relevant cell types is critical to understand the role of epigenetics to disease development (e.g. analysis of renal epithelial cells to understand kidney disease). Because the cell type is the key determinant of the epigenome, it is unlikely that much mechanistic understanding would be gained by analyzing cells that are not functionally relevant to disease development. Cytosine methylation is one of the most stable epigenetic marks, it can be studied even in archived samples because genomic DNA is stable [241-243].
**DNA Methylation:** There are three major approaches to studying cytosine methylation, based on restriction enzyme digestion, affinity enrichment, and sodium bisulfite conversion. Each method has its strengths and caveats. Currently, all of these methods are available for both locus-specific and genome-wide studies when coupled with whole-genome-covering microarrays or sequencing technologies.

The restriction enzyme digestion method is based on isoschizomer enzymes that recognize the same nucleotides but differentially digest DNA based on cytosine methylation status. The most frequently used isoschizomers for DNA methylation studies are HpaII and MspI, which recognize the same sequences but digest differently according to the methylation levels [244]. MspI is methylation-insensitive whereas HpaII is sensitive to methylation level. The methylation-insensitive enzymes are not able to cleave the methylated-cytosine residues, leaving methylated DNA intact. By comparing the fragments after restriction-enzyme treatments, site-specific methylation differences can be determined (Figure 1-4A). The digested DNA can be subjected either to polymerase chain reaction (PCR) for locus-specific study, or to either microarray or sequencing for genome-wide screening.

Smal and its neoschizomer Xmal is another favored enzyme pair. Xmal and Smal both recognize the sequence CCCGGG, but they digest it at different sites. Smal is sensitive to cytosine methylation and is blocked by CG methylation, leaving a blunt end 5'-GGG. Xmal is methylation-insensitive enzyme that cleaves at sites with methylated CGs. Xmal leaves 5'-CCGG overhangs (Figure 1-4B). By comparing the two enzymes or use the restriction enzymes in serial, we can detect the differentially methylated regions [245-247].
The advantage of restriction enzyme digestion is its sensitivity to hypomethylated regions (e.g., CGIs). The drawback of this method is that the coverage and resolution is limited to regions that restriction enzymes recognize. One of the most commonly used restriction enzymes methods is HpaII tiny fragment enrichment by ligation-mediated PCR assay, developed by Suzuki et al [248, 249] and Thompson et al [250].

Affinity enrichment methods use antibodies against m5C, or methyl group binding protein domains (MBD) that target m5C, such as MeCP2 or MBC. Five proteins with mCpG-binding motifs have been identified in the MBD family: MBD1, MBD2, MBD3, MBD4, and MeCP2 [251]. To capture m5C with antibodies and proteins, the genomic DNA is randomly sheared into small fragments by sonication. The fragments then subject to immunoprecipitation with antibodies targeting methylated cytosines in DNA (MeDIP) [252, 253]. The advantage of the MeDIP method is a true genome-wide coverage, because it is not limited by restriction enzyme sites. However,

![Diagram of restriction enzyme digestion](image)

**Figure 1-3** Schematic diagram of the restriction enzyme based method to identify differentially methylated regions.

(A) MspI and HpaII digestion. HpaII is methylation-sensitive enzyme and MspI is methylation-insensitive enzyme. Comparing the fragmentation of the enzyme digestion results, we can discover the differentially methylated regions. (B) methylated DNA and reduction of genome complexity were achieved by serial digestion with Smal (methylation sensitive) and Xmal (methylation insensitive) restriction enzymes. Both (A) and (B) methods are suitable for whole genome analysis using microarray or sequencing platform.
the limited resolution of MeDIP is a key disadvantage. The binding of the antibody is determined by the cumulative methylation level of a region, and the resolution is limited by the degree of fragmentation from the sonication process.

Determination of cytosine methylation by sodium bisulfite conversion is an unbiased and sensitive method, and is currently the "gold standard" for determining cytosine methylation [254-257]. The method is based on the principle that unmethylated cytosines will be converted to uracil after bisulfite treatment, but methylated cytosines will not. Fragments can later be subjected either to cloning and sequencing or to mass spectrometry (e.g. the Sequenom EpiTYPER-type equipment, Sequenom, San Diego, CA) to interrogate locus-specific cytosine methylation. Genome-wide quantification of bisulfite-converted cytosines can be performed by either microarray- or sequencing-based platforms. The Infinium HumanMethylation450 BeadChip arrays (bisulfate conversion-based assays, Illumina, San Diego, CA) gained popularity as an easy and cost-effective method for methylation analysis. These arrays appear to perform quite well for epigenome-wide association studies. One key disadvantage is that they only analyze the methylation of 480,000 preselected loci, representing less than 1% of the whole genome. The reduced representation bisulfite sequencing is another bisulfite-conversion-based method that combines restriction enzyme-based size selection followed by next-generation sequencing. This method provides excellent coverage for CGI and promoter regions. Genome-wide bisulfite sequencing provides the best and highest resolution method to examine cytosine methylation status. The main drawback of this method is the current cost associated with whole-genome sequencing and data analysis.
**Chromatin Immunoprecipitation** Chromatin immunoprecipitation (ChIP), followed by either microarrays or next-generation sequencing (ChIP-chip/ChIP-Seq), is the most frequently used method to map transcription factor binding and histone modifications. Here, genomic DNA and proteins in either live cells or fresh tissues are cross-linked with formaldehyde, and the chromatin is fragmented into 200 base pair size. The sheared DNA/protein complex is then subjected to immunoprecipitation with an antibody against the selected transcription factor or histone tail modification. Once the genomic DNA is reverse cross-linked and released from the protein complex, it can be analyzed using quantitative PCR, microarrays, or sequencing.

Although ChIP has revolutionized our understanding of the epigenome, the method is not without limitations. First, the method does not provide an epigenetic map with base pair resolution, because the resolution is limited by the size of the sheared chromatin. To circumvent this issue, an enzyme digestion-based method, the exo-ChIP, was developed [258]. The key limitation of the ChIP-seq relates mainly to the specificity of the antibodies used for the studies. To address this issue, several antibody validation databases were created to evaluate antibody quality [259-261].
Next-generation sequencing platforms and data analysis: For genome-wide bisulfite sequencing, high genome coverage (usually at least 30X) and long reads are needed. Sometimes, 500-to-600 bp paired-end sequencing (e.g. MiSeqV2 System generates paired 250 bp or MiSeqV3 that generates paired 300 bp reads) is used [262]. For ChIP-seq, generally 100-bp single-end sequencing is used. ChIP-seq analysis begins by aligning each read to the reference genome and then determining the number of reads that aligned within each genomic region. A frequently used aligner is BWA. After reads are aligned, we can define protein binding sites using sequence reads enrichment in the genome, which is termed peak calling. Several peak calling algorithms are available including SPP [263], PeakSeq [264], and MACS [265], which were used in the ENCODE consortium [266]. Other available tools include SICER [267] and Genome-Wide Event Finding and Motif Discovery (GEM) [268]. Most of the methods target peaks that are around 150-300 bp, e.g. MACS and PeakSeq, whereas SICER identifies rather diffused histone marks that can span kilobases to megabases of the genome. Another special feature of these method is that GEM improves the resolution of peaks by combining peak calling and motif analysis.

Because the accessibility of each genomic region is slightly different, an enrichment score is developed to show the enrichment fraction normalized to input DNA (that is, sonicated DNA without antibody enrichment). Both the degree of enrichment and the size of the region that is marked by an individual histone mark are important, because several different modifications can be present on an individual region. To address this challenge, several algorithms have been developed using statistical methods that incorporate histone mark information to identify functional elements from high-throughput genomic data sets. Ernst and Kellis et al [269] developed a method that is based on the hidden Markov Model, which can be used to generate gene regulatory region annotation based on a panel of histone ChIP-seq experiments performed on the same sample. The method takes into consideration the presence of multiple different histone marks and their reference sequence-based location. This method is very valuable for
comparing multiple different histone marks, either between different cell types, or within the same cell type but expressing different phenotypic characteristics, to highlight key similarities and differences in datasets.
CHAPTER 2: Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease

2.1. Abstract

Chronic kidney disease (CKD) affects close to 10% of the US population and is associated with a 3- to 5-fold increase in risk of death. Genome-wide association studies (GWASs) have identified sequence variants, localized to non-coding genomic regions, associated with kidney function. Despite these robust observations, the mechanism by which variants lead to CKD remains a critical unanswered question. Expression quantitative trait loci (eQTL) analysis is a method to identify genetic variation associated with gene expression changes in specific tissue types. We hypothesized that an integrative analysis combining CKD GWAS and kidney eQTL results can identify candidate genes for mechanistic and causal studies.

We performed eQTL analysis by correlating genotype with RNAseq-based gene expression levels in 96 human kidney samples. Applying stringent statistical criteria, we identified 1,886 genes (eGenes) associated with sequence variants. Using direct overlap and Bayesian methods, we identified new potential target genes for CKD. With respect to one of the target genes, Lysosomal Beta A Mannosidase (MANBA), we observed that SNPs associated with MANBA expression in the kidney showed statistically significant colocalization with SNPs identified in CKD GWAS studies. The results indicate that MANBA is a likely causative factor in CKD. The expression of MANBA was significantly lower in kidneys of subjects with risk alleles. Suppressing MANBA expression in zebrafish resulted in pericardial edema, a phenotype typically induced by kidney dysfunction.

Our analysis shows that gene-expression changes driven by genetic variation in the kidney can highlight potential new target genes for CKD development.
2.2. Introduction

The key function of the human kidney is to remove waste from the body. The kidney filters more than 180 liters of plasma every day and selectively reabsorbs or secretes specific electrolytes. The kidney excretes most of the endogenous and exogenous waste products that the body can modify into hydrophilic compounds. Ten percent of the body’s basal energy requirement is used by the kidney to perform its transport function [5, 270]. Chronic kidney disease (CKD) is characterized by at least a 40% decrease in the filtering function or abnormal leakiness to albumin. CKD is a common condition affecting close to 20 million people in the US and 10% of the population worldwide [6, 271]. Impaired kidney function is associated with a 3- to 5-fold increase in risk of death [272-274].

CKD is a gene-environment disease, and diabetes and hypertension are responsible for more than 75% of the CKD cases in the US [275, 276]. The genetic underpinning of CKD has been examined by genotyping and phenotyping large numbers of subjects with CKD and conducting GWAS studies. In most GWAS studies, creatinine-based estimates of kidney function (eGFRcrea) have been used to define kidney disease. Our review of the literature identified 110 loci that exhibited significant genome-wide association with CKD status. The association between genetic variants and disease has been replicated by several studies, indicating the robustness of the relationship.

Despite the remarkable progress of the GWAS experiments, these results have not translated into improved understanding of the mechanisms of CKD development. One critical bottleneck is that more than 90% of the CKD-associated SNPs are localized to non-coding genomic regions [34]. It is not clear, though, which genes and which cell types are affected by the genetic variants. Data from other complex-trait studies indicate that disease-causing genetic variants are localized to regulatory regions in disease-relevant cell types, altering transcription factor binding and downstream quantitative expression of target genes [69].

Most GWAS studies have proposed the nearest gene as the causal gene for disease
development; however, mechanistic research indicates that the nearest gene is often not causal, and identification of GWAS targets remains difficult [277, 278]. The expression quantitative trait loci (eQTL) method has been developed to associate genetic variation with gene expression changes. The method requires genotype- and tissue-specific gene-expression data from large numbers of healthy individuals, and controlled complex computational methods, to identify genotype-driven gene expression changes. Several tissue types have been difficult to collect in large enough numbers to perform eQTL analysis. The largest eQTL datasets have therefore been generated from easily accessible cell types, for example blood [97]. The Genotype Tissue Expression Project (GTEx) Consortium was established to circumvent tissue-procurement problems, and has been successful in creating a large number of eQTL datasets.[92] Unfortunately, the number of kidney samples in GTEx (n=26) is limited due to allocation and organ transplantation; hence GTEx has not published kidney-specific eQTL maps.

The aim of the current study was to generate a kidney-specific eQTL dataset and integrate this information with CKD-associated GWAS variants, in order to identify target genes for chronic kidney disease and thereby better understand kidney function and disease development.

2.3. Results

2.3.1 eQTL analysis of human kidney tissues
We obtained genotype and gene expression data for 99 normal European descent human kidney cortical samples. Genotype data was obtained from Genome-Wide Human SNP Array 6.0 and expression data by RNA sequencing. We excluded samples that did not pass quality control (QC), and to avoid detecting genetic variations that are caused by population genetic structure differences, we further removed samples using principal component analysis (PCA) (Figure 2-1) [279-281]. There were 96 individual samples that passed QC and were used for downstream analysis (see Methods). These samples were phased and imputed from the 1,000 Genome Phase 3 haplotype CEU reference panel using SHAPEIT and IMPUTE2 [62, 282]. Gene expression data were normalized using quantile normalization followed by inverse normal transformation. We then added sex and age as covariables for probabilistic estimation of expression residuals (PEER) before the final transformation of the residuals (see Methods).

We determined the association between 5.1 million genetic variants and expression of 17,388 genes using Matrix eQTL (Figure 2-2) [283]. The analysis was limited to cis-eQTLs whose SNPs and gene pairs were within 1MB distance. We followed the protocol described in GTEx to
correct for multiple testing and establish significance [92]. In summary, we obtained permutation-adjusted p-values for each gene using FASTQTL [284], which yielded the most significant gene-SNP pairs. After deriving the empirical p-value from this analysis, we calculated the q-value using the Storey method [285], with a significance threshold of less than or equal to 0.05. This method identified 1,886 eGenes and 124,612 eQTL SNPs (eSNPs) (Appendix 2). In summary, we have generated human kidney eQTL data from 96 CEU samples.

![Diagram of eQTL analysis](image)

**Figure 2.2 | Cis-eQTL analysis of human kidney genome.**

(A) Flowchart of eQTL analysis (B) Global significance of eQTL SNPs across the genome. The y-axis shows $-\log_{10}$ p-value; the significance threshold used in our analysis is determined by the corrected permutation test p-value.

### 2.3.2 Validation and replication of kidney cis-eQTL signals
To evaluate the kidney eQTL signals, we first compared the number of eGenes identified by our analysis to 32 other tissue samples published in GTEx (v6). The analysis was limited to tissues from GTEx (v6) were used as comparison, with the selection criteria that each tissue type has to have more than 90 samples. The number of eGenes ranged from 1628 in liver to 9937 in thyroid. Kidney sample (n=96) is shown in red.

**Figure 2-3** | **Positive correlation between sample number and identified eGenes.**

The number of eGenes discovered is strongly correlated positively to the number of samples used for the analyses. Tissues from GTEx (v6) were used as comparison, with the selection criteria that each tissue type has to have more than 90 samples. The number of eGenes ranged from 1628 in liver to 9937 in thyroid. Kidney sample (n=96) is shown in red.

**Figure 2-4** | **Positive correlation between sample number and replicated significant SNP-eGene pairs.**

Significant eQTL SNP-gene pairs are overlapped between kidney and the 32 samples from GTEx (v6). We also observed a strong positive correlation of sample number and SNP-gene pair replication. The x-axis shows the sample number of each tissue type; the y-axis is the number of eSNP-eGene pair overlapped between GTEx samples with kidney eQTL.

To evaluate the kidney eQTL signals, we first compared the number of eGenes identified by our analysis to 32 other tissue samples published in GTEx (v6). The analysis was limited to
tissue samples where more than 90 samples were analyzed. The number of significant eGenes (n=1,886) in the kidney eQTL followed the expected linear relation between sample size and number of detected eGenes observed by GTEx (Figure 2-3). We also surveyed the overlap of eSNP-eGene association between kidney eQTL in the 32 tissue types in GTEx (v6). Consistent with previous studies, larger sample size increased the power of association detection, and we observed higher eQTL eSNP-eGene pair overlap among samples at larger sample sizes (Figure 2-4).

Next, given that eSNPs influence gene expression, we analyzed whether eSNPs are enriched on regulatory regions. We plotted the significance of association between eSNPs and their distance to transcription start sites (TSSs). We found that the strength of association was greater around the TSSs (Figure 2-5A), indicating that eSNPs were enriched around transcription start sites, likely promoter regions.

To establish that the eSNPs were in fact enriched on regulatory regions, we used histone tail modification-based chromatin immunoprecipitation sequencing (ChIP-seq) to identify kidney-

Figure 2-5 | Significant SNPs in the eQTL are enriched at regulatory regions.
(A) Density plot of eSNPs and their distance relative to TSS. (B) eSNPs overlap with 6 different histone marks in human kidney proximal tubule epithelial cells, compared with the 1000 null sets of variants matched for number, distance to TSS and frequency. The vertical line represents the 95% confidence interval. Histone marks with significance level of p-value < 1.0E-2 are highlighted with red asterisks. (C) eSNP enrichment in ENCODE TFBSs. Odds ratios of eSNPs are shown in y-axis and the vertical line represents 95% confidence interval. We marked the TFBSs reaching significance level p < 1.0E-2 with red asterisks. Control SNPs were the 1000 null sets of variants that are matched to eSNPs for number, distance to TSS and allele frequency.
specific regulatory regions. We performed this analysis using different histone tail mark-specific antibodies and human kidney proximal tubule cells (HKC8) [166]. We overlapped eQTL SNPs with each histone mark, quantifying the enrichment ratio by comparing it with 1,000 randomly generated SNP sets that were matched for frequency and distance. To avoid enrichment bias, we removed variants in the same linkage disequilibrium (LD) (see Methods). Kidney eQTL SNPs were statistically significantly enriched (~4-fold) on active promoters, as they showed highest enrichment with active promoter marks H3K4me2 and H3K4me3 (p-value < 0.01). There was no statistically significant enrichment for repressed promoter and gene-body-specific histone marks (Figure 2-5B). These studies indicate that kidney eQTL SNPs are enriched on promoter/enhancer regions in the kidney.

Another feature of regulatory regions is that they are enriched for transcription factor binding sites. We next examined the overlap of eSNPs with the transcription factor binding ChIP-seq data from ENCODE [186, 286, 287]. We found that eSNPs were statistically significantly enriched for transcription factor binding sites. Specifically, we found enrichment for CBX3, HDAC1, KDM5b, NRF1, ZEB1, IRF1, SIX5, and ZEB1 binding compared to 1,000 randomly generated SNP sets that were matched for frequency and distance (Figure 2-5C). Many of these transcription factors have been shown to be important in kidney development or kidney disease [288-290]. An example of a kidney-specific eQTL is shown in Figure 2-6A-B, where rs946213 variant influences the expression of RRP15. The SNP is located at a promoter and the variant interrupted the GATA5 binding site.
In summary, kidney eQTLs were enriched on kidney-specific regulatory regions, indicating the robustness of the approach. Our results indicate enrichment of the eSNPs on kidney-specific active promoter regions and transcription factor binding areas.

2.3.3  **eGenes are enriched for metabolism-associated genes**
According to GTEx, large numbers of eQTLs are conserved between different organs [92]. After overlapping kidney eGenes with 9 different tissue types in GTEx (v4) eGenes, we found 847 eGenes unique to kidney eQTL, indicating that our analysis identified both shared and unique eQTLs. To understand the functional role of the eGenes, and therefore the eQTL regulatory mechanism, we performed gene ontology (GO) analysis using all 1,886 eGenes and the Database for Annotation, Visualization and Integrated Discovery method (DAVID v6.7), which includes eGene-enriched pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG)

Figure 2-7 | Network of eGenes.
(A) The most significant eGene pathway is enriched in metabolic pathways and xenobiotic and drug metabolism. The KEGG pathways displayed here are those below the Benjamini-Hochberg corrected p-value of 5.0E-2. (B) IPA analysis of eGenes in the pathway related to metabolism of xenobiotics. Genes presented in the significant eGene list are in orange, and the rest of the genes in the network were supplied algorithmically. Genes enclosed in the box are kidney-specific eGenes.

[291]. As shown in Figure 2-7A, we found that the most significantly enriched pathway was related to metabolism, specifically drug and xenobiotic metabolism (p-value = 2.30E-06), which is a key function of the kidney.

To further understand the xenobiotic-metabolic eGene properties, we performed network analyses using Ingenuity Pathway Analysis (IPA). IPA highlights the xenobiotic pathway (p-value = 1.57E-23) in Figure 2-7B. In summary, many of the kidney eGenes show enrichment for drug,
xenobiotic metabolism, and pharmacogenomics, which is highly consistent with the function of the kidney and could be related to the higher expression these genes in the kidney (Figure 2-8).

2.3.4 Kidney eQTL highlights the genetics of disease traits

A critical goal of eQTL analysis is to provide a framework for biological interpretation of disease-related variants. To evaluate the relevance of the eQTLs identified in disease-mapping studies, we tested kidney eQTL SNPs with disease-associated SNPs using the National Human Genome Research Institute (NHGRI) catalog (www.ebi.ac.uk/gwas, accessed 9/30/2015) [292]. We analyzed statistical enrichment for GWAS SNPs in our kidney eQTL database (see Methods). We found that kidney-specific eQTL SNPs are enriched for CKD-associated traits (Figure 2-9). When compared to other traits (digestive, nervous, immune system diseases, cardiovascular, body measurement, biological processes, and other diseases) we found a significantly greater overlap between kidney eSNPs and polymorphisms that are associated with kidney function [35, 36, 293, 294]. These results indicate that genetic variants associated with kidney function and CKD are enriched in our eQTL studies and likely drive gene expression changes in the kidney.

![Figure 2-8](image_url)

**Figure 2-8 | Density distributions of gene expression of eGenes and all genes in kidney samples.**

Gene expression distributions are shown for both eGenes and all genes. A shift toward higher expression of eGenes can be seen with the overlap of the translucent plots. All genes are shown in light red and eGenes are in light blue.
2.3.5 Defining the genetic network underlying CKD

A key objective of our study was to identify potential target genes for CKD-associated GWAS variants. We manually curated a list of SNPs that both passed genome wide significance

![Figure 2-9 | Disease trait GWAS Enrichment analysis.](image)

SNPs associated with different traits were taken from National Human Genome Research Institute (NHGRI) database, with SNPs categorized based on disease ontology. Enrichment was calculated using Fisher's exact test; the significance is shown in the y-axis with -log₁₀ p-value.

![Figure 2-10 | Integrative analysis of GWAS and eQTL SNPs.](image)

(A) We used eSNPs that are also present in CKD-associated GWAS as a means to identify target genes that affect kidney disease. (B) Boxplot of rs1719246 and SPATA5L1, rs6429746 is within the LD block with rs2467853, which is one of the CKD-associated GWAS-identified SNPs (D'= 0.95; r² = 0.89).
and were associated with kidney disease traits. To acquire independent loci, we removed any SNPs having $r^2 \geq 0.2$. In total, we analyzed 110 leading SNPs and 2,357 tagging SNPs with $r^2 \geq 0.8$. We performed a direct overlap between CKD-associated GWAS SNPs and eQTL SNPs (Figure 2-10).

Overlapping SNPs were identified between the two datasets. These SNPs passed genome-wide significance both in GWAS and eQTL studies. They mapped to seven loci (eight genes) and regulated the expression of CST9, SPATA5L1, LOC654433, ACCS, CFHR1, ALMS1P, PGAP3 and TPRKB (Table 2-1).

The direct overlap method has a few limitations. First, eQTL and GWAS variants were not genotyped on the same platform, and a large number of variants were missing in the eQTL genotypes. Second, the method is very stringent, as it requires independent genome-wide significance for the SNPs in both studies, even though the GWAS and eQTL studies can gain significance from each other as they are independent methods.

To address the latter issue, we also implemented a Bayesian method (coloc [295, 296]) to formally test the hypothesis of association of a given region with eQTL and GWAS signals. Due

**Table 2-1 | SNPs identified in human kidney eQTL that are in tight LD with CKD GWAS SNPs**

<table>
<thead>
<tr>
<th>Trait</th>
<th>GWAS P-value</th>
<th>CKD GWAS SNP</th>
<th>CKD GWAS suggested target Genes</th>
<th>eGene identified eQTL</th>
<th>eSNP Best P-value</th>
<th>eSNP to target gene TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUN</td>
<td>3.30E-10</td>
<td>rs111231</td>
<td></td>
<td>PAx8</td>
<td>2.67E-07</td>
<td>9,927 bp upstream of TSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
<td></td>
<td>GATM/SPATA5L1</td>
<td>3.75E-08</td>
<td>41,152 bp upstream of TSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs246785</td>
<td></td>
<td>ALMS1/NAT8</td>
<td>2.63E-06</td>
<td>25,029 bp upstream of TSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-14</td>
<td>rs13538</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD</td>
<td>1.05E-42</td>
<td>rs110789</td>
<td></td>
<td>CDK12</td>
<td>1.28E-10</td>
<td>339,337 bp downstream of TSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>03</td>
<td></td>
<td>PGAP3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs130383</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>05</td>
<td></td>
<td>CSTE9</td>
<td>2.68E-09</td>
<td>9,027 bp upstream of TSS</td>
</tr>
<tr>
<td>eGFRcraea</td>
<td>9.00E-13</td>
<td>rs207403</td>
<td></td>
<td>ACCS</td>
<td>3.01E-14</td>
<td>510 bp downstream of TSS</td>
</tr>
<tr>
<td>eGFRcraea</td>
<td>2.00E-88</td>
<td>rs667760</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eGFRcraea</td>
<td>3.93E-09</td>
<td>rs207403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eGFRcraea</td>
<td>2.96E-10</td>
<td>rs667760</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IgA-NP</td>
<td>3.93E-09</td>
<td>rs207403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IgA-NP</td>
<td>2.96E-10</td>
<td>rs667760</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The direct overlap method has a few limitations. First, eQTL and GWAS variants were not genotyped on the same platform, and a large number of variants were missing in the eQTL genotypes. Second, the method is very stringent, as it requires independent genome-wide significance for the SNPs in both studies, even though the GWAS and eQTL studies can gain significance from each other as they are independent methods.

To address the latter issue, we also implemented a Bayesian method (coloc [295, 296]) to formally test the hypothesis of association of a given region with eQTL and GWAS signals. Due
to the nature of the analyses, we could only import one study from GWAS analyses instead of using our curated list with a multiple cohort of studies. As input to coloc, we used the largest GWAS of the kidney function measure eGFR from the CKDGen Consortium with the best statistical power, and its known summary statistics and sample sizes [69]. The five hypotheses used in the coloc analyses were: H₀: No association with either trait (trait 1 for eGFR GWAS and trait 2 for eQTL associations or vice versa); H₁: Association with trait 1, not with trait 2; H₂: Association with trait 2, not with trait 1; H₃: Association with trait 1 and trait 2, two independent SNPs; H₄: Association with trait 1 and trait 2, one shared SNP. As sufficient support for significant results, we set PP₄ > 0.8 and PP₄/PP₃ > 5 (posterior probability from H₃ and H₄) as criteria [297]. Here we used only SNPs that are associated with the eGFRcrea trait in the GWAS studies. As expected, this method identified SNPs that passed genome-wide significance in both methods, including Post-GPI Attachment To Proteins 3 (PGAP3), Spermatogenesis Associated 5 Like 1 (SPATA5L1), and Alstrom Syndrome 1 Pseudogene (ALMS1P). Furthermore, several additional potential target genes were identified, including PIGU, METTL10, MANBA, and others (Table 2-2).

PGAP3 encodes phospholipase for the glycolipid complex (glycosylphosphatidylinositol, or GPI) in the Golgi apparatus. The modification is crucial for linking GPI-anchored proteins to lipid rafts, which plays an important role in protein sorting and trafficking. One of the major phenotypes of PGAP3 knockout mice is that they present with progressive enlarged renal glomeruli with deposition of immune complexes and matrix expansion upon aging.[298] Furthermore, humans with PGAP3 loss-of-function mutations present with neurological and renal phenotypes [299, 300]. These experimental observations nicely validate our computational methods.
2.3.6 Integrative analysis identifies MANBA as a potential gene for CKD

To prioritize the validation of target genes, we performed network analysis for all eGenes using IPA. IPA highlights the metabolic pathway with a score of 38 (the negative log p-value of the odds ratio). MANBA was the only gene predicted in coloc that is presented in the pathway (Figure 2-11).

The GWAS study revealed a significant association between chromosome 4 SNPs and

<table>
<thead>
<tr>
<th>Gene</th>
<th>PP_H3</th>
<th>PP_H4</th>
<th>H4/H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. PGAP3</td>
<td>0.012715756</td>
<td>0.985344385</td>
<td>77.49</td>
</tr>
<tr>
<td>2. SPATA5L1</td>
<td>0.015059344</td>
<td>0.979123189</td>
<td>65.02</td>
</tr>
<tr>
<td>3. METTL10</td>
<td>0.009878945</td>
<td>0.978529109</td>
<td>99.05</td>
</tr>
<tr>
<td>4. PIGU</td>
<td>0.040348343</td>
<td>0.955231251</td>
<td>23.67</td>
</tr>
<tr>
<td>5. KNG1</td>
<td>0.013444728</td>
<td>0.955009245</td>
<td>71.03</td>
</tr>
<tr>
<td>6. MANBA</td>
<td>0.063418013</td>
<td>0.913297935</td>
<td>14.40</td>
</tr>
<tr>
<td>7. ALMS1P</td>
<td>0.002501869</td>
<td>0.887791453</td>
<td>354.85</td>
</tr>
<tr>
<td>8. C9</td>
<td>0.004801323</td>
<td>0.884400778</td>
<td>184.20</td>
</tr>
<tr>
<td>9. WHAMM</td>
<td>0.004698951</td>
<td>0.882157986</td>
<td>187.74</td>
</tr>
</tbody>
</table>

Figure 2-11 | Integrative analysis of GWAS and eQTL SNPs.
(A) We used eSNPs that are also present in CKD-associated GWAS as a means to identify target genes that affect kidney disease. (B) Boxplot of rs1719246 and SPATA5L1, rs6429746 is within the LD block with rs2467853, which is one of the CKD-associated GWAS-identified SNPs (D' = 0.95; r² = 0.89).
eGFR. While the genetic variants were located in the non-coding genomic regions, the proposed likely causal gene for kidney function in the GWAS study was listed as NFKB1. On the contrary, our analysis indicates MANBA is a potential target gene for kidney-function associated GWAS variants. We examined the association of gene expression based on the GWAS variants, and found that while MANBA shows statistically significant changes based on the genotype, expression of nearby genes, including NFKB1, was unaffected.

We found that the SNPs associated with kidney function also influenced the expression of MANBA. While the top GWAS SNP was not present in our eQTL database, we observed that

Figure 2-12 | Genetic and epigenetic annotation of MANBA-locus SNPs.
The lower-left part of the square is the analysis of LD between all of the chromosome 4:103.4-103.8Mb region using Haploview version 4.2 (http://www.broad.mit.edu/mpg/haploview). The strength of the LD is shown in the grey gradient as indicated in the figure. Pair-wise D’ values were calculated and plotted using Haploview v4.1. Dashed lines indicate the annotated genes in the locus. The upper-right corner of the square shows chromosome conformation capture (Hi-C) interactions contact probabilities in a human lymphoblastoid cell line (GM12878). The plot shows the 200 kb topologically associating domain at 1kb resolution. Locations of both CKD and eQTL SNPs are labeled. (B) The zoom-in locus of MANBA region at chr4:103.5-103.64 Mb. CKD GWAS SNP, MANBA-associated best eSNP, and the overlapped SNP between two datasets are labeled on the plot. (C) The genome browser display of the MANBA region. Tracks from top to bottom are: UCSC gene annotation; H3K4me1 and H3K27ac overlays of 9 cell lines from ENCODE; human kidney histone mark H3K4me1, H3K4me3, and H3K27ac from Epigenome RoadMap; and ENCODE chromatin states (ChromHMM) showing the regulatory properties of the sequence. (D) Sequence-predicted TF motifs were shown in the bottom for rs227361 (overlapped SNP) and rs170563 (eSNP of MANBA). SNP rs227361 disrupts transcription factor IRF2 motif and rs170563 disrupts FOXP2 motif.
the eGFR GWAS SNPs and eQTL variants were in strong LD. Figure 2-12 shows the LD structure of the human MANBA region. The top eGFR (rs228611) and eQTL (rs227361) SNPs in the chr4: 103.5 Mb-103.6 Mb region were within the same LD block (r²=0.88, D’=0.95). Upon examining the interaction by extended chromosome conformation capture (Hi-C) in the human lymphoblastoid cell line (GM12878), we found that the eQTL and GWAS regions were likely interacting with each other, which could explain our results.

Figure 2-13 | Locuszoom plots in MANBA region at chr4:103.4-103.8Mb.
(A) CKDGen Phase III GWAS SNPs and the corresponding -log10 p-value. (B)-(D) are eQTL association of MANBA in kidney, whole blood, and transformed fibroblasts in GTEx (v6).
Next, we examined whether the eQTL signal can be replicated in other organs. Upon examining the GTEx database we found that transformed fibroblasts showed a similar association between genotype and gene expression as we observed in the kidney (Figure 2-13). Unfortunately, the GTEx Browser does not contain kidney eQTL, which prevented us from performing the comparison. However, only the pattern of the transformed fibroblasts eQTL signal generated by GTEx closely resembled the kidney eQTL signal.
To further identify the causal variant (region), we also analyzed histone ChIP-Seq-based regulatory annotation maps. Human-kidney-specific annotation maps indicated that the most enriched histone marks are H3K4me1 and H3K27ac (active enhancer), which overlapped with the eSNPs Figure 2-14A. Examining the ENCODE database, we confirmed that several cell types also showed H3K27ac enrichment of this region. There was no H3K27ac enrichment around the top GWAS variant in human kidneys or other ENCODE cells, although fibroblasts showed enrichment for the latent enhancer mark H3K4me1. The eSNP rs170563 was enriched for the H3K27ac histone mark and disrupted the FOXP2 transcription factor binding site, suggesting that rs170563 eSNP potentially has a regulatory function. Detailed examination of the MANBA locus is shown in Figure 2-14B-C.

In summary, the observed eQTL signal overlapped with active enhancer regions both in the kidney and other organs, while the GWAS signal has no enrichment of the histone marks, suggesting that the causal variant is most likely in the region of eQTL-identified region and not in the peak GWAS region.

### 2.3.7 Analysis of MANBA gene expression

Next, we examined the gene expression of MANBA in different organs. Using the protein atlas [301] and Illumina Body Map (GSE30611), we found that the reproductive tract and the kidney both have intermediate expression of MANBA (not shown), and MANBA is expressed in lysosome of kidney tubule cells (Figure 2-15A).[302] Segment-specific expression data obtained from rat kidneys [303] indicate that MANBA is expressed in the long descending limb of the loop of Henle in the inner medulla, with the highest expression level in the inner medullary collecting duct segment.
Next, we examined the direction of regulation by the risk allele. We found that expression of MANBA was much lower (close to 50%) in healthy human kidney tissue samples obtained from subjects with risk alleles when compared to reference allele (Figure 2-15B).

We also analyzed MANBA expression in patients with chronic kidney disease. We have examined MANBA expression analyzed by Affymetrix microarrays on 95 human kidney samples.[290] We found that MANBA expression is lower in patients with CKD and kidney fibrosis (Figure 2-15C). Similarly, expression of MANBA was lower in a mouse model of fibrosis induced by unilateral obstruction (unpublished data) fibrosis (Figure 2-15D). In summary, MANBA expression was lower in risk allele samples.

2.3.8 Knockdown of Manba results in kidney defect in zebrafish

Figure 2-15 | Gene expression analyses of MANBA.
(A) Immunohistochemistry staining in human kidney tissues (Protein Atlas). (B) Boxplot showing association between rs170563 and MANBA. (C) Gene correlation with eGFR, the leading indicator of kidney function, using microarray and clinical parameters collected in a different independent cohort of 95 human kidneys. (D) Manba expression is significantly lower in mouse UUO model compare with sham (n=3 in each group). The statistical analysis results from DESeq are presented in the table under the plot.
As healthy subjects with risk alleles had lower MANBA expression, we wanted to understand the functional significance of lower MANBA expression. We conducted genetic knockdown experiments in zebrafish (Danio rerio). We used two different forms of morpholino oligonucleotides, translation blocking, and splice blocking. We confirmed the MANBA knockdown efficiency by quantitative reverse transcription polymerase chain reaction (qRT-PCR) analysis (Figure 2-17F). Zebrafish with MANBA knockdown presented with significantly increased pericardial edema and lethality. Pericardial edema is a commonly observed kidney defect phenotype in zebrafish (Figure 2-16A), as the fish is unable to excrete salt and water upon kidney dysfunction. Careful quantification of the phenotype is shown in Figure 2-17A, B. To confirm that the pericardial edema was caused by renal defect, we turned to Tg(enpep:GFP) zebrafish[304], which express GFP in pronephric ducts. GFP expression was reduced in the pronephros of MANBA knockdown embryos compared to controls. GFP expression was significantly reduced, and indeed often absent, in the Manba knockdown morphants (Figure 2-17C, D).
Because heart developmental defects could also contribute to edema formation, we analyzed heart morphology on hematoxylin- and eosin-stained sections. From the histological analysis, we did not observe severe heart abnormalities in zebrafish with severe edema (Figure 2-17E), which indicates that reduced MANBA levels are likely associated with kidney function defect in the zebrafish model. Knockdown of Nfkb1, which was suggested as the causal gene by GWAS studies, did not develop kidney related phenotypes in zebrafish (Figure 2-16B). In summary,

**Figure 2-17 | Kidney defect phenotype observed in zebrafish after knockdown MANBA.**

(A) Edema in zebrafish: Pericardial edema after gene knockdown is shown on the right panel; the control-injected zebrafish has no edema at 200 μM concentration. (B) Edema rate of Nfkb1 and Manba knocked-down zebrafish; p-values were calculated for each concentration group using Fisher's exact test. We tested the suggested target gene from GWAS study to confirm the eQTL analysis result.

Genetic knockdown of eQTL target gene MANBA, but not the proposed GWAS target NFKB1, was associated with kidney development defect in zebrafish.
2.4. Discussion and Conclusion

Genetic variation has significant impact on gene expression regulation, downstream organ function, and phenotype development. Here we present the first comprehensive RNA sequencing-based gene expression analysis for genetically driven gene-expression changes in human kidney samples. While large-scale efforts are in progress to generate comprehensive eQTL analysis for the human body, kidney-specific maps have not been generated. Our project aims to fill that gap by generating kidney-specific maps.

Our results indicate that while a large percentage of genotype-driven gene-expression variation is conserved between different tissue types, kidney-specific maps have highlighted variation in gene levels that play a key role in kidney function, indicating the critical need for tissue-specific maps. In prior work, an algorithm has been developed to impute expression levels for unavailable tissue types on the basis of eQTL. However, the sample size available today does not carry enough power to detect tissue-specific eQTLs [305].

We further found that kidney eSNPs showed tissue-specificity and are enriched on kidney-specific regulatory regions: promoters, enhancers and transcription factor binding sites. We used the kidney eSNPs we discovered to interpret disease-related variants discovered in the GWAS studies. We found four major loci where genome-wide significant eSNPs and eGFR-associated GWAS SNPs directly overlapped. We also employed Bayesian methods to confirm that eSNPs and GWAS-associated SNPs are present at the same locus and contribute to both phenotypes and gene expression levels. This analysis highlighted several genes that are likely responsible for the association observed between genotype and kidney disease associated traits in large-scale GWAS. For example, Cystatin 9 (CST9), variants showed association with serum cystatin levels. Another example is PGAP3, the phospholipase that attaches GPI to proteins, which is another likely causal gene for CKD development.

We have further focused our analysis on MANBA as a likely causal gene for genetic
variants on chromosome 4 shown to be associated with CKD. Variants that are associated with CKD traits also altered the expression of MANBA in human kidney samples. The risk genotype was associated with lower levels of MANBA levels in human kidney samples. Interestingly, this association can be recapitulated in the fibroblast cell line in the GTEx dataset, potentially indicating that the signal could come from fibroblasts. A similar direction on gene expression was observed in patients with kidney disease and mouse models with kidney fibrosis. We have used the zebrafish model to show that lower MANBA expression is causally related to kidney disease development. Morpholino knockdown of Manba resulted in pericardial edema formation, indicating the functional role of MANBA in kidney disease development. MANBA is a lysosomal enzyme that participates in the post-translational modification of proteins by hydrolysis of the beta-D-mannose residues. Genetic deletion of Manba in mouse models was associated with neurological defects and tissue deposits such as observed in the kidney.

Recent reports highlight the role of lysosomes and lysosomal dysfunction in CKD development [306]. Lysosomes are acidic organelles that play an important role in macromolecule degradation and recycling. Fabry’s disease, cystinosis and neuronal ceroid lipofuscinosis are primary genetic diseases caused by loss-of-function mutations in lysosomal enzymes [306-308]. While many lysosomal enzyme mutations present with primary central nervous system defects, these mutations are also associated with altered kidney function and ultimately lead to CKD. Renal tubules reabsorb more than 100 liters of fluid and large amounts of organic and inorganic matter to maintain homeostasis. Renal tubule cell lysosomes seem to play an important role in recycling and uptake of urinary substances [307]. Accumulation of toxic substances have been observed in cytinosis and Fabry’s disease that likely result in tubule cell death and dysfunction, further fibroblast activation, and structural defect. Further studies will be needed to study the role of Manba in renal tubule cells.

While our present study mostly focused on functional annotation of CKD GWAS hits, the
kidney-specific eQTL maps will likely be essential to the understanding of other traits linked to the kidney, for example metabolite levels and hypertension. Our work nicely illustrates the critical role of the kidney in xenobiotic metabolism as one of the key pathways identified by the eQTL analysis. The genetic regulation of drug metabolism-associated genes is likely important to understand drug effectiveness, toxicity and metabolism. Future studies shall aim to define these connections.

There are several limitations of our work, importantly the sample size. While 100 samples obtained from subjects with a single genetic background can identify genotype-driven gene-expression changes, they have a limited power to identify smaller gene-expression changes. Furthermore, the data was obtained from TCGA database with limited clinical characterization, including kidney function and structural information. Another limitation of this study is that we used the public dataset form TCGA where the kidney cortex has multiple cell types. We are currently implementing a new method, called multi-tissue meta-analysis (METASOFT) that was developed in GTEx [98] to identify tissue-specific eGenes. METASOFT borrows the statistical power from an additional 44 tissues to identify shared and tissue-specific eGenes. For future studies, we will need to perform cell-type-specific analyses with single-cell resolution.

Despite these potential deficiencies, the dataset will be critically important for interpretation of current and future GWAS studies and traits that are likely related to kidney function, such as metabolite levels and blood pressure-associated genes, to name a few. While proximity-based analysis has been popular in the past, our first pass eQTL analysis does not seem to support previously proposed CKD target genes, but highlights the potential role of a set of different genes and steers the field to refocus our attention. This analysis identified MANBA as a gene that has not previously been implicated in kidney disease development.

2.5. Materials and Methods

Subjects
99 normal human kidney cortex samples including genotype (Affymetrix Genome-Wide Human SNP Array 6.0) and RNA-seq expression profiles were obtained from The Cancer Genome Atlas (TCGA) through the TCGA Data portal at https://tcga-data.nci.nih.gov/tcga/.

**Genotype quality control and imputation**

Quality control was performed as follows: Individuals with call rate less than 95%, with ambiguous gender, were removed. Samples that have excessive and reduced proportion of heterozygosity rate were removed for DNA sample contamination or inbreeding. To further examine sample contamination, we examined genome identity-by-descent (IBD), cryptic relationships, and sample duplication. IBD was computed pairwise between all samples using genome-wide genotype data. To avoid genotyping calling error, we excluded SNPs with missing rates > 0.05, minor allele frequency (MAF) < 0.05, Hardy-Weinberg equilibrium (HWE, p-value < 1.0E-3). We then performed PCA against 1000 Genomes samples that have European ancestry and removed samples in our analyses that are deviated from the cluster. After QC, we had 96 individual samples that we used for downstream eQTL analyses.

SNPs that passed quality control were further subject to haplotype phasing using SHAPEIT2 [282], and genotype imputation using IMPUTE2 [62], with 1000 Genomes Project phase 3 haplotypes as a reference. The imputed genotypes were subjected to second round quality control with missing rates <0.05, MAF > 0.05, HWE P-value > 1.0E-6 using Plink.[309]

**Gene expression data processing**

We performed RNA-seq data QC and normalization following the standards in the GTEx project.[92] We excluded genes whose mean RNA-Seq by Expectation Maximization (RSEM) value was less than 1, and then transformed the RSEM values as follows: (1) quantile normalization across samples; (2) rank values across samples followed by mapping to normal standard (namely, inverse normal transformation); (3) probabilistic estimation of expression residuals (PEER) with available information from samples (gender and age), because these two
covariates could have an effect on gene expression; (4) transformation of the residuals of each gene to standard normal distribution $N(0,1)$ for each.

**cis-eQTL analysis, identification of eSNPs and eGenes, and network analyses**

We used 1 Mb from transcription start site (TSS) to define *cis*-eQTL window for each gene. The nominal $p$-values for the SNP-gene pairs in this window were generated using Matrix eQTL[283], which performed linear regression on the transformed residuals with the corresponding imputed genotypes. To address multiple comparisons of many SNPs in LD for a given gene, we applied the permutation procedure using FastQTL[284], which generated an empirical $p$-value for each gene. We then used the Storey approach to calculate $q$-values[285], and defined genes with a significant $q$-value ($\leq 0.05$) as eGenes. In addition, we generated a list of all significantly associated SNP-gene pairs by calculating the threshold of significant nominal $p$-value for each eGene based on the permutation method.

The eGene network analyses were generated using of QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity).

**Histone Mark and Transcription Factor Binding Site Enrichment Analysis**

Each histone tail modification marks a different regulatory region of the genome. We use this information to discover which regulatory regions the eSNPs were localized to, and to interpret what their function could be. The enrichment analysis considered the odds ratio of significant SNPs compared to the 1000 null sets of variants. The null sets of variants were matched for number, distance to TSS, and allele frequency of eSNPs. To avoid enrichment bias, we pruned the eSNPs so that no pair of SNPs had an $r^2$ greater that 0.2 using PLINK (--exclude range --indep-pairwise 50 5 0.2).

We acquired the histone marks from human kidney proximal tubule cells that we previously published[166], as well as from human kidney tissues at the Roadmap Epigenome Project (GSM621634, GSM670025, GSM621648, GSM772811, GSM621651, GSM1112806,
GWAS SNPs enrichment in kidney cis-eQTLs

To determine whether disease traits in our SNPs were enriched, we obtained eGFR-associated GWAS SNPs from the CKDGen Consortium (http://ckdgen.imbi.uni-freiburg.de) and several other studies (Table S6). This yielded 67 leading SNPs. We obtained other GWAS-identified SNPs from the National Human Genome Research Institute catalog (NHGRI) (www.ebi.ac.uk/gwas; accessed date: 9/30/2015).

For the NHGRI GWAS SNPs, we randomly sampled a matched number of SNPs (n=67) from each trait: digestive system disease, nervous system disease, immune system diseases, cardiovascular disease, biological processes, other traits, and body measurement. We took SNPs from traits that have comparable number of SNPs to kidney diseases (n=110). We expanded the list of GWAS SNPs by including tagging SNPs that are in $r^2 \geq 0.8$. We later removed SNPs that are in the same LD before we performed cis-eQTL analysis using Matrix eQTL. We set the cis-eQTL discovery p-value threshold to 1.0E-2 to compromise for the number of SNPs overlapped at each group and the weak multiple hypothesis effect (Table S1). The enrichment significance was computed using Fisher’s exact test, and we compared each trait against all other traits whether or not they are in the cis-eQTL. We ran the test for all groups as mentioned above.

Integrative statistical analysis of kidney eQTL and kidney function GWAS

To integrate the kidney eQTL and kidney function GWAS results, we applied the Bayesian colocalization approach coloc.[295, 296] In short, coloc used summary statistics to test for colocalization of GWAS and eQTL signals with the LD information. Coloc supports multiple hypotheses and can distinguish insufficient association from colocalization or distinct
associations. *Coloc* uses posterior probabilities to measure the likelihood of each hypothesis, with the prior beliefs of association and colocalization, and the data on genetic association with disease and gene expression.

The five hypotheses used in the analyses were: $H_0$: No association with either trait; $H_1$: Association with trait 1, not with trait 2; $H_2$: Association with trait 2, not with trait 1; $H_3$: Association with trait 1 and trait 2, two independent SNPs; $H_4$: Association with trait 1 and trait 2, one shared SNP. As criteria for significant results, we used both $PP_5 > 0.8$ and $PP_4/PP_3 > 5$ (posterior probability from $H_3$ and $H_4$) as criteria [297].

**Gene Expression correlation with eGFR**

Correlations for the expression of target gene with the eGFR in human kidney samples were examined using microdissected-tubule microarray data published last year from 95 human cortices.[290] Data were uploaded to Array Express and can be accessed under accession code E-MTAB-2502. Pearson’s correlation was applied to the normalized gene expression and the eGFR level in R (version 3.2.1.).

**Zebrafish functional experiment**

Zebrafish were maintained according to established Institutional Animal Care and Use Committee protocols. Briefly, we injected zebrafish embryos with Manba splice-blocking morpholino (MO, GeneTools, Philomath OR) at the one-cell stage at various doses. We compared the number of abnormal morphant embryos to control embryos, injected with a standard control MO designed by GeneTools, using the chi-square test. We documented the development of gross edema at 4 days post-fertilization in live embryos.
CHAPTER 3: Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development

3.1 Abstract

One in eleven people is affected by chronic kidney disease, a condition characterized by kidney fibrosis and progressive loss of kidney function. Epidemiological studies indicate that adverse intrauterine and postnatal environments have a long-lasting role in chronic kidney disease development. Epigenetic information represents a plausible carrier for mediating this programming effect. Here we demonstrate that genome-wide cytosine methylation patterns of healthy and chronic kidney disease tubule samples obtained from patients show significant differences. We identify differentially methylated regions and validate these in a large replication dataset. The differentially methylated regions are rarely observed on promoters, but mostly overlap with putative enhancer regions, and they are enriched in consensus binding sequences for important renal transcription factors. This indicates their importance in gene expression regulation. A core set of genes that are known to be related to kidney fibrosis, including genes encoding collagens, show cytosine methylation changes correlating with downstream transcript levels. Our report raises the possibility that epigenetic dysregulation plays a role in chronic kidney disease development via influencing core pro-fibrotic pathways and can aid the development of novel biomarkers and future therapeutics.

3.2 Introduction

CKD clinical retrospective data indicate that altered nutrient availability during development can have a long lasting effect on the development of adult diseases, a phenomenon called programming. Hypertension and chronic kidney disease (CKD) present one of the highest sensitivities to intrauterine programming [310]. Epigenetic changes caused by altered intrauterine nutrient availability have been proposed as the mechanistic link for hypertension and CKD
development [311]. Epigenetic modifications are inherited during cell division, thus solidifying “the memory or programming” effects of the environment [312].

The epigenome, which includes the covalent modifications of DNA and its associated proteins, and defines DNA accessibility to the transcriptional machinery, is the key determinant of outcome after transcription factor binding. At the root of epigenetic modification is the direct chemical modification of cytosines by methylation [313]. In different cancer types, hypermethylation of tumor suppressor gene promoters has been observed [314]. Increased promoter methylation can interfere with transcription factor binding, causing loss of tumor suppressor expression and thereby contributing to the malignant transformation [315, 316]. Agents that reduce cytosine methylation, for example azacytidine, are now in clinical use and associated with improvements in clinical outcomes, especially for patients with myelodysplastic syndrome [317]. In addition, mutations of different chromatin-modifying enzymes have been described in various cancer types, contributing to alterations in the cancer epigenome [318].

However, little was known about the epigenome of chronic human diseases as opposed to cancers. Most previous studies have been performed on cultured cells, animal models, or surrogate cell types (mostly circulating mononuclear cells) [319]. As the epigenome is cell-type-specific, little mechanistic information can be drawn from cultured cells and surrogate cell types [320]. To understand whether epigenetic changes occur and potentially contribute to CKD development in human samples, we performed genome-wide cytosine methylation profiling of tubule epithelial cells obtained from CKD and control kidneys. We found that core fibrosis-related genes show cytosine methylation changes in their gene regulatory regions. In vitro studies indicate that cytosine methylation differences play a role in regulating transcript expression. Examining the CKD epigenome can be an important first step in understanding the role of epigenetics outside the cancer field [321].
### 3.3 Results

#### 3.3.1 CKD kidneys show distinct cytosine methylation profile

For our study, human kidney samples were collected from healthy living transplant and surgical nephrectomies and categorized based on their clinical and pathological characteristics (Table 3-1). In the initial dataset, we combined samples that are hypertensive and diabetic CKD as cases, as the clinical, histological and gene expression profiles of these samples were very similar. In the replication dataset, only diabetic CKD (DKD) samples were used. In both datasets, the criteria for controls were an estimated glomerular filtration rate (eGFR) greater than 60 cc/minute/1.73m$^2$, absence of significant proteinuria, and less than 10% fibrosis on histology. Samples with significant hematuria or other signs of glomerulonephritis (HIV, hepatitis, or lupus) were excluded from the analysis. In summary, 26 samples were used for the initial discovery phase, the phenotype analysis was significant for racial diversity, and it included subjects with and without diabetes both as cases and controls (Table 3-1).

To avoid cell-type heterogeneity, we microdissected each renal cortical sample and used the tubular epithelial cell portion for the initial analysis. Our and other labs have previously published that this fraction represents mainly the proximal tubule portion of the human kidney [322]. Genome-wide cytosine methylation analysis was performed on each sample using methylation-sensitive and -insensitive isoschizomer enzymes (HpaII and MspI) followed by (HpaII) fragment enrichment by ligation-mediated PCR (HELP) [248]. Samples were hybridized on NimbleGen whole genome-covering microarrays (1.3 million loci). Focusing on loci that showed more than 50% difference in their methylation ratio and a P-value <0.01, we identified 4,751 differentially methylated regions (DMRs) between control and diseased tubule samples (Figure 3-1A; complete list provided in Appendix 3). The volcano plot analysis (fold change of methylation plotted against the negative log2 of P-value) indicated that 70% of the DMRs showed lower methylation levels in CKD (Figure 3-1A). We found that cytosine methylation differences suffice for proper clustering and supervised classification of control and CKD kidney samples (Figure 3-
The computational annotation identified a total of 1,535 unique genes in the vicinity of the DMRs.

**Table 3-1 | Demographic, clinical and histological characteristics of the samples**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Diseased</th>
<th>Healthy</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>12</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Age (years) mean ± SD</td>
<td>68.0 ± 10.81</td>
<td>61.14 ± 11.2</td>
<td>0.11</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian, Pacific Islander</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Other and unknown</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Height (cm) mean ± SD</td>
<td>165 ± 8.69</td>
<td>166.5 ± 8.63</td>
<td>0.6</td>
</tr>
<tr>
<td>Weight (kg) mean ± SD</td>
<td>78.0 ± 22.02</td>
<td>88.32 ± 15.93</td>
<td>0.2</td>
</tr>
<tr>
<td>BMI (kg/m²) mean ± SD</td>
<td>27.85 ± 6.41</td>
<td>31.25 ± 5.58</td>
<td>0.18</td>
</tr>
<tr>
<td>Diabetes</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Proteinuria (dipstick)</td>
<td>3.0 ± 1.83</td>
<td>0.36 ± 0.81</td>
<td>1.80E-04</td>
</tr>
<tr>
<td>Serum BUN (mg/dL) mean ± SD</td>
<td>35.0 ± 14.7</td>
<td>17.71 ± 5.85</td>
<td>4.60E-04</td>
</tr>
<tr>
<td>Serum creatinine (mg/dL) mean ± SD</td>
<td>3.0 ± 1.61</td>
<td>1.08 ± 0.18</td>
<td>2.00E-03</td>
</tr>
<tr>
<td>eGFR (ml/minute/1.73 m²) mean ± SD</td>
<td>29.0 ± 13.68</td>
<td>70.94 ± 8.35</td>
<td>1.06E-09</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glomerulosclerosis (%)</td>
<td>31.0 ± 31.35</td>
<td>3.31 ± 5.52</td>
<td>4.00E-03</td>
</tr>
<tr>
<td>Mesangial matrix expansion</td>
<td>1 ± 0.91</td>
<td>0.17 ± 0.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Tubular atrophy (%)</td>
<td>34.0 ± 24.94</td>
<td>9.82 ± 15.76</td>
<td>6.00E-03</td>
</tr>
<tr>
<td>Interstitial fibrosis (%)</td>
<td>34.0 ± 25.15</td>
<td>5.68 ± 5.07</td>
<td>4.00E-04</td>
</tr>
<tr>
<td>Vascular sclerosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intima</td>
<td>2.0 ± 0.78</td>
<td>0.9 ± 1.1</td>
<td>1.50E-03</td>
</tr>
<tr>
<td>Arterioles</td>
<td>2.0 ± 0.78</td>
<td>0.29 ± 0.62</td>
<td>4.00E-04</td>
</tr>
</tbody>
</table>

Gene ontology annotation showed that genes around the DMRs are enriched for cell adhesion and development-related functions including collagen, fibronectin, TGF-β and Smad proteins (Figure 3-1C), and many of these genes are known to play a critical role in CKD development. In summary, microdissected kidney tubule cells showed distinct differences in their cytosine methylation patterns in CKD.
3.3.2 Validation and external replication of the results

Internal validation of the results was performed using site-specific primer-based amplification of bisulfite-converted genomic DNA and Sequenom MassArray Epityper quantification of modified cytosines [323]. This method is based on mass spectrometry that allows us to determine absolute methylation levels. We correlated these mass array based absolute methylation levels with the HpaII/MspI relative ratios (Appendix 4).

Figure 3-1 | Statistically significant cytosine methylation differences in chronic kidney disease.

(A) Volcano plot analysis of cytosine methylation differences. The x-axis represents the relative cytosine methylation difference of control (CTL) versus CKD samples, the y-axis represents the negative log₂ of the P-value of that locus. The mean P-value and mean difference of 1.3 million loci present on the chips are plotted on the graph. The green and red lines represent the statistical criteria used for further analysis (P-value and fold change, respectively). (B) Hierarchical cluster analysis of the differentially methylated regions. Each column represents changes from one individual kidney sample; blue indicates hypermethylation in CKD, while red represents hypomethylation in CKD. The chart below shows the clinical parameters of the samples: glomerular filtration rate, diabetes status (DM, diabetes mellitus), sex, and age (aged >65 years or <65 years (C) Gene Ontology analysis of the 1,535 DMRs mapped to unique genes using DAVID gene ontology annotation groups (biological process level 1 annotation).
External validation was performed on 87 microdissected human kidney tubule epithelial samples, 21 from patients with DKD and 66 controls including hypertension (n = 22), diabetes mellitus (n = 22) or none (n = 22)) (SYH and KS, unpublished observation). Genome-wide methylation profiling of the validation set was performed using Illumina Infinium 450K methylation-sensitive bead arrays. This method uses site-specific probes for bisulfite-converted DNA, which is fundamentally different from the restriction enzyme-based method used in the HELP analysis.

From the 1,535 unique genes found around DMRs in the initial dataset, we examined 1,092, as these genes were present also on the Illumina Infinium (and Affymetrix for gene expression) arrays (Figure 3-2A). Significant methylation differences were detected for 1,061 genes, corresponding to 98% of the genes in the original dataset) (Figure 3-2A). The complete list of DMRs in the original and the replication dataset can be found in Appendix 4.

Locus-specific validation was performed for six different genes, including COLIVA1. COLIV4A1/A2 are critical basement membrane proteins synthesized by epithelial cells. Increased expression is known to be responsible for increasing the thickness of the basement membrane and it is considered to be an early change in progressive kidney fibrosis [324]. The COLIVA1 and COLIVA2 transcripts are transcribed from a single promoter (Figure 3-2B). This locus showed significantly lower cytosine methylation of CKD samples (Figure 3-2C). We examined the absolute methylation level of COLIV4A1/2 by MassArray Epityper analysis (Figure 3-2D) in control and CKD samples and confirmed the methylation differences between healthy and diseased tubule epithelial cells. Next we examined COLIVA1/2 methylation in the validation dataset (Infinium arrays from 66 control and 21 DKD samples). Using this dataset, we also confirmed the predominant (2 to 12%) hypomethylation of this locus (Figure 3-2E). The methylation differences correlated with increased COLIVA1 transcript (Figure 3-2F) and protein levels (Figure 3-2G). Using the MassArray Epityper, we also validated the methylation status of additional loci (Appendix 5). In summary, the methylation differences appear to be highly consistent between the original and validation experiments using multiple different methods.
3.3.3 Differentially methylated loci are enriched in kidney-specific gene regulatory regions

Cytosine methylation of promoters is critically important, because it can interfere with TF
binding and thereby modulate transcription (You and Jones, 2012). The number of DMRs localized to RefSeq annotated promoters and 5' UTRs was significantly lower (about 50%) than the expected ratio (Figure 3-3A). On the other hand, more than half of the DMRs were in gene body-related regions. Only a few DMRs localized to exons (approximately 200); the majority of the differences we observed were in the intronic regions (Figure 3-3A). We also examined the RefSeq annotated genomic distribution of the hypo- and hypermethylated regions (Appendix 6). The percentage of hypermethylated regions was similar in the different RefSeq-based annotation groups. We found that more loci showed increased methylation at the 3' UTR. In summary, the genomic regions that showed differences in their cytosine methylation pattern in CKD were not promoters, but intronic and transcription termination regions and 3' UTRs.

To further understand the functional significance of the DMRs, we generated genome-wide chromatin annotation maps using cultured human proximal tubular epithelial cells (HKC8). First, we performed chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) for a panel of important histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, and H3K36me3. Next, we generated gene regulatory annotation maps from the panel of ChIP-seq data using the hidden Markov model-based ChromHMM chromatin segmentation program [325, 326]. Consistent with the RefSeq-based annotation, there were very few DMRs localized to ChromHMM-annotated kidney promoter regions (Figure 3-3B). The analysis indicated that 30% of the DMRs localized to enhancer regions, which was the most significant enrichment. Similar results were obtained when we generated adult kidney cortex ChromHMM maps from published ChIP-seq data (Figure 3-3B) [1].
Next, we compared CKD-specific DMRs with chromatin annotation maps of other, different cell types using the publicly available ENCODE database (Figure 3-3C). We found that CKD-specific DMRs localized mostly to repressed chromatin regions, while transcription and enhancer regions showed the second-highest enrichment. The results indicated that DMRs in CKD were enriched in kidney-specific gene regulatory regions, mainly (intrinsic) enhancers.

Gene regulatory regions are usually characterized by DNase I hypersensitivity (DHS) [327], because DNA is usually histone-free in gene regulatory regions so transcription factors can bind to these regions. Therefore, we overlapped the DMRs with human fetal kidney and human proximal tubule epithelial cell DHS-seq data (Gene Expression Omnibus (GEO) accession
GSM530655). The statistical analysis confirmed that DMRs are enriched in DHS sites in both the fetal kidney epithelial dataset and the cultured tubule epithelial cell dataset. In addition, we examined whether DMRs that overlap with DHS sites show similarities, by identifying the top 10 consensus sequences using the MEME software package [328]. To search for TF binding motifs among the top 10 sequences, we mapped the sequences to the JASPAR, UniProbe and Transfac databases [329]. This analysis highlighted that the DMRs contain consensus-binding sequences for transcription factors that play important roles in proximal tubule development, including SIX2, HNF, and TCFAP. The list of computationally identified transcription factor consensus motifs is shown in Figure 3-4A.

Figure 3-4 | Chronic kidney disease differentially methylated regions are enriched for kidney-specific transcription factor binding sites. (A) The DMR and DHS sites contain consensus sequences. The transcription factor binding site motifs and their statistical enrichment from the de novo searched consensus sequences in DMR and DHS sites. (B) A specific example of an intronic DMR (within the EZR gene). The genomic location of the DMR is at the top, followed by the RefSeq representation of EZR; fetal kidney (FK)-specific DHS tracks (in blue); HKC8 cell-specific H3K4me1 and H3K4me3 tracks; HKC8 cell specific ChromHMM annotation of the locus (yellow, enhancer; red, promoter; green, transcription-associated region; the full color coding key is shown bottom right) - the sequences contain consensus-binding sites for the key kidney transcription factor SIX2/3, with the SIX2/3 binding motif illustrated as a sequence logo plot below; and adult kidney (AK)-specific H3K4me1 (blue) and H3K4me3 (green) tracks [1].
Figure 3-4B illustrates our motif analysis. Here a DMR is located in the intronic region of the EZR (ezrin) gene. The DMR overlapped with adult kidney and renal tubular epithelial cell-specific H3K4me1 histone modification, but not with H3K4me3 enrichment. H3K4me1 is a specific histone tail modification for enhancer regions, while H3K4me3 is a marker of promoters. The ChromHMM-based gene regulatory region annotation confirmed that this region is an enhancer in renal epithelial cells (yellow region in the genome).

These results indicate that by multiple different approaches this DMR is located in a gene regulatory region, an enhancer (Figure 3-4B). In addition, this region contained a consensus-binding sequence for SIX2, further confirming that this is a gene regulatory region. In summary, our results indicate that CKD-specific DMRs are located in non-promoter gene regulatory regions, mainly enhancers, and contain consensus-binding motifs for renal-specific transcription factors.

3.3.4 Differentially methylated regions are functionally relevant and correlate with transcript levels

To study the functional significance of DMRs, we first, we examined whether they correlate with downstream transcript levels. Gene expression changes were analyzed using RNA samples extracted from the same microdissected tubule samples used in the methylation assay. Individual RNA samples were hybridized to Affymetrix U133 arrays and the data were normalized and analyzed using established pipelines [322]. From the 1,092 transcripts in close proximity to the DMR regions (Figure 3-5A), we identified 415 (approximately 40%) genes showing significantly differential expression in the CKD samples (Figure 3-5A).
As most DMRs were in non-promoter regions, most transcript changes correlated with intronic DMRs (Supplementary file 8). Gene ontology and network analyses highlighted differences in cell adhesion (collagens and laminins) and development-related pathways (Figure 3-5B, C). Specifically, we observed significant enrichment for differential expression and methylation in the TGF-β pathway, especially in TGFBR3, SMAD3, SMAD6 and other targets.

**Figure 3-5 | Differentially methylated regions correlate with transcript changes.**

(A) The 4,751 DMRs mapped to 1,092 unique genes that were present in the Affymetrix arrays. There were 415 transcripts that showed differences both in their methylation status and their expression in CKD samples. The RefSeq-based locations of the DMRs are also shown. While most differentially methylated regions localize to gene body regions, they also show correlation with the expression of many of those genes. Not only are the 415 transcripts differentially expressed, they also show differences in their cytosine methylation profiles as well. (B) DAVID-based gene ontology annotation of the 415 transcripts. (C) Network chart of the genes that are both differentially expressed and methylated (pathways related to development are highlighted in orange). (D) Methylation and gene expression level of key molecules (*RUNX3, RARB, SMAD6*) identified by the network analysis in control (CTL) and CKD samples.
(Figure 3-5D). These genes are known to be critical in CKD development [330, 331]. In summary, cytosine methylation changes showed correlation with gene expression differences and identified concordant changes in the TGF-β pathway, a well-known regulator of kidney fibrosis development.

3.3.5 Methylation in differentially methylated regions affect gene expression in vitro

To further examine the relationship between cytosine methylation and transcript level changes, we analyzed gene expression and cytosine methylation patterns of tubule epithelial cells at both baseline and 9 days after treatment with a DNA methyltransferase inhibitor, decitabine (5-aza-2-dexoycytidine). We used AffymetrixST1.0 arrays to compare gene expression changes and the Infinium 450K arrays to examine cytosine methylation changes in control (n = 3) and decitabine-treated cells (n = 4).

We tested whether there is correlation between in vivo (CKD) and in vitro (decitabine treatment) of the DMR and gene expression changes. Decitabine is a cytosine analogue; therefore, we hypothesize that after decitabine treatment, the cytosine methylation changes are

<table>
<thead>
<tr>
<th>GOTERM/DMR (5mC)</th>
<th>p-value</th>
<th>GOTERM/Transcripts (RNA)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cellular process</td>
<td>5.90E-24</td>
<td>developmental process</td>
<td>8.90E-03</td>
</tr>
<tr>
<td>developmental process</td>
<td>1.20E-15</td>
<td>cellular component biogenesis</td>
<td>4.30E-02</td>
</tr>
<tr>
<td>localization</td>
<td>9.00E-10</td>
<td>multi-organism process</td>
<td>2.90E-02</td>
</tr>
<tr>
<td>establishment of localization</td>
<td>1.10E-09</td>
<td>death</td>
<td>2.90E-02</td>
</tr>
<tr>
<td>biological adhesion</td>
<td>7.00E-08</td>
<td>biological adhesion</td>
<td>1.20E-01</td>
</tr>
<tr>
<td>cellular component organization</td>
<td>9.30E-08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3-6 | Regulation of transcripts by a DNA methyltransferase inhibitor in in vitro cultured human tubular epithelial cells.

Gene ontology terms of transcripts showing differential expression in the decitabine-treated cells. (A) Illustration of regions that showed differential methylation of cultured HKC8 cells treated with 0.5 µM decitabine (5'DAC), CTL, control. (B) The interconnected network analysis highlighted the differential expression of cell adhesion and developmental pathways. These genes are also differentially expressed and methylated in the original CKD dataset. GO, gene ontology.
the primary cause of transcript level changes. A limitation of the experiment is that decitabine induces demethylation of genomic loci that could be different from the CKD DMRs. Large numbers of loci showed concordant differential methylation and gene expression changes in CKD in vivo as well as after (0.5 µM) decitabine treatment in vitro, indicating that cytosine methylation changes in CKD might be the functional drivers of transcript level changes (Supplementary file 9).

Genes related to cell adhesion (for example, collagen molecules) showed differential methylation following decitabine treatment as well (Figure 3-6A, B). In addition, just as we observed before, we found that genes related to development and cell adhesion were also differentially expressed following decitabine treatment (Figure 3-6B, C).

SMAD3 appears to be one of the most important mediators of the pro-fibrotic effect in the TGF-β and angiotensin II pathways [331]. The SMAD3 locus contained DMRs in both the initial and validation datasets. SMAD3 expression levels were lower in both the original and confirmation datasets. Decitabine changed the methylation of this locus and subsequently also changed SMAD3 transcript levels.

To illustrate our findings, while RUNX1 clearly plays an important role in leukemia development, it is expressed in both mice and humans in the developing and adult kidneys [332]. RUNX1 was also shown to be differentially expressed in CKD tubules [322]. Both the original and replication dataset showed differential methylation of this locus (Figure 3-7A, C) and RUNX1 transcript levels were increased in both datasets (Figure 3-7B, D). RUNX1 DMRs clustered in ChromHMM annotated enhancer regions (Figure 3-7G) (with H3K4me1 and DHSs) in vitro treatment with decitabine changed the cytosine methylation of this locus, and the changes overlapped with the enhancer DMRs (Figure 3-7E).

Subsequent to the DMR change of this locus, we also observed an increase in RUNX1 transcript levels both in vivo and in vitro (Figure 3-7F). As decitabine did not change the methylation of the RUNX1 promoter and affected only the methylation levels of the enhancer site, the result potentially indicates a causal relationship between enhancer-related DMRs and gene...
expression changes. The concordant changes in cytosine methylation and gene expression in CKD and in vitro (following DNA methyltransferase inhibitors) indicate that DMRs are potential drivers of critical CKD gene expression.

**Figure 3-7 | Gene body cytosine methylation changes drive gene expression differences.**

*RUNX1* methylation and gene expression were examined. (A,B) In the original discovery dataset, the gene body region of *RUNX1* was hypomethylated (A) and the corresponding transcript level was increased (B) in the CKD (discovery) dataset. CTL, control. (C,D) The differential methylation (C) and expression (D) of *RUNX1* in the DKD replication dataset. (E,F) Transcript levels are increased (F) in vitro in cultured tubules after decreasing the methylation level of the locus following 0.5 µM decitabine (DAC) treatment (E). (G) Genomic representation of the *RUNX1* locus showing DMRs in the DKD dataset and in the CKD dataset. Different tracks are shown for the *RUNX1* locus, including RefSeq gene, DMRs in the DKD dataset, DMRs in the CKD cells, and histone ChIP-seq data for H3K4me1 and H3K4me3 for adult kidney cortex and DHS sites from fetal kidneys. In addition, ENCODE-based transcription factor binding sites are also shown.

### 3.4 Discussion and Conclusion

While epigenetic dysregulation has been suggested as a mechanism for the
development of many diseases, little is known about the epigenome of normal and diseased human cells and organs. Here we describe cytosine methylation differences in tubule cells obtained from patients with CKD. We found that CKD DMRs have many special features. First, most loci showed consistent cytosine methylation differences in CKD. These changes were smaller than what has been described in the cancer literature previously. While the absolute differences were modest, the identified loci showed highly consistent changes even across different datasets and platforms. Unexpectedly, we found that most methylation differences were localized outside of promoter areas, with promoter regions markedly spared from cytosine methylation differences.

Our results indicate that the differentially methylated regions were located mainly at candidate enhancers. We found that the DMRs contain consensus-binding motifs for key renal transcription factors (HNF, TCFAP, SIX2). Furthermore, cytosine methylation levels correlated with baseline gene expression changes. These epigenetically distinct but morphologically similar cells also showed differences in their cytokine response. We illustrated our findings in a model hypothesizing that enhancer DMRs might modify TF binding and thereby modify downstream transcript levels. Based on our results, we propose that cytosine methylation changes are causally linked to transcript levels and phenotype development.

As hypertensive and diabetic tubule samples showed similarities both in cytosine methylation and gene expression changes, the observed changes are likely to be part of a common mechanism of progression. This may be expected, as phenotypically the tubulointerstitial fibrosis of DKD and hypertensive CKD is similar. In addition, we found that DMRs were enriched for genes related to kidney development, many of them no longer expressed in the adult kidney. The DMR regions also contained binding sites for key kidney developmental factors such as SIX2, HNF, and TCFAP.

One possible interpretation of our findings is that the epigenetic differences are established during development. This is the time when the cell-type-specific epigenome is
established and these genes and transcription factors play functional roles. Therefore, they can possibly provide the mechanistic link between fetal programming and CKD development - the Brenner-Barker hypothesis put forward many decades ago [333, 334], proposed that nutrient availability during development has a longlasting programming role in hypertension and CKD development.

In addition, reactivation of the developmental pathways is also needed during organ injury repair [335]. We can also speculate that the altered developmental wiring of these pathways continues to play a role later on as alterations observed after repair. Indeed, control and CKD kidney epithelial cells showed not only cytosine methylation differences but also different responses to cytokine treatment.

A limitation of our results remains that our samples were collected in a single center. Furthermore, base pair resolution results will likely help to refine the more precise location of DMRs and the methylation differences in the future. While microdissection is an excellent separation method to generate a homogenous tubular epithelial cell population from the kidney, the potential risk for increased cell type heterogeneity in CKD remains. As isolated and cultured cells continued to show many of the epigenetic and transcriptional differences, it is more likely that the observed differences are not related to cell type heterogeneity. In summary, while it has long been speculated that epigenetic dysregulation might occur in non-cancerous diseases, including CKD, here we provide experimental evidence for cytosine methylation changes in human kidney tissue samples, opening the possibility that they play a role in CKD development.

3.5 Materials and Methods

Ethics statement

The clinical study used the cross-sectional design. Kidney samples were obtained from routine surgical nephrectomies. Samples were de-identified and the corresponding clinical information was collected by an individual who was not involved in the research protocol. The study was approved by the Institutional Review Boards of the Albert Einstein College of Medicine
Montefiore Medical Center (IRB#2002-202) and the University of Pennsylvania. Histological analysis was performed by an expert pathologist (IRB#815796).

**Tissue handling and microdissection**

Tissue was placed into RNAlater and manually microdissected at 4C for glomerular and tubular compartments as described earlier. Dissected tissue was homogenized and RNA was prepared using RNAeasy mini columns (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. RNA quality and quantity was determined using Lab-on-Chip Total RNA PicoKit (Agilent BioAnalyzer, Santa Clara, CA, USA). Only samples without evidence of degradation were used. Genomic DNA was extracted by phenol chloroform protocol for HELP analysis and the DNAeasy kit was used for the Infinium platform.

**DNA methylation analysis by HELP**

The HELP assay was carried out as previously published [336]. Intact DNA of high molecular weight was corroborated by electrophoresis on 1% agarose gels in all cases. One microgram of genomic DNA was digested overnight with either HpaII or MspI (NEB, Ipswich, MA, USA). The digested DNA was used to set up an overnight ligation of the HpaII adapter using T4 DNA ligase. The adapter-ligated DNA was used to carry out the PCR amplification of the HpaII- and MspI-digested DNA as previously described [248]. Both amplified fractions were submitted to Roche NimbleGen Inc. (Madison, WI, USA) for labeling and hybridization onto a human hg18 high-density custom-designed oligonucleotide array (50-mers) containing 2.6 million loci. HpaII-amplifiable fragments are defined as genomic sequences contained between two flanking HpaII sites found within 200 to 2,000 bp of each other. All microarray hybridizations were subjected to extensive quality control using the following strategies. First, uniformity of hybridization was evaluated using a modified version of a previously published algorithm [337] adapted for the NimbleGen platform, and any hybridization with strong regional artifacts was discarded and repeated. The raw data can be accessed under GSE49557.

**HELP data processing and analysis**
Signal intensities at each HpaII amplifiable fragment were calculated as a robust (25% trimmed) mean of their component probe-level signal intensities. Any fragments found within the level of background MspI signal intensity, measured as 2.5 mean absolute differences (MAD) above the median of random probe signals, were categorized as “failed”. The failed loci therefore represent the population of fragments that did not amplify by PCR, whatever the biological or experimental cause (for example, genomic deletions and other sequence errors). On the other hand, “methylated” loci were so designated when the level of HpaII signal intensity was similarly indistinguishable from background. PCR-amplifying fragments (those not flagged as either “methylated” or “failed”) were normalized using an intra-array quantile approach wherein HpaII/MspI ratios are aligned across density-dependent sliding windows of fragment size-sorted data. The log2 (HpaII/MspI) was used as a representative for methylation and analyzed as a continuous variable. For most loci, each fragment was categorized as either methylated, if the centered log HpaII/MspI ratio was less than zero, or hypomethylated if the log ratio was greater than zero.

Statistical analysis of HELP data was performed using the statistical software R version 2.13.1 [337]. A two-sample t-test was used for each gene or locus to summarize methylation differences between the two clinical groups (cases and controls). Genes were ranked based on the magnitude of this test statistic and a set of differentially methylated loci with P-value <0.01 and a fold change >0.5 was identified.

**Quantitative DNA methylation analysis by MassArray epityping**

Validation of HELP microarray findings was carried out by matrix-assisted laser desorption/ionisation-time of flight (MALDI-TOF) mass spectrometry using EpiTyper by MassArray (Sequenom, San Diego, CA, USA) on bisulfite-converted DNA as previously described (Zhou, Opalinska et al. 2011). MassArray primers were designed to cover the flanking HpaII sites for a given HpaII-amplifiable fragments (HAF), as well as any other HpaII sites found up to 2,000 bp upstream of the downstream site and up to 2,000 bp downstream of the upstream
site, in order to cover all possible alternative sites of digestion. HAF is defined by those fragments where two HpaII sites are located 200-2000 bp apart with at least some unique sequence between them and selected those located at gene promoters and imprinted regions.

**Gene expression analysis using Affymetrix arrays**

Transcript levels were analyzed using Affymetrix U133A and 1.0ST arrays. Probes were prepared using an Affymetrix 3’ IVT kit. After hybridization and scanning, raw data files were imported into Genespring GX software (Agilent Technologies). Raw expression levels were normalized using the RMA16 summarization algorithm. Genespring GX software was then used for statistical analysis; the data were above the 20th percentile when filtered by expression. We used a Benjamini-Hochberg multiple testing correction with a P-value <0.05. Heatmaps of methylation data and gene expression data were generated using an unsupervised hierarchical clustering method calculated by squared Euclidean distances. Methylation data used in clustering have a P-value <0.00015 and a fold change ≥0.5. The raw data can be accessed through accession GSE48944.

**Gene ontology and transcription factor binding sites**

The Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics package was used for gene ontology and pathway analysis. In addition, Ingenuity Pathway Analysis (IPA, Redwood City, CA, USA) was used to generate networks.

Sequences of DMRs (n = 4,751) were lifted over from hg18 to hg19 using UCSC Genome Browser Utilities. The regions were then intersected with fetal kidney or human kidney epithelial-specific DHS peaks (data from GEO GSM530655); a total of 364 overlapping regions were used. Motif weight matrices overrepresented in the overlapped sequences were identified using MEME version 4.8.0 [328] on the 364 regions with parameter -oc -nmotifs 10 -minw 8 -maxw 50.
Adult kidney ChIP-seq data were acquired from the Roadmap database (GEO accessions number GSM670025, GSM621638). The overlap was set to be a minimum of 1 bp in length.

**Motif searching**

We compared de novo motifs to motifs available as part of various databases, including Transfac, version 2011.1, Jaspar Core, and UniPROBE using TOMTOM software [329], version 4.8.1. TOMTOM parameters were set to their default values during motif comparisons. When partitioning the de novo motifs, assigning each to a single category, the order of match assignment preference was to Transfac, Jaspar Core, UniPROBE, and then to the novel motif category.

**Cell lines**

HKC8 cells were kindly provided by Lorainne Racusen (Johns Hopkins University) and were cultured in DMEM/F12 medium supplemented with 2.5% fetal bovine serum, antibiotics and insulin, transferrin and selenium. Cells were incubated with 0.5 µM decitabine on days 2, 4, 6, and 8 and harvested on day 9. RNA was isolated using a Qiagen RNeasy kit labeled using an Ovation transcript labeling kit and hybridized onto Affymetrix Human ST1.0 arrays.

**Chromatin immunoprecipitation sequencing**

HKC8 cells were harvested and crosslinked with 1% formaldehyde when they reached 80% confluency on culture plates. Chromatin was sheared using a Bioruptor and immunoprecipitated using H3K4me1 (Abcam ab8895, Cambridge, MA, USA), H3K4me2 (Abcam ab11946), H3K4me3 (Abcam ab8580), H3K36me3 (Abcam ab9050), H3K27ac (Abcam ab4729) and H3K27me3 (Millipore 07-499, Billerica, MA, USA) marks. ChIP was performed as described in the manual of MAGnify™ Chromatin Immunoprecipitation System (Invitrogen, Grand Island, NY, USA). Quantitative real-time PCR was performed to ensure the quality of the ChIP product. The ChIP product was assessed for size, purity, and quantity using an Agilent 2100 Bioanalyzer (Agilent Technologies). Library preparation and sequencing were performed at the Einstein
Epigenome Center. Sequence reads (100 bp) were generated from Illumina HiSeq 2000 [338]. Reads were aligned to the reference genomes (NCBI build 37, hg19) using Bowtie (v 0.12.7) [339]. Repetitively mapped and duplicate reads were excluded. The data can be accessed using accession GSE49637.

**ChIP-seq data analysis**

We used the MACS version 1.4.1 (model-based analysis of ChIP-Seq) peak-finding algorithm to identify regions of ChIP-Seq enrichment over background [340]. A false discovery rate threshold of enrichment of 0.01 was used for all datasets. The resulting genomic coordinates in BED format were further used in ChromHMM v1.06 for chromatin annotation. The following parameters were used: -Xmx1600M -jar ChromHMM.jar BinarizeBed hg19 -Xmx2000M -jar ChromHMM.jar LearnModel 10 hg19.

**DNase I hypersensitive site analysis**

Human kidney DHS sequencing data (GEO GSM530655) was analyzed with MACS (v.1.4.1). The resulting peaks were overlapped with the differentially methylated regions. The control random genomic loci were generated using Regulatory Sequence Analysis Tools [341]. Based on the data property of differentially methylated regions, we used the same number of fragments (4,751) and the same average fragment size (443 bp) as parameters for the random loci.

**Illumina infinium 450K BeadChip arrays**

Genomic DNA (200 ng) was purified using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. Purified DNA quality and concentration were assessed with a NanoDrop ND-1000 (Thermo Fisher Scientific, Waltham, MA, USA) and by Quant-iT™ PicoGreen® dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) prior to bisulfite conversion. Purified genomic DNA was bisulfite converted using the EZ DNA Methylation Kit (Zymo Research, Orange, CA, USA) following the manufacturer’s protocol. Bisulfite DNA quality and concentration were assessed, following the Illumina 450K array protocol, the bisulfite
converted sample was whole-genome amplified, enzymatically digested, and hybridized to the array, and then single nucleotide extension was performed.

Chips were scanned using an Illumina HiScan on a two-color channel to detect Cy3-labeled probes on the green channel and Cy5-labeled probes on the red channel. Illumina GenomeStudio Software 2011.1 Methylation Module 1.8 was used to read the array output and conduct background normalization. The level of DNAm for 428,216 probes in our sample dataset was intersected with the expanded annotation for further analyses. All samples were run together to eliminate the batch effect according to the pipelines established by Illumina Genome Studio. The full dataset can be accessed in GEO under GSE50874.
CHAPTER 4: Conclusion

4.1. Summary of results and discussion

The work described in the dissertation aimed to understand genetic and epigenetic factors related to CKD and CKD development. Our eQTL analysis opens the door to considerable future research in gene regulation in the human kidney. While a human-body eQTL map has been generated through GTEx, a kidney-specific eQTL map was not available through the public database due to organ allocation and transplantation. In addition, published results showed that eQTL performed from imputed gene expression for unavailable tissue types such as the human kidney does not have enough power to detect tissue-specific eQTLs [305]. For these reasons, our kidney eQTL analysis using RNA sequencing-based gene expression in human kidney samples is an important resource for the research community.

The eQTL results showed that kidney eSNPs are localized to kidney-specific regulatory regions: promoters, enhancers and TF binding sites, which suggested regulatory functions for these eSNPs. We used the kidney eSNPs to interpret CKD-associated SNPs using both direct overlap and Bayesian colocalization methods. We identified a short list of genes that are likely responsible for the association observed between genotype and kidney disease traits in large-scale GWAS. Among the 9 identified genes, a few target genes have presented with kidney phenotypes in knockout transgenic mice, including Cystatin 9 (CST9) and PGAP3.

These examples validated that the colocalization method can identify kidney-relevant target genes. We validated one of the target gene, MANBA, as a likely causal gene for genetic variants on chromosome 4 that are associated with CKD. We found that the risk allele is associated with lower MANBA gene expression level in human kidney samples.

From zebrafish experiments, we observed that lower MANBA expression is causally related to kidney disease development. Morpholino knockdown of Manba resulted in pericardial edema, indicating a functional role of MANBA in kidney disease development. MANBA is a lysosomal enzyme that participates in the post-translational modification of proteins by hydrolysis.
of the beta-D-mannose residues. Manba knockout mice were associated with neurological defects and presented with cellular deposits in the kidney.

Lysosomal dysfunction in CKD development has been observed through several loss-of-function mutations in lysosomal enzymes that causes genetic diseases with altered kidney function and ultimately lead to CKD. For example, accumulation of toxic substances has been observed in Fabry’s disease and cystinosis that likely result in tubule cell death and dysfunction, further fibroblast activation, and structural defect. Further studies will be needed to study the role of Manba in kidney tubule cells.

As GWAS explain little of the heritability of CKD, another possible mediator could be epigenetics. Little was known about the epigenome of normal and diseased human kidneys when we performed our cytosine methylation experiments and analysis in 2012. We were the first researchers to describe genome-wide cytosine methylation differences in tubule cells obtained from patients with CKD.

We found that while DMRs associated with CDK do not have the scale of difference observed in cancer studies, the differences are very stable and consistent in patient samples across different sample cohorts and different assay platforms. To our surprise, we found that CKD DMRs are depleted from promoter regions and enriched in enhancer regions. These enhancer regions are enriched with consensus-binding motifs for key renal transcription factors (HNF, TCFAP, SIX2). We observed consistent methylation and gene expression correlation in different cohorts as well.

We illustrated our findings in a model hypothesizing that enhancer DMRs modify TF binding and thereby modify downstream transcript levels. Based on our results, we propose that cytosine methylation changes are causally linked to transcript levels and phenotype development.

4.2. Future direction

Our present study mostly focused on functional annotation of CKD GWAS variants, but
kidney-specific eQTL maps would be useful to the understanding of other phenotypes that are associated with kidney, including metabolite levels and hypertension. Our work illustrates the critical role of the kidney in xenobiotic metabolism as one of the key pathways identified by the eQTL analysis. The genetic regulation of drug metabolism-associated genes is likely important to understand drug effectiveness, toxicity and metabolism. Further experiments are needed to validate these connections.

One remaining question for eQTL study is how to identify the causal variant among a cluster of variants that all present similar significance association with the target gene. To perform the experimental validation and to understand how many variants at each locus function together, a high-throughput method including massively parallel reporter assay (MPRA) [342], or a modified version that shows improved throughput and sensitivity [343], can perform systematic examination. For in vivo analysis, expression of the identified regulatory regions in zebrafish can elucidate the cell-type specificity of the regions. This can be performed by attaching the regulatory regions to the green fluorescence protein construct, and then using the transposon system to integrate the sequence to the zebrafish genome.

There remain some limitations on our work, importantly sample size. While 100 samples obtained from subjects with a single genetic background can identify genotype-driven gene-expression changes, they have a limited power to identify smaller gene-expression changes. Furthermore, the genotyping and RNA-seq data were obtained from the TCGA database with limited clinical characterization of kidney function and structural information. Greater sample size as well as clear clinical parameters will increase the power to detect kidney-specific gene expression changes associated with variants.

For the cytosine methylation study, we also had the limitation that our sample size was small and the samples were collected in a single center. In addition, our analysis was performed using array-based method, but in order to refine the DMR location, we can apply bisulfite sequencing to the samples, which will increase the resolution (location and methylation level).
With current genotyping results available for some of the human kidney samples, it is also possible to perform methylation QTL. With the advance of experimental tools, site-specific manipulation will allow us to modify DNA methylation and differentiate the functional consequences using CRISPR/Cas9 system [344].

For both studies, we encountered issues arising from the cell type heterogeneity. Although the majority of the cell types in the cortex are proximal, distal and convoluted tubules, there are other cell types in the cortex, including mesangial cells, podocytes, and endothelial cells. Ideally, we would have to perform microdissection and separate the tubule and glomeruli compartments, but it was not feasible using the publicly available dataset that we used to perform eQTL analysis. Future studies using single-cell resolution transcriptome analysis in the kidney will be essential to define the transcriptome signature of different cell types and the expression of cell-type-specific genes.

In summary, an integrative analysis performed using clinical parameters, genotype, gene expression and methylation datasets may advance the identification of causal genomic loci that regulate gene expression and cell phenotypes involved in the development of CKD.
APPENDIX

Appendix 1 | List of CKD associated SNPs. List of SNPs curated manually from all kidney-function-associate GWAS.

<table>
<thead>
<tr>
<th>Gene Symbols</th>
<th>SNP</th>
<th>Chr</th>
<th>Pos (b37)</th>
<th>Freq</th>
<th>Effect Size 95%CI/SE</th>
<th>p-value</th>
<th>Referencer</th>
<th>Association parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYPL2</td>
<td>rs1213606</td>
<td>3</td>
<td>110014170</td>
<td>0.3</td>
<td>-0.0045</td>
<td>0.0008</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>SDCCAG8</td>
<td>rs2802729</td>
<td>1</td>
<td>243501763</td>
<td>0.43</td>
<td>-0.0046</td>
<td>0.0008</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>SARS, CELSR2, PSRC1, MYBPHL, SORT1, PSMA5, SYPL2, ATXN7L2, CYB561D1, AMIGO1, GPR61, GNAI3, AMPD2, GSTM2, GSTM4, GSTM1, GNT1</td>
<td>rs1933182</td>
<td>1</td>
<td>109999588</td>
<td>0.3</td>
<td>-0.008</td>
<td>0.001</td>
<td>Kottgen 2010</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>VAV3</td>
<td>rs1701960 2</td>
<td>1</td>
<td>108188858</td>
<td>0.19</td>
<td>1.17</td>
<td>1.11-1.24</td>
<td>Kyr'yuk 2014</td>
<td>IgA-NP</td>
</tr>
<tr>
<td>CFHR1,3</td>
<td>rs6677604 1</td>
<td>1</td>
<td>19668918</td>
<td>0.073</td>
<td>0.68</td>
<td>2.96E-10</td>
<td>Gharavi 2011</td>
<td>IgA-NP</td>
</tr>
<tr>
<td>CACNA1S</td>
<td>rs3850625 1</td>
<td>1</td>
<td>201016296</td>
<td>0.12</td>
<td>0.0083</td>
<td>0.0013</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>MTX1-GBA</td>
<td>rs2049805 1</td>
<td>1</td>
<td>153461604</td>
<td>0.17</td>
<td>0.0072</td>
<td>0.0013</td>
<td>Kottgen 2010</td>
<td>BUN</td>
</tr>
<tr>
<td>ANXA9/LASS2</td>
<td>rs267734 1</td>
<td>1</td>
<td>150951477</td>
<td>0.2</td>
<td>0.01</td>
<td>0.002</td>
<td>Kottgen 2010</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>DNAJC16</td>
<td>rs1212407 8</td>
<td>1</td>
<td>15869899</td>
<td>0.3</td>
<td>0.06</td>
<td>0.0011</td>
<td>Pattaro 2012</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>SOX11</td>
<td>rs1686417 2</td>
<td>1</td>
<td>5907880</td>
<td>0.05</td>
<td>0.011</td>
<td>0.003</td>
<td>Kottgen 2012</td>
<td>CKD</td>
</tr>
<tr>
<td>DDX1</td>
<td>rs6431731 2</td>
<td>1</td>
<td>15863002</td>
<td>0.06</td>
<td>0.0127</td>
<td>0.0023</td>
<td>Pattaro 2012</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>CDCA7</td>
<td>rs4972593 2</td>
<td>1</td>
<td>174462854</td>
<td>0.14</td>
<td>1.18</td>
<td>1.47-2.24</td>
<td>Sandholm 2013</td>
<td>T1DK- FemaleOnly</td>
</tr>
<tr>
<td>LRP2</td>
<td>rs4667594 2</td>
<td>1</td>
<td>170008506</td>
<td>0.47</td>
<td>0.0044</td>
<td>0.0008</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>AFF3, TSC22D4, NYAP1</td>
<td>rs7583977 2</td>
<td>1</td>
<td>100460654</td>
<td>0.289</td>
<td>1.29</td>
<td>1.18-1.40</td>
<td>Sandholm 2012</td>
<td>T1-ESRD</td>
</tr>
<tr>
<td>CPS1</td>
<td>rs7422339 2</td>
<td>1</td>
<td>211540507</td>
<td>0.32</td>
<td>1.11</td>
<td>1.07-1.15</td>
<td>Pattaro 2015</td>
<td>CKD</td>
</tr>
<tr>
<td>RND3</td>
<td>rs7560163 2</td>
<td>1</td>
<td>151637936</td>
<td>0.14</td>
<td>0.75</td>
<td>0.68-0.87</td>
<td>Palmer 2012</td>
<td>T2-ESRD</td>
</tr>
<tr>
<td>PAX8</td>
<td>rs1112317 2</td>
<td>1</td>
<td>113694511</td>
<td>0.35</td>
<td>0.0551</td>
<td>0.0006</td>
<td>Okada 2012</td>
<td>BUN</td>
</tr>
<tr>
<td>IGFBP5</td>
<td>rs2712184 2</td>
<td>1</td>
<td>217682779</td>
<td>0.42</td>
<td>0.0053</td>
<td>0.0009</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>ALMS1/NAT8</td>
<td>rs13538 2</td>
<td>1</td>
<td>73868328</td>
<td>0.23</td>
<td>0.009</td>
<td>0.002</td>
<td>Kottgen 2010</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>GCKR</td>
<td>rs1260326 2</td>
<td>1</td>
<td>27730940</td>
<td>0.41</td>
<td>0.009</td>
<td>0.001</td>
<td>Kottgen 2010</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>PLA2R1</td>
<td>rs4664308 2</td>
<td>1</td>
<td>160917497</td>
<td>0.43</td>
<td>0.43359649</td>
<td>0.38-0.51</td>
<td>Stanescu 2011</td>
<td>IMN</td>
</tr>
<tr>
<td>WNT7A</td>
<td>rs795744 3</td>
<td>1</td>
<td>13906850</td>
<td>0.15</td>
<td>0.006</td>
<td>0.0011</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>SKIL</td>
<td>rs9682041 3</td>
<td>1</td>
<td>170091902</td>
<td>0.13</td>
<td>0.0068</td>
<td>0.0012</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>ETV5</td>
<td>rs1051380 3</td>
<td>1</td>
<td>185822353</td>
<td>0.13</td>
<td>-0.0072</td>
<td>0.0012</td>
<td>Pattaro 2015</td>
<td>eGFRcrea</td>
</tr>
<tr>
<td>ST6GAL1</td>
<td>rs7634389 3</td>
<td>1</td>
<td>186738421</td>
<td>0.44</td>
<td>1.13</td>
<td>1.09-1.18</td>
<td>Li 2015</td>
<td>IgA-NP</td>
</tr>
<tr>
<td>Gene</td>
<td>rsID</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
<td>Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAP1/2/Psm8/2</td>
<td>rs347685</td>
<td>141807137</td>
<td>0.28</td>
<td>0.009</td>
<td>3.00E-11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mecom</td>
<td>rs1685372</td>
<td>170633326</td>
<td>0.29</td>
<td>0.0064</td>
<td>2.70E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lrig1-Kbtbd8</td>
<td>rs1306900</td>
<td>66881640</td>
<td>0.17</td>
<td>0.0093</td>
<td>1.40E-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bcl6-Lpp</td>
<td>rs1093732</td>
<td>189196412</td>
<td>0.69</td>
<td>0.01</td>
<td>8.80E-30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fam47e,</td>
<td>rs1314635</td>
<td>77412140</td>
<td>0.21</td>
<td>-0.0117</td>
<td>6.60E-11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hla-Dpb2</td>
<td>rs1883414</td>
<td>33086448</td>
<td>0.24</td>
<td>0.78</td>
<td>4.84E-09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hla-Dqa/A</td>
<td>rs1794275</td>
<td>32671248</td>
<td>0.15</td>
<td>1.3</td>
<td>3.43E-13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vegfa</td>
<td>rs881858</td>
<td>43806609</td>
<td>0.28</td>
<td>0.011</td>
<td>9.40E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hla-Dqb1</td>
<td>rs9275424</td>
<td>32702799</td>
<td>NA</td>
<td>NA</td>
<td>4.00E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hla-Dqa1</td>
<td>rs1129740</td>
<td>32609105</td>
<td>0.49</td>
<td>2.11</td>
<td>2.14E-19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hla-Drb1</td>
<td>rs660895</td>
<td>32577380</td>
<td>0.24</td>
<td>1.34</td>
<td>4.13E-20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rnf32</td>
<td>rs645680</td>
<td>15625856</td>
<td>0.26</td>
<td>0.0055</td>
<td>1.07E-09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncx</td>
<td>rs1027711</td>
<td>1285195</td>
<td>0.35</td>
<td>-0.0058</td>
<td>1.00E-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prkag2</td>
<td>rs780574</td>
<td>1514708</td>
<td>0.24</td>
<td>1.13-1.25</td>
<td>4.20E-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stc1</td>
<td>rs1010941</td>
<td>23753115</td>
<td>0.42</td>
<td>-0.008</td>
<td>1.00E-08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klf10/Odf1</td>
<td>rs2933562</td>
<td>103547739</td>
<td>0.48</td>
<td>0.88495575</td>
<td>1.41E-09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defa</td>
<td>rs1008656</td>
<td>6900336</td>
<td>0.33</td>
<td>1.16</td>
<td>1.11E-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defa</td>
<td>rs934614</td>
<td>6679731</td>
<td>0.36</td>
<td>1.14</td>
<td>1.09E-18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defa</td>
<td>rs2738048</td>
<td>6822765</td>
<td>0.32</td>
<td>0.79</td>
<td>3.80E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>rsID</td>
<td>Minor Allele</td>
<td>Minor Allele Frequency</td>
<td>Major Allele Frequency</td>
<td>Associated Parameter</td>
<td>Study Year</td>
<td>Location</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>--------------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>----------------------</td>
<td>------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>MPPED2</td>
<td>rs4077515</td>
<td>9</td>
<td>139266496</td>
<td>0.4</td>
<td>1.16</td>
<td>Kyryluk</td>
<td>2014</td>
<td></td>
</tr>
<tr>
<td>SH2B3</td>
<td>rs4744712</td>
<td>9</td>
<td>71434707</td>
<td>0.39</td>
<td>-0.008</td>
<td>Köttgen</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>ATXN2</td>
<td>rs1079472</td>
<td>10</td>
<td>1156165</td>
<td>0.08</td>
<td>-0.014</td>
<td>Köttgen</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>FAM109A</td>
<td>rs1801239</td>
<td>10</td>
<td>16919052</td>
<td>0.098</td>
<td>0.0835</td>
<td>Büger</td>
<td>CUBN</td>
<td></td>
</tr>
<tr>
<td>SLC6A13</td>
<td>rs1099486</td>
<td>10</td>
<td>52645424</td>
<td>0.19</td>
<td>0.0077</td>
<td>Pettaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>TNFSF13</td>
<td>rs2074038</td>
<td>11</td>
<td>44087989</td>
<td>0.29</td>
<td>1.14</td>
<td>Li</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>ACAD10</td>
<td>rs1631160</td>
<td>11</td>
<td>2789955</td>
<td>0.18</td>
<td>-0.0065</td>
<td>Pettaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>C12orf51</td>
<td>rs4014195</td>
<td>11</td>
<td>65506822</td>
<td>0.36</td>
<td>-0.0055</td>
<td>Pettaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>NFATC1</td>
<td>rs653178</td>
<td>12</td>
<td>112007756</td>
<td>0.5</td>
<td>0.003</td>
<td>Köttgen</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>MPPED2-DCDC5</td>
<td>rs1076787</td>
<td>11</td>
<td>30725254</td>
<td>0.69</td>
<td>0.0068</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>INHBC</td>
<td>rs1106766</td>
<td>11</td>
<td>57809456</td>
<td>0.22</td>
<td>0.0061</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>C12orf51</td>
<td>rs2074356</td>
<td>11</td>
<td>111129784</td>
<td>0.23</td>
<td>0.0064</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>SLCEA13</td>
<td>rs1077402</td>
<td>12</td>
<td>349298</td>
<td>0.36</td>
<td>0.008</td>
<td>Köttgen</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>CUX2, FAM199A, SH2B3, ATXN2, BRAP, ACAD10, ALDH2</td>
<td>rs7956534</td>
<td>12</td>
<td>15321194</td>
<td>0.19</td>
<td>0.0068</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>PTPRO</td>
<td>rs1049196</td>
<td>12</td>
<td>3368093</td>
<td>0.1</td>
<td>-0.0095</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>TSPAN9</td>
<td>rs1262777</td>
<td>13</td>
<td>73247696</td>
<td>0.4</td>
<td>0.009</td>
<td>Köttgen</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>IN2D2</td>
<td>rs2928148</td>
<td>15</td>
<td>41401550</td>
<td>0.48</td>
<td>-0.0051</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>no gene in &lt;250 kb distance</td>
<td>rs1243785</td>
<td>15</td>
<td>94114833</td>
<td>0.038</td>
<td>1.6</td>
<td>Sandholm</td>
<td>T1-ESRD</td>
<td></td>
</tr>
<tr>
<td>WDR72</td>
<td>rs491567</td>
<td>15</td>
<td>53946593</td>
<td>0.22</td>
<td>0.009</td>
<td>Kyryluk</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>UBE2Q2</td>
<td>rs1394125</td>
<td>15</td>
<td>76158983</td>
<td>0.35</td>
<td>-0.009</td>
<td>Köttgen</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>GATM/SPATA5L</td>
<td>rs2467853</td>
<td>15</td>
<td>45698793</td>
<td>0.62</td>
<td>0.0126</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>DPEP1</td>
<td>rs164748</td>
<td>16</td>
<td>89708292</td>
<td>0.47</td>
<td>-0.0046</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>PDILT</td>
<td>rs1186490</td>
<td>16</td>
<td>20400839</td>
<td>0.81</td>
<td>-0.0062</td>
<td>Okada</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>ITGAM-ITGAX</td>
<td>rs1115061</td>
<td>16</td>
<td>31357760</td>
<td>0.36</td>
<td>1.18</td>
<td>Kyryluk</td>
<td>IgA-NP</td>
<td></td>
</tr>
<tr>
<td>UMOD</td>
<td>rs1291770</td>
<td>16</td>
<td>20367960</td>
<td>0.18</td>
<td>0.8</td>
<td>Köttgen</td>
<td>CKD</td>
<td></td>
</tr>
<tr>
<td>ITGAM-ITGAX</td>
<td>rs1157453</td>
<td>16</td>
<td>31368874</td>
<td>0.18</td>
<td>0.75757575</td>
<td>Kyryluk</td>
<td>IgA-NP</td>
<td></td>
</tr>
<tr>
<td>SLC47A1</td>
<td>rs2453580</td>
<td>17</td>
<td>19438321</td>
<td>0.41</td>
<td>-0.0059</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>BCAS3</td>
<td>rs1186844</td>
<td>17</td>
<td>56594003</td>
<td>0.75</td>
<td>0.0059</td>
<td>Okada</td>
<td>BUN</td>
<td></td>
</tr>
<tr>
<td>MPDU1</td>
<td>rs4227</td>
<td>17</td>
<td>7491177</td>
<td>0.24</td>
<td>1.23</td>
<td>Yu</td>
<td>IgA-NP</td>
<td></td>
</tr>
<tr>
<td>TNFSF13</td>
<td>rs3803800</td>
<td>17</td>
<td>7462969</td>
<td>0.35</td>
<td>0.12</td>
<td>Yu</td>
<td>IgA-NP</td>
<td></td>
</tr>
<tr>
<td>CDK12</td>
<td>rs1107890</td>
<td>17</td>
<td>37631924</td>
<td>0.24</td>
<td>0.0096</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>BCAS3</td>
<td>rs985661</td>
<td>17</td>
<td>59456589</td>
<td>0.19</td>
<td>-0.011</td>
<td>Köttgen</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>NFATC1</td>
<td>rs8091180</td>
<td>18</td>
<td>77164243</td>
<td>0.44</td>
<td>0.006</td>
<td>Pattaro</td>
<td>eGFRcrea</td>
<td></td>
</tr>
<tr>
<td>Gene Symbols, Official gene symbol; SNP, SNP ID; Chr, chromosome number; Pos (b37), chromosome position; Freq, for case-control studies, the minor allele frequency in controls is presented. If available, the combined frequency (discovery+replication) is used; otherwise, the frequency in the discovery group is reported; Effect Size, the effect estimate is reported with respect to the minor allele; Combined effect estimates (beta/OR, SE/95% CI, p-value) from discovery + replication were used when available. Otherwise, effect estimates for the discovery sample are shown; 95%CI, 95% confidence interval; p-value, raw p-value; Reference, GWAS Publication where it is found, all papers are cited in Chapter 2; Association parameter, disease that was reported to have a GWAS hit for the disease.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Symbols</th>
<th>Chr</th>
<th>Pos (b37)</th>
<th>Freq</th>
<th>Effect Size</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC14A2</td>
<td>18</td>
<td>41441128</td>
<td>0.8</td>
<td>0.0083</td>
<td>Okada 2012</td>
</tr>
<tr>
<td>SIPA1L3</td>
<td>7</td>
<td>38464262</td>
<td>0.18</td>
<td>-0.0058</td>
<td>Patiaro 2015</td>
</tr>
<tr>
<td>SLC7A9</td>
<td>8</td>
<td>33356891</td>
<td>0.39</td>
<td>0.008</td>
<td>Köttgen 2010</td>
</tr>
<tr>
<td>MKKS, SLX4IP, JAG1</td>
<td>6</td>
<td>60400555</td>
<td>0.39</td>
<td>-0.017</td>
<td>Köttgen 2009</td>
</tr>
<tr>
<td>GNAS</td>
<td></td>
<td>56902468</td>
<td>0.30</td>
<td>0.0052</td>
<td>Okada 2012</td>
</tr>
<tr>
<td>TP53INP2</td>
<td>1</td>
<td>33285053</td>
<td>0.47</td>
<td>-0.0049</td>
<td>Patiaro 2015</td>
</tr>
<tr>
<td>BCAS1</td>
<td>7</td>
<td>52732362</td>
<td>0.21</td>
<td>0.0077</td>
<td>Patiaro 2015</td>
</tr>
<tr>
<td>NAPB, CSTL1, CST11, CST8, CST9L, CST9, CST3, CST4, CST1, CST2, CST5</td>
<td>5</td>
<td>13038330</td>
<td>0.21</td>
<td>0.076</td>
<td>Köttgen 2009</td>
</tr>
<tr>
<td>APOL1</td>
<td>22</td>
<td>36669095</td>
<td>0.79</td>
<td>NA</td>
<td>Genovese 2010</td>
</tr>
<tr>
<td>MYH9</td>
<td>22</td>
<td>36311616</td>
<td>0.91</td>
<td>NA</td>
<td>Genovese 2010</td>
</tr>
<tr>
<td>APOL2</td>
<td>22</td>
<td>36220399</td>
<td>0.8</td>
<td>NA</td>
<td>Genovese 2010</td>
</tr>
<tr>
<td>MYH9</td>
<td>2</td>
<td>36329925</td>
<td>0.56</td>
<td>NA</td>
<td>Genovese 2010</td>
</tr>
<tr>
<td>MTMR3</td>
<td>2</td>
<td>30423460</td>
<td>0.16</td>
<td>0.78</td>
<td>Yu 2011</td>
</tr>
<tr>
<td>APOL1</td>
<td>22</td>
<td>36265284</td>
<td>0.94</td>
<td>NA</td>
<td>Genovese 2010</td>
</tr>
</tbody>
</table>

This table is restricted to single nucleotide polymorphisms (SNPs) that reached genome-wide significance (p<5E-08). Associated loci are reported from the study in which they were identified with lowest p-value. The same locus is only reported from a subsequent study if the index SNP in the later study is in low LD with the previously reported index SNP (r2<0.2 in the corresponding 1000 Genomes population). If SNPs within an associated locus showed independent evidence for association by conditional analyses, all independent index SNPs within the locus are reported.

Appendix 2 | All significant eGenes

Appendix 3 | Differentially methylated loci and the corresponding differentially expressed transcripts

Appendix 4 | Infinium DKD dataset and HELP CKD dataset differentially methylated loci

Appendix 5 | Technical validation of methylation level using MassArray Epityper

Appendix 6 | RefSeq annotated genomic distribution of the hypo- and hypermethylated regions

Appendix 7 | List of differentially methylated regions with p-value methylation level, genomic locus and nearest annotated transcript
57. 1000-Genomes-Project (2016).
103. Nicolae, D.L. et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6 (4), e1000888.


106
260. Helsby, M.A. et al. (2014) CiteAb: a searchable antibody database that ranks antibodies by the number of times they have been cited. BMC Cell Biol 15, 6.
309. Gavin Band, J.M.