

https://dspace.org/

Identifying the transcriptional regulators of PXDN in the kidney and determining the impact of genetic variation in PXDN and its transcriptional regulators in chronic kidney disease in a South African cohort

Flisher, Khaleila

2025-06

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/47634

Flisher, Khaleila. (2025). Identifying the transcriptional regulators of PXDN in the kidney and determining the impact of genetic variation in PXDN and its transcriptional regulators in chronic kidney disease in a South African cohort. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/47634

Downloaded from a DSpace repository. https://dspace.org/



Identifying the transcriptional regulators of *PXDN* in the kidney and determining the impact of genetic variation in *PXDN* and its transcriptional regulators in chronic kidney disease in a South African cohort

by

Khaleila Flisher

(2924052)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Prof. Demetra Mavri-Damelin

Co-supervisor: Dr. Mahtaab Hayat

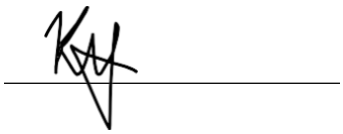
June 2025

Declaration

I, Khaleila Flisher (2924052), am a student registered for the degree of Master of Science (Dissertation) in the academic year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other university.
- The information used in the thesis has not been obtained by me while employed by, working under the aegis of, any person or organisation other than the university.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of ideas or words in my writing.



Khaleila Flisher

9 June 2025

Abstract

Chronic kidney disease (CKD), characterised by impaired glomerular filtration, is a global heath challenge that disproportionately affects African populations. South Africa exhibits high CKD incidence (8.7%), with many cases observed in individuals lacking common risk factors. *PXDN* is expressed in kidney tissue, and its dysregulated expression leads to excessive PXDN deposition, contributing to renal fibrosis, a key driver of CKD progression. However, regulation of *PXDN* in the kidney and the impact of genetic variation in *PXDN* and its regulators on kidney function is largely unknown.

Firstly, ChIP-seq data from ChIP-Atlas, comprising over 380 experiments across 36 transcriptional regulators, were analysed. Putative TF binding sites in the promoter and intron 1 region of *PXDN* were identified using ChIP-seeker. High-confidence binding sites were identified using FIMO with position-weight matrices from JASPAR, HOCOMOCO and Factorbook, and mapped using NCBI Genome Data Viewer. Thirteen high-confidence TF binding sites were mapped, and TFs were characterised based on their potential role in kidney function. Several TFs, such as *SIX2* and *CTCF*, have established roles in kidney function, while others, including *EGR2, ZEB1* and*TWIST1*, are implicated in kidney dysfunction.

Secondly, to investigate whether variation in *PXDN*, its identified TFs and key kidney disease-associated genes are implicated in kidney dysfunction in a black South African cohort, we conducted an association analysis using estimate glomerular rate (eGFR) as the phenotype. Identified variants were annotated and characterised. No variants remained statistically significant after correcting for multiple testing, likely due to the small sample size used. However, identified variants may still contribute to low eGFR levels, as variants in *ZNF423*, *KLF4* and *PXDN* in the South African cohort are predicted to be damaging, and these genes have established roles in kidney dysfunction.

These findings identify transcriptional regulators of *PXDN* in the kidney and highlight their potential roles in kidney dysfunction. Experimental validation and functional assays are required to confirm identified binding sites and determine the regulatory role of these TFs. While a *PXDN* variant was associated with low eGFR levels, larger-scale studies are required to confirm this association and further investigate its impact.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Demetra Mavri-Damelin and Dr. Mahtaab Hayat, for their unwavering support, optimism, and kindness. Their invaluable guidance has provided me with a strong foundation for future research and learning, and I am very grateful for their mentorship.

I would also like to acknowledge the Functional Genetics Lab, with special thanks to Shiven Naidoo for his assistance at the start of my project, as well as everyone who has contributed to this research.

To my friends and family, both near and far, thank you for your encouragement and strength throughout this journey. Your moral support and excitement have been a constant source of motivation.

Finally, I am grateful for funding provided by the National Research Foundation (NRF) and the WITS Postgraduate Merit Award, which made this research possible.

Table of Contents

CHAPTER 1. Introduction 13

1.1 CHRONIC KIDNEY DISEASE 13
1.1.1 Overview of kidney function 13
1.1.2 Progression of CKD 15
1.1.3 Fibrosis in CKD 15
1.1.4 Systemic impact of CKD on other organs 17
1.1.5 Diagnostic criteria for CKD 18
1.1.6 CKD prevalence 20
1.1.7 Treatment 21
1.1.8 Risk factors 22
1.1.9 Transcriptional regulation in disease progression 26
1.1.10 Investigating the genetics of CKD in an African context 27

1.1 PEROXIDASIN 29
1.2.1 PXDN structure 29
1.2.2 Expression and regulation 30
1.2.3 Mechanism of action 31
1.2.4 Roles of PXDN in disease 32

1.3 Rationale for the current study 34

1.4 Aims and objectives 35

CHAPTER 2. Identifying the transcriptional regulators of PXDN using publicly available ChIP-seq data 36

2.1 INTRODUCTION 36

2.2 METHODOLOGY 38
2.2.1 ChIP-Atlas data collection and analysis 38
2.2.2 Data pre-processing and peak detection 39
2.2.3 Data annotation, analysis and visualisation 41
2.2.4 Expression analysis 42
2.2.4 TF characterisation and mapping 43

2.2 RESULTS 45
2.3.1 Data pre-processing and peak detection 45
2.3.2 Data analysis, annotation and visualisation 45
2.3.3 Expression analysis 52
2.3.4 TF characterisation and mapping 53

2.3 DISCUSSION 63
2.4.1 Exploration of TF binding patterns 63
2.4.2 Characterisation of mapped TFs 68
2.4.3 Limitations 72
2.4.4. Conclusion 73
2.4.5 Future Research 74

CHAPTER 3. Investigating genetic variation in PXDN, PXDN regulators and CKD-associated genes linked to eGFR levels in a SA cohort 75

3.1 INTRODUCTION 75

3.2 METHODOLOGY 77
3.2.1 Dataset 77
3.2.2 Phenotype 78

3.2.3 Quality control 78
3.2.4 Association analysis 79
3.2.5 Variant annotation and prioritisation 80
3.2.6 Functional Annotation and gene-set analysis 82
3.3 RESULTS 83
3.3.1 Dataset 83
3.3.2 Phenotype 83
3.3.3 Quality control 83
3.3.4 Association analysis 85
3.3.5 Variant annotation and prioritisation 86
3.3.6 Gene-set analysis 92
3.4.1 Interpretation of association analysis findings 94
3.4.2 Limitations and future research 99
3.4.3 Conclusion 99

CHAPTER 4. Overall Discussion and conclusions 101

REFERENCES 104

APPENDICES 130
Appendix A 130
Appendix B 131
Appendix C 134
Appendix D 136
Appendix E 137
Appendix F 141
Appendix G 143
Appendix H 145

List of Figures

Figure 1. Diagram of the kidney 14
Figure 2. The structure of Peroxidasin (PXDN) and its constituents..... 30
Figure 3. Overview of ChIP-seq data analysis..... 39
Figure 4. Pie chart of the percentages of *PXDN* binding sites according to peak location…..... 46
Figure 5. Distribution of binding sites relative to the TSS..... 47
Figure 6. Bar graph showing TFs binding to the promoter of the *PXDN* gene in kidney cells, as identified from ChIP-Atlas..... 48
Figure 7. Bar graph showing TFs binding to the first intronic region of *PXDN* in kidney cells, as identified from ChIP-Atlas 48
Figure 8. Bar graph showing the different cell types used in each TF experiment from ChIP-Atlas for the promoter region… 51
Figure 9. Bar graph showing the different cell types used in each TF experiment from ChIP-Atlas for the first intronic region 52
Figure 10. Bar graph showing expression levels of significant *PXDN* regulators using RNA-sequencing data generated by the Human Protein Atlas (HPA) Genotype-Tissue Expression (GTEx) portal and CAGE data generated by The Functional Annotation of Mammalian Genomes 5 (FANTOM5) consortium..... 54
Figure 11. Mapping of TF binding sites for the PXDN gene..... 63
Figure 12. Principal component analysis (PCA) of AWI-Gen samples with 1000 Genomes reference populations..... 85
Figure 13. Figure 13. Number of variants removed during quality control 86

List of Tables

Table 1. ChIP-Atlas labels of cell types and their corresponding classifications..... 46
Table 2. TFs identified in the promoter region of *PXDN*.....49
Table 3. TFs identified in the first intronic region of *PXDN*..... 50
Table 4. Binding sites and motifs for TFs identified in the promoter region 55
Table 5. Binding sites and motifs for TFs identified in the intron 1 region..... 55
Table 6. Binding sites of non-TF regulators..... 58
Table 7. Transcription factor binding sites identified in the *PXDN* promoter region..... 59
Table 8. Transcription factor binding sites identified in intron 1 of *PXDN*..... 59
Table 9. Description of variant prediction tools and the scoring system applied 81
Table 10. Top linear regression analysis results (p-value < $5x10^{-4}$) 89
Table 11. Selected annotation of top variants and normalised scores, ranked in descending order 90
Table 12. *ZNF423* variants in LD that met functional significance threshold.....92
Table 13. Gene set analysis results with adjusted p-value < 0.05….....93

List of Abbreviations

3C	Chromosome conformation capture
ACE	Angiotensin-converting enzyme
AGT	Angiotensinogen
APOL1	Apolipoprotein L1
ARBs	Angiotensin II receptor blockers
BMP6	Bone morphogenic protein 6
CADD	Combined Annotation Dependent Depletion
CAGE	Cap Analysis of Gene Expression
CGP	Chemical and Genetic Perturbation
ChIP-seq	Chromatin immunoprecipitation sequencing
CKD	Chronic kidney disease
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
CRP	C-reactive protein
CST9	Cystatin 9
dbGaP	The database of Genotypes and Phenotypes
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
eGFR	Estimated glomerular filtration rate
EMT	Epithelial-mesenchymal transitions
EPO	Eosinophil peroxidase
ESRD	End stage renal disease
FANTOM5	Functional Annotation of Mammalian Genomes 5
FDR	False discovery rate
FIMO	Find Individual Motif Occurrences
GDNF	Glial cell line-derived neurotrophic factor
GFR	Glomerular filtration rate
GO	Geno ontology
GSK-3	Glycogen synthase kinase 3
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association studies
HEK-293	Human embryonic kidney 293 cells

Hh	Hedgehog
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HOBr	Hypobromous acid
HPA	Human Protein Atlas
HUVECs	Human umbilical cord vein endothelial cells
Ig	Immunoglobin
IL	Interleukin
Kb	kilobase
kDa	Kilodaltons
KDIGO	Kidney Disease: Improving Global Outcomes
KGP	The Thousand Genomes Project
LD	Linkage disequilibrium
LRRs	Leucine-rich repeats
MACS	Model-based analysis of ChIP-seq
MDRD	Modification of Diet in Renal Disease
MHC I	Major histocompatibility complex I
MPO	Myeloperoxidase
mRNA	Messenger ribonucleic acid
miRNA	MicroRNA
MSigDB	Molecular Signatures Database
MYH9	Myosin heavy chain 9
NCBI	Nation Centre of Biotechnology Information
NRF2	Nuclear erythroid-2 related factor 2
nTPM	Normalised tags per million
PCA	Principal component analysis
PCR	Polymerase chain reaction
PCs	Principal components
PKD	Polycystic kidney disease
POX	Peroxidase domain
PWMs	Position-weight matrices
PXDN	Peroxidasin
QUAL	Quality score

RPM	Reads per million
RNA	Ribonucleic acid
RPTECs	Renal proximal tubule epithelial cells
RT-PCR	Reverse transcription-polymerase chain reaction
SA	South Africa
SAC	South African Coloured
SDI	Social-demographic index
SNAI1	Snail family transcriptional repressor 1
SNPs	Single nucleotide polymorphisms
SRF	Serum response factor
SSA	Sub-Saharan Africa
SUMO2	Small ubiquitin-like modifier 2
TB	Tuberculosis
TF	Transcription factor
TGF	Transforming growth factor
TNF	Tumour necrosis factor
TPM	Tags per million
TPO	Thyroid peroxidase
TSS	Transcription start site
UTR	Untranslated region
UACR	Urinary albumin-to-creatine ratio
UMOD	Uromodulin
UUO	Unilateral ureteral obstruction
VPO1	Vascular Peroxidase-1
VWC	von Willebrand factor type C domain

List of symbols

α	Alpha
β	Beta
κ	Kappa

CHAPTER 1. Introduction

1.1 CHRONIC KIDNEY DISEASE

Chronic Kidney Disease (CKD) is a progressive condition characterised by kidney damage that impairs the ability to filter blood effectively (Chen et al., 2019). This decline in kidney function results from structural and functional transformations, such as tubular atrophy, reduced blood flow and interstitial fibrosis (Webster et al., 2017). Over time, these irreversible changes cause the accumulation of fluid and waste, leading to further complications in the body. CKD is associated with a range of health issues, including cardiovascular problems and disruptions in blood mineral concentrations (Chen et al., 2019).

1.1.1 Overview of kidney function

The kidneys are comprised of several key structures, including an outer cortex, an inner medulla, various veins and arteries, and nephrons (Figure 1). Each nephron consists of a tubule and glomerulus, which filter blood and remove waste, respectively. The nephron contains Bowman's capsule, which encases Bowman's space, the loop of Henle, the proximal convoluted tubule, and the distal convoluted tubule, which onloads into collecting ducts before urine is excreted from the kidneys (Wallace, 1998). Normal kidney activity involves the coordinated function of nephron segments and the glomerulus (Scholz et al., 2021). Along with endothelial cells and podocytes, the glomerular basement membrane, a specialised extracellular matrix (ECM), constitutes the glomerular filtration barrier (Miner, 2012). The glomerular basement membrane, among other components, contains collagen IV (Miner, 2012). Collagen IV is crucial for kidney function as mutations affecting collagen IV cause affected filtration and renal disease (Van Agtmael et al., 2005). The structural components of the kidney work together to perform key functions of the kidney: glomerular blood filtration, reabsorption of vital filtered substances, and excretion of waste products, all of which are necessary to maintain homeostasis in the body (Wallace, 1998).

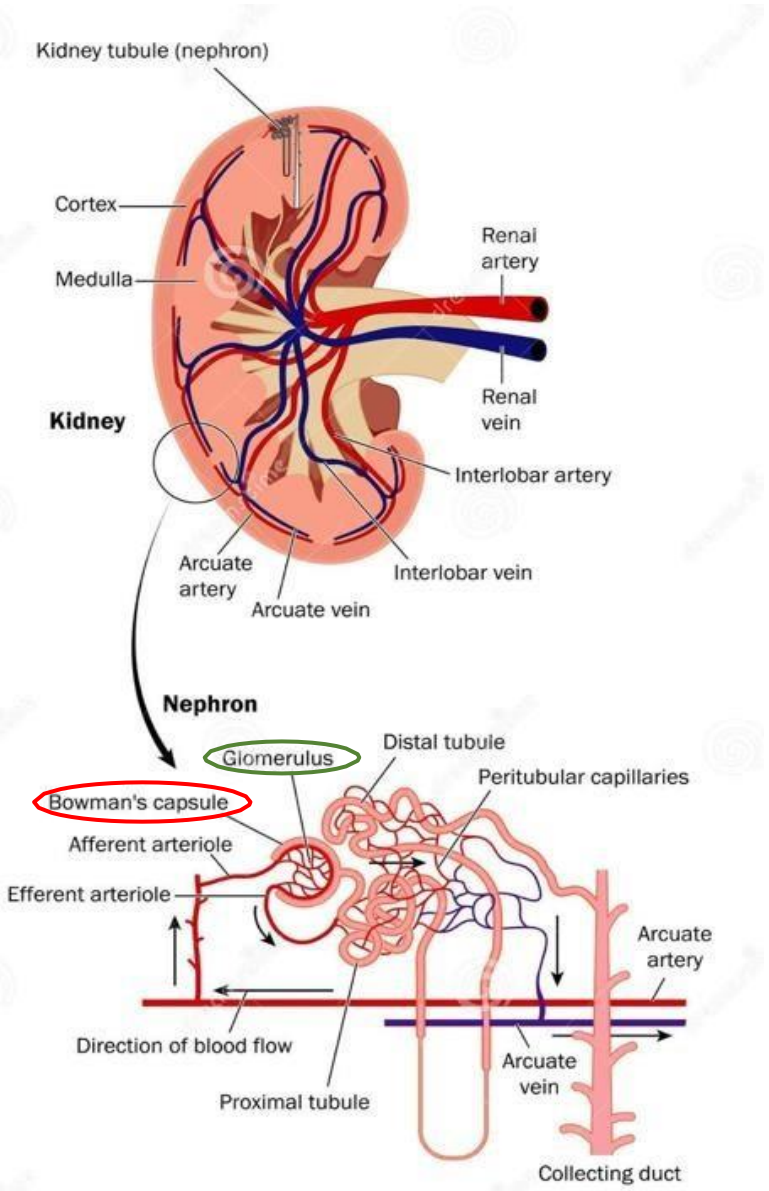


Figure 1. Diagram of the kidney. The lower portion illustrates the detailed anatomy of a nephron, highlighting the glomerulus (in green), Bowman's Capsule (in red) and Bowman's space. This is the site where ultrafiltration occurs, allowing filtrate to enter the nephron and begin its journey through the renal tubules. Figure adapted from dreamstime.com.

Markers of kidney function include glomerular filtration rate (GFR) and the presence of albuminuria. Glomerular filtration, a key kidney function, involves ultrafiltration of plasma in Bowman's space of Bowman's capsule, indicated in red in Figure 1 (Levey et al., 2015). This space allows filtrate to enter the nephron and proceed to the proximal tubule for reabsorption (Falkson & Bordoni, 2020). The estimated glomerular filtration rate (eGFR) represents the volume of fluid filtered through functioning nephrons per unit time. A decline in eGFR reflects a decrease in kidney function (Levey et al., 2015). Secondly, a urinary albumin-to-creatine ratio (UACR) measures albuminuria, a marker of kidney damage, which occurs when albumin leaks into urine due to a compromised glomerular filtration barrier (Heerspink & Gansevoort,

2015). Albumin is a blood plasma protein essential for ligand transport and blood pressure regulation (Moman et al., 2023). Higher levels of albuminuria are associated with a faster decline in renal function (Van Velde et al., 2009).

1.1.2 Progression of CKD

While acute kidney disease is characterised as structural and functional kidney abnormalities lasting less than three months, CKD is characterised by the same traits persisting for three months or longer (Lameire et al., 2021). When the kidneys lose their ability to function entirely, end-stage renal disease (ESRD) occurs (Lameire et al., 2021). CKD progression is characterised by a gradual decline in kidney function, exacerbated by fibrosis (Yuan et al., 2022). As CKD advances, it leads to various health complications, including heart disease, strokes, and weakened immunity, as well as systemic issues, such as metabolic disturbances, lung diseases, and neurological complications (Chen et al., 2019; Zoccali et al., 2017). Fibrosis, defined as the accumulation of ECM components, plays a critical role in CKD progression by disrupting essential filtration and reabsorption processes (Nørregaard et al., 2023). Although fibrosis initially serves as a repair mechanism in response to kidney damage, excessive matrix deposition damages kidney structure, reduces blood flow, and impairs overall kidney function (Panizo et al., 2021).

1.1.3 Fibrosis in CKD

Fibrosis is the final common pathway of CKD and ESRD due to its disruption of both kidney structure and function (Yamashita & Kramann, 2024). Fibrosis is driven by ECM deposition, which disrupts tissue architecture, promoting fibrotic remodelling (Bonnans et al., 2014). The fibrogenic niche, a specialised microenvironment characterised by fibrotic lesions and fibroblast activation, is formed by an ECM network of proteins and contains inflammatory cells, ECM network, metabolites, extracellular vesicles and soluble factors (Li et al., 2022). There are four categories of ECM proteins: structural, matricellular, matrix-modifying proteins and proteoglycans, but matricellular proteins, including tenascin-C, fibrilin-1 and periostin, are the most abundant in kidney fibrosis (Li et al., 2022). These proteins impact several cellular

processes including ECM assembly, apoptosis, migration, wound healing, inflammation, and fibrosis. Fibrogenic niche constituents recruit extracellular soluble factors, such as hedgehog (Hh) and transforming growth factor-beta (TGF)-β, which causes elevated profibrotic levels in the microenvironment, further promoting fibrosis and CKD pathological features. Both pro-fibrotic growth factors, like TGF-β, and pro-inflammatory cytokines, like tumour necrosis factor (TNF)-α and interleukin (IL)-6, cause an increase in ECM deposition in fibrosis (Panizo et al., 2021).

Renal proximal tubule epithelial cells (RPTECs) play a pivotal role in fibrosis pathogenesis, as they are highly susceptible to injury due to their significant role in reabsorption and high enzymatic demands (Chevalier, 2016). In response to damage, RPTECs undergo epithelial-mesenchymal transition (EMT), a process wherein epithelial cells in contact with the basement membrane undergo biochemical changes, adopting a mesenchymal cell phenotype (Kalluri & Weinberg, 2009). Epithelial cells no longer possess their cell-to-cell adhesion properties, associated with downregulated epithelial marker like E-cadherin, and instead obtain mesenchymal characteristics including migration and infiltration capability, associated with upregulated mesenchymal markers like fibronectin (Kalluri & Weinberg, 2009; L. Li et al., 2022). This myofibroblast-like phenotype results in elevated ECM deposition and actin cytoskeleton reorganisation, exacerbating fibrosis. Targeting EMT can therefore be regarded as a strategy to reduce fibrosis. EMT is triggered by hypoxia, senescence and, most importantly, inflammation, as observed in fibrosis (Di Gregorio et al., 2020; Gonzalez-Gonzalez et al., 2017; Wei et al., 2020). Proinflammatory cytokines activate inflammatory cells, for example macrophages and neutrophils, which then release proinflammatory cytokines, as mentioned previously, stimulating EMT (Li et al., 2022). EMT contributes to tissue scarring and impaired kidney function, playing a prominent role in CKD progression (Li et al., 2016). Importantly, *in vivo* studies have shown that inhibiting EMT-related pathways can attenuate fibrosis, underlining EMT as a therapeutic target (Li et al., 2016).

A newly identified protein, peroxidasin (PXDN), has been found to be implicated in both the ECM and EMT, and to potentially play a role in kidney fibrosis (Bhave et al., 2012; Sitole & Mavri-Damelin, 2018). PXDN is a part of the peroxidase family of enzymes, which have roles in immune responses, disease development and defence mechanisms (Cheng & Shi, 2022).

PXDN is an extracellular protein located in kidney tissue which contains ECM motifs and stabilises collagen IV (Pöschl et al., 2004). Animal models have shown that PXDN contributes to renal inflammation and tubulointerstitial fibrosis (Colon et al., 2019), likely due to its role in the ECM remodelling and EMT.

1.1.4 Systemic impact of CKD on other organs

In addition to its effect on the kidneys, CKD has widespread systemic consequences, significantly increasing morbidity and mortality, particularly due to hypertension-related cardiovascular complications (Zoccali et al., 2017). CKD affects both cardiac structure and function, often leading to chronic cardiac dysfunction (Zanoli et al., 2019). Notably, cardiovascular disease is the leading cause of death among CKD patients (Saritas & Floege, 2020; Weiner et al., 2006). An American-based study found that 41% of CKD-related deaths were due to cardiovascular complications, highlighting the need for early and effective management to reduce both disease progression and mortality (Hoefield et al., 2010). This occurs because CKD causes a persistent inflammatory state, resulting in myocardial and vascular remodelling (Jankowski et al., 2021). This results in vascular calcification, atherosclerotic lesions and myocardial fibrosis which put increased pressure on the renal system, for example, with increased hypertension. CKD and cardiovascular diseases also share risk factors, namely, diabetes, hypertension and smoking (Li et al., 2016). In addition to CKD impacting the cardiovascular system, the cardiovascular system is also greatly impacted by CKD. For example, one of the primary cardiovascular consequences, arterial calcifications, is caused by reduced GFR (Zanoli et al., 2019). Additionally, lower GFR is strongly associated with an increased hazard ratio for cardiac events, such as ischemia, where insufficient blood supply to the heart is cause by vessel blockages (Go et al., 2004; Mahmoodi et al., 2014).

Beyond cardiovascular effects, CKD influences various other body systems. It disrupts immunity and metabolism, as demonstrated by increased insulin resistance and elevated innate immunity biomarkers, such as IL-6 and C-reactive protein (CRP) seen in advanced CKD, resulting in hyperglycaemia and inflammatory diseases, respectively (Zoccali et al., 2017). Inflammation, driven by mediators like TNF, is also closely linked to deteriorating renal function (Gupta et al., 2012). CKD has also been associated with an elevated risk of obstructive

lung diseases due to the presence of neuroendocrine cells, which modulate immune responses and inflammation in the lungs (Zoccali et al., 2017). Additionally, CKD affects the gut-kidney axis by compromising intestinal wall integrity, leading to endotoxin leakage and bacterial translocation. This exacerbates both renal and cardiovascular damage (Cha et al., 2016). Neurological complications are another hallmark of advanced CKD, manifesting as conditions like dementia, cognitive impairments, and cerebrovascular disease, which affect brain blood vessels and flow (Zoccali et al., 2017). These complications arise because CKD manifestations, such as low eGFR and albuminuria, are independently linked to structural and functional brain changes in the brain. Given its wide-ranging effects, CKD is classified as a systemic disease, impacting not only the kidney but the entire body, as evidenced by its influence on cardiovascular, immune, metabolic, pulmonary, and neurological systems.

1.1.5 Diagnostic criteria for CKD

CKD is classified into five stages, with kidney function deteriorating progressively from stage 1 (mild) to stage 5 (kidney failure). Diagnosis relies on laboratory measures of kidney function, primarily eGFR or UACR (Lameire et al., 2021). CKD is defined as reduced kidney function (eGFR <60 mL/min/1.73m^2), the presence of albuminuria (UACR ≥3 mg/mmol), or both, persisting for more than three months (Lameire et al., 2021; George et al., 2019; Stevens & Levin, 2013). While both diagnostic methods are useful, an eGFR of <60 mL/min/1.73m^2 primarily indicates moderate to severe CKD (CKD stages 3-5). In contrast, when UACR is available, CKD can also be identified in its earlier stages (CKD stages 1-5), making it a valuable tool for early identification. One of the key mechanisms underlying CKD involves the basement membrane (shown in green in Figure 1), which is essential for maintaining filtration function, and damage or structural abnormalities in this membrane impair glomerular filtration and albumin leakage (Chew & Lennon, 2018).

While albuminuria measurements are associated with risk of ESRD, eGFR is the marker used to define ESRD (Carrero et al., 2017). When eGFR falls below 15mL/min per 1.73m^2, CKD is considered to have progressed to ESRD, at which point the kidney is not able to sustain life (Webster et al., 2017). At this terminal stage, treatment typically involves dialysis or kidney transplantation to manage kidney failure (Saran et al., 2020).

1.1.5.1 Assessment of kidney function in African populations

To effectively investigate CKD in African populations, it is important to use appropriate markers that are accessible and reliable for these populations. A number of equations with different variable weightings are used for GFR estimation, kidney function staging and CKD diagnosis. Commonly used equations include the Modification of Diet in Renal Disease (MDRD), CKD Epidemiology Collaboration (CKD-EPI) and Cockcroft-Gault equations (Appendix A). However, the accuracy and reliability of these equations can vary depending on the population in which they are applied, and GFR estimation has proven particularly challenging in African populations.

Initially in South Africa (SA) the Cockcroft-Gault and MDRD formulas were recommended for GFR estimation, although increasing support for the CKD-EPI formula has emerged (Fabian, Gondwe, et al., 2022). A study conducted on mixed ancestry South African populations evaluating concordance between kidney function estimators found the highest agreement between MDRD and CKD-EPI equations (Matsha et al., 2013). The highest concordance was found with MDRD and CKD-EPI when applying ethnicity correction, followed by the same two metrics without ethnicity adjustments. However, an updated race-neutral CKD-EPI 2021 equation was adopted after the re-evaluation of race inclusion in diagnostic algorithms. This equation was comparable to the non-race-adjusted CKD-EPI 2009 equation when tested in asymptomatic black Africans (Omuse et al., 2024), but the same cannot be said for all populations. A multicentre cohort study, which included a black SA population, and populations from Malawi and Uganda, evaluated multiple equations for estimating GFR and found that none achieved accuracy within 30% of the measured GFR for at least 75% of participants (Fabian, Kalyesubula, et al., 2022). Despite their widespread use in South African clinical settings, many of the tested GFR equations are not appropriate for black South African populations, as demonstrated by Fabian *et al.* (2022). More accurate methods are needed to effectively assess glomerular filtration rate in individuals with impaired kidney function.

UACR is widely accepted as the standard method for measuring albuminuria and may serve as a more reliable indicator of kidney function. A study in SA showed that albuminuria was a strong predictor of CKD and may even precede reductions in GFR (Khoza et al., 2024). In a

sub-Saharan African (SSA) study that included three out of six South African collection sites, albuminuria was found to have the largest relative contribution to CKD prevalence, compared to eGFR. This demonstrates the importance of albuminuria measurements in classifying CKD in SSA populations, including those in SA (George et al., 2019). Additionally, a novel variant in an intronic region of the Bone morphogenic protein 6 (*BPM6)* is associated with UACR, indicative of affected kidney function and kidney disease traits, such as fibrosis, were identified in SSA populations, exemplifying the utility of UACR for kidney disease classification (Brandenburg et al., 2024).

1.1.6 CKD prevalence

CKD is a global health challenge, claiming over a million lives in 2017 alone (Bikbov et al., 2020). Between 1990 and 2017, the global all-age mortality rate for CKD increased by 41.5%, with the heaviest burden experienced by countries in the lowest socio-demographic index (SDI) quantiles (Bikbov et al., 2020). Significant disparities exist in CKD prevalence based on gender, race, and geography. In African populations, women are disproportionately affected, with a prevalence of 12% compared to 9.5% in men, as assessed during 2013-2016 (CKD defined using CKD-EPI without race factor correction, corresponding to CKD stages 3-5, or UACR ≥3 mg/mmol, or both) (George et al., 2019). Racial disparities are particularly pronounced in ESRD, the final stage of CKD. An American study found that black men had a 7.3% risk of developing ESRD, compared to 2.5% in white men, while black women had a risk of 7.8%, compared to 1.8% in white women (Kiberd & Clase, 2002). In Europe, CKD prevalence varies widely across countries. For instance, a prevalence of 10.2% was recorded in Spain, 6.0% in Italy, and as low as 3.3% in Norway (CKD stages 1-5, using isotope dilution mass spectrometry (IDMS)-traceable creatine, adjusted for age and sex) (Brück et al., 2016). In Africa, the overall prevalence for CKD stages 1-5 is 15.8%, while the prevalence for stages 3-5 its 4.6% (Kase et al., 2018). Among African regions, Southern African populations are more affected, with a prevalence of 14%, compared to 6.6% in Burkina Faso and 8.0% in Ghana (CKD defined using CKD-EPI without ethnicity correction, corresponding to CKD stages 3-5, or UACR ≥3 mg/mmol, or both) (George et al., 2019). Additionally, the CKD burden in SSA is disproportionately high relative to the region's level of development (Bikbov et al., 2020).

SA has experienced a significant increase in CKD-related mortality, with deaths rising by 67% from 1999 to 2006, reaching 15 deaths per 100,000 population (Mayosi et al., 2009; Moosa et al., 2015). This is largely attributed to a rise in relevant risk factors in rural and urban areas, an ageing population and non-communicable prevention and treatment being undervalued (Alberts et al., 2005; Mayosi et al., 2009). However, studies in SA show that a large proportion of individuals with CKD do not present with associated risk factors (Fabian et al., 2022). A 2023 study investigating CKD prevalence across various South African populations reported rates of up to 8.7% (CKD stages 3-5, CKD-EPI without ethnicity correction) (Hariparshad et al., 2023). These statistics highlight the urgent need for targeted interventions and enhanced healthcare strategies to address the growing CKD burden in SSA, particularly in high-prevalence areas like SA.

1.1.7 Treatment

Although CKD cannot be reversed, treatment can slow its progression (Schaefer & Wühl, 2012). Without treatment, CKD often leads to kidney failure (Hanafusa et al., 2015). Treatment includes medications aimed at preserving glomerular filtration rate and addressing complications arising from the disease. Angiotensin-converting enzyme (ACE) inhibitors and angiotensin II receptor blockers (ARBs) are commonly prescribed for managing hypertension and slowing nephropathy, a condition involving the progressive deterioration of kidney function (Brenner et al., 2001; Lewis et al., 1993; Turner et al., 2012). Additionally, medications targeting key pathological processes caused, such as fibrosis, are utilised to mitigate kidney damage (Klinkhammer et al., 2017; Turner et al., 2012). Other complications caused by CKD, such as dyslipidaemia, anaemia, and proteinuria, can also be managed with appropriate pharmacological interventions (Levin et al., 2008).

Preventative and lifestyle measures are equally important in managing CKD and the associated health risks. Regular exercise and dietary modification help prevent the development of complications like hypertension (Turner et al., 2012). Lifestyle changes, including weight reduction, smoking cessation, controlled protein intake, and reduced alcohol consumption, are essential for slowing CKD progression and reducing the likelihood of additional health threats (Klahr et al., 1994; Levin et al., 2008). These combined strategies aim to improve overall outcomes and quality of life for individuals living with CKD. If left untreated, CKD is fatal.

1.1.8 Risk factors

CKD has various well-established risk factors that are also valid in SA, including diabetes, smoking, hypertension, heart disease, hyperlipidaemia, older age, obesity, and infectious diseases like human immunodeficiency virus (HIV) and tuberculosis (TB) (Atun et al., 2017; Kalyesubula et al., 2014; Katz, 2005; Mills et al., 2015). Additionally, non-traditional risk factors, such as kidney stones, infections, foetal and maternal factors like low or high birth weight, preeclampsia and gestational diabetes, have also been shown to contribute to CKD development in non-SA populations (Garcia-Garcia & Jha, 2015; Luyckx et al., 2017). CKD has also been associated with malnutrition and poverty (Gutiérrez, 2015).

Regarding the prevalence of CKD risk factors in African populations, older age, hypertension and diabetes are most commonly associated conditions (George et al., 2019). Additional factors that increase CKD risk in specific African regions include positive HIV statuses, smoking, alcohol use, low socioeconomic status, and high body mass index (BMI) (George et al., 2019). In SA, the prevalence of these risk factors is exceptionally high, with hypertension affecting 30% of the population (Peer et al., 2021); type 2 diabetes mellitus affecting up to 22% (Grundlingh et al., 2022; Levitt et al., 1999), and HIV prevalence around 12% (South African HIV Clinicians Society, https://sahivsoc.org). SA has the highest TB incidence according to the World Health Organisation (2020). Moreover, the association between hypertension, type 2 diabetes mellitus, and ESRD is most pronounced in black ethnic groups (Moosa et al., 2015).

1.1.8.1 The genetics of CKD

In addition to environmental risk factors, genetic susceptibility plays a significant role in CKD (Obrador et al., 2017). European populations have a heritability of 21%, highlighting the substantial genetic contribution to disease risk (Gorski et al., 2017). Variation in deoxyribonucleic acid (DNA) sequences among individuals can influence disease aetiology by disrupting normal molecular and cellular functions (Vivante, 2024). Variants in coding regions typically alter protein structure and function, while variants in non-coding regulatory regions, such as enhancers, promoters and transcription factor (TF) binding sites, have been shown to alter expression patterns (Rojano et al., 2019; Vihinen, 2021). While a subset of

kidney diseases is monogenic, driven by a single causative mutation, CKD is a complex, multifactorial condition influenced by genetic, environmental and lifestyle factors (Jefferis et al., 2024). CKD risk variants range from low to high penetrance, with low penetrance variants frequently working with environmental factors to drive CKD development (Vivante, 2024). Overall CKD risk and susceptibility is determined by the cumulative effect of multiple risk variants alongside other contributing factors. Given the heterogeneity of CKD, a genetic diagnosis helps to determine the underlying cause and gives insight into disease development and progression. This can guide more effective treatment strategies than those based on symptoms alone (Vivante, 2024).

In African American populations, the apolipoprotein L1 (*APOL1*) gene has been strongly linked to CKD. APOL1 is involved in the innate immune system and circulates in the bloodstream as a part of the high-density lipid particles which contribute to trypanosome lytic factor, involved in lysing *Trypanosoma brucei* parasites (Weckerle et al., 2016). *APOL1* is regulated by antiviral pathways and upregulated by interferons, such as interferon (IFN)-γ, which are key players in immune response (Nichols et al., 2015).

APOL1 risk variants occur in more than 30% of African Americans and have been associated with diseases such as focal segmental glomerulosclerosis, HIV nephropathy and hypertension (Freedman, 2013; Genovese et al., 2010; Wuttke & Köttgen, 2016). *APOL1* variants were initially thought to provide resistance against the parasite *Trypanosoma brucei brucei*, found in African countries, leading to positive selection of protective *APOL1* alleles (Beckerman & Susztak, 2018). In addition to this, *APOL1* variants are now recognised as a significant risk factor in CKD, as they promote kidney damage–resulting in a genetic trade-off (Beckerman & Susztak, 2018).

The two significant most notable *APOL1* gene variants, G1 and G2, are coding variants located in the last exon of the gene and have been linked to CKD. The G1 variant contains two non-synonymous single nucleotide polymorphisms (SNPs), rs73885319 and rs60910145, which cause an amino acid change, while the G2 variant, rs71785313, has a six-base pair in frame deletion, which causes the loss of two amino acid residues, N388 and Y389. Using allele frequencies form The Thousand Genomes Project (KGP) Phase 3 (Byrska-Bishop et al., 2022; The 1000 Genomes Project Consortium, 2015), no variation was observed at these loci for European, East Asian and South Asian populations. African populations exhibited

variation, with a minor allele frequency (MAF) of 0.259 for rs73885319 and rs60910145, and a MAF of 0.129 for rs71785313. As described by the 1000 Genomes Project Consortium (2015), the included African populations featured both African diaspora and resident East and West African populations. These variants confer resistance to trypanosomal infections, which may have contributed to positive selection for this chromosomal locus in individuals of African ancestry, as reflected in population allele frequency data (Madhavan & O'Toole, 2014). Similarly, these variants have also been found to increase risk for non-diabetic kidney disease (Genovese et al., 2010; Madhavan & O'Toole, 2014; Tzur et al., 2010). Individuals with two risk variants have a three-fold higher risk of developing ESRD compared to those with one or no risk alleles (Fine et al., 2012). In African American populations, *APOL1* variants contribute to approximately 70% of the excess risk of CKD progression and severity (Freedman, 2013; Obrador et al., 2017).

The myosin heavy chain 9 (*MYH9*) gene, located on chromosome 22, has also been implicated in CKD in African Americans. *MYH9* encodes the non-muscle myosin heavy chain IIA, which is involved in cellular movement (Vicente-Manzanares et al., 2009). Variants of *MYH9* are associated with focal segmental glomerulosclerosis, a cause of CKD (Kiffel et al., 2011; Kopp et al., 2008). *MYH9* and *APOL1* variants are in linkage disequilibrium, meaning they are inherited together, and the presence of both variants significantly increases the risk of CKD (Tzur et al., 2010). Another study found that *MYH9* risk alleles are more common in African American populations, with susceptible alleles ranging from 29 to 91% for African Americans, than in European American populations, where susceptible alleles ranged from three to 52% (Kopp et al., 2008). This provides genetic reasoning for the increased risk of focal segmental glomerulosclerosis in African American populations compared to European American populations.

While these genetic association are particularly relevant to African American populations, similar studies in other populations have identified different genetic variants associated with kidney dysfunction. Examples of genes implicated in non-African populations include Major histocompatibility complex I (*MHC I*), Uromodulin (*UMOD*), Angiotensinogen (*AGT*) and Cystatin 9 (*CST9*), as described below. A meta-analysis of East Asian populations found the major histocompatibility complex region, containing human leukocyte antigen (HLA) genes, to be implicated (Okada et al., 2012). MHC class I proteins are vital for adaptive immunity, allowing for recognition and elimination of both infected and abnormal cells (Wieczorek et al.,

2017). The SNP, rs3828890, located in *MHC I*, was found to be implicated in serum creatine concentration and creatine eGFR, two traits impacting kidney function. *UMOD*, a gene encoding uromodulin, a protein speculated to protect against kidney stones, urinary tract infection and kidney injury, was identified in the East Asian study as well as in a European meta-analysis, with rs12917707 found to be implicated in overall eGFR changes (Devuyst et al., 2017; Gorski et al., 2015). For South Asian populations, the *AGT* gene, encoding the angiotensinogen protein which regulates inflammation and blood pressure, was found to be implicated in kidney function (Lu et al., 2016). The *AGT* M235T polymorphism has been significantly associated with the deterioration of kidney function in multiple studies (Anbazhagan et al., 2009; Buraczynska et al., 2006). Alternatively, a Middle Eastern study found that a variant in the *CST9* gene, which encodes a secreted protein thought to be involved in inflammation and hematopoietic differentiation, affects kidney function. The rs13037490 variant was found to be protective of kidney function (Mohamed et al., 2022).

In terms of the genetics of CKD in SSA populations, a study on Yoruba individuals from south-west Nigeria found two *APOL1* variants (rs73885319 or rs60910145) strongly associated with non-diabetic CKD ($p = 0.006$) (Tayo et al., 2013). A separate study of Igbo individuals from south-eastern Nigeria also found a significant between the two *APOL1* risk alleles with CKD ($p = 5.2x10^{-3}$) (Ulasi et al., 2013). However, studies in South Africa have not found these *APOL1* risk alleles to be significant (Matsha et al., 2015; Nqebelele et al., 2019). It is worth noting that *APOL1* variants associated with CKD in African populations are not found to have the same prevalence or effect in non-African populations (Genovese et al., 2010; Thomson et al., 2014). Further research on mixed ancestry in South Africans has shown an association between *MYH9* variants and renal traits, particularly in diabetic individuals (Matsha et al., 2012).

These findings highlight the complex genetic landscape of CKD, with ethnic and regional differences playing a critical role in susceptibility. The presence of genetic risk variants in different populations emphasises the need for tailored approaches to understanding CKD and its treatment in SSA populations, as variants known to contribute to kidney dysfunction in non-African populations cannot be assumed to apply to SSA populations, due to the complexity of African genetics.

1.1.9 Transcriptional regulation in disease progression

TFs are known to be major drivers of a wide range of diseases, including cardiovascular disease, cancer, and diabetes, through their role in transcriptional regulation (Lee & Young, 2013). Transcriptional regulation is the control of gene expression at the transcriptional level using TFs, chromatin remodelers and co-regulators, to name a few. It is well-established that TFs regulate gene expression by binding to specific DNA sequences where they recruit transcription machinery and modulate chromatin structure. Precise regulation of gene expression is critical for normal cellular function and development; dysregulation can lead to pathological conditions (Lee & Young, 2013).

Transcriptional dysregulation may arise from various mechanisms, such as mutations or altered regulation of key transcription elements, leading to altered gene expression levels. One major contributor is the dysregulation of TFs, which occurs through mutations in either regulatory DNA sequences or in TFs themselves. These disruptions impair the normal control of gene expression, altering expression levels and contributing to disease development (Lee & Young, 2013). For example, mutations in multiple TFs have been implicated in diabetes (Maestro et al., 2007), while the overexpression of an oncogenic TF has been shown to drive acute lymphoblastic leukaemia (Sanda et al., 2012). Such findings emphasise the role of TFs not only in disease progression but also as valuable targets for disease prognosis and diagnosis (Dupont & Wickström, 2022; Lee & Young, 2013).

An example of transcriptional regulation in kidney disease is the TF KLF4, which is highly expressed in glomerular podocytes and has emerged as a potential therapeutic target. Reduced KLF4 expression has been observed in a unilateral ureteral obstruction (UUO) renal fibrosis model, suggesting antifibrotic activity (Xiao et al., 2015). Transient restoration of KLF4 expression in nephropathic model podocytes led to increased nephrin expression and lowered albuminuria (Hishikawa et al., 2018). Mechanistically, KLF4 regulates nephrin by reducing DNA methyltransferase 1 (DNMT1) binding to the nephrin promoter, lowering DNA methylation, and increasing histone acetylation, thereby promoting its activation (Hishikawa et al., 2018). These findings illustrate how TF dysregulation contributes to kidney disease and further reinforce their therapeutic relevance.

TF-target gene relationship analysis contributes novel insights into gene regulation and is useful in studying disease mechanism. There is no single universal method or resource used to identify targets of all TFs across the full spectrum of cellular conditions. However, there are four main methods currently used to evaluate human TF activities as described by Garcia-Alonso *et al.* (2019). Firstly, utilising literature-curated resources whereby high-quality regulatory TF-target interactions supported by experimental evidence. There are multiple databases storing these resources, for example TRRUST (Han et al., 2018) and ORegAnno (Lesurf et al., 2016). Importantly, coverage on such databases is skewed towards TFs that have been widely investigated. Secondly, chromatin immunoprecipitation sequencing (ChIP-seq) can be used to identify transcriptional regulators (Mundade et al., 2014). ChIP-seq has been proven to have high coverage and accuracy, as first shown by Roberton *et al.* (2007). ChIP-seq offers high resolution targets of DNA-binding regions. However, a large proportion of binding events can be non-functional interactions which require further validation (Fisher et al., 2012). ChIP-seq information is rapidly accumulating, hence the use of multiple public databases coming into effect. An example of this is ChIP-Atlas (https://chip-atlas.org), a web service containing high-throughput ChIP-seq datasets, which was developed to visualise TFs' genome-wide binding data (Zou et al., 2022). Thirdly, *in silico* prediction of TF-target interactions making using of TF binding motifs to scan regulatory sequences (Sinha et al., 2006). For this method, many false positives arise due to binding not occurring *in vivo* and, similarly to the previous method, many interactions can be non-functional requiring further confirmation. Lastly, inference from gene expression profiles from a specific condition which assumes that TF transcript levels correlate with expression levels of targeted genes (Margolin et al., 2006). This method favours the investigation of a specific phenotype, which can be useful for investigating disease mechanisms.

1.1.10 Investigating the genetics of CKD in an African context

Characterising genetic and phenotypic variation is vital for identifying genes involved in CKD disease susceptibility. However, genetic variants found in populations of African ancestry, such as Africans living in Europe and America, cannot be directly extrapolated to Africans residing in Africa due to genetic admixture, as well as environmental, cultural, and lifestyle differences (George et al., 2018; Tishkoff & Williams, 2002). While resident African populations are greatly impacted by CKD, the complexity and underrepresentation of resident African genetic diversity contribute to disparities in understanding of risk factors for CKD.

African regions, although understudied, are among the most genetically diverse in the world, making it challenging to characterise genetic contributions to disease due to high levels of population structure and limited representation of African genomes in genetic studies (Choudhury et al., 2020). An example admixture in Africa can be seen in the migration history of Southern Africa, which has contributed to the high genetic variation observed in present-day African populations as a result of gene flow (O'Connell et al., 2021). Significant events include the Bantu migration into KhoeSan-occupied regions 1,500 years ago and the arrival of European colonists into the Cape around 1,000 years ago (Berniell-Lee et al., 2009; Petersen et al., 2013; Sengupta et al., 2021). The establishment of ports, such as the Cape of Good Hope, which linked European and Eastern trade routes, further increased unions among different populations, affecting genetic variation in Southern African peoples (Petersen et al., 2013). This admixture had led to greater genetic diversity and a more complex population structure (Sengupta et al., 2021). A notable example is seen in the South African Coloured (SAC) population, which displays five-way admixture, with gene flow from African San, African non-San, East Asian, South Asian, and European ancestries (De Wit et al., 2010). Therefore, identifying genetic loci associated with CKD in resident African populations is crucial to advancing our understanding of CKD genetics specific to African individuals. Data from other non-African populations cannot be assumed to be applicable to South Africans (George et al., 2018).

Another challenge in characterising African genetics and disease lies in the design of association studies. Genome-wide association studies (GWAS) are among the most powerful tools to investigating gene-disease interactions. However, less than two percent of individuals involved in GWAS are African (Sirugo et al., 2019). This underrepresentation is concerning, as GWAS findings vary significantly across ethnic groups, meaning that result may not be applicable to all populations. The limited African representation in GWAS hampers our ability to fully understand the genetic underpinnings of disease, exacerbating health disparities between African and European populations, the latter of which have been extensively studied (Sirugo et al., 2019).

1.1 PEROXIDASIN

Haem-peroxidases are haem-containing enzymes involved in immune reactions and defence mechanisms, functioning by catalysing the reduction of hydrogen peroxide (Cheng & Shi, 2022; Ingraham & Meyer, 1985; O'Brien, 2000). Among the eight haem-peroxidases found in humans are myeloperoxidase (MPO), thyroid peroxidase (TPO), eosinophil peroxidase (EPO) and PXDN, our protein of interest, along with several others (Cheng et al., 2008). PXDN, also called Vascular Peroxidase-1 (VPO1), was first discovered in *Drosophila melanogaster* by Nelson *et al.* (1994), but it was later found that *PXDN* is expressed in various mammals as well, including humans (Cheng et al., 2008).

1.2.1 PXDN structure

The human *PXDN* gene, occurring in the 2p25.3 telomeric region, is 110 kilobases (kb) in size with 25 exons and 22 introns (Cheng et al., 2008). The PXDN protein has a molecular weight of 165 kilodaltons (kDa). As shown in Figure 2, PXDN consists multiple structural domains including the peroxidase domain (POX), which remains the catalytic site conserved in mammalian peroxidases that consists of a covalently linked haem group which increases enzyme catalytic activity. Other domains include five leucine-rich repeats (LRRs), LRR N-terminal or C-terminal domain, four immunoglobin (Ig) C-2 type domains and a von Willebrand factor type C domain (VWC) at the extreme C-terminus, which are all predicted to facilitate protein-protein or protein-ligand interaction (Cheng et al., 2008; Zederbauer et al., 2007). Due to these domains appearing in ECM proteins, it points to PXDN being involved in both protein-protein matrix interactions and cell adhesion. PXDN contains four highly glycosylated sites as well as six partially glycosylated sites (Soudi et al., 2015). PXDN oligomerises intracellularly into a homotrimer before being secreted (Paumann-Page et al., 2020; Soudi et al., 2015). Recent literature has indicated that PXDN could have important roles in various diseases (Cheng & Shi, 2022).

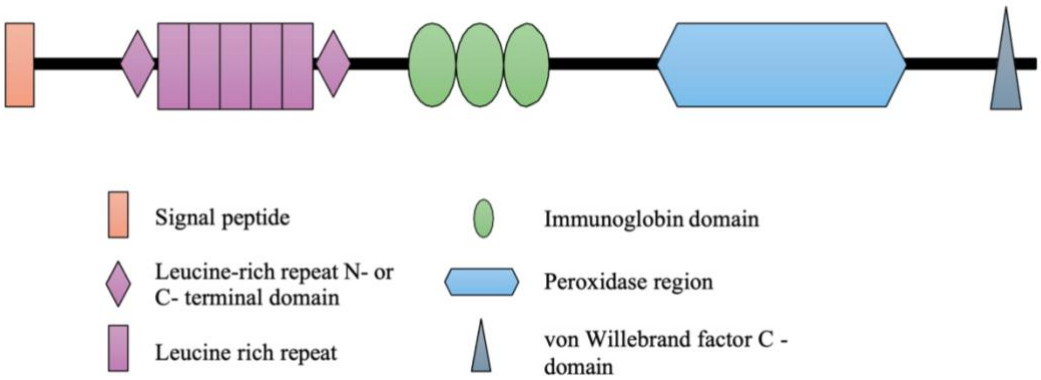


Figure 2. The structure of Peroxidasin (PXDN) and its constituents. PXDN consists of a signal peptide, five leucine-rich repeats, two leucine-rich N- or C- terminal domains, a peroxidase region and one von Willebrand factor C- domain.

1.2.2 Expression and regulation

PXDN messenger ribonucleic acid (mRNA) has been detected in multiple cells and tissues (Cheng et al., 2008; Shi et al., 2018), while PXDN protein has been identified in multiple tissues, including both embryonic and adult tissues, including the lung, brain and skeletal muscle (Cheng et al., 2008; Shi et al., 2018). PXDN is expressed and secreted in the kidney and myofibroblasts (Péterfi et al., 2009), with high expression observed in the eyes (Khan et al., 2011; Yan et al., 2014). Expression of PXDN has also been observed in the cardiovascular system, with high expression in the vascular wall and heart (Cheng et al., 2008).

PXDN expression is controlled by several mechanisms, including epigenetic regulation, mRNA transcript regulation and transcriptional regulation. Epigenetics refers to heritable changes in gene expression that do not involve alterations to the underlying DNA sequence. The most well-documented epigenetic mechanism is methylation, which has been found to downregulate *PXDN* expression in various cancers (Zhou et al., 2022). Additionally, in human umbilical cord vein endothelial cells (HUVECs), folic acid–which prevents oxidative damage–has been shown to activate *PXDN* methylation (Cui et al., 2018), demonstrating how methylation can regulate *PXDN* expression.

mRNA transcript regulation, the targeting of mRNA for degradation, is another mechanism that regulates *PXDN*. For example, miR-203A has been identified as a repressor of *PXDN*. In breast epithelial cells, overexpression of miR-203A led to a decrease in mesenchymal

characteristics, linking miR-203A to the regulation of *PXDN* expression and its function (Briem et al., 2019).

Although there is a need for more investigation into transcriptional regulation of *PXDN*, a few transcriptional regulators have been identified. *PXDN* expression has been shown to be regulated by nuclear erythroid-2 related factor 2 (NRF2) which is a redox sensitive TF (Hanmer & Mavri-Damelin, 2018). Increased NRF2 nuclear translocation resulted in greater PXDN protein expression in response to H_2O_2 This means that PXDN, may take part in redox-related ECM modifications and disease states where oxidation- reduction regulation is included, such as fibrosis as oxidative stress promotes fibrotic processes in the body (Hanmer & Mavri-Damelin, 2018).

Further, Snail family transcriptional repressor 1 (SNAI1), the EMT master TF, has been identified as a regulator of *PXDN* (Sitole & Mavri-Damelin, 2018). TGF-β1-induced EMT decreased expression of *PXDN* and increased SNAI1. Therefore, PXDN affects processes that include EMT, for example fibrosis, and SNAI1 is a regulator of *PXDN* expression.

Recent work has shown that serum response factor (SRF), through the upregulation of PXDN, regulates hepatic stellate cells (Guo et al., 2023). Hepatic stellate cells are a source of liver myofibroblasts, which produce ECM proteins known to promote liver fibrosis. Additionally, over expression of PXDN protected against senescence, a process which reduces liver fibrosis, while the opposite was seen when there was a depletion of PXDN. Further, reduced liver fibrosis was seen in *PXDN* knockout mice. This demonstrates that SRF is able to regulate *PXDN* influencing hepatic stellate cells, contributing to liver fibrosis (Guo et al., 2023).

1.2.3 Mechanism of action

PXDN has been evolutionarily conserved in Metazoa, highlighting its vital function (Fidler et al., 2014). PXDN plays a central role on the formation of basement membranes, which consist of a network of large glycoproteins providing structural support to tissues and influencing cellular activities (LeBleu et al., 2007). This process consists of collagen IV forming a structure of oligomerised triple helical promotors, characteristic of a mesh (Khoshnoodi et al., 2008). The C-terminal non-collagenous domains of the promoters then connect in a head-to-head fashion forming a hexamer, reinforced by sulfilimine bonds between hydroxylysine residues

and methionine (Vanacore et al., 2009). This is done through hydrogen peroxide ($H_2O_2^-$) and bromide (Br^-) ions being utilised by PXDN to form the reactive intermediate, hypobromous acid (HOBr) (Bhave et al., 2017; McCall et al., 2014).

1.2.4 Roles of PXDN in disease

PXDN has been shown to affect development in both animal and human studies. An example of this seen in *Drosophila* PXDN mutants which showed malformed and torn collagen IV networks as well as impaired muscle basement membranes, showing how development is affected by PXDN sulfilimine crosslinks conserving tissue integrity through strengthening basement membranes (Bhaveet al., 2012). Another example of affected development seen in the *Xenopus tropicalis* orthologue of *PXDN*, expressed during neural tube early development, which was shown to potentially alter morphogenesis ECM components (Tindall et al., 2005).

Human *PXDN* mutations affect eye development in microphthalmia and anterior segment dysgenesis, as a result of loss of basement membrane integrity in the developing eye (Choi et al., 2015; Zhu et al., 2021). Additionally, a study by Marutha *et al.* (2025) utilised a South African cohort of individuals with anterior segment dysgenesis to screen for pathogenic variants in *PXDN* where variants of uncertain significance were identified.

As well as PXDN's role in development, PXDN is also involved in cardiovascular diseases (Cheng & Shi, 2022; Yang et al., 2013). PXDN is induced by proinflammatory stimulators and metabolic factors affecting cardiovascular disease (Cheng & Shi, 2022). Hypertension has been shown to upregulate expression of *PXDN* highlighting a potential regulatory force on PXDN expression (Yang et al., 2013). A GWAS investigating variants related to blood pressure regulation in multiple different ancestries identified PXDN as a potential modulator of vascular structure and function, specifically in the African American population (Sung et al., 2018).

Increased PXDN expression has been found in multiple cancers (Zheng & Liang, 2018). Compared to normal tissues, higher expression was observed in various tumour tissues, including kidney renal clear cell carcinoma and kidney renal papillary cell carcinoma (Li et al., 2024). High PXDN levels promote proliferation of cells and metastasis and are associated with shorter disease-free survival (Li et al., 2024)

PXDN has many roles in cancer with the main implication thought to be its role of cross-linking collagen IV, providing structural integrity, as loss of basement membrane integrity is common in cancer (Banerjee et al., 2022; Glentis et al., 2017). Reduced basement membrane structure and changes in biomechanical properties result in the possibility for cell invasion through the basement membrane which is a process associated with metastatic spread of cancer (Banerjee et al., 2022; Chang & Chaudhuri, 2019; Glentis et al., 2017). Additionally, PXDN contributes to cancer progression by inhibiting oxidative stress, which leads to decreased apoptosis (Dougan et al., 2019). PXDN could therefore be a biomarker and therapeutic target for different cancers it has been shown to play a role in, for example, prostate cancer, pancreatic cancer, bladder cancer, thyroid cancer and stomach cancer, to name a few (Di et al., 2019; Li et al., 2023; Zhou et al., 2022). In addition to a strong role in cancers, PXDN is implicated in fibrosis, particularly in the kidney (Péterfi et al., 2009).

1.2.4.1 PXDN in fibrosis and CKD

As mentioned, by producing HOBr to form sulfilimine cross-links in collagen IV, PXDN provides mechanical stability of basement membranes which, if dysregulated, could contribute to fibrotic processes (Bhave et al., 2017). In a murine UUO model which represents pathological features of CKD such a fibrosis and tubular atrophy, a great amount of PXDN deposition was observed in the accumulating interstitial matrix suggesting that after injury, PXDN may contribute to renal fibrosis (Péterfi et al., 2009). Additionally, Colon *et al.* (2019) found that loss of PXDN lessens renal fibrosis. This therefore means that PXDN could potentially be used as a therapeutic target for both renal fibrosis and CKD. Two key factors linking PXDN to CKD include renal fibrosis and TGF-β.

PXDN is proven to respond to TGF-β, which has a substantial impact in fibrosis with fibroblasts being a main target (Péterfi et al., 2009; Sitole & Mavri-Damelin, 2018). In response to TGF-β, myofibroblasts, developed from fibroblasts or through EMT, secrete PXDN into the ECM, where it assists in ECM formation, thereby worsening fibrosis, as excessive ECM deposition is characteristic of fibrosis (Nogueira et al., 2017; Péterfi et al., 2009). Smad-dependent signalling activated by TGF-β promotes ECM protein synthesis (Shi et al., 2020). TGF-β1 increases PXDN mRNA expression in fibroblasts and, after twenty-four hours, a

twofold increase in PXDN protein was detected (Péterfi et al., 2009). After seventy-two hours, PXDN was revealed in the form of a dense network extracellular deposition. It therefore makes sense that TGF-β is proven to be increasingly upregulated as the extent of renal fibrosis worsens (Murakami et al., 1997; Yamamoto et al., 1996). Additionally, blocking of TGF-β leads to lessening extent of kidney fibrosis (Miyajima et al., 2000; Petersen et al., 2008; Sanderson et al., 1995). Because of this link, TGF-β is known to play a main role in CKD as well as for its function of inducing EMT, another key process in CKD (Meng et al., 2016). Since PXDN is responsive to TGF-β signalling, elevated PXDN expression is likely to play a direct role in kidney fibrosis.

TGF-β signalling contributes to fibrosis through several mechanisms. Firstly, TGF-β affects ECM remodelling through managing cell proliferation, survival, migration and gene expression of fibroblasts (Hinz, 2015). Secondly, TGF-β activates transcription of collagens and plays a role in posttranslational collagen modification through increasing its stability by enhanced cross-linking (McAnulty et al., 1991; Raghow et al., 1987). Lastly, TGF-β supresses matrix-degrading protease activity through inducing inhibitors, for example, tissue inhibits of metalloproteinases and plasminogen activator inhibitor 1 (Mauviel, 2005; Schiller et al., 2004).

1.3 Rationale for the current study

The high CKD incidence in SA, coupled with the absence of risk factors to explain this prevalence, emphasises the need for further investigation to bridge this gap in knowledge. Despite the potential link between PXDN and renal fibrosis, the regulatory mechanisms of *PXDN* are largely unexplored. Identification of these mechanisms may provide insights into the molecular mechanisms underlying kidney function and disease. Additionally, by investigating if variation in *PXDN* and its transcriptional regulators are implicated in kidney dysfunction, this foundational study seeks to uncover potential genetic factors involved in kidney dysfunction that may contribute to the high CKD incidence in SA. Discovering genetic variants and transcriptional regulators associated with *PXDN* in African populations may assist in potentially identifying therapeutic targets, contributing to work that will eventually reduce the prevalence as well as the number of deaths caused by PXDN-related conditions, such as kidney fibrosis and potentially CKD.

1.4 Aims and objectives

Aim 1

To identify the transcriptional regulators of *PXDN* using publicly available ChIP-Seq data. The objectives of Aim 1 are to:

1. identify transcriptional regulators of *PXDN* in kidney cells using ChIP-Atlas,
2. map significant TFs, and
3. determine how *PXDN* transcriptional regulators identified may impact kidney dysfunction.

Aim 2

To investigate the genetic variation found in *PXDN*, its transcriptional regulators (identified in Aim 1) and genes previously implicated in kidney dysfunction, in a South African cohort, to determine whether any changes at the DNA level may be associated with the kidney dysfunction seen in these populations. The objectives of Aim 2 are to:

1. perform quality control and account for population substructure in our data using principal component analysis (PCA),
2. conduct an association analysis to identify whether genetic variation in *PXDN*, its transcriptional regulators, and previously discovered genes which play a role in kidney dysfunction, are associated with kidney dysfunction in a South African population, using eGFR levels as a measure, and
3. determine whether changes at the DNA level may impact kidney dysfunction using different bioinformatic tools.

CHAPTER 2. Identifying the transcriptional regulators of *PXDN* using publicly available ChIP-seq data

2.1 INTRODUCTION

PXDN plays crucial roles in synthesis of collagen IV in basement membrane development, contributing to ECM stability, and homeostasis (Cheng et al., 2011). Dysregulated PXDN is implicated in impaired host defence, vascular disease pathogenesis and fibrosis, among other conditions (Cheng & Shi, 2022). In the kidney, PXDN is involved in structural maintenance and has been linked to fibrotic diseases, for example CKD (Cheng & Shi, 2022; Péterfi et al., 2009). Despite clear evidence for the significant role PXDN plays in normal and disease processes, the regulation of *PXDN* is still poorly understood. Some mechanisms of regulation recently identified include epigenetics regulation, whereby modifications alter DNA expression without changing the DNA sequence as seen with DNA methylation downregulating *PXDN* expression (Zhou et al., 2022), micro ribonucleic acid (miRNA)-mediated regulation, whereby the binding of miRNA leads to mRNA degradation as seen with miR203a being identified as *PXDN* repressor (Briem et al., 2019), and our main focus, transcriptional regulations, where the binding of TFs to the *PXDN* promoter causes activation or repression of gene expression. While further investigation is needed, a few transcriptional regulators have been identified, as discussed in section 1.2.2. *PXDN* expression has been shown to be regulated by NRF2, which is involved in redox-related ECM modifications and disease (Hanmer & Mavri-Damelin, 2018), SNAI1, which affects processes related to EMT (Sitole & Mavri- Damelin, 2018), and SRF, where PXDN has been shown to protect against senescence (Guo et al., 2023). To further explore the transcriptional regulation of *PXDN*, we utilised publicly available ChIP-seq data.

ChIP-based methods can be used to profile epigenomes in different types of cells in order to gain information about gene regulation in both normal and disease states and developmental processes (Park, 2009). ChIP-seq, which combines chromatin immunoprecipitation with sequencing, is a powerful technique to identify DNA binding sites of TFs and additional

DNA-binding proteins such as chromatin modifiers, providing insight into the regulation of PXDN (Barski et al., 2007; Johnson et al., 2007). ChIP works by enriching fragments of DNA, which are bound to specific proteins, such as TFs and histones, and have thereby been immunoprecipitated, primarily by using protein-specific antibodies (Park, 2009). This process involves cross-linking proteins to their interacting DNA using formaldehyde, fragmenting the chromatin through sonication or enzymatic digestion, and then performing chromatin immunoprecipitation with antibodies targeting the protein of interest (Wells & Farnham, 2002). Lastly, reversal of the crosslinks occurs, and the DNA released is assayed to identify sequences where the specific protein was bound. For ChIP-seq, high throughput next generation sequencing of DNA occurs after purification. This allows for alignment to the genome to locate regions enriched in the sample in order to distinguish TF binding sites across the genome (Park, 2009).

ChIP-seq experiments tend to have a single protein target and all DNA bound to that protein is immunoprecipitated and sequenced, thereby generating large amounts of data, which can be placed in repositories such as ChIP-Atlas (Zou et al., 2022). ChIP-Atlas is a web-based data mining tool of omics datasets, including a library of ChIP-seq datasets for a variety of species, from several public databases (Oki et al., 2018). ChIP-Atlas compiles, processes and visualises data, enabling a wide range of applications in gene regulation studies and TF analysis.

This chapter aimed to identify transcriptional regulators of *PXDN* in the kidney using publicly available ChIP-seq data from ChIP-Atlas, focusing on the promoter and first intronic regions. The objectives included identifying transcriptional regulators of *PXDN* in kidney cells using ChIP-Atlas, mapping significant TFs, and determine the degree to which *PXDN* transcriptional regulators identified may impact kidney dysfunction.

2.2 METHODOLOGY

2.2.1 ChIP-Atlas data collection and analysis

Publicly available ChIP-seq datasets on ChIP-Atlas were used to identify potential TFs of *PXDN* in kidney and kidney diseases. ChIP-Atlas, accessible at https://ChIP-Atlas.org, is a data-mining tool that provides information about transcription regulation through integration of data sets from high-throughput ChIP-seq and DNA-seq (Zou et al., 2022). This study utilised ChIP-Atlas due to its extensive collection of over 300,000 experiments compared to other ChIP-seq databases, such as Cistrom DB and ReMap, which have approximately 20,000 and 55,000 fewer experiments, respectively (Oki et al., 2018; Zou et al., 2022). Additionally, ChIP-Atlas contains a substantial amount of kidney cell data with approximately 7,500 kidney cell types (Oki et al., 2018; Zou et al., 2022). ChIP-Atlas is updated monthly, and results are standardised through a uniform data process pipeline applied to all ChIP-seq, ATAC-seq and DNase-seq data experiments (Oki et al., 2018). Following processing, analysis is conducted using model-based analysis of ChIP-seq 2 (MACS2) (Oki et al., 2018; Zhang et al., 2008), which is utilised for peak detection using ChIP-seq data to identify significantly enriched loci (peaks) in the genome where TFs bind (Oki et al., 2018; Zhang et al., 2008). MACS2 provides detailed information for each peak, including genome coordinates, false discovery rate (FDR), p-value, fold enrichment and peak centre (Zhang et al., 2008). For the ChIP-Atlas enrichment analysis, the focus will be transcriptional regulators that bind within *PXDN* in kidney cells.

The following flow chart (Figure 3) provides an overview of the methodology used in this study. Starting with data pre-processing, it outlines the steps carried out, from ChIP-Atlas processing uploaded data to peak detection done by MACS2. For this, filtration was done to include data relevant to our aim, such as the selection of TFs in kidney cells. This was followed by data annotation, analysis and visualisation using ChIPseeker tools in order to successfully investigate TFs located in the promoter and intron 1 regions. An expression analysis using Protein-Atlas data was then done to investigate the expression levels of identified TFs in kidney tissue. Finally, Find Individual Motif Occurrences (FIMO) was used to identify putative DNA binding sites and National Centre for Biotechnology Information (NCBI) Genome data viewer was used map putative binding sites of significant TFs, TFs that met the FIMO p-value and q-value thresholds of 0.05.

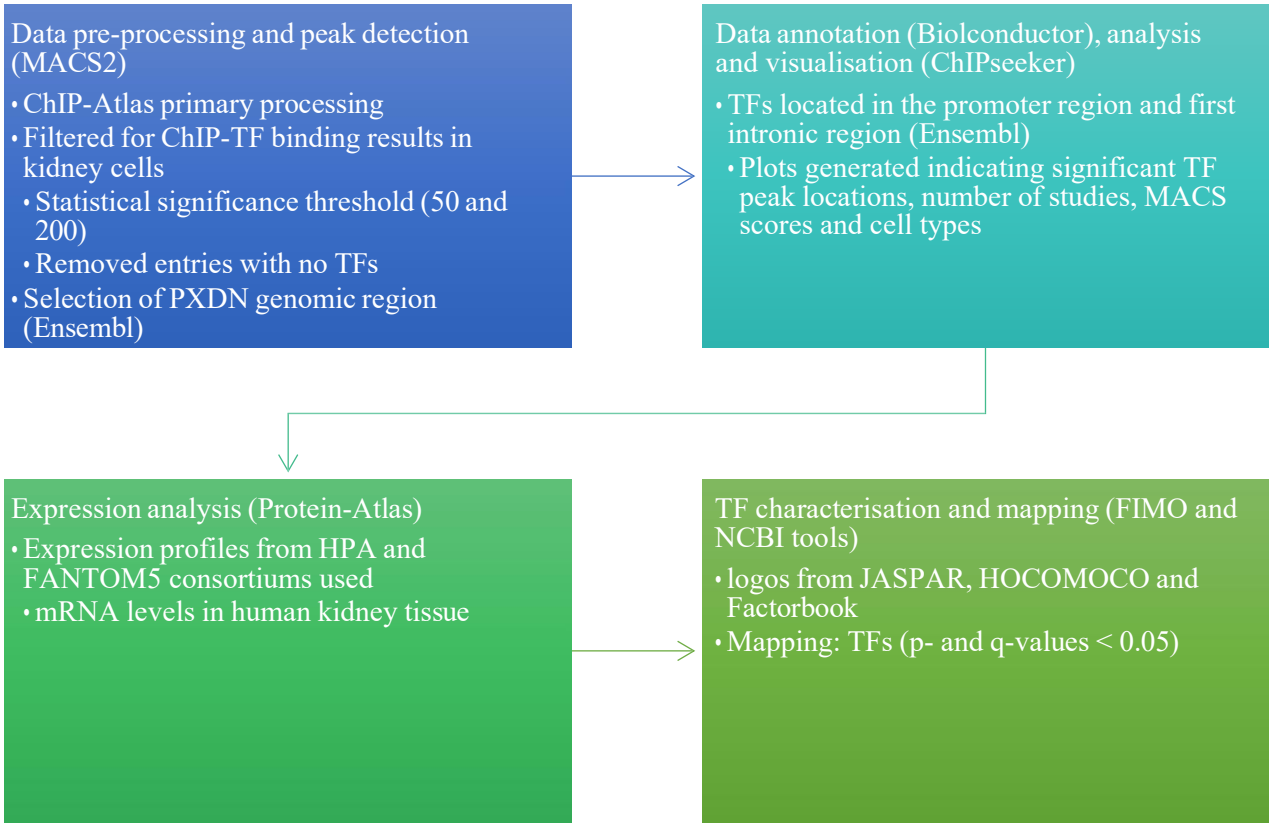


Figure 3. Overview of ChIP-seq data analysis. Primary processing is done by ChIP-Atlas. This is followed by peak detection, selection of region of interest, motif analysis and annotation and visualisation. Following data interpretation, each transcription factor (TF) was characterised and mapped to their respective binding sites in the PXDN gene, focusing on both the promoter and first intronic regions.

2.2.2 Data pre-processing and peak detection

ChIP-Atlas performs primary processing of raw sequencing data from ChIP-Seq experiments uploaded to the database with the following methods, as described in their documents (https://github.com/inutano/ChIP-Atlas/wiki). Raw sequence data were downloaded and converted to Fastq format using the SRA Toolkit (version 2.3.2-4) using default mode and concatenated into a single file. The sequence data were aligned to reference genomes using Bowtie2 (version 2.2.2) in default mode. For paired end reads, the -1 option was used to specify 'forward' reads, and the -2 option for the 'reverse' reads. Reads were mapped to the reference genome in order to identify perfect or near perfect matches to the read. Bowtie was chosen for this task due to its demonstrated high-quality throughput in comparison to other short sequence mapping tools (Hatem et al., 2013; Langmead et al., 2009). The genome assembly used was GRCh38 (*H. sapiens*) as it is the most current version (Lee et al., 2022). The output SAM files were then binarized into BAM format using SAMtools (version 0.1.19), then sorted and

polymerase chain reaction (PCR) duplicates were removed. Bedtools (version 2.17.0) was used to calculate BedGraph-formatted coverage scores, expressed in reads per million (RPM) mapped reads. These BedGraph files were then binarized into BigWig format with USCS tools (version 4). The BAM files, now containing BedGraph scores, were used for peak call using MACS2 (version 2.1.0) in BED4format. The significance threshold for MACS ranged from 50 to 500 (Langmead et al., 2009). Lastly, the BED4 files were converted into BigBed format using UCSC tools. ChIP-Atlas Peak browser (https://ChIP-Atlas.org/peak_browser) was used to obtain peaks found in *Homo sapiens* (GRCh38) which have undergone processing by ChIP- Atlas, as described above. Peaks are regions in the genome where TFs or other proteins bind to DNA, indicating regulatory activity. While MACS2 was employed for peak detection, subsequent filtration of results was not performed by ChIP-Atlas.

The peak detection results were refined to include entries categorised as 'ChIP: TF and others' along with 'kidney cells. To evaluate the impact of significance thresholds, two thresholds, 50 and 200, were applied. The statistical significance of peaks, represented by the Q value, is calculated by MACS2 using the formula: -10 * Log10(MACS2 Q-value). This means that if a MACS threshold of 50 is applied, peaks with a Q-value less than $1x10^{-5}$ will be selected for analysis. A threshold of 50 was used to include a greater number of peaks ensuring that potentially relevant binding sites, including those with weaker detection, were considered in the analysis. This broader threshold increased the sensitivity of the analysis, allowing for the capture of a wider range of TFs. On the other hand, a higher threshold of 200 was applied to refine the analysis and focus on TFs with stronger, more statistically significant binding patterns, providing high confidence in results. By using these two thresholds, we aimed to balance a comprehensive exploration of potential binding sites with the identification of the most reliable TFs. The resulting output was in .bed file format, containing TF binding peak data mapped across the entire genome specifically derived from kidney cell types. The .bed file containing TF peak data from ChIP-Atlas was modified to ensure compatibility with R (version 4.4.0). Entries with "Epitope" or "GFP" in the TF column were removed, as these are not TFs. Lastly, the most current genomic coordinates from Ensembl (https://www.ensembl.org) were used to define the full *PXDN* gene region on for GRCh38: Chromosome 2, 1631887-1744852 on the reverse strand (GRCh38:CM000664.2, ENSG00000130508.11) (Dyer et al., 2025).

2.2.3 Data annotation, analysis and visualisation

The ChIPseeker package, utilised for ChIP-seq data analysis in R, was employed for motif analysis, peak annotation, comparison, and visualisation. This approach was used to assess peak coverage and profiles of peaks (Yu et al., 2015). Other R packages used for subsequent analyses included Tidyverse, a data science package used for data manipulation and visualisation, Bioconductor, used for genomic data analysis, and RColorBrewer, used for creating colour palettes for data visualisation.

The Bioconductor transcript database, 'TxDb.Hsapiens.UCSC.hg38.knownGene', was used for annotation. The database provides comprehensive annotations of *Homo sapiens* genomic features based on the UCSC Genome Browser hg 38 (GRCh38) assembly, including information on genes, transcripts, exons, coding sequences and other transcript features. Peaks obtained from ChIP-Atlas were annotated based on their overlap with entries in this transcript database. Peak Annotation was performed using the 'annotatePeak' function. The transcription start site (TSS) region was defined with a wide radius of ±5 kb (kilobase) to include distal regulatory elements which may influence transcriptional activity of the *PXDN*. The intron 1 coordinates for the *PXDN* gene, as obtained from Ensembl and the UCSC Genome Browser, are from 1,693,135 to 1,744,255 on chromosome 2. Peak locations were annotated to specify whether a peak was found in the TSS, exon, 5′ or 3′ untranslated region (UTR) and intronic or intergenic regions. The two key regions of the *PXDN* gene, the promoter and first intronic regions, were the focus of this study. These two regions play important roles in transcription, with the first intronic region containing vital regulatory elements, for example enhancers and silencers, which affect transcription and the promoter region being responsible for the initiation of transcription as well as being the typical binding site for TFs regulating transcription (Haberle & Stark, 2018; Li et al., 2012).

Analysis and visualisation were carried out using ChIPseeker to annotate TF binding peaks in relation to their location, classify peaks based on their genomic locations and generative visual representations of the data, as done by Naidoo (2024). A pie chart was generated to demonstrate the distribution of TFs identified. Additionally, a stacked bar plot was created to visualise the distribution of TF-binding loci relative to the TSS. For both the promoter and intron 1 regions bar plots were generated to visualise the distribution of regulators identified across the various cell types. Tables were also created to show experiment details, including cell type

and treatment, if any. This was included as TF binding is context-dependent, hence cell treatments have the ability to influence binding to *PXDN*. Examples of this include targeting gene expression, chromatin accessibility, and post-translational modifications (Weidemüller et al., 2021). For instance, chromatin modifications can influence TF binding by altering chromatin accessibility (Sönmezer et al., 2021).

2.2.4 Expression analysis

Protein-Atlas, a publicly available database providing information on human tissue-specific protein expression, was used to interpret results obtained from ChIPseeker in order to investigate TF expression levels in the human kidney. Protein-Atlas was used to obtain mRNA expression profiles for each TF with MACS scores greater than 200. Additionally, potentially significant TFs, identified using a MACS threshold of 50, were included in the analysis based on their potential implication in kidney disease pathogenesis, as described in current literature (Appendix B).

The mRNA expression levels obtained are based on ribonucleic acid (RNA)-seq data generated by the Human Protein Atlas (HPA), Genotype-Tissue Expression (GTEx) portal and Cap Analysis of Gene Expression (CAGE) data generated by the Functional Annotation of Mammalian Genomes 5 (FANTOM5) consortium. The HPA data consists of tissues collected from the Uppsala Swedish Biobank and includes 1 206 cell lines, 40 human tissues and 18 blood cell types (Uhlén et al., 2015). The GTEx project consists of multiple human postmortem tissues (Aguet et al., 2020). The project comprises samples from up to 54 non-diseased tissue sites across approximately 1,000 deceased individuals. RNA-seq data from 36 of their tissue types was mapped based on RNA-Seq by Expectation-Maximisation (RSEM version 1.3.0), a computational tool for quantifying gene expression levels from RNA-seq data for different genes and transcripts. The FANTOM5 project featured thorough expression profiles and functional annotation of mammalian cell- type specific transcriptomes using CAGE (Lizio et al., 2019). CAGE data was acquired from the FANTOM5 repository (https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_v2/extra/) and mapped to Ensembl (version 109) (Takahashi et al., 2012).

2.2.4 TF characterisation and mapping

In order to characterise TFs and gain insight into putative binding motifs, the following databases were utilised to obtain both TF logos and position-weight matrices (PWMs) for each TF: JASPAR (https://jaspar2020.genereg.net), Factorbook (https://www.factorbook.org) or HOCOMOCO (https://hocomoco11.autosome.org). These three databases were used as no single database contained information for all TFs identified in this study.

FIMO, Find Individual Motif Occurrences (https://meme-suite.org/meme/tools/fimo), was utilised to identify putative TF binding sites within the peak region as identified in the data from ChIP-Atlas. FIMO is a software tool which is used for scanning protein or DNA sequences with position-specific scoring matrices which describe motifs (Grant et al., 2011). FIMO calculates a log-likelihood ratio score for each motif based on its position in the sequence. Dynamic programming is then used to transform the log-likelihood ratio scores into p-values (Staden, 1994). This transformation relies on a zero-order null model with sequences generated at random with per-letter background frequencies used. A bootstrap method is then used to calculate FDRs (Storey, 2002). However, since the FDR does not consistently increase or decrease with the p-value, a q-value is provided. The q-value represents the minimal FDR threshold at which a p-value is considered significant (Storey, 2003).

The coordinates of the ChIP-seq enriched region provided by ChIP-Atlas, the peak region, were converted into FASTA format using NCBI tools for FIMO analysis (Meme Suite 5.5.7). The NCBI *H. sapiens* chromosome 2 GRCh38.p14 primary assembly (NC_000002.12) was used for this to allow for consistency. FIMO was then used to scan the binding site range for individual matches to the TF PWMs (Grant et al., 2011).

The sequence matches of putative binding sites from FIMO were used to search for binding sites within *PXDN* regions using NCBI Genome Data Viewer (NC_000002.12). Sequences found within the enriched peak region provided by ChIP-Atlas were recorded. TFs binding to positive or negative DNA strands were included in the analysis, TFs can regulate transcription irrespective of the strand to which they bind. A p-value threshold of 0.05 was applied in FIMO, and only significant TFs with both p-values 0.05 and a q-value below 0.05 were retained for TF mapping. As this was exploratory research, a p-value threshold of 0.05

was selected to capture potentially meaningful results only, while a q-value threshold of 0.05 was used to effectively balance the control of false positives with the ability to identify significant findings.

2.3 RESULTS

2.3.1 Data pre-processing and peak detection

After obtaining ChIP-seq data, R was used for motif analysis, annotation and visualisation. *PXDN* ChIP-TF binding results that underwent primary processing (30,929 experiments) were filtered based on significance thresholds and those derived from kidney cell experiments (2,535 experiments) in ChIP-Atlas. A significance threshold of 200 was applied and this data was analysed further below, with additional results using the less stringent threshold of 50 shown in Appendix C. After removal of problematic entries with no TFs and selection of the *PXDN* genomic region using Ensembl coordinates, 382 sites within *PXDN* remained.

2.3.2 Data analysis, annotation and visualisation

The analysis focused on genomic distribution (Figures 4 and 5), genomic location (Figures 6 and 7) and cell type specificity (Figures 8 and 9). Table 1 shows the different kidney cell types used in the study, while Tables 2 and 3 summarise ChIP-Atlas experiment details for regulators identified in the promoter and intron 1 regions, respectively. ChIP-Atlas defined the promoter region as both 2 kb upstream and downstream of the TSS. Consequently, TFs classified as 'promoter TFs' follows this same definition. Table 1 provides an overview of kidney cell types used in this study and their labels used in the data visualisation.

Using a MACS threshold of 200, equivalent to a q-value of $1x10^{-10}$, the genomic locations of the 382 *PXDN* binding sites were annotated (Figure 4). The annotated regulators showed a wide distribution pattern with the first intron being responsible for 22% of binding sites while the promoter region contained three times fewer binding sites, with a total percentage of 7.6% (Figure 4). Overall, 10 sites (2.6.%) were detected 1 kb from the TSS and 19 sites (5%) were located 1-2 kb from the TSS. 24 sites (6.2%) were in the 3′ UTRs, 7 sites (1.8%) were found in the first exon, 84 sites (22%) were in the first intron, 120 sites (31.4%) were found in remaining introns. The remaining 118 sites (30.9%) were found in the distal intergenic regions. These results demonstrated that the majority of regulator binding sites were identified in the intronic regions (excluding the first intron), followed by distal intergenic regions, the

first intron, and to a smaller extent, the promoter regions, the 3′ UTR and the first exon.

Table 1. ChIP-Atlas labels of cell types and their corresponding classifications.

Cell type description	Label
Human embryonic kidney 293 cells (HEK-293)	293
Human renal cell carcinoma cell line	786-O
Human proximal tubular epithelial cells	HCK8
Human renal epithelial cells from the cortex and glomerular region	HRE
Cortex cells of the kidney	KidneyCortex
Polycystic kidney disease (PKD) cells	Polycystickidney
Renal cell carcinoma cells	Renalcellcarcinoma
Renal proximal tubule epithelial cell line	RPTEC
Rhaboid tumour cells of the kidney	TTC-1240
Retroperitoneal tumour of renal cell carcinoma cell line	FU-UR-1

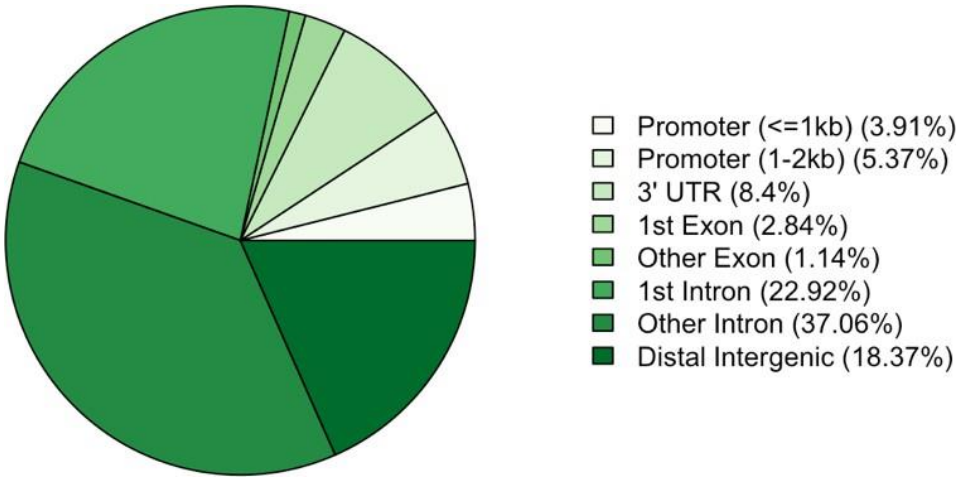


Figure 4. Pie chart of the percentages of *PXDN* binding sites according to peak location. Different shades of green represent various genomic regions of the *PXDN* gene. 3′ UTR refers to 3′ untranslated regions. For the promoter region, 3.91% of binding sites were found up to 1kb from the TSS while 5.37% were found 1-2kb from the TSS. 8.4% were found in the 3′ untranslated region, 2.84% were found in the first exon and 1.14% was found in remaining exons. 22.92% were found in the first intron and 37.06% were found in remaining introns. The remaining 18.37% accounted for distal intergenic binding sites.

The distribution of binding sites in relation to the TSS is illustrated in Figure 5. Most binding

sites (approximately 80%) were present 10-100 kb from the TSS. Approximately 13% of binding sites were found 5-10 kb upstream of the TSS. The remaining binding sites were distributed downstream of the TSS: 5% were shown to be 1-3 kb downstream and 2.5% were found within 1 kb downstream of the TSS.

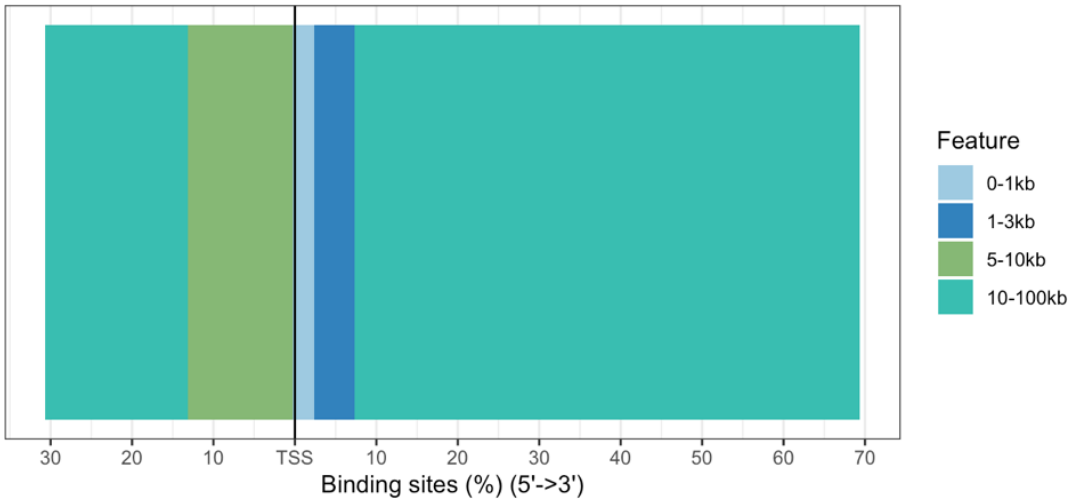


Figure 5. Distribution of binding sites relative to the transcription start site (TSS). The TSS is indicated by the black vertical line. Different colours were used to denote different distances from the TSS. Light blue indicates up to 1 kb from the TSS, dark blue shows 1-3 kb from the TSS, green shows 5-10 kb from the TSS and turquoise shows 10-100 kb from the TSS.

Figures 6 and 7 illustrate the number of studies in which each regulator was identified for the promoter and intron 1 regions, respectively. The figure also indicates the number of studies included which had MACS scores greater than 500, the most stringent threshold for binding peaks on ChIP-Atlas. For both the promoter and first intronic regions, more than half of the regulators had MACS scores greater than 500. For the promoter region, the TFs *CTCF, KAT6B, MEF2B* and *RAD21* were identified in the greatest number of studies (Figure 6). Conversely, three regulators were identified in only one experiment in the promoter region. For the first intronic region, *AR, NR3C1* and *RELA* were identified in the greatest number of experiments, with seven regulators identified in only one experiment (Figure 7).

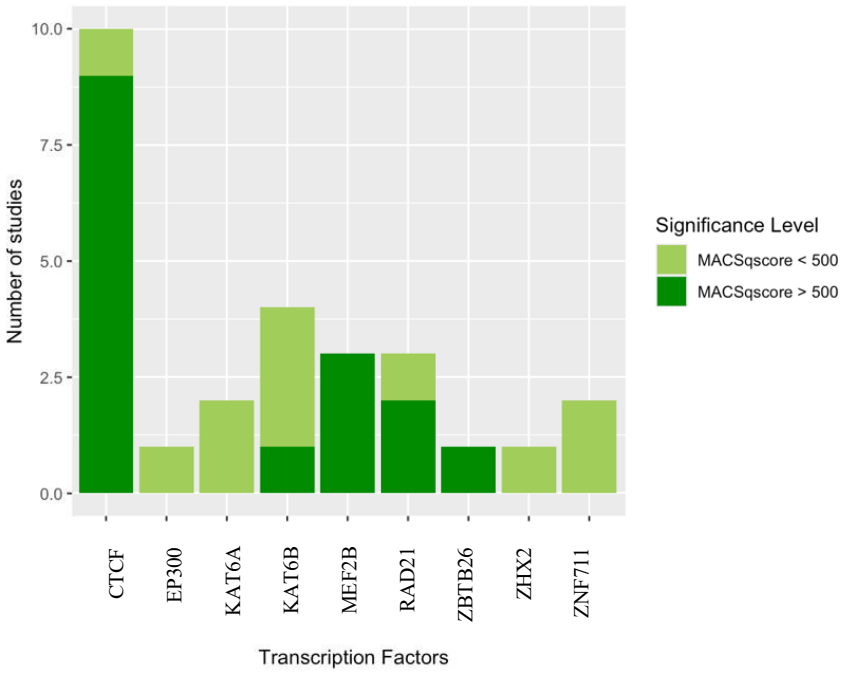


Figure 6. Bar graph showing regulators binding to the promoter of *PXDN* in kidney cells, as identified from ChIP-Atlas. The graph displays the number of studies supporting each TF, with MACS scores indicated. Dark green bars represent TFs with a MACS score greater than 500, while lighter green bars indicate TFs with a MACS score less than 500.

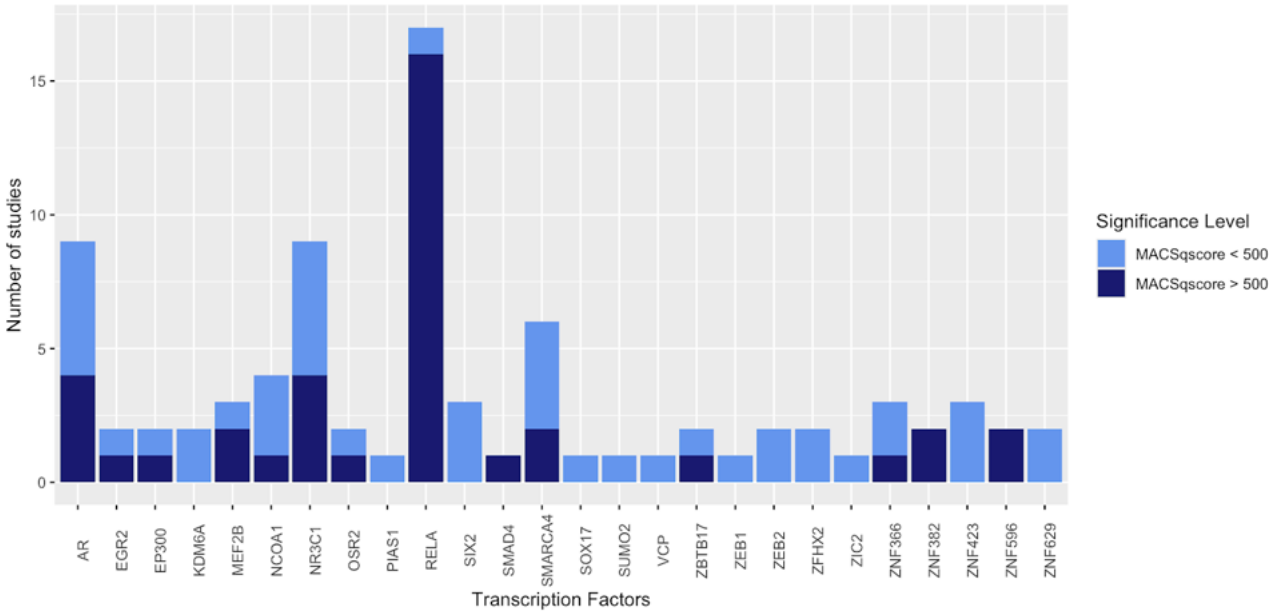


Figure 7. Bar graph showing regulators binding to the first intronic region of *PXDN* in kidney cells, as identified from ChIP-Atlas. The graph displays the number studies each regulator was identified in, with MACS scores indicated in different shades of blue. Dark blue bars represent regulators with a MACS score greater than 500, while lighter blue bars indicate regulators with a MACS score less than 500.

Tables 2 and 3 summarise experiments identifying regulators in the promoter and intron 1 regions with MACS scores of at least 200, detailing the associated cell types and any treatments administered. Additionally, certain regulators identified using the least stringent threshold of 50 (Figures 1 and 2 in Appendix C) were included after a literature review highlighted some of these, including *ARNT*, *CREB1* and *TWIST1*, as having clear roles in kidney function and disease. Cell treatments mentioned in Tables 2 and 3 are described in Appendix D.

Human embryonic kidney 293 cells (HEK293) were used in the majority of experiments in which TFs were identified for both the promoter and first intronic regions. Most experiments used untreated cells. However, treated cells were included in three experiments for TFs in the promoter region, all involving *CTCF*, and in eleven experiments for TFs in the first intronic region, including *AR, NCOA1, NR3C1, PIAS1, RELA, SMARCA4, SOX17 and SUMO2*.

Table 2. TFs identified in the promoter region of *PXDN*.

TF	Cell type	Treatment
CTCF	786-O	Normoxia
	RPTEC	Untreated
	TTC-1240	Untreated
	HRE	Untreated
	293	Untreated
		Expression of CpG-free plasmid expressing β-galactosidase
		Expression of BORIS from CpG-free plasmid
EP300	Kidney cortex	Untreated
KAT6A	293	Untreated
KAT6B	293	Untreated
MEF2B	293	Untreated
RAD21	293	Untreated
ZBTB26	293	Untreated
ZHX2	Renal cell carcinoma	Untreated
ZNF711	293	Untreated

Cell types include 786-O (human renal cell carcinoma cell line), RPTEC (renal proximal tubule epithelial cell line, TTC-1240 (Rhaboid tumour cells of the kidney, HRE (human renal epithelial cells from the cortex and glomerular region), 293 (human embryonic kidney 293 cells), kidney cortex cells and renal cell carcinoma cells.

Table 3. TFs identified in the first intronic region of *PXDN*.

TF	Cell Type	Treatment
AR	293	10 nM R1881
ARNT	HCK8	Untreated
	786-O	Untreated
CREB1	Kidney Cortex	Untreated
EGR2	293	Untreated
EP300	786-O	Untreated
	Renal cell carcinoma	Untreated
KDM6A	293	Untreated
MEF2B	293	Untreated
NCOA1	293	Dexamethasone
NR3C1	293	Dexamethasone
		ML-792 and dexamethasone
OSR2	293	Untreated
PIAS1	293	Dexamethasone
RELA	Renal cell carcinoma	Untreated
	786-O	DMSO
		Bortezomib 5nM for 9hr
		Untreated
		PBRM1 knockout
SIX2	Kidney Cortex	Untreated
SMAD4	786-O	Untreated
SMARCA4	786-O	PBRM1 knockout
		Untreated
SOX17	293	10 um CHIR for 24 hours
SUMO2	293	Dexamethasone
TWIST1	293	TWIST1 and ALX4 transfection
VCP	FU-UR-1	Untreated
ZBTB17	293	Untreated
ZEB1	293	Untreated
ZEB2	293	Untreated
ZFHX2	293	Untreated
ZIC2	293	Untreated
ZNF366	293	Untreated
ZNF382	293	Untreated
ZNF423	293	Untreated
ZNF596	293	Untreated
ZNF629	293	Untreated

Cell types include 293 (human embryonic kidney 293 cells), HCK8 (human proximal tubular epithelial cells), kidney cortex cells, 786-O (human renal cell carcinoma cell line), renal cell carcinoma cells and FU-UR-1 (retroperitoneal tumour of renal cell carcinoma cell line).

Figures 8 and 9 provide a visual representation of the distribution of experiments conducted across different cell types. For the promoter region, most experiments were conducted in the HEK293 cell line (Figure 8). CTCF was found in five different cell types, whereas other TFs were associated with only one cell type. In the first intronic region, experiments conducted in 293-HEK cells were also predominant, with only six out of 26 experiments not conducted in this cell type (Figure 9). In a total of 23 experiments, four different TFs were identified in experiments conducted in 786-O human renal cell carcinoma cell line. Other cell types, including FU-UR-1, kidney cortex cells and renal cell carcinoma cells, were used in more than three studies each.

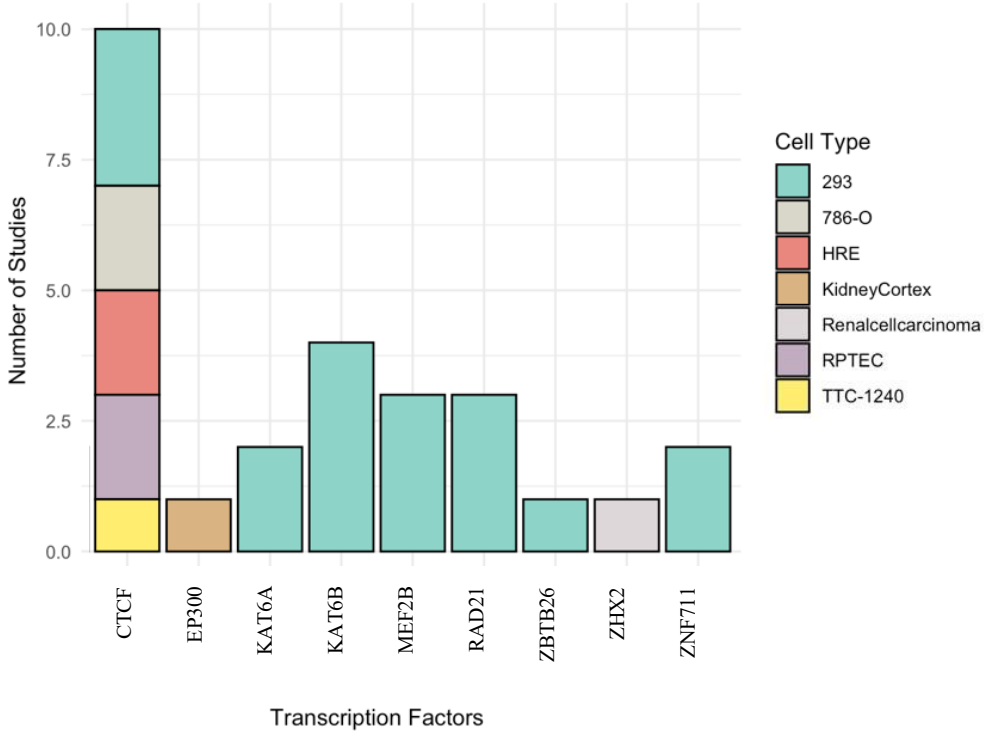


Figure 8. Bar graph showing the number of different cell types used in each TF experiment from ChIP-Atlas for the promoter region. Cell types of each experiment are shown including 293 (human embryonic kidney 293 cells), 786-O (human renal cell carcinoma cell line), HRE (human renal epithelial cells from the cortex and glomerular region), kidney cortex cells, renal cell carcinoma cells, RPTEC (renal proximal tubule epithelial cell line and TTC-1240 (Rhaboid tumour cells of the kidney).

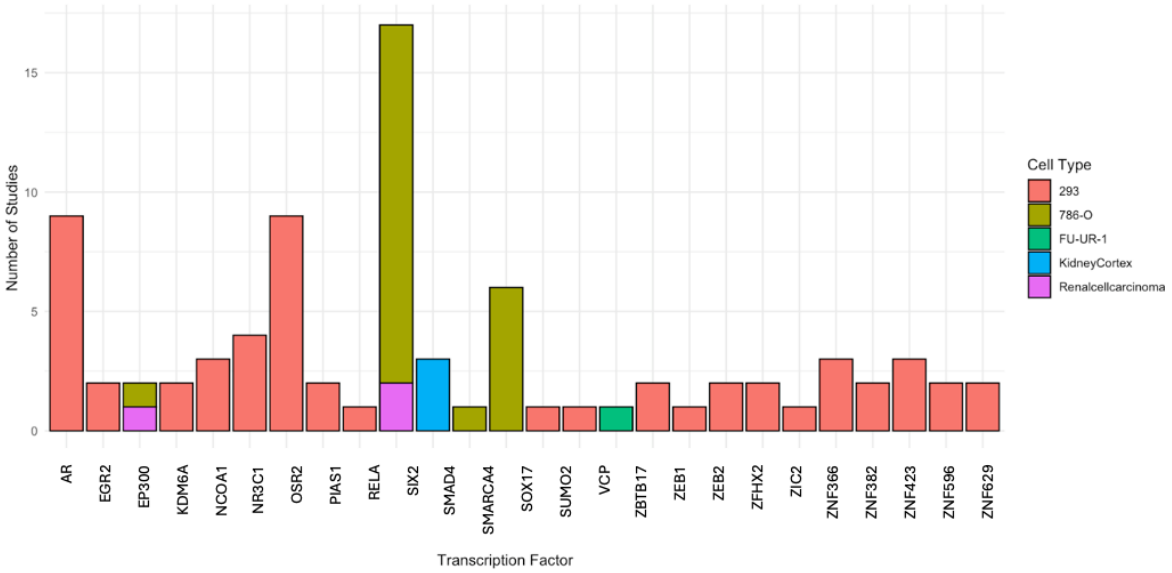


Figure 9. Bar graph showing number of different cell types used in each TF experiment from ChIP-Atlas for the first intronic region. Cell types of each experiment are shown including 293 (human embryonic kidney 293 cells), 786-O (human renal cell carcinoma cell line), FU-UR-1 (retroperitoneal tumour of renal cell carcinoma cell line), kidney cortex cells and renal cell carcinoma cells.

2.3.3 Expression analysis

After identifying regulators that bind to the *PXDN* promoter and intron 1 regions, further analysis was performed to evaluate their expression levels in kidney tissue. mRNA expression levels, based on RNA-seq data generated by the HPA GTEx portal and CAGE data generated by the FANTOM5 consortium, are shown in Figure 10. The FANTOM5 data is presented as scaled tags per million (TPM), a normalised value quantifying gene expression levels across different samples. The normalised transcripts per million (nTPM) values for GTEx and HPA represent the number of transcripts detected for a given TF, adjusting for gene length and total reads in the sequencing experiment, allowing comparison within and between samples.

FANTOM5 expression levels were higher compared to GTEx and HPA expression levels, which are reported in nTPM. TFs with the highest FANTOM5 expression levels included *SUMO2, VCP, RELA, RAD21, SMAD4* and *PIAS1,* all exceeding 100 scaled TPM. For GTEx, the highest expression levels were observed for *SUMO2, VCP, RELA* and *RAD21*, each with nTPM values greater than 25. In HPA data, *SUMO2, RAD21* and *VCP* showed the highest expression levels (nTPM > 25), followed by *NR3C1, NCOA1, RELA, ZEB2, CTCF* and *ARNT*, all with nTPM values greater than 17. Overall, *SUMO2* demonstrated the highest expression levels, with the highest FANTOM5 expression and more than twice the expression level of

other TFs for both GTEx and HPA datasets. Several TFs, including *ZIC2, SIX2, OSR2, ZNF423, ZFHX2* and *TWIST1*, showed minimal expression across FANTOM5, GTEx and HPA data. Specifically, ZIC2 and SIX2 had the lowest expression levels across all datasets, not exceeding 0.5 in FANTOM5, GTEx and HPA.

2.3.4 TF characterisation and mapping

After determining the expression levels of TFs in the kidney, further analysis and characterisation was conducted using JASPAR, HOCOMOCO and Factorbook to obtain PWMs and sequence logos. Since not all TFs were available on JASPAR, either HOCOMOCO or Factorbook were used to supplement missing TF data. Using these logos, the precise TF binding sites in *PXDN* were identified.

Of the 36 regulators shown in Figure 10, 30 are TFs, while six are non-TF regulators. Table 4 and 5 summarise TF binding peak information and sequence logos for promoter and intron 1 TFs, respectively. Logos are provided to visualise the general preferred DNA sequence motifs that TFs tend to bind. Since this analysis is qualitative, frequency scores are not provided. The Database column indicates the source of the binding site data, either JASPAR, HOCOMOCO or Factorbook. The Peak Location column specifies the genomic region TF binding was detected, while the Peak Region column provides the approximate coordinates of where TFs were detected to bind in ChIP-seq experiments on ChIP-Atlas. Logo colours vary depending on the database used. Table 6 lists the six non-TF transcriptional regulators, for which sequence logos are not provided.

A total of 30 TFs identified in the *PXDN* promoter, intron 1 or both regions, were evaluated for subsequent analysis, with eight TFs found in the promoter region and 24 in intron 1. Two TFs appeared in both regions, these being *MEF2B* and *EP300*. The logos illustrate nucleotide conservation at various positions within the binding site. Most logos are long, with only two being less than 9 nucleotides in length (*EP300* and *ZNF596).* The diversity of logos and the range of sequence lengths reflects the wide variety of binding preferences among the *PXDN* TFs.

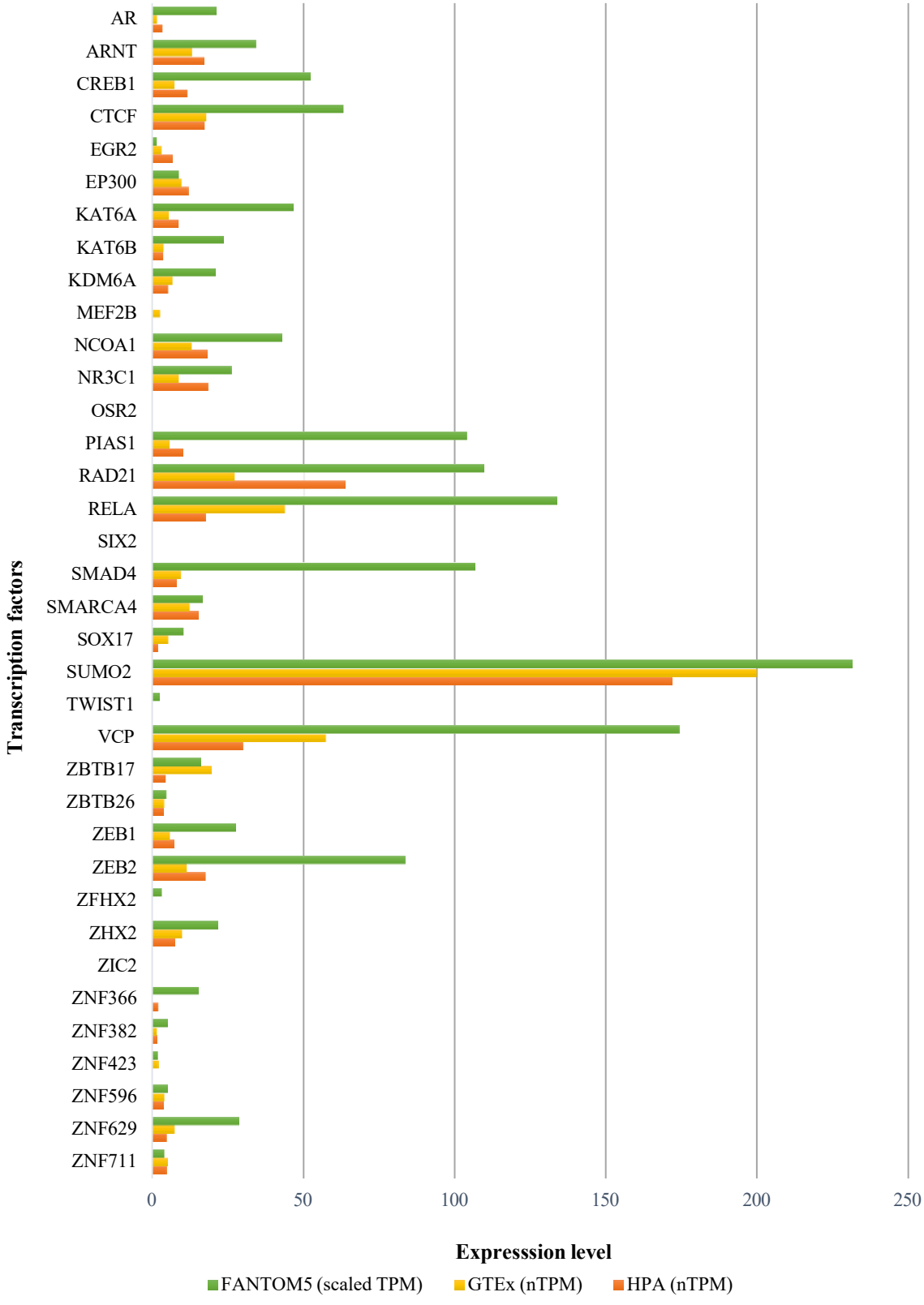


Figure 10. Bar graph showing expression levels of PXDN regulators using RNA-sequencing data generated by the Human Protein Atlas (HPA) Genotype-Tissue Expression (GTEx) portal and CAGE data generated by The Functional Annotation of Mammalian Genomes 5 (FANTOM5) consortium. Expression levels are indicated with green representing scaled tags per million (FANTOM5) as well as yellow and orange representing normalised transcripts per million (nTPM) for GTEx and HPA, respectively. Expression level bars not present indicate either no expression or expression less than 0.5.

Table 4. Binding sites and motifs for TFs identified in the promoter region.

TF Name	Peak Location	Peak Region	Logo	Database
CREB1	P (1-2kb)	02:1592016-1592474		JASPAR
CTCF	P (1-2kb)	02:174301 1-1743313		JASPAR
EP300	P(<=1kb)	02:1743877-1744216		Factorbook
MEF2B	P (1-2kb)	2:1742834-1743240		JASPAR
RAD21	P (1-2kb)	2:1742893-1743388		Factorbook
ZBTB26	P(<=1kb)	2:1743886-1744441		JASPAR
ZHX2	P(<=1kb)	02:1744154-1744684		Factorbook
ZNF711	P(<=1kb)	02:1744082-1744648		HOCOMOCO

Table 5. Binding sites and motifs for TFs identified in the intron 1 region.

TF Name	Peak Region	Logo	Source
AR	02:1705433-1728896		JASPAR
ARNT	2:1736928-1737243		HOCOMOCO

EGR2	2:1710528-1710890		JASPAR
EP300	2:1721769-1722150		Factorbook
MEF2B	2:1707605-1708102		JASPAR
NCOA1	2: 1707910-1728886		Factorbook
NR3C1	2:1708163-1728895		JASPAR
OSR2	2:1708012-1708341		JASPAR
RELA	2:1721816-1722096		JASPAR
SIX2	2:1727076-1727356		JASPAR
SMAD4	2:1710161-1710434		Factorbook
SMARCA 4	02:1721160-1722286		Factorbook
SOX17	2:1707929-1708100		HOCOMOCO
TWIST1	2:1732023-1732369		JASPAR
ZBTB17	2:1707645-1708044		HOCOMOCO

ZEB1	2:1704169-1704500		JASPAR
ZEB2	2:1704169-1704657		Factorbook
ZFHX2	2:1710535- 1710886		Factorbook
ZIC2	02:1708083-1708349		Factorbook
ZNF366	2:1720123-1723736		Factorbook
ZNF382	02:1707668-1708071		JASPAR
ZNF423	2:1710572-1711448		Factorbook
ZNF596	2:1719988-1720418		Factorbook
ZNF629	2:1707779-1708025		Factorbook

Table 6 summarises the binding sites ranges for potential regulators of *PXDN* which are not classified as TFs. *KAT6A* and *KAT6B* bind within the promoter region, while *KDM6A, PIAS1*, *VCP* and *SUMO2* bind within intron 1.

Table 6. Binding sites of non-TF regulators.

Regulator	Type of Regulator	Region	Peak Region
KAT6A	Histone acetyltransferase	P(<=1kb), P(1-2kb)	2:1743302-1744085
KAT6B	Histone acetyltransferase	P(<=1kb), P(1-2kb)	2:1743083-1745058
KDM6A	Histone demethylase	Intron 1	2:1707941-1708333
PIAS1	Transcriptional coregulator	Intron 1	02:1707974-1708144
VCP	Chromatin modelling regulator	Intron 1	2:1732230-1732490
SUMO2	Post-translational modifier	Intron 1	2:1707897-1708264

After obtaining the TF PWMs, TF mapping using putative binding sites was performed. FIMO was used to scan the peak regions identified by ChIP-Atlas (shown in Tables 4 and 5) for putative binding sites, using PWMs from JASPAR, HOCOMOCO or Factorbook. NCBI Genome Data Viewer was employed to locate these sequences within the respective *PXDN* promoter or intron 1 regions. Only TFs with putative binding sites exhibiting FIMO p-values and q-values below 0.05 were included in the final TF mapping diagram (Figure 11).

Table 7 and 8 present the putative *PXDN* binding sites identified for promoter and intron 1 TFs, respectively. For instances where a TF had various binding sequences identified by FIMO, the best-scoring sequence–defined as the one with the lowest p-value – which was identified on NCBI Genome Data Viewer included in the analysis. In the tables, Strand indicates the DNA strand, antisense (-) or sense (+), where the binding site is located. The p-value represents the statistical significance of the binding motif, and the q-value reflects the FDR-adjusted p-value. TFs in bold indicate TF binding sites with a p-value and q-value < 0.05 selected for mapping (Figure 11).

All putative binding sites output by FIMO were found within the respective peak regions. This included binding sites from eight TFs identified in the promoter region and 24 in intron 1. All promoter TFs are located on the negative DNA strand, with the exception of *ZBTB26*. In contrast, the TFs that bind within intron 1 are distributed across both the positive and negative DNA strands, although the negative strand is favoured, with 15 TFs binding to the negative strand, eight binding to the positive strand and one binding to both strands.

In the promoter region, each TF's putative binding site was identified once within the peak range, whilst many intron 1 TFs had multiple binding sites. Notable TFs include *EGR2* and *RELA*, which bind twice; *ZEB2* and *ZNF423*, which binds thrice; *ZEB1*, which binds 15 times; *ZFHX2*, which binds seven times; and *ZNF596*, which binds 16 times. All promoter TF putative binding sites had p-values less than 0.05. Among these, six TFs are highly significant, with p-values less that 0.001: *CTCF*, *MEF2B*, *RAD21*, *ZBTB26*, *ZHX2* and *ZNF711*. Three promoter TF binding sites, *CTCF*, *MEF2B,* and *ZNF711,* met the q-value threshold of 0.05. *CREB1* and *EP300* had q-values of 1, suggesting that these binding sequences may represent false positives.

Table 7. Transcription factor binding sites identified in the *PXDN* promoter region.

TF	Binding Site	Strand	Putative binding sequence	p-value	q-value
CREB1	02:1592335-1592347	-	ATCTGTTGTGAGA	0.003	1
CTCF	02:1743139-1743157	-	ATGCCGCCAGAGTTCGCTG	8.03×10^{-5}	0.041
EP300	02:1744116-1744123	-	GGACGCAG	0.003	1
MEF2B	02:1743016-1743027	-	GCCAAAAATAGT	1.79×10^{-5}	0.014
RAD21	02:1743125-1743139	-	ACACAGGAGGCCTGC	6.37×10^{-4}	0.245
ZBTB26	02:1744289-1744303	+	GCCTCCAGCAGCAGA	3.4×10^{-4}	0.369
ZHX2	02:1744269-1744287	-	GTGCCCGCCGTGGCGCCGC	1.51×10^{-4}	0.147
ZNF711	02:1744284-1744296	-	CTGGAGGCCGTGC	1.14×10^{-5}	0.005

TFs in bold indicate a p-value and q-value < 0.05.

All TF putative binding sequences were found to bind within intron 1 had p-values below than 0.05, with 19 having a p-value less than 0.001, indicating they are highly significant. Ten TFs in intron 1 met the 0.05 q-value threshold.

Table 8. Transcription factor binding sites identified in intron 1 of *PXDN*.

TF	Binding site	Strand	Putative binding sequence	p-value	q-value
AR	02:1722810-1722824	-	AGGCACACCGTGCCC	2.3×10^{-5}	0.315
ARNT	02:1737082-1737090	+	AGACGTGCA	3.86×10^{-4}	0.201
EGR2	02:1710646-1710656	+	ACACCCACTCT	7.2×10^{-5}	0.012
	02:1710693-1710703	+			
EP300	02:1721880-1721887	-	TGATTCAT	2.12×10^{-4}	0.154

MEF2B	02:1707849-1707860	-	CTCTTAAATGGT	0.013	1
NCOA1	02:1708364-1708377	-	GGGGCAAACTCCAG	$4.53x10^{-6}$	0.183
NR3C1	02:1728799-1728815	-	AAGGACACTTTGTTTCT	$3.58x10^{-5}$	0.606
OSR2	02:1708301-1708312	-	ATCCAGGAGCTT	$4.15x10^{-4}$	0.22
RELA	02:1721954-1721963	-	GGGAATTCCC	$9.53x10^{-6}$	0.003
	02:172154-1721963	+			
SIX2	02:1727310-1727325	-	AACTGAAATATGAGGC	$2.04x10^{-5}$	0.012
SMAD4	02:1710289-1710303	-	AGTTCCTGACCTTCA	$5.11x10^{-4}$	0.178
SMARCA4	02:1721723-1721733	-	TGCCTCAGTCT	$6.05x10^{-5}$	0.098
SOX17	02:1707943-1707953	-	CGCTTTGTGTA	0.002	0.192
TWIST1	02:1732168-1732180	+	AAAACAGATGTAG	$7.4x10^{-5}$	0.049
ZBTB17	02:1707674-1707692	+	GAGGACAAAAAAGGGGAAG	$1.73x10^{-4}$	0.094
ZEB1	02:1704185-1704195	-	TTCACCTGGAG	$9.28x10^{-5}$	0.003
	02:1704208-1704218	-			
	02:1704229-1704239	-			
	02:1704249-1704259	-			
	02:1704271-1704281	-			
	02:1704293-1704303	-			
	02:1704314-1704324	-			
	02:1704336-1704346	-			
	02:1704357-1704367	-			
	02:1704378-1704388	-			
	02:1704400-1704410	-			
	02:1704420-1704430	-			
	02:1704443-1704453	-			
	02:1704463-1704473	-			
	02:1704485-1704495	-			
ZEB2	02:1704263-1704270	+	GGGGACAG	0.012	1
	02:1704285-1704292	+			
	02:1704582-1704589	+			
ZFHX2	02:1710598-1710608	-	GAGAGTGGGTG	$1.72x10^{-5}$	0.001
	02:1710615-1710625	-			
	02:1710647-1710657	-			
	02:1710694-1710704	-			
	02:1710771-1710781	-			
	02:1710818-1710828	-			
	02:1710865-1710875	-			
ZIC2	02:1708277-1708287	-	CCCAGCAGGAC	$2.72x10^{-6}$	0.001
ZNF366	02:1723068-1723082	+	ATGGATGGATGGATG	$9.56x10^{-10}$	$1.24x10^{-6}$

ZNF382	02:1707783-1707806	+	AGCGACGATCTGCAGATATCCTTC	0.001	0.625
ZNF423	02:1710636-1710650	-	GGTGTTCATAGAGTG	0.003	0.462
	02:1710683-1710697	-			
	02:1710933-1710947	-			
ZNF596	02:1719990-1719997	+	GAGAGAGA	$1.47x10^{-5}$	$2.74x10^{-4}$
	02:1720015-1720022	+			
	02:1720023-1720030	+			
	02:1720045-1720052	+			
	02:1720073-1720080	+			
	02:1720097-1720104	+			
	02:1720121-1720128	+			
	02:1720145-1720152	+			
	02:1720173-1720180	+			
	02:1720127-1720134	+			
	02:1720255-1720262	+			
	02:1720263-1720270	+			
	02:1720283-1720290	+			
	02:1720311-1720318	+			
	02:1720365-1720372	+			
	02:1720390-1720397	+			
ZNF629	02:1707892-1707900	-	CTGACGTAA	$7.61x10^{-5}$	0.035

TFs in bold indicate a p-value and q-value < 0.05.

Figure 11 illustrates the significant TFs (p-value < 0.05, q-value < 0.05) identified for the *PXDN* gene, as listed in bold in Tables 7 and 8. The figure begins with an overview of chromosome 2, highlighting the 2p25.3 region where *PXDN* is located. It then zooms in on the *PXDN* gene, indicating the promoter, with the TSS, and intron 1 regions. Black rectangles represent the approximate locations of TF binding sites, with their binding sequences shown to the right. The strand direction is marked with (+) for the sense strand and (-) for the antisense strand.

All promoter-associated TFs (marked with asterisks) cluster at the start of intron 1. As ChIP-Atlas defined the promoter region as up to 2kb from the TSS, promoter TFs identified were found to bind at the start of intron 1. Clustering is seen at 1708 kb (*ZNF629* and *ZIC2*), 1 710 kb (*EGR2* and *ZFHX2*), 1743 kb (MEF2B, CTCF and ZNF711). Five TFs were found between 1720 kb and 1732 kb *(ZNF596*, *RELA*, *ZNF366*, *SIX2* and *TWIST1*), although they are not as closely clustered as the previously described groups.

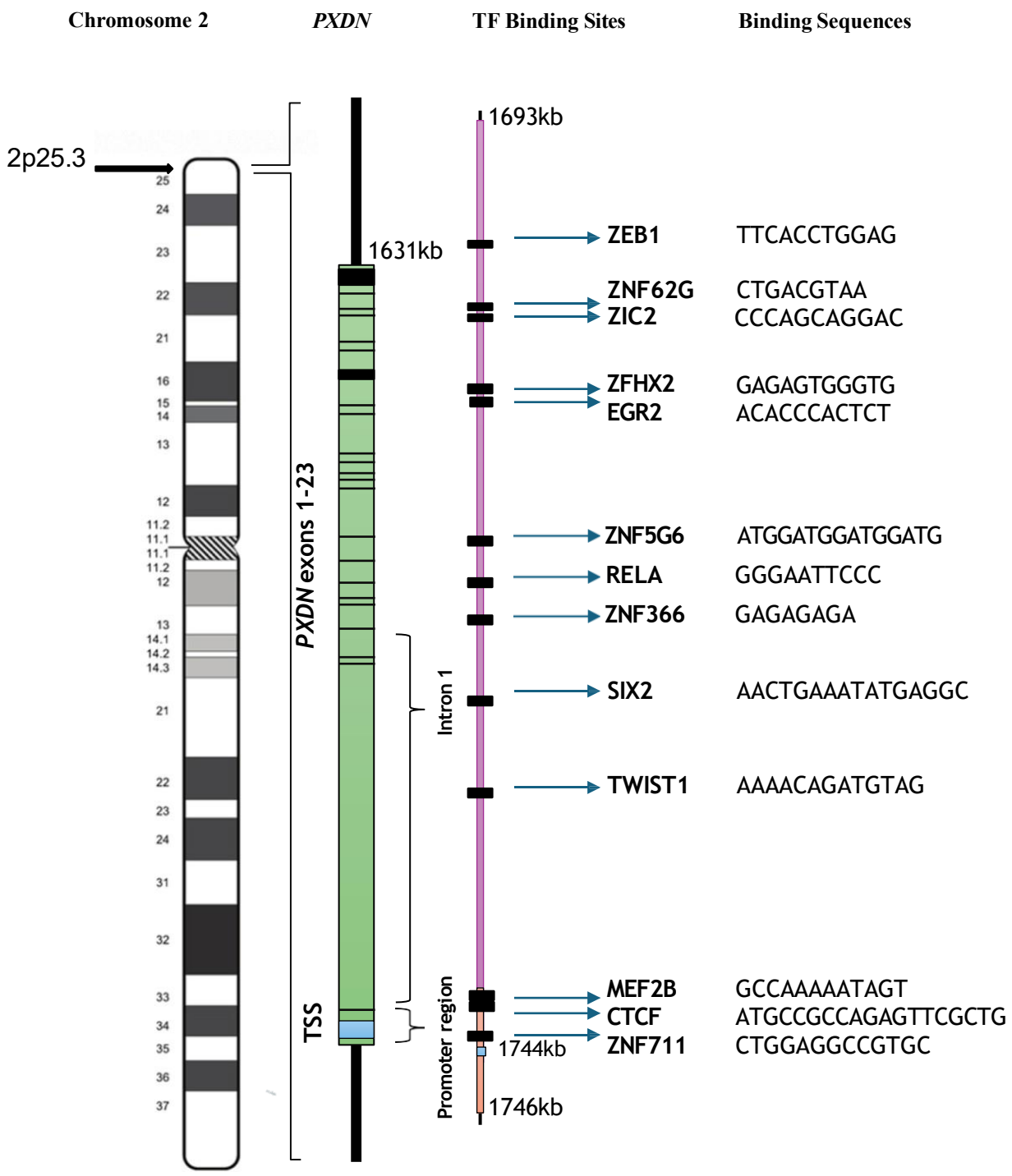


Figure 11. Mapping of TF binding sites for the PXDN gene. This figure illustrates the distribution of TFs binding to the promoter and intron 1 regions of the PXDN gene. The gene is depicted from the 5′ end to the 3′ end at the bottom of the figure. The transcription start site (TSS) is marked in blue, and exons are represented using black lines of varying widths, indicating their different sizes. The promoter region is highlighted in orange, and intron one is indicated in purple. Black rectangles denote the locations of different TFs that meet the chosen significance threshold (p-value < 0.05, q-value < 0.05). The exact binding sequence of each TF is displayed to the right of each TF name. A plus (+) sign indicates that the binding site is on the sense strand while a minus (-) sign indicates site is on the antisense strand.

2.4 DISCUSSION

PXDN is an extracellular peroxidase that facilitates the formation of sulfilimine crosslinks in collagen IV, which are essential for basement membrane integrity, ECM assembly, and tissue development (Bhave et al 2012; Fidler et al., 2014; Lee et al., 2020; McCall et al., 2014). Dysregulated *PXDN* expression has been implicated in various diseases linked to basement membranes defects and collagen IV abnormalities, including developmental disorders, cancer and fibrosis (Cheng & Shi, 2022). In the kidney, PXDN is particularly important because the basement membrane plays a critical role in structural integrity and function of the kidney. Notably, *PXDN* dysregulation has been associated with renal fibrosis, where its overexpression has been observed in response to TGF-β signalling (Péterfi et al., 2009). Given the established roles of PXDN in disease and the predominant influence of transcriptional regulation on expression, understanding the regulatory mechanisms governing *PXDN* is essential. This study has identified several TFs that may regulate *PXDN* expression in the kidney, potentially contributing to kidney function and CKD, as well as identifying potential PXDN binding patterns, previously unknown.

This chapter aimed to investigate the transcriptional regulation of *PXDN* in the kidney. ChIP-Atlas data was used to identify TFs which may regulate expression in kidney cells. *In silico* methods were then implored to identify putative binding sites and map them to *PXDN*. TF binding patterns are explored followed by characterisation of mapped TFs.

2.4.1 Exploration of TF binding patterns

A large proportion of TF binding peaks were enriched in the promoter and intron 1 regions of *PXDN* (Figure 4), suggesting that these regions may have important roles in transcriptional regulation. A total of thirteen TFs had high confidence putative binding sites, having met the FIMO p-value and q-value thresholds (see Methods), and were successfully mapped (Figure 11). These TFs are therefore more likely to bind to *PXDN* and play a role in its regulation, with three binding in the promoter region and ten in intron 1. Transcription machinery, including RNA polymerase II and associated TFs, binds to the core promoter, which includes the TSS, where transcription is initiated (Haberle & Stark, 2018; Hampsey, 1998). ChIP-Atlas defined the promoter as the region spanning 5 kb upstream and downstream of the TSS, facilitating the

identification of TFs binding within this regulatory region. These TFs may regulate *PXDN* expression by recruiting RNA polymerase II and essential cofactors needed for transcription (Zabidi & Stark, 2016). Furthermore, introns play crucial roles in gene regulation through mechanisms such as the binding of transcription regulatory elements, including enhancers, and splicing (Le Hir et al., 2003). Intronic binding is also essential for intron-mediated enhancement (IME), in which introns enhance gene expression independently of promoter activity, with the first intron having the most significant effect (Rose, 2008).

Within identified TFs, the diversity of logos and the varying sequence lengths reflects the broad range of TFs that regulate *PXDN* (Tables 4 and 5), whereas TFs within the same structural family often recognise and bind to similar binding sequences (Luscombe et al., 2000). This lends support to a wide range of roles for PXDN rather than relying on a single, highly conserved mechanism. All but two TF binding sites were longer than 10 nucleotides long, which is often associated with conserved regulation of gene expression. Short binding sites tend to evolve into longer sequences more rapidly over time, while longer binding sites, under stronger selective pressure, typically maintain their stable length, usually ranging from five to 20 nucleotides (Stewart & Plotkin, 2012). These longer binding sites are more likely to be evolutionarily conserved, as they may play a critical role in regulating *PXDN* expression. An example of this is CTCF, which has a binding site of 19 nucleotides (Table 4). CTCF is a highly conserved, multifunctional protein in gene expression and genome regulation. Among its many roles, it is able to act as a repressor or insulator and has the ability to recruit other TFs (Kim et al., 2015; Klenova et al., 1993). In mammalian genomes, CTCF binds up to 65,000 sites indicating its pervasive role in genome organization and the significance of longer binding site sequences (Chen et al., 2012).

Although *PXDN* is located on the negative strand, differences in strand-specific binding were observed for promoter and intron 1 TFs (Tables 7 and 8). Since TFs bind regardless of strand orientation, those binding to the positive strand may still influence *PXDN* transcription. TFs identified as a part of the promoter region all bind on the negative strand, which serves as the template strand for RNA synthesis in *PXDN*. This suggests that the transcription regulation of *PXDN* is strand-specific, with these TFs likely playing a critical role in transcription initiation by interacting with RNA polymerase during gene transcription. Their recognition of promoter elements is essential for transcription initiation, highlighting their importance in regulation of *PXDN* expression. In contrast, TFs that bind to both the negative and positive

strands are observed in intron 1, suggesting broader regulatory mechanisms that may influence various aspects of gene expression, such as alternative splicing or chromatin remodelling, extending beyond transcription initiation (Miyazaki & Miyazaki, 2021). Additionally, similar to promoter region binding, the orientation of TF binding sites within intron 1 may influence enhancer activity, as support by models of cis-regulatory syntax. One model proposes that motif orientation can vary without affecting enhancer function, while another posits that specific orientation is essential (Kulkarni & Arnosti, 2003; Panne, 2008). This large-scale study on TF binding site orientation found that 12 out of 17 non-palindromic sites exhibited differences, with up to 21% higher activity in the template strand for a TF called *PPARA* (Kulkarni & Arnosti, 2003; Panne, 2008). This variability suggests further investigation is needed to confirm whether higher activity occurs on the template strand of *PXDN*. Although *in silico* tools in this study have predicted *PXDN* TF preferences, *in vitro* studies, such as comprehensive ChIP-seq experiments in kidney tissue, are needed to confirm strand-specific binding patterns.

Heterotypic clustering of TF binding sites was observed among *PXDN* TFs (Figure 11). Clustering refers to the presence of more than one TF binding in close proximity within a specific genomic region, as seen with ZNF629, KDM6A and ZIC2 clustering around 1708 kb, EGR2 and ZFHX2 clustering around 1710 kb, and MEF2B, CTCF and ZNF711 clustering around 1743 kb (Figure 11). Additionally, ZNF596, RELA, ZNF366, SIX2 and TWIST were located between 1720 kb and 1732 kb, though they are not as tightly grouped as the aforementioned clusters. Heterotrophic clustering suggests the presence of regulatory regions, such as promoters or enhancers, as these arrangements are unlikely to have occurred by chance alone (Berman et al., 2002; Wagner, 1999). Promoters are typically located near the TSS and regulate transcription initiation, whereas enhancers can act at greater distances to modulate gene expression. Although no TFs fell within the enhancer and promoter regulatory feature regions defined by Ensembl, this clustering pattern suggests the possible existence of uncharacterized tissue-specific promoter or enhancer elements in these regions. Alternatively, the TF cluster in the *PXDN* promoter region–CTCF, MEF2B and ZNF711–falls within an open chromatin region documented on Ensembl. Open chromatin is often associated with regulatory elements that influence gene expression in a copy number-dependent and tissue-specific fashion, for example, enhancers, insulators and silencers (Cockerill, 2011; Gross & Garrard, 1988). This would provide insight into the functional significance of TF clustering in either enhancing or repressing *PXDN* expression, highlighting its potential role in PXDN regulation.

Alongside heterotypic clustering, several intron 1 TFs exhibited homotypic clustering, with putative binding sites being identified multiple times within respective peak regions, including two each for EGR2 and RELA, seven for ZFHX2, 15 for ZEB1, and 16 for ZNF596 (Figure 11). This suggests that, while promoter TFs have more specific interactions, intron 1 TFs may play a more complex role in regulating *PXDN* expression. Gotea *et al.* (2010) highlighted that homotypic clusters in human promoters and enhancers enable TFs to fine-tune transcription in response to various signals. Ezer et al. (2014) described three mechanisms by which homotypic clusters influence gene regulation: (1) Non-interacting TFs, where regulatory effects depend on the number of binding sites occupied; (2) Direct cooperativity, where TF concentration modulates gene expression (Giorgetti et al., 2010); and (3) indirect cooperativity, where TF binding influences one another's binding capacity (Vashee et al., 1998). The differential frequency of binding sites likely reflects varying regulatory influence, with more frequent binders potentially exerting a stronger effect (Cusanovich et al., 2014). Homotypic clustering regulates expression through mechanisms like lateral TF movement, functional redundancy, and cooperative binding (Coleman & Pugh, 1995; Hertel et al., 1997; Papatsenko et al., 2002; Somma et al., 1991; Weingarten-Gabbay & Segal, 2014). Lateral TF allows TFs to interact with multiple sequences, fine-tuning gene expression. Functional redundancy ensures that if one binding site it disrupted, others can compensate. Cooperative binding occurs enable synergistic regulation of gene expression. Homotypic clusters are present in approximately half of known enhancers and promoters (Gotea et al., 2010) and conserved homotypic clusters often regulate TF-encoding genes, suggesting an ancestral mechanism of TF expression (Gotea et al., 2010). These findings support the potential for EGR, RELA, ZFHX2, ZEB1, and ZNF596 clusters to enhance *PXDN* expression. As above, future studies could employ high-resolution ChIP-seq and chromosome conformation capture (3C) to investigate these clusters further.

Several TFs clustering in Figure 11 are known to interact with enhancer regions or modulate enhancer activity, supporting the clustering regions as possible regulatory regions. Enhancers are distal DNA regions containing TF binding sites that upregulate gene expression. TFs regulate enhancer activity by recruiting co-repressors and co-activators, while enhancer-promoter looping facilitates gene activation. Several TFs withing the identified clusters are involved in kidney function and enhancer activity. For example, SIX2 exhibits enhancer activity in kidney development by activating glial cell line-derived neurotrophic factor (GDNF) expression, essential for nephron progenitor cell survival and proliferation (Brodbeck et al.,

2004). SIX2 also interacts with chromatin remodelers, targeting enhancers regulating nephron progenitor cell maintenance (Li et al., 2021). Further studies could use gene knockdown to assess their impact on *PXDN* expression, 3C to explore enhancer-promoter interactions, and histone modification analysis, such as H3K27ac and H3K4me1, to confirm active enhancer regions.

In addition to the identified TFs in the promoter and intron 1 regions of *PXDN*, approximately 80% of binding peaks were found 10-100 kb upstream and downstream of the TSS (Figure 5), suggesting regulation by distal elements, potentially through enhancer-like mechanisms (Gasperini et al., 2020). The large size of intron 1 likely contributes to this distribution; however, the prevalence of distal binding sites suggests the involvement of enhancers, silencers, locus control regions or insulators. Additionally, the substantial enrichment of binding sites in intron 1 compared to the promoter region (Figure 4) further supports the role of long-range chromatin interactions in *PXDN* expression. These regulatory elements can modulate transcription through DNA looping, where the enhancer is brought into close proximity to transcription machinery at the promoter (Maston et al., 2006; Vilar & Saiz, 2005). Given this, further investigation of the long-range interactions in *PXDN* is warranted, as distal enhancers may influence its expression under specific conditions. Future work should aim to experimentally confirm these long-range interactions using 3C or 4C-seq techniques, which enable the study of DNA-DNA interactions and chromatin interactions at specific loci.

Another mechanism of PXDN regulation was identified, with six non-TF regulators identified (Table 6), though their roles in the kidney remain largely unexplored. Among these, *KAT6A* and *KAT6B,* found binding to the *PXDN* promoter in untreated HEK293 cells (Table 2), are histone acetyltransferases that regulate transcription by modifying chromatin structure through acetylation of histones (Wiesel-Motiuk & Assaraf, 2020). All mapped promoter TFs–*CTCF*, *MEF2B* and *ZNF711* (Table 7)–were found within *KAT6A* and *KAT6B* binding regions (Table 6), suggesting these TFs may rely on KAT6A/KAT6B for chromatin remodelling to facilitate binding and regulate *PXDN* expression. However, since these regulators were identified in separate ChIP-seq experiments, further studies are needed to confirm this tentative relationship. If validated, this could suggest a synergistic mechanism where KAT6A/B-mediated histone acetylation, in conjunction with other histone modifiers, enables TF binding to regulate *PXDN* transcription. Similar mechanisms have been observed, such as where KAT6A acetylating H3K23 to recruit TRIM24 and activate PIK3CA transcription (Lv et al., 2017), KAT6A and

KAT6B cooperatively promoting transcription of genes like *ERG* and *PXB1* (Bergamasco et al., 2024).

The remaining non-TF regulators identified bind to intron 1 (Table 6). Of these, protein inhibitor of activated STAT1 (PIAS1) and small ubiquitin-like modifier (SUMO2) are both involved in SUMOylation, a post-translational modification which modifies other proteins by conjugating the SUMO protein to specific lysine residues (Hilgarth et al., 2004; Queiroz et al., 2024), and identified in HEK293 cells. PIAS1 is a coregulator of multiple cellular pathways, including the WNT signalling pathway involved in kidney fibrosis (Schmidt & Müller, 2003). *KDM6A* promotes EMT, and its inhibition in diabetic nephropathy models increases E-cadherin expression, suppressing EMT in high-glucose conditions (Bai et al., 2021). Lastly, VCP, identified binding in untreated kidney cell line retroperitoneal tumour cells (FU-UR-1), is involved in chromatin remodelling (Torrecilla et al., 2017). Following ischemia-reperfusion in rats, VCP has been found to improve kidney structure and function (Ghebremariam et al., 2006). Further studies, such as reporter gene assays to assess the regulatory relationship between the mapped TFs and non-TF regulator activity, are necessary to confirm their role in regulating TFs in the kidney.

2.4.2 Characterisation of mapped TFs

Functional characterisation of mapped TFs based on existing literature suggests that several are implicated in kidney-related process. The 13 mapped TFs discussion in this study are *SIX2*, *CTCF*, *RELA*, *TWIST1*, *ZEB1*, *EGR2*, *ZIC2*, MEF2B, *ZNF366*, *ZNF596*, *ZNF629*, *ZNF711*, and *ZFHX2*. PXDN is involved in several processes, from development to tissue fibrosis (Cheng & Shi, 2022). The TFs identified here may provide insight into the potential functions of PXDN and its transcription regulation in the kidney, though further studies should be used to clarify these roles.

TFs involved in normal kidney function were identified, including *SIX2* and *CTCF*, which are known to play roles in kidney development. SIX2, observed binding to *PXDN* in kidney cortex cells, plays a role in embryonic kidney development, as supported by its low expression in non-diseased adult kidney tissue (Figure 10). It generates nephron progenitor cells during nephrogenesis (Kobayashi et al., 2008), and its haploinsufficiency leads to renal hypoplasia,

reduced nephron number and ECM accumulation (Fogelgren et al., 2009); however, whether similar mechanisms are at play in adult kidney function required further investigation. Secondly, *CTCF*, a master regulator of chromatin organisation and TF activity (Kim et al., 2015), shows high expression in non-disease adult kidney tissue (Figure 10) and bound *PXDN* in several cell types, including renal proximal tubule epithelial cells, renal epithelial cells from glomerular regions, HEK293 cells and cancerous 786-O and TTC-1240 cells. *CTCF* regulates renin-producing cells, essential for blood pressure control, with its deletion leading to hypotension, severe albuminuria, kidney abnormalities, and failure, supporting the crucial role of *CTCF* in normal kidney function (Christov et al., 2018; Martinez et al., 2020; Ohlsson et al., 2001). Although treatments like CpG-free plasmid transfection in HEK293 cells and normoxic conditions in 786-O cells were applied, they likely did not affect *PXDN*-CTCF binding. Unlike SIX2, CTCF may regulate *PXDN* expression more broadly across multiple cell types, including proximal tubule and glomerular epithelial cells, supporting kidney function and homeostasis, however the exact mechanism is unconfirmed. Additionally, CTCF's binding to *PXDN* in renal cancer cells warrants further investigation into CTCF-regulated PXDN within tumorigenic contexts.

Another TF involved in normal kidney function is *RELA*, a subunit of NF-κB. RELA was identified as binding to *PXDN* in renal cell carcinoma and showed high expression in non-disease kidney tissue (Figure 10). NF-κB is essential for normal immune responses to infection (Jayab et al., 2024; Zhang & Sun, 2015). *PXDN* expression is observed in both renal clear cell carcinomas and healthy kidney cells, with higher levels in the former, highlighting its role in both diseased tissue and healthy tissue (Tang et al., 2017; Wyllie et al., 2023). In experiments using treated 786-O cells, bortezomib was likely used to supress the NF-κB pathway and induce oxidative stress, contributing to cancer cell death. Dimethyl sulfoxide (DMSO) may have served as a control, and knockout of the tumour suppressor gene Polybromo1 (*PBRM1*) was, and dysregulated NF-κB signalling has been linked to inflammation in renal disease as well as renal cell carcinoma (Jayab et al., 2024). *PXDN* expression is observed in both renal clear cell carcinomas and healthy kidney cells, with higher levels in the former, highlighting its role in both diseased tissue and healthy tissue (Tang et al., 2017; Wyllie et al., 2023). In experiments using treated 786-O cells, bortezomib was likely used to supress the NF-κB pathway and induce oxidative stress, contributing to cancer cell death. Dimethyl sulfoxide (DMSO) may have served as a control, and knockout of the tumour suppressor gene Polybromo1 (*PBRM1*) was likely used to assess its effect on renal cancer progression. Our data

suggest a potential role for RELA in the regulation of *PXDN* in cancerous kidney tissue, as evidenced by multiple independent experiments demonstrating *PXDN-RELA* interactions. However, the precise mechanism remains to be elucidated.

In addition to TFs involved in normal kidney function, several TFs are known to play roles in processes associated with kidney dysfunction, such as EMT. TWIST1, known for its role in fibrosis, was identified as binding to *PXDN* in HEK293 cells following *TWIST1* and *ALX4* transfection (Table 3). *TWIST1*, a TF that contains a basic helix-loop-helix (bHLH) domain, is crucial for mesoderm specification and differentiation and has been implicated in promoting renal fibrosis through multiple mechanisms. One such mechanism involves *TWIST1* enhancing the binding of *PRRX1*, a gene that activates kidney fibroblasts, contributing to the development of interstitial fibrosis. Due to its role in fibrosis, *TWIST1* has been proposed as a therapeutic target for CKD (Sun et al., 2024). Furthermore, *TWIST1* is implicated in various cancers, where it promotes EMT by repressing of E-cadherin (Qin et al., 2012; Yang et al., 2004). Notably, *TWIST1* shows minimal expression in healthy kidney tissue (Figure 10), highlighting its potential role in disease processes in the kidney, including fibrosis and cancer, as observed in ccRCC (Yang et al., 2023). Similarly, the experiment containing *ALX4*, another TF associated with EMT in cancer cells (Yuan et al., 2015), suggests the possibility of a cooperative regulatory mechanism with TWIST1 in modulating *PXDN* expression, potentially under fibrotic or EMT conditions.

ZEB1, ZIC1 and EGR2 were observed binding to *PXDN* in untreated HEK293 cells and, according to the literature, are also involved in EMT. These TFs also exhibit low expression in adult healthy kidney tissue (Figure 10). *ZEB1*, a zinc finger TF essential for embryonic development (Vandewalle et al., 2009), is a driver of EMT (Zhang et al., 2015). In ccRCC, elevated ZEB1 is linked to E-cadherin downregulation–a marker of EMT (Harb et al., 2018). TFG-β1-induced EMT has been shown to decrease *PXDN* expression (Sitole & Mavri-Damelin, 2018), suggesting that increased ZEB1 during EMT may contribute to reduced PXDN levels, potentially affecting basement membrane stability. However, as our data only demonstrate ZEB1 binding in healthy kidney cells, further investigation is needed to clarify its role in regulating *PXDN*. Similarly, ZIC2, linked to EMT through the TGF-β pathway and associated with ccRCC progression (Huang et al., 2022; Lv et al., 2023), requires further investigation to determine its exact role. Dysregulation of *EGR2*, a member of the immediate

early gene family involved in immune regulation (Taefehshokr et al., 2017), has been implicated in renal fibrosis, with overexpression promoting ECM deposition via partial EMT in both UUO models and CKD patient samples (Song et al., 2024). Inhibition of EGR2 reduces EMT and ECM accumulation in TGF-β-treated kidney cells (Song et al., 2024). *PXDN* could be regulated by EGR2, with inhibition potentially downregulating ECM-modifying proteins like PXDN, while overexpression may increase ECM deposition and fibrosis, processes associated with elevated PXDN levels in the kidney (Péterfi et al., 2009). Thus, further investigation is required to determine the exact role of EGR2 in *PXDN* regulation.

For the remaining TFs, *MEF2B*, *ZNF366*, *ZNF596*, *ZNF629*, *ZNF711*, and *ZFHX2*, many do not have strong documented roles in the kidney. These TFs were all identified as binding to PXDN in untreated HEK293 cells and show low expression levels in normal kidney tissue (Figure 10), although this does not rule out their potential importance in kidney function. MEF2B, a TF that regulates the de-differentiation of vascular smooth muscle cells (Rodríguez et al., 2015) and is suggested as an oncogene (Pon & Marra, 2016), does not appear to have a role in kidney function. *ZNF366*, *ZNF596*, *ZNF629*, *ZNF711*, and *ZFHX2* are all zinc fingers, which interact with RNA, DNA and proteins, regulating various cellular processes such as transcriptional regulation, DNA repair, and signal transduction (Cassandri et al., 2017). While some zinc finger proteins are well-characterised, many, including the ones listed, remain less explored. *ZNF711* is a transcriptional regulator in breast cancer and neuroblastoma cell lines (Rhie et al., 2018), but its activity in the kidney has not been studied. Similarly, *ZNF596*, part of the ZNF/EZH2/STAT3 signalling pathway, is upregulated in glioblastoma, correlating with poor survival outcomes (Tang et al., 2019), and *ZNF366* is involved in repressing oestrogen-responsive genes in breast cancer cells, though no links in kidney cancer have been reported (Lopez-Garcia et al., 2006). *ZNF629* and *ZFHX2* are involved in kidney function, but their roles are not strongly established. ZFHX2 promotes the upregulation of somatostatin, a hormone that affects several cellular processes, leading to a large reduction in glomerular filtration rate (Ginès et al., 1992; Habib et al., 2018). *ZNF629*, though not widely studied, has been identified as a binding site for *VWF* in a recent study on sickle cell disease nephropathy, which contains a proteinuria-associated SNP (Garrett et al., 2023). The limited literature on these TFs leaves ample opportunity to explore their roles in the kidney and their potential impact on regulating *PXDN*.

The characterisation of mapped *PXDN* TFs suggests a complex regulatory system, with multiple TFs being identified and several pathways involving EMT, potentially indicating a link between EMT and PXDN in the kidney, although this remains unconfirmed. While these TFs were characterised in the context of kidney function based on existing literature, future studies–such as cellular assays–should aim to elucidate the functional roles of these TFs in *PXDN* regulation, specifically within the kidney. Understanding their involvement in pathways like EMT, which play a critical role in disease progression, may provide insights into kidney biology and gene regulation.

2.4.3 Limitations

Although this study provides valuable insights into regulation of PXDN in the kidney, several limitations must be acknowledged. This research relied on publicly available ChIP-seq data from ChIP-Atlas, which aggregates datasets submitted by various collaborators. As such, our analysis was constrained by the data available within ChIP-Atlas at the time of the study. Further, the inclusion of HEK293 cells as a proxy for kidney transcription is debatable. Although the HEK293 human cell line is extensively utilised across various areas of cell biology, genomic and transcriptomic analyses have shown that these cells do not represent typical kidney cells. Despite this limitation, HEK293 cells were included for exploratory purposes to provide preliminary insight into potential transcription mechanisms. Further validation using more kidney-representative cell lines-–such as those derived from renal cortex and renal epithelial cells– is recommended to confirm the relevance of these findings in kidney tissue. There is a potential challenge in replicating this study, as the ongoing influx of new data submissions to ChIP-Atlas could lead to variations in the datasets available at different times. Another limitation is the lack of standardisation in the data used for the expression analysis. For instance, GTEx expression data originates from individuals with varying health statuses, and while inclusion and exclusion criteria were applied, the presence of underlying conditions may introduce confounding variables. This could impact gene expression levels and introduce difficulty with generalising findings to healthy populations. Additionally, variation in sampling, processing, and sequencing across datasets limits the comparability of the results. The use of multiple tools to obtain TF information, such as JASPAR, HOCOMOCO and Factorbook, introduces discrepancies in terms of motif prediction algorithms and annotation standards, which can influence both TF predictions and scoring, impacting the consistency of results. Furthermore, despite thorough quality control

practices, the reliance on third-party data introduces potential issues, such as errors in the datasets, which could impact the reliability of the findings. Lastly, ChIP-seq experiments have inherent limitations, particularly with regard to the potential for false-positive or false-negative results, which can arise due to factors such as chromatin accessibility and antibody specificity.

2.4.4. Conclusion

To our knowledge, this is the first study to investigate transcriptional regulation of *PXDN* in the kidney. Our analysis identified three significant TFs in the promoter region and ten in intron 1, highlighting an intricate regulatory framework for *PXDN* in kidney tissue. Among these TFs, several have documented roles in EMT, an important process contributing to disease in the kidney, and their binding to PXDN suggests a potential regulatory link that warrants further investigation. Notably, mechanisms potentially governing *PXDN* regulation, including the involvement of non-TF regulators and distinct TFs binding preferences, were described. Additionally, we observed clustering of TFs–both homotypic and heterotypic–at specific intervals along intron 1.

The findings of this study hold significant theoretical implications in our understanding of *PXDN* regulation in the kidney. For the first time, potential kidney-specific TFs regulating *PXDN* have been identified, offering a foundation for exploring tissue-specific gene expression mechanisms. This research provides valuable insights into transcriptional regulation, that expand the current understanding of how PXDN is regulated in both healthy and pathological kidney conditions. Furthermore, the study highlights practical implication including the utility of integrating publicly available data, such as ChIP-seq datasets and TF binding predications, to uncover novel regulatory interactions. This approach supports the potential of bioinformatic-driven methodologies in advancing gene regulation research, particularly less-studies gene like PXDN.

Our findings provide a foundation for the regulatory landscape of *PXDN* in the kidney, highlighting TFs potentially implicated in mechanisms that may contribute to dysfunction in the kidney. Overall, this work provides a deeper understanding of *PXDN* regulation and its potential involvement in disease mechanisms in the kidney, offering potential pathways for

improving diagnosis and treatment of kidney disease.

2.4.5 Future Research

As a foundational study, the research provides a basis for a wide range of future investigations. In terms of novel TFs identified, further validation is required to decipher the functional implication of these TFs. Thorough future studies, including both *in vitro* and *in vivo* studies, are advised. This involves experiments to validate the physical interactions between the identified TFs and *PXDN*, as well as to clarify how these interactions influence biological function in both healthy and diseased contexts. Further, given the exploratory nature of this study and the use of publicly available data, there is substantial opportunity for more in-depth investigations. This includes making used of high quality, newly acquired data and more stringent research methodologies. Addressing the limitations outlined above will allow for future research to enhance our understanding of kidney function, aid in the development of more targeted therapies for kidney disease and provide deeper insights into the regulatory mechanisms of *PXDN*. This study provides new avenues for further investigations into the role of PXDN in kidney function and disease.

CHAPTER 3. Investigating genetic variation in *PXDN*, *PXDN* regulators and CKD-associated genes linked to eGFR levels in a SA cohort

3.1 INTRODUCTION

The Kidney Disease Improving Global Outcomes (KDIGO) organisation has emphasised the significant role of genetics in the management and classification of CKD, advocating for genetic testing in attempt to better diagnostic accuracy (Köttgen et al., 2022). While precision medicine has advanced globally in kidney disease research, African populations remain underrepresented in large-scale studies aimed at identifying susceptibility variants. Given the disproportionately high burden of CKD in individuals of African ancestry, further investigation into genetic risk factors is essential (Hariparshad et al., 2023).

However, the most recent review of CKD genetic risk factors in African populations found insufficient evidence to suggest that specific variants posed a greater risk in these populations (George et al., 2018). Despite these challenges, ensuring adequate representation of African populations in genetic research remains crucial, as hundreds of risk-associated genes have been identified in other populations.

Association studies test genetic variants either within specific genes, as in gene association studies, or across the genome, as in GWAS, to note genotype-phenotype associations. The aim of these studies is to identify novel disease susceptibility variants and biological pathways that may eventually inform clinical care (Tam et al., 2019). As discussed in Chapter 1, several key genes have been linked to CKD susceptibility, many of which influence critical pathways such as kidney development, fibrosis and EMT. While genetic risk factors for kidney disease are increasingly studied in African populations (Fatumo et al., 2021; Kintu et al., 2023), research in SA remains limited (Hariparshad et al., 2023). Furthermore, existing studies often yield conflicting results. For instance, while one study found no association between *APOL1* kidney risk variants and hypertension-attributed CKD (Nqebelele et al., 2019), another reported an

association between *APOL1* variants and albuminuria, a key characteristic of CKD (Brandenburg et al., 2022).

Additionally, studies in SA indicate a significant proportion of CKD patients do not present with traditional risk factors. In a cohort of black South Africans, one third of participants lacked diabetes, hypertension, or HIV infection, conditions commonly associated with CKD (Fabian et al., 2022). Given the high prevalence in SA (8.7%) (Hariparshad et al., 2023; Matsha et al., 2013), further research is needed to investigate alternative contributors, particularly genetic factors, that may underlie in CKD individuals without known risk factors.

Given the involvement of multiple TFs and signalling pathways in kidney dysfunction, is it essential to consider genes like *PXDN*, which contribute to both structural maintenance and fibrotic processes in the kidney (Cheng & Shi, 2022). PXDN has been implicated in various disorders of the kidney as well as fibrotic disease, for example, CKD (Péterfi et al., 2009; Zhong et al., 2023). Transcriptional regulators play a major role in gene expression, therefore investigating genetic variation in the transcriptional regulators of *PXDN* in the kidney may provide insight to potential PXDN pathways involved in kidney dysfunction.

This chapter aimed to investigate variants associated with eGFR levels in a SA cohort, with the goal of expanding SA representation in CKD research. The objectives included performing quality control on genes extracted from whole genome sequence (WGS) data, carrying out an association analysis to determine whether *PXDN* transcriptional regulators, identified in Aim 1, and genes previously implicated in kidney dysfunction in non-SA populations were associated with kidney dysfunction in a SA population, and to determine whether identified variants may play a role in impaired kidney function.

3.2 METHODOLOGY

3.2.1 Dataset

We used data from the Wits Rural Public Health and Health Transitions Research Unit involving a rural South African cohort in Bushbuckridge, Mpumalanga, South Africa, as described by Fabian *et al.* (2022). This cohort was recruited between 2017 and 2018 and includes 2,021 adults ranging from 20-79 years old. The data consisted of WGS data of 100 samples from black South Africans accompanied by patient data including sex, age at collection, and kidney function data, including UACR and eGFR (CKD EPI without adjustment for race). This dataset was sequenced with a coverage of 30x and referenced to GRCh38. The TruSeq Nano DNA Library Prep Kits were used to prepare samples before they underwent sequencing using the Illumina TenX (150 bp) as described by da Rocha *et al.* (2021). This forms a part of the publicly available AWI-Gen data set containing approximately 8 200 individuals (Ramsay et al., 2016).

For the study, *PXDN* and all identified *PXDN* TFs identified in Chapter 2 were included to assess whether *PXDN* or its regulators affect kidney function in the study sample. Additionally, 20 genes previously implicated in kidney dysfunction were analysed to determine whether findings from other populations are applicable to the black South African cohort. Some of these genes have been linked to CKD, such as *JUN*, *PAX2* and *PAX8* (Pillebout et al., 2003; Zhang et al., 2018; Zivotic et al., 2022); others are associated with disorders affecting the kidney, including *EYA1* and *HNF4A* (Marable et al., 2018; Wang et al., 2024); while others, like , *EGR1*, *KLF4* and *RELB*, have been implicated in renal fibrosis (Ai et al., 2022; Sun et al., 2021; Wen et al., 2019). Refer to Appendix D for the list of included genes alongside relevant literature supporting their involvement in kidney dysfunction. A list of extracted regions for these genes, including 4 000bp flanking regions, is available in Appendix F.

3.2.2 Phenotype

As low eGFR is a key indicator of impaired kidney function in kidney disease, eGFR levels were used as the primary phenotype to identify associated genetic variants. As previously discussed, eGFR represents the total amount of fluid that is filtered through the functioning nephrons per unit of time, meaning that eGFR decreases as kidney function declines (Levey et al., 2015). According to the Kidney Disease Global Outcomes criteria, low eGFR is commonly defined as less than 60 mL/min/1.73m^2 (Stevens & Levin, 2013) or 90 mL/min/1.73m^2 (Inker et al., 2014). However, the CKD-EPI equation has been found to overestimate GFR in black South African and other African populations (Niang & Luyckx, 2022). Studies suggest that a threshold of 90 mL/min/1.73m^2 is a more appropriate predictor kidney function for black South Africans, particularly females (Moodley et al., 2018). This adjusted threshold has been applied in recent studies of kidney function in South African populations (Navise et al., 2023).

The second phenotype chosen was UACR as it has been considered a strong marker for kidney disease in African populations (Brandenburg et al., 2024). It is a measure of the amount of albumin and creatine in one's urine, with measurements exceeding 30 mg/g indicating that the kidneys may not be filtering blood effectively (Inker et al., 2014). Due to the lack of recorded UACR values for most individuals in our sample, we were unable to include this parameter in our study, limiting the association analysis to eGFR. The phenotype data for AWI-Gen samples is shown in Appendix G.

3.2.3 Quality control

The AWI-Gen data had already undergone partial quality control where variants were filtered based on quality score (QUAL) and reads were filtered based on read depth. A quality score threshold of 20 was applied to retain variants with a 99% confidence of being true variants. A read depth threshold of 25 was implemented to ensure that only variants with a minimum read depth of 25 were included.

A PCA was done to assess the population substructure of the dataset. Principal components (PCs) were generated using PLINK (version 1.9) with reference populations from the KGP,

including Northern and Western European (CEPH), Han Chinese (CHB) classified as East Asian, and Yoruba Nigerian (YRI) populations (Purcell et al., 2007; The 1000 Genomes Project Consortium, 2015). No additional quality control was applied to populations used in the PCA in order to maximise genetic information, given that AWI-Gen data only covered selected regions. The PCA was visualised in R using the Foreign, Ggplot2, Ggfortify and Dplyr packages.

Following the PCA, PLINK (version 1.9) was used to perform further quality control for the association analysis. Given our relatively small sample size, we opted for less stringent thresholds while still maintaining their utility. This included filtering out variants with a MAF less than 0.01, filtering out variants with missing genotype call rates per marker exceeding 10% and filtering out variants which have Hardy-Weinberg Equilibrium (HWE) exact test p-values below 0.0005. Linkage disequilibrium (LD) pruning was then done using the indep-pairwise command with a window size of 100 variants, a step size of 20 and an R^2 threshold of 0.2.

3.2.4 Association analysis

PLINK2 was used on the command line to identify genetic variants associated with eGFR levels, as it offers improved performance and additional features compared to PLINK1.9, including advanced statistical analysis (Purcell et al., 2007; Steimle et al., 2011). Linear regression was used due to the continuous nature of eGFR values, consistent with other association studies that have used eGFR as a phenotype (Gorski et al., 2022). This approach assumes that the data follows a normal distribution to ensure valid statistical inferences (Casson & Farmer, 2014). Therefore, the distribution of raw eGFR values was visualised using a histogram in R, which indicated that normalisation of the raw values via log-transformation was necessary. A one sample t-test for age and a Student's t-test (equal variances) for sex were done in R to investigate whether age and/or sex should be included as covariates. Significant variable(s) were incorporated as covariates in the linear regression analysis. To account for multiple testing, a Bonferroni correction was performed by dividing 0.05 by the total number of SNPs tested. As no SNPs met the Bonferroni-corrected significance threshold, an arbitrary, less stringent p-value threshold of less than $5x10^{-4}$ was used to select top variants to test further.

Quanto (v1.2.4) was used to perform a power calculation for the 97 individuals included in the study (Gauderman, 2002). A continuous trait, additive genetic model of independent individuals was done. The significance level (α) applied was 0.05 (two-sided), the effect size (βG) ranged from 0.1 to 0.5, and the allele frequencies tested were 0.01 to 0.1, in increments of 0.2.

3.2.5 Variant annotation and prioritisation

Making use of a filtered .vcf file containing these variants, Combined Annotation Dependent Depletion (CADD) Annotation (GRCh38 version 1.7) was used to gain insight into the top variants, including potential deleteriousness and pathogenic variants. The output included several scores including GERP N scores, b statistics, raw CADD scores and primate phastCons scores. Additionally, MAFs were obtained from the database of Genotypes and Phenotypes (dbGaP) using ALFA to compare allele frequencies between European and African populations and the study population, providing insight into the variant distribution across populations (Tryka et al., 2014). Primate phastCons, GERP N, GERP S and CADD scores were selected to be used to rank potentially important indels (Table 9).

Primate phastCons generates conservation scores through the use of a hidden Markov model-based method (Siepel et al., 2005). It measures evolutionary conservation at specific genomic positions in primates. Scores range from zero to one, with higher scores representing higher evolutionary constraint. GERP++ makes use of rigorous algorithms and genomes from multiple specifies to estimate neutral evolution (GERP N or GERP ++), which help identify evolutionarily constrained sites (Cooper et al., 2005). Scores range from -12 to 7 approximately (Sakthikumar et al., 2020), with scores over 2 representing functionally constrained sites (Davydov et al., 2010). Lastly, CADD, a comprehensive scoring system designed to evaluate the deleteriousness of SNVs, indels and substitutions (Rentzsch et al., 2019), was used. CADD uses a machine learning model and data containing mutation and functional annotation to create scores that depict the predicted deleteriousness of variants (Kircher et al., 2014). Raw (unscaled) CADD scores were used. Higher scores present a higher likelihood that the variant is deleterious or damaging to gene function.

Each tool score was normalised to a range of 0 to 1, with higher scores indicating greater predicted functional importance. The weighted scores from each tool were then summed to create a cumulative score for each variant, ranging from 0 to 3. Variants were ranked according to their cumulative scores and scores which met the threshold of 1.2 were prioritised.

Table 9. Description of variant prediction tools and the scoring system applied.

Tool	Function	Score details	Scoring system
	Neutral evolution score: Negative scores indicate that a site is evolving at a rate faster than expected, indicating less functional constraint. The opposite is true for positive scores.	Range from -12.3 to 7	(GERP N score – minimum score) / (maximum score – minimum score)
CADD	Higher Combined Annotation Dependent Depletion scores represent variants more likely to have deleterious effects.	No range specified	(CADD score of variant – minimum score) / (maximum score – minimum score)
Primate phastCons	Primate phastCons conservation score (excl. human) provides the probability of negative selection indicating functionally important regions.	Range of 0 to 1	PhastCons score

Additionally, variant characterisation was performed using Ensembl (https://www.ensembl.org/index.html), to assess potential variant consequences, SpliceAI, to predict the impact of variants on RNA splicing (https://spliceailookup.broadinstitute.org) (de Sainte Agathe et al., 2023), and the LDpair tool (https://ldlink.nih.gov/?tab=home), to determine whether variants in the same region, for example variants in the same intron, were in LD using data from all populations (Machiela & Chanock, 2015). LDpair uses the KGP and outputs four key measures to evaluate LD. D′ indicates the degree of allelic segregation between two variants, with values ranging from 0 (no linkage) to 1 (complete linkage). For two variants, R^2 measures correlation between two alleles, where a value of 0 suggests independence and a value of 1 means that one variant allele perfectly predicts the other allele. Lastly, chi-square statistics and p-values are used to test whether the observed haplotype frequencies deviate significantly from expected values, suggesting the presence of LD.

3.2.6 Functional Annotation and gene-set analysis

Gene-set analysis was performed using GENE2FUNC on FUMA to identify overrepresented gene sets among the top variants in the analysis using Gene Set Enrichment Analysis (GSEA) (https://www.gsea-msigdb.org/gsea/index.jsp). Hypergeometric tests were performed to determine if genes that top variants were identified in, genes of interest, were overrepresented in pre-defined gene sets, including Molecular Signatures Database (MSigDB), which contains curated gene sets from published research, WikiPathways and genes present in the GWAS catalogue. The gene set was described along with the pre-defined gene set category of GSEA (details available on https://www.gsea-msigdb.org/gsea/index.jsp). This analysis aimed to explore the biological significance of the identified genes, investigating whether they were enriched in certain pathways, molecular functions or biological processes which may influence kidney function. For instance, common pathways identified provided insight into potential mechanisms contributing to kidney dysfunction, potentially affecting eGFR levels. MAGMA was used to perform the gene-set analysis and uses regression to analyse continuous gene properties, for example p-values, and study multiple gene sets and their characteristics at the same time (de Leeuw et al., 2015). MAGMA was selected as it has been shown to outperform other gene-set analysis tools with regards to both power and speed. Enrichment of input genes in gene sets is provided according to various grouping systems, for example, GO biological processes. Benjamini-Hochberg FDR was used for multiple testing correction was applied to each category to account for increased risk of Type 1 errors for the gene-set enrichment test. The minimum number of overlapping genes was set to two and maximum p-value of 0.05 was used to exclude gene sets based on statistical significance.

3.3 RESULTS

3.3.1 Dataset

AWI-Gen GRCh38 WGS were obtained in .bcf format after being filtered for our genes of interest, including 4 000bp flanking regions, as described in Appendix F. Of the 100 WGS samples, 51 were females and 49 were males.

3.3.2 Phenotype

Three samples lacked eGFR values, resulting in a final sample size of 97 for the analysis. Refer to phenotype data in Appendix G which also includes sex and age of collection of all samples in the cohort.

3.3.3 Quality control

PCA was performed using CEPH, CHB and YRI populations from the KGP as reference populations. While PC1 primarily separated African (YRI) and non-African (CEPH and CHB) populations, with admixed SA clustering between the two (Appendix H), PC2 vs PC3 better represented the population differences (Figure 12). Importantly, no population stratification is observed within the AWI- Gen cluster, demonstrating that PCs did not need to be included as covariates in the association analysis.



Figure 12. Principal component analysis (PCA) of AWI-Gen samples with 1000 Genomes reference populations. The included populations are South African AWI-Gen samples (orange), Northern and Western Europeans (CEPH) (green), Han Chinese (CHB) (blue) and Yoruba Nigerians (YRI) (purple).

Quality control resulted in 98 262 variants removed from an initial 124 268 variants (Figure 13), leaving a total of 26 006 variants to be included in the following analysis. The LD pruning step was excluded due to the significant proportion of variants that would have been removed (> 21 000). Including this step would have left only 5 000 variants in the analysis, thereby impacting the robustness and power of the study. While LD pruning allows for the inclusion of independent variants and reduced redundancy, it may also exclude variants with weaker LD, potentially limiting the discovery of certain genetic associations. Given the large number of variants that would be excluded, this step was excluded from the analysis.

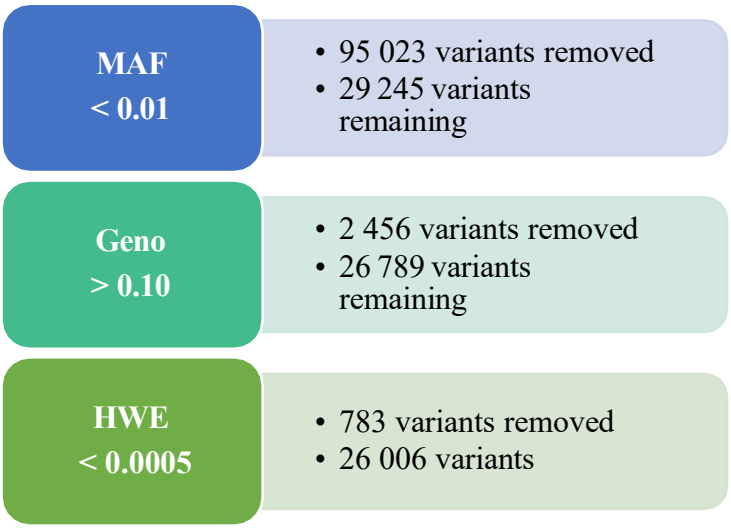


Figure 13. Number of variants removed during quality control. Minor allele frequency (MAF), missing genotype data per SNP (Geno), and Hardy-Weinberg Equilibrium (HWE) thresholds were applied.

3.3.4 Association analysis

To assess whether the eGFR data followed a normal distribution required for linear regression analysis, the distribution of eGFR values was visualised. The histogram revealed that the data was left-skewed which shows that more eGFR values clustered towards the higher end with fewer low eGFR values, thereby indicating the need for normalisation through log-transformation, which was performed in R and used for subsequent analysis.

Making use of the normalised eGFR values, *t*-Tests were carried out. The p-value indicates the probability that the observed difference is due to chance, with p-values less than 0.05 being considered statistically significant. A one sample *t*-test was conducted to assess whether the mean age in the sample differed significantly from zero. This type of *t*-test was chosen as age is a continuous variable. The results indicated that there was a significant difference, with a p-value of less than 2.2×10^{-16}. Therefore, age was included as a covariate in the association analysis. For sex, a Student's *t*-test was done, as sex is a categorical variable, to determine if there was a significant difference in sex distribution for eGFR levels. The p- value of 0.307 suggested that the difference was not statistically significant and therefore sex was not included as a covariate.

Following this, a linear regression analysis, with age included as a covariate, was conducted. To correct for multiple testing, the Bonferroni correction method was used by dividing 0.05 by the total number of variants included in the study. This resulted in a threshold $p = 1{,}9x10^{-6}$. No variants met this threshold, indicating that none were statistically significant. As this threshold was calculated using the total number of variants and not the number of unlinked variants, it is likely to be overly conservative. As no variants reached this Bonferroni threshold, the top variants and the genes they reside as associated with will be discussed. An arbitrary p-value of $5x10^{-4}$ was used as the threshold to select top variants for variant annotation, variant prediction and further analyses. Table 10 shows the top 21 variants that met this threshold, hereafter referred to as "top variants". Of the top variants, two were indels while the remaining 19 were SNPs. Beta values, shown in column 7, represent effect sizes describing the relationship between each identified variant and the investigated trait–in this study, eGFR (Marees et al., 2018). As eGFR is a quantitative trait, the beta value indicates the change in mean eGFR per allele, measured in units of standard deviation. The direction of association is shown by the sign of the beta value: positive values indicate an increase in eGFR, while negative values indicate a decrease. Interestingly, all of the top variants were associated with lower eGFR values, as indicated by negative beta values. The power calculation revealed that the power was approximately 5% across all parameters selected.

3.3.5 Variant annotation and prioritisation

Table 10 presents the functional annotation results for the top 21 variants (p-value < $5x10^{-4}$) from the eGFR association analysis. Consequence describes the nature of the variant, with variants classified as occurring in intronic regions (Intronic, intron number in brackets), affecting regulatory regions (Regulatory), and located upstream of the gene in an intergenic region (Upstream) while Gene indicates the specific gene the variant lies within. EUR and AFR MAF refer to European and African aggregate allele frequencies from dbGaP. The MAF reported here refers to the alternate allele frequency. The AFR sample includes Ghana and Yoruba populations, while the EUR sample includes unspecified European population(s). The 21 top variants are distributed across several genes: *ZNF423*, *MRPS21P7*, *MRPS21P8*, *KLF4*, *PXDN*, *EYA1* and *ZEB1*. Sixteen of the 21 variants were identified in *ZNF423*, with two overlapping in *MRPS21P7* and *MRPS21P8*, two variants were found in *EYA1*, and one variant each was identified in *KLF4*, *PXDN* and *ZEB1*. The lowest MAF in the study

population was observed in *ZNF423* variants, each with a MAF of 0.015. This was followed by *ZEB1* (MAF = 0.021), *PXDN* (MAF = 0.026) and *KLF4* (MAF = 0.036) variants. The highest MAF in the study populations was seen in *EYA1* variants (MAF = 0.170).

In addition to MAFs in the study population, MAFs were obtained from dbGaP using ALFA to compare allele frequencies between European and African populations and the study population (Table 10). N/A denotes MAFs that were unavailable for the effect allele in the study. Most variant MAFs were 0 for both European and African populations, however, a higher number of MAFs greater than 0 were observed in African populations.

Each of these top variants was then scored using CADD, PhastCons and GERP N scores to assess their predicted functional significance. The majority of *ZNF423* variants are intronic, with three in regulatory regions. Variants in the third intron of *ZNF423* showed highest CADD scores, especially in comparison to *ZNF423* variants in the second intron. The highest primate PhastCons scores are also seen in *ZNF423* intron 3 variants. Two *ZNF423* variants were also identified as upstream for *MRPS21P7* (rs1596888340) and *MRPS21P8* (rs2033816838), which are mitochondrial ribosomal pseudogenes, likely identified because of the 4 000 bp flanking regions included in the analysis.

The highest GERP N score was seen in a *KLF4* variant (rs75509972). While one *PXDN* variant (rs1375881180) was identified, it had the second highest observed GERP N score, followed by two *EYA1* variants (rs80082766 and rs79409161). The *ZEB1* variant (rs183670771) was identified in intron 1, although its annotation scores are low.

Following variant annotation, CADD, PhastCons and GERP N scores were normalised to allow for fair comparison across all tools used (refer to Table 9 in Methods). Normalised scores for each tool are also shown in Table 11 in the last few columns, with the summed normalised scores for each variant indicated in bold. Total scores, with a maximum possible score of 3, are ranked in descending order, with higher scores predicted to have a greater functional effect. While Primate PhastCons scores already ranged from 0 to 1, GERP N and CADD scores were transformed into normalised scores that scaled between 0 and 1 to enable consistent comparison across variants.

As one score was unavailable for the *ZEB1* variant due to PhastCons not generating an output, the total is based on two scores instead of three, as in the other variants. An arbitrary threshold of 1.2 was used to select variants for further characterisation (shown in bold in Table 11), as they are predicted to have a greater functional effect. Twelve variants met this threshold, located in *ZNF423*, *MRPS21P7, KLF4* and *PXDN*, while the other variants in *ZNF423*, *EYA1*, *MRPS21P8* and *ZEB1* fell below the threshold.

ZNF423 had the highest number of variants that exceeded the 1.2 threshold for functional significance, with a total of nine variants (including one also associated with *MRPS21P7*) showing a range of scores from 1.07 to 2.69. The top variant (rs1596888340) had the highest score of 2.69, indicating a greater predicted functional impact. *KLF4* had one variant (rs75509972) with a score of 1.45, suggesting a moderate functional effect. *PXDN* also contributed one variant (rs1375881180) with a score of 1.423. Along with seven *ZNF423* variants, the *EYA1* gene had two variants, rs79409161 and rs80082766, with scores of 0.828 and 1.107, respectively, both falling below the threshold. Similarly, *ZEB1* had one variant (rs183670771) with the lowest total score (0.281), indicating a lower predicted functional impact.

Table 10. Top linear regression analysis results (p-value < $5x10^{-4}$)

rsID	Position	A1	A2	Gene	MAF	Beta	p-value	EUR MAF	AFR MAF
rs1234699363	16:49746764	A	G	*ZNF423*	0.015	-0.007	$3.07x10^{-5}$	0.0	0.0
rs2033560851	16:49748108	AT	A	*ZNF423*	0.015	-0.402	$3.07x10^{-5}$	N/A	N/A
rs1447977091	16:49752913	A	G	*ZNF423*	0.015	-0.402	$3.07x10^{-5}$	0.0	0.0
rs1596888340	16:49705898	G	A	*ZNF423, MRPS21P7*	0.015	-0.402	$3.07x10^{-5}$	0.0	0.0
rs368569835	16:49723516	T	C	*ZNF423*	0.015	-0.402	$3.07x10^{-5}$	0.0	0.0
rs1402577150	16:49771952	C	T	*ZNF423*	0.015	-0.402	$3.07x10^{-5}$	$1x10^{-4}$	0.0
rs2033816838	16:49760835	CG	C	*ZNF423, MRPS21P8*	0.015	-0.402	$3.07x10^{-5}$	N/A	N/A
rs371937482	16:49757962	A	G	*ZNF423*	0.015	-0.402	$3.07x10^{-5}$	0.0	0.0
rs1261756052	16:49799448	T	C	*ZNF423*	0.010	-0.484	$3.89x10^{-5}$	N/A	N/A
rs1375881180	2:1638996	A	G	*PXDN*	0.026	-0.305	$5.7x10^{-5}$	0.0	0.0
rs183670771	10:31409602	G	A	*ZEB1*	0.021	-0.338	$6.43x10^{-5}$	0.0	0.002
rs8062007	16:49682839	G	T	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	0.0	$4x10^{-4}$
rs146746521	16:49685577	G	A	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	0.0	0.019
rs142854276	16:49675916	G	C	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	$1x10^{-4}$	0.018
rs8055361	16:49679142	G	A	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	0.0	0.019
rs144234411	16:49680240	G	C	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	0.0	0.019
rs147790435	16:49680373	G	A	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	$1x10^{-4}$	0.018
rs8055110	16:49679418	T	C	*ZNF423*	0.015	-0.365	$1.78x10^{-4}$	0.0	0.019
rs75509972	9:107480883	T	G	*KLF4*	0.036	-0.236	$3.02x10^{-4}$	$2x10^{-4}$	0.023
rs80082766	8:71193700	T	C	*EYA1*	0.170	-0.114	$3.27x10^{-4}$	0.002	0.097
rs79409161	8:71193879	A	G	*EYA1*	0.170	-0.114	$3.27x10^{-4}$	$6.3x10^{-4}$	0.025

A1= effect allele for which the effect size (Beta) and standard error (SE) using log-transformed eGFR values is reported. A2 = reference allele, the represents the other allele at the locus. MAF = A1 frequency indicating how common the effect allele is in the study population. EUR MAF = MAFs of the variant in European populations from dbGaP. AFR MAF = MAFs of the variant in African populations from dbGaP. N/A = MAFs unavailable

Table 11. Selected annotation of top variants and normalised scores, ranked in descending order.

rsID	Position	Consequence	Gene	priPhCons	GERP N	CADD	n priPhCons	n GERP N	n CADD	Total
rs1596888340	**16:49705898**	**Intronic (3/7), Upstream**	***ZNF423*, *MRPS21P7***	**0.929**	**4.790**	**1.630**	**0.929**	**0.761**	**1.000**	**2.69**
rs368569835	**16:49723516**	**Intronic (3/7)**	***ZNF423***	**0.761**	**5.200**	**0.997**	**0.761**	**0.826**	**0.800**	**2.387**
rs146746521	**16:49685577**	**Intronic (3/7)**	***ZNF423***	**0.091**	**5.400**	**-0.515**	**0.091**	**0.858**	**0.324**	**1.717**
rs8062007	**16:49682839**	**Intronic (3/7), Regulatory**	***ZNF423***	**0.336**	**4.750**	**-0.072**	**0.336**	**0.775**	**0.463**	**1.574**
rs75509972	**9:107480883**	**Downstream, Regulatory**	***KLF4***	**0.027**	**6.290**	**-0.187**	**0.027**	**1.000**	**0.427**	**1.450**
rs1375881180	**2:1638996**	**Intronic (20/22)**	***PXDN***	**0.052**	**6.130**	**-0.336**	**0.052**	**0.990**	**0.381**	**1.423**
rs8055361	**16:49679142**	**Intronic (3/7)**	***ZNF423***	**0.228**	**0.271**	**0.751**	**0.228**	**0.430**	**0.723**	**1.381**
rs371937482	**16:49757962**	**Intronic (2/7)**	***ZNF423***	**0**	**4.990**	**-0.101**	**0**	**0.793**	**0.454**	**1.247**
rs1447977091	**16:49752913**	**Intronic (2/7)**	***ZNF423***	**0.003**	**3.800**	**0.474**	**0.003**	**0.604**	**0.635**	**1.242**
rs147790435	**16:49680373**	**Intronic (3/7), Regulatory**	***ZNF423***	**0.160**	**0.271**	**0.500**	**0.160**	**0.430**	**0.644**	**1.234**
rs1261756052	**16:49799448**	**Intronic (1/7), Regulatory**	***ZNF423***	**0.020**	**3.880**	**0.319**	**0.020**	**0.616**	**0.587**	**1.223**
rs144234411	16:49680240	Intronic (3/7)	*ZNF423*	0.099	0.271	0.536	0.099	0.430	0.655	**1.184**
rs142854276	16:49675916	Intronic (3/7)	*ZNF423*	0.146	2.700	0.320	0.146	0.429	0.587	**1.162**
rs2033560851	16:49748108	Intronic (2/7)	*ZNF423*	0.039	3.820	-0.021	0.039	0.607	0.481	**1.127**
rs80082766	8:71193700	Intronic (3/5), Downstream	*EYA1*	0.007	5.190	-0.670	0.007	0.825	0.275	**1.107**
rs2033816838	16:49760835	Intronic (2/7), Upstream	*ZNF423*, *MRPS21P8*	0.001	4.190	-0.406	0.001	0.666	0.359	**1.026**
rs8055110	16:49679418	Intronic (3/7)	*ZNF423*	0.001	0.271	0.045	0.001	0.430	0.500	**0.931**

rs79409161	8:71193879	Intronic (3/5), Downstream	*EYA1*	0.002	5.200	-1.547	0.002	0.826	0.00	**0.828**
rs1234699363	16:49746764	Intronic (2/7)	*ZNF423*	0.001	1.730	-0.312	0.001	0.275	0.388	**0.664**
rs1402577150	16:49771952	Intronic (2/7)	*ZNF423*	0.040	0.249	0.150	0.04	0.039	0.533	**0.612**
rs183670771	10:31409602	Intronic (1/8)	*ZEB1*	N/A	0	0.245	N/A	0	0.563	**0.281**

Bold rows = variants that passes the annotation threshold of 1.2. Consequence = variant locations (intron number shown in brackets). Annotation scores: Primate PhastCons (priPhCons), GERP N, and Combined Annotation Dependent Depletion (CADD) scores. N/A was used when a tool did not output a score for a specific variant. n = normalised scores. Total is the sum of all available tools for each variant.

For further characterisation of variants that met the threshold, Ensembl was used to further characterise variant consequences and revealed that the *KLF4* variant resides within an enhancer region, however, no additional information on other regulatory variants was available. SpliceAI results showed that none of the identified variant were predicted to affect RNA splicing. Lastly, the LDpair tool revealed that multiple *ZNF423* intron 3 variants were suggested to be in LD (Table 12). In intron 3 rs8062007, rs146746521, rs8055361 and rs147790435 were predicted to be in LD with all D′ values greater than 0.967 and all R^2 values greater than 0.904 The remaining variant in intron 3, rs1596888340 and rs368569835, were not present in the KGP reference panel so unfortunately could not be evaluated. For intron 2, rs1447977091 was not present in the KGP reference panel this LD between rs1447977091 and rs371937482 could not be evaluated.

Table 12. *ZNF423* variants in LD that met functional significance threshold.

Variants	Haplotypes				D′	R^2	Chi-square	p-value
rs8062007, rs146746521	T_A (0.994)	G_G (0.29)	T_G (0.0)	G_A (0.0)	0.997	0.934	4677.753	$< 1x10^{-4}$
rs8062007, rs147790435	T_A (0.994)	G_G (0.006)	T_G (0.0)	G_A (0.0)	1.0	1.0	5008.0	$< 1x10^{-4}$
rs8055361, rs147790435	A_A (0.994)	G_G (0.006)	G_A (0.0)	A_G (0.0)	1.0	0.968	4845.478	$< 1x10^{-4}$
rs146746521, rs8055361	A_A (0.994)	G_G (0.006)	A_G (0.0)	G_A (0.0)	1.0	0.968	4845.478	$< 1x10^{-4}$
rs8062007, rs8055361	T_A (0.994)	G_G (0.29)	T_G (0.0)	G_A (0.0)	0.967	0.904	4525.885	$< 1x10^{-4}$
rs146746521, rs147790435	A_A (0.994)	G_G (0.006)	A_G (0.0)	G_A (0.0)	1.0	1.0	5008.0	$< 1x10^{-}$

3.3.6 Gene-set analysis

Gene set analysis revealed commonalities among top genetic variants at a gene-set level for three sets of genes (Table 13). The gene sets included pathway-related gene sets, gene ontology (GO)-based gene sets, and regulatory gene sets. N refers to the total number of genes present in a gene family, a collection of proteins which have common features, while n refers to the genes of interest which were identified as a part of the gene set. The adjusted p-value for multiple testing for the gene- set analysis was calculated using Benjamin-Hochberg (FDR).

The first gene set belongs to the Chemical and Genetic Perturbation (CGP) category (MSigDB) in GSEA, which includes genes that show differential expression following drug treatment or genetic modifications. The identified gene set includes genes upregulated in in rhabdomyosarcoma cells by expression of PAX3-FOXO1 or PAX7-FOXO1 (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/DAVICIONI_TARGETS_OF_PAX_FOXO1_FUSIONS_UP) identified by the Davicioni *et al.* (2006) Enriched genes from this set include *PXDN*, *EYA1* and *KLF4*, with a p-value of 3.87×10^{-2}. Secondly, a gene set belonging to the GO Biological Process category (MSigDB) identified *EYA1* and *KLF4* to be enriched in the differentiation process by which a cell becomes committed to the mesoderm lineage (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/GOBP_MESODERMAL_CELL_FATE_COMMITMENT), with a significance level of 4.01×10^{-2}. The last gene set belonged to the WikiPathways category for the white fat cell differentiation gene set (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/WP_WHITE_FAT_CELL_DIFFERENTIATION). *KLF4* and *ZNF423* were identified with a significance level of 1.23×10^{-2}.

Table 13. Gene set analysis results with adjusted p-value < 0.05

Genes	Gene set	Gene set category	N	n	p-value	Adjusted p-value
PXDN*, *EYA1*, *KLF4	Davicioni targets of PAX FOXO1	Chemical and genetic perturbation (MSigDB)	256	3	1.14×10^{-5}	3.87×10^{-2}
EYA1*, *KLF4	GOBP mesodermal cell fate commitment	GO biological processes (MsigDB)	18	2	5.17×10^{-6}	4.01×10^{-2}
KLF4*, *ZNF423	WP white fat cell differentiation	WikiPathways	32	2	1.68×10^{-5}	1.23×10^{-2}

3.4 DISCUSSION

3.4.1 Interpretation of association analysis findings

The association analysis of variants linked to eGFR levels, including those in kidney-related genes, *PXDN*, and all its potential transcriptional regulators identified in Chapter 2, revealed no statistically significant results after applying the stringent Bonferroni correction for multiple testing. Additionally, the low power detected indicates a high probability of a Type II error, where a true association is not detected. This is likely due to the small sample size, which limits the ability to detect the small genetic effects that contribute to eGFR variation. However, when examining the top variants that met a less conservative arbitrary p-value threshold ($5x10^{-4}$), 21 variants were identified in genes associated with kidney function. This suggests that these variants may potentially influence eGFR levels, warranting further investigation into their role in kidney health.

Of the 21 top variants submitted for functional annotation, 16 were located in *ZNF423*, with one in the first intron, six in intron 2, and nine in intron 3. Two of the variants identified in *ZNF423* overlapped with variants identified upstream of the two pseudogenes, *MRPS21P7* (rs1596888340) and *MRPS21P8* (rs2033816838). One variant was identified in intron 20 of *PXDN* (rs1375881180), one in intron 1 of *ZEB1* (rs183670771), one variant in a downstream and regulatory region of *KLF4* (rs75509972), and two variants identified in intron 3 of *EYA1* (rs80082766 and rs79409161). Interestingly, all identified variants were associated with low eGFR levels, as indicated by their negative beta values.

The *ZNF423* variants had the lowest MAFs in the study population, each with a MAF of 0.015, meaning that they were present in 1.5% of the alleles studied in this cohort. This suggests that these variants are relatively rare and could potentially have a stronger functional effect in the black South African populations, as risk alleles are often enriched in alleles with low MAFs (<0.1) (Kido et al., 2018). Variants with low MAFs may be under negative selection, where harmful alleles are reduced in frequency due to their negative effects on fitness and survival (Chan et al., 2014). Except for the *EYA1* variants (MAF=0.170), all other variants had MAFs less than 0.1, indicating their rarity and potential risk status (Kido et al., 2018).

When comparing European and African MAFs for the identified alleles, six were absent in

European populations (MAF = 0) but present in African populations (MAF > 0). This suggests population-specific variation in individuals of African ancestry, as these alleles were observed in both our black South African cohort and broader African populations but not in European populations. This highlights variants in African populations that may be associated with a phenotype, low eGFR, that have not been identified in non-African populations. A similar case has been observed with genetic variation in the *LEP* gene associated with higher eGFR levels in a black African cohort (Okpechi et al., 2010). It is also noteworthy that some variants had a MAF = 0 in the African populations as these populations are sampled from the African diaspora which originate predominantly from West Africa. This indicates that populations of African Ancestry cannot be used as a proxy for all African populations.

Variants were ranked using normalised scores from three prediction tools, CADD, PhastCons and GERP N, with an arbitrary threshold of 1.2 applied to select those predicted to be more functionally significant. Combining scores from multiple tools enhances robustness and reduces bias, as each tool employs different algorithms and prioritises different features. This approach allows for a more reliable and consistent assessment, ensuring that variants are fairly compared using multiple methods rather than relying on just one. Variants meeting the threshold were predicted to be more functionally significant (variants in bold in Table 11) and, as such, are more likely to contribute to lower eGFR values observed in the association analysis.

Of the 11 variants that met the functional significance threshold, nine were located in *ZNF423*, one was located in *PXDN* and one was associated with *KLF4*. Notably, none of the identified top variants, including the 11 mentioned, were linked to any known disease phenotypes. The top SNP (rs1596888340, $p = 3.07x10^{-5}$, $\beta = -0.402$), located in intron 3, had the highest CADD score (1.63) and a high primate PhastCons score (0.929), suggesting that this variant could be damaging and resides in a highly evolutionarily conserved region, respectively. This observation is consistent with literature showing that SNPs in highly conserved regions tend to have lower allele frequencies due to negative selection, making them more likely to be functionally significant and potentially damaging (Levenstien & Klein, 2011).

The following variants, rs368569835 ($p = 3.07x^{-5}$, $\beta = -0.402$), rs146746521 ($p = 1.78x10^{-4}$, $\beta = -0.365$), and rs8062007 ($p = 1.78x10^{-4}$, $\beta = -0.365$), along with the second-to-last variant that met the threshold, rs147790435 ($p = 1.78x10^{-4}$, $\beta = -0.365$), are also located in intron 3 of *ZNF423*. While primate PhastCons scores vary, these variants are likely in LD, as evidenced by all D′ values exceeding than 0.967, indicating strong linkage, and all R^2 scores being above 0.904, signifying allele dependence (Table 13). Some D′ and R^2 values reach a perfect score of 1, indicating complete linkage and that one allele perfectly predicts the other, respectively. Variants in LD are useful because they may be linked to a disease locus, and by assessing the association between these variants and the disease, the disease locus may be detected (Ardlie et al., 2002). However, the presence of LD among variants was expected, as the LD pruning step was excluded from the quality control process in attempt to retain more variants for the association analysis. For these variants, GERP N scores, which measure evolutionary constraint under neutral evolution, range from approximately 3.8 to 6.4, placing them among the higher GERP N scores obtained. This suggests greater functional constraint, meaning that changes at these sites are more likely to be deleterious, as supported by studies showing that variants with GERP scores of above four are associated with strong deleterious effects (Henn et al., 2016; van der Valk et al., 2021).

One *ZNF423* variant was present in intron 1 (rs1261756052, $p = 3.89x10^{-5}$, $\beta = -0.484$), although it ranked the lowest among the variants that met the threshold. The remaining variants were located in intron 2 (rs8055361, $p = 1.78x10^{-4}$, $\beta = -0.365$; rs371937482, $p = 3.07x10^{-5}$, $\beta = -0.402$; rs1447977091, $p = 3.07x10^{-5}$, $\beta = -0.402$), all of which had low primate PhastCons scores but moderate GERP N and CADD scores. In addition to being identified in intronic regions, three variants (rs1261756052, rs8062007 and rs147790435) were also identified to reside regulatory regions, suggesting they may influence TF binding or other regulatory mechanisms (Rojano et al., 2019). However, none have been identified to overlap with regulatory elements by Ensembl, indicating further investigation is needed.

ZNF423 encodes a protein in the Kruppel-like C_2H_2 zinc finger protein family and functions as a TF. ZNF423 has been implicated in various conditions, including oestrogen receptor-positive breast cancer, through BRCA1 trans-activation (Bond et al., 2018), and neurodevelopmental disorders (Deshpande et al., 2020) and Joubert Syndrome, which results from cerebellar vermis underdevelopment (Casoni et al., 2017; Warming et al., 2006).

Interestingly, mutations in *ZNF423* have also been linked to nephronophthisis-related ciliopathies, degenerative, single-gene recessive diseases affecting the kidney, retina, and brain (Chaki et al., 2011; Hildebrandt et al., 2011). Nephronophthisis (NPHP), is a cystic kidney disease, is the leading cause of ESRD in individuals under 30 (Chaki et al., 2012). Recessive loss of function mutations in NPHP lead to severe congenital abnormalities, including malformation and dysplasia in polycystic dysplastic kidneys, microphthalmia/coloboma (eyes), Joubert syndrome (cerebellum), cysts and ductal plate malformation (liver), while partial loss of function mutations cause adult-onset degenerative conditions, including tubular degeneration of the kidney with fibrosis, Senior-Loken syndrome (degeneration of the retina) and liver fibrosis (Chaki et al., 2011; Hildebrandt et al., 2011). The *ZNF423* variants identified in this study have no known disease associations. Additionally, despite occurring in intronic regions, they may still influence gene expression through mechanisms such as splicing, intron-mediated enhancement (IME) and the efficiency of mRNA translation (Shaul, 2017). Notably, many SNPs in deep intronic regions have been implicated in kidney disease through GWAS, including intronic variants in *SHROOM3* associated with renal fibrosis and in *MYH9* associated with glomerular disease (Kopp et al., 2010; Menon et al., 2015).

One variant was identified in *KLF4* (rs75509972, $p = 3.02 \times 10^{-4}$, $\beta = -0.236$), a downstream regulatory variant associated with low eGFR. While this variant had weaker primate PhastCons (0.027) and CADD (-0.186) scores, it had the highest GERP N score (6.29), suggesting it occurs in an evolutionary constrained and highly conserved region (Henn et al., 2016). This indicates potential functional importance, further supported by its identification in a regulatory region and in an enhancer by Ensembl. However, it is not predicted to be deleterious by CADD and has not been associated with any disease phenotypes.

Like *ZNF423*, *KLF4* is a zinc finger TF part of the Kruppel-like factor family. KLF4 regulates a range of cellular processes, including cell differentiation, growth and proliferation (Ghaleb & Yang, 2017). Among the roles of KLF4 in various organs and diseases, KLF4 has been implicated in the kidney (Ghaleb & Yang, 2017). KLF4 regulates kidney processes, for example fibrosis progression, acting as a negative regulator of stiff matrix-induced proliferation and cell spreading, both involved in renal fibrosis pathogenesis (Chen et al., 2015). Additionally, in diabetic nephropathy, KLF4 is thought to have a therapeutic potential

by suppressing TGF-β1-induced inflammation (Mreich et al., 2015).

Although the identified variant is downstream of *KLF4*, its presence in an enhancer region, suggests it could modulate *KLF4* expression through regulatory mechanisms (Rojano et al., 2019). Further studies are required to determine its role in kidney disease processes, such as fibrosis.

The *PXDN* variant, located in intron 20 (rs1375881180, $p = 5.73 \times 10^{-5}$, $\beta = -0.305$), was also associated with lower eGFR levels, although no current data support this association or any disease phenotype. The lack of data could be attributed to the fact that the allele frequency for the A allele is 0 in most non-African and African Ancestry populations. This variant had weaker primate PhastCons (0.052) and CADD scores (-0.336) and had the second highest GERP N score (6.13), indicating that it is located in a region under evolutionary constraint and therefore could be functionally important (Henn et al., 2016). Mutations in regions under evolutionary constraint could affect gene function, potentially influencing eGFR. *PXDN* intronic variants have been previously identified as damaging in various anterior segment dysgenesis of the eye, as seen in intron 17, where the c.3609-1307G>A variant associates with anophthalmias and microphthalmies and introduces a premature termination codon (Basharat et al., 2023). This variant is predicted to lead to either a truncated PXDN protein lacking the peroxidase domain or degradation via nonsense-mediated decay (Basharat et al., 2023). This emphasises the impact of non-coding *PXDN* variants in relation to disease phenotypes and the importance of considering them in research. As discussed previously, PXDN plays a key role in stabilising collagen IV in the basement membrane and, because of this, plays a role in structural integrity of the kidney (Cheng & Shi, 2022). *PXDN* dysregulation is associated with renal fibrosis, as it is overexpressed in response to TGF-β signalling (Péterfi et al., 2009). The identified variant could result in splicing regulation (Joynt et al., 2020) and, although unlikely due to its location in intron 20, in improper regulation of transcription potentially affecting *PXDN* expression, which could activate pathways, such as TGF-β, resulting in a negative effect on the structural integrity of the kidney, causing decreased eGFR. Further work is needed to explore the potential impact of this variant.

Lastly, the gene set analysis identified three sets of genes overrepresented in predefined gene set categories. While two gene set do not appear directly relevant to processes affecting eGFR in the kidney, the mesodermal cell fate commitment gene set, responsible for differentiation

processes leading to mesoderm-derived cells, includes *EYA1* and *KLF4*.

EYA1, a member of the EYA protein family, plays a critical role in kidney development (Tian et al., 2024). Variants in *EYA1* have been identified to cause Branchio oto renal (BOR) syndrome, a condition characterised by kidney abnormalities (Orten et al., 2008; Zhang et al., 2024). Similarly, as previously discussed, KLF4 is implicated in kidney function, particularly in fibrosis regulation (Chen et al., 2015; Mreich et al., 2015). Given their toles in mesodermal differentiation, variants in *EYA1* and *KLF4* may influence kidney development and fibrosis, potentially affecting eGFR levels. However, further investigation is needed to confirm these effects.

3.4.2 Limitations and future research

This study was limited by the small sample size and limited regions used in the association analysis, likely impacting the lack of statistically significant results and low powered analysis. Additionally, less stringent quality control steps were applied in order to increase the number of variants included in the analysis, which may have introduced noise into the results. This study used cross-sectional data, limiting the ability to assess causality of identified associations over time. Due to the majority of UACR values being missing for individuals in our sample, we were limited in our ability to investigate this phenotype. Further, we did not have access to additional patient data such as diabetes status, HIV infection and hypertension status, which may have provided valuable context for interpreting our findings or identifying potential confounding factors. Lastly, this study had a limited understanding of the functional impact of variants, as the analysis focused primarily on association.

Future studies should address the limitations mentioned above. For example, a larger sample size should be employed, and more stringent quality control should be applied. Since no statistically significant variants were identified in this study, *in vivo* studies investigating identified variants are not recommended for future research at this stage. Larger sample sizes and other adjustments outlined are necessary before considering such approaches. As a small-scale, candidate gene study this work can be expanded on by performing by increasing the sample size and conducting a GWAS, rather than a candidate gene study.

3.4.3 Conclusion

This study is the first to investigate genetic variation in *PXDN*, its transcriptional regulators, and key kidney disease-associated genes in a black South African cohort, aiming to identify potential associations with eGFR levels. Although no variants were found to be statistically significant after correcting for multiple testing, the identified variants may still contribute to low eGFR levels, given the known roles of the implicated genes in in kidney dysfunction. Further investigation is needed to confirm these association and determine the precise mechanisms underlying the low eGFR levels observed in our cohort. Bioinformatic annotations and analyses revealed variants that may be deleterious and damaging. Interestingly, a variant in *PXDN* was identified as potentially associated with eGFR levels, suggesting a possible role for PXDN in kidney function. Finally, gene set enrichment analysis highlighted potentially important pathways associated with the genes containing the top variants, providing insight for future research.

CHAPTER 4. Overall Discussion and conclusions

This study set out with two main aims. Firstly, it aimed to identify potential TFs regulating the *PXDN* gene in kidney tissue. This objective was successfully achieved, with several TFs identified. Secondly, the study aimed to identify whether genetic variants within *PXDN*, the identified TFs, or previously reported TFs are associated with eGFR in a black South African cohort. While this analysis did not reveal any statistically significant variants, the findings contribute to the growing understanding of genetic variation linked to kidney function in understudied populations.

In Chapter 2, we identified multiple TFs that may regulate *PXDN* in the kidney, mapping those that met our p and q-value thresholds and characterising their potential roles in kidney dysfunction. Thirteen high-confidence TF binding sites were mapped, several of which are implicated kidney dysfunction, providing insight into the regulatory mechanisms driving *PXDN* expression.

Chapter 3 consisted of an association analysis in a black South African cohort to examine the relationship between variants in *PXDN*, *PXDN* transcriptional regulators identified in Chapter 2, and kidney disease-associated genes with eGFR levels. Although no statistically significant variants were detected, several potentially damaging variants were identified in genes with known roles in kidney dysfunction.

PXDN was a major focus for this project due to its roles in renal fibrosis (Colon et al., 2019; Péterfi et al., 2009). PXDN plays an important role in stabilising collagen IV which is essential for ECM assembly, basement membrane integrity and tissue development (Bhave et al., 2012; Fidler et al., 2014; Lee et al., 2020; McCall et al., 2014). Dysregulated *PXDN* expression has been linked to diseases characterised by basement membrane defects and abnormalities in collagen IV (Cheng & Shi, 2022). In line with these findings, this study identified both potential regulators of *PXDN* in the context of kidney function and disease as well as variants in the *PXDN* gene that associate with low eGFR. The kidneys rely on the coordinated function of nephron segments and the glomerulus for blood filtration and waste removal (Scholz et al., 2021). The glomerular basement, containing collagen IV, plays a crucial role in maintaining

filtration integrity (Miner, 2012). Impaired filtration leads to decreased kidney function, which is reflected in reduced eGFR.

The association analysis identified an intronic variant in *PXDN*, suggesting that mutations in the gene may influence eGFR levels in the kidney. This variant had a high GERP N score (6.130), signifying that it is located in a region under evolutionary constraint and therefore may be functionally important (Henn et al., 2016). Despite this variant not reaching stringent statistical significance for this, the variant warrants further investigation using a larger cohort to achieve greater statistical power.

Variants associated with low eGFR were found in two *PXDN* TFs, *ZNF423* and *ZEB1*, both of which were identified as binding to intron 1 in untreated HEK293 cells (Chapter 2). Although the putative binding site for ZNF423 did not meet the FIMO q-value threshold, *ZNF423* was still included in the association analysis in Chapter 3 because it was identified as a potential regulator of *PXDN*. Notably, the majority of top variants meeting the $5x10^{-4}$ threshold (16 out of 21) were *ZNF423* variants, suggesting that this gene may play a significant role in the genetic variation associated with low eGFR levels in our cohort. Further, multiple variants had strong functionally significance scores indicating potentially damaging consequences of variants. Examples of this include rs1596888340, ranked as the top variant, which had the highest CADD score. However, the role of ZNF423 in regulating *PXDN* remains to be confirmed, and its potential interaction with *PXDN* and relationship to low eGFR should be investigated further.

A variant was identified in *ZEB1*, which has also been characterised as a potential regulator of *PXDN* in kidney cells (Chapter 2). In Chapter 2, ZEB1's putative *PXDN* binding site met both FIMO p and q-value thresholds, allowing for its binding site to be mapped. ZEB1 is known as a driver of EMT and promotes the transcription of gene responsive to TGF-β (Zhang et al., 2015) In the kidney, increased ZEB1 expression is associated with EMT of renal cell carcinoma cells (S. Zhang et al., 2017). Despite the ZEB1 variant ranking last among the top variants based on prediction from three functional significance tools, this does not exclude its potential relevance to kidney function. Further studies are needed to draw definitive conclusions. Additionally, the role ZEB1 in regulating *PXDN*, and whether their interaction might contribute to low eGFR levels, remains unclear.

While the study was limited by the reliance on third-party data for investigating transcriptional regulators of *PXDN* and the modest sample size of the association study, this work provides a foundation for understanding both the genetic factors affecting kidney function and regulation of *PXDN* in the kidney. This study also advances our understanding of the unexplained susceptibility to CKD in South African populations, although further research, utilising larger populations cohorts, is needed to fully address the existing gap in understanding. Furthermore, the need for future studies to define the comprehensive role of identified TFs, and genetic variants observed in the association study, in the mechanisms underlying impaired kidney function is emphasised.

In summary, this study provides insight into the regulatory mechanisms of *PXDN* in the kidney, as well as TFs which could be key players in regulating *PXDN*. Further, variants in *PXDN* TFs were found to be associated with low eGFR in our sample, suggesting a potential link between these TFs and impaired kidney function. Whether *PXDN* itself is involved remains to be elucidated. Overall, these findings suggest that both transcriptional regulators of *PXDN* and genetic variation in *PXDN* may contribute to kidney dysfunction. Further investigation is required to validate the regulation of *PXDN* by identified TFs, the putative binding sites of TFs identified, and variants identified in the association analysis. Experimental validation and functional assays are needed to confirm identified binding sites and determine the regulatory role of these TFs. Further, larger cohort studies are needed to confirm identified variants and explore their implications in kidney dysfunction.

REFERENCES

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., … Volpi, S. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509). https://doi.org/10.1126/SCIENCE.AAZ1776

Ai, K., Li, X., Zhang, P., Pan, J., Li, H., He, Z., Zhang, H., Yi, L., Kang, Y., Wang, Y., Chen, J., Li, Y., Xiang, X., Chai, X., & Zhang, D. (2022). Genetic or siRNA inhibition of MBD2 attenuates the UUO- and I/R-induced renal fibrosis via downregulation of EGR1. *Molecular Therapy Nucleic Acids*, *28*. https://doi.org/10.1016/j.omtn.2022.02.015

Alberts, M., Urdal, P., Steyn, K., Stensvold, I., Tverdal, A., Nel, J. H., & Steyn, N. P. (2005). Prevalence of cardiovascular diseases and associated risk factors in a rural black population of South Africa. *European Journal of Preventive Cardiology*, *12*(4). https://doi.org/10.1097/01.hjr.0000174792.24188.8e

AlOkda, A., & Van Raamsdonk, J. M. (2022). Effect of DMSO on lifespan and physiology in C. elegans : Implications for use of DMSO as a solvent for compound delivery. *MicroPublication Biology*, *2022*. https://doi.org/10.17912/micropub.biology.000634

Anbazhagan, K., Sampathkumar, K., Ramakrishnan, M., Gomathi, P., Gomathi, S., & Selvam, G. S. (2009). Analysis of polymorphism in Renin Angiotensin System and other related genes in South Indian chronic kidney disease patients. *Clinica Chimica Acta*, *406*(1–2). https://doi.org/10.1016/j.cca.2009.06.003

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. In *Nature Reviews Genetics* (Vol. 3, Issue 4). https://doi.org/10.1038/nrg777

Atun, R., Davies, J. I., Gale, E. A. M., Bärnighausen, T., Beran, D., Kengne, A. P., Levitt, N. S., Mangugu, F. W., Nyirenda, M. J., Ogle, G. D., Ramaiya, K., Sewankambo, N. K., Sobngwi, E., Tesfaye, S., Yudkin, J. S., Basu, S., Bommer, C., Heesemann, E., Manne-Goehler, J., … Werfalli, M. (2017). Diabetes in sub-Saharan Africa: from clinical care to health policy. In *The Lancet Diabetes and Endocrinology* (Vol. 5, Issue 8). https://doi.org/10.1016/S2213-8587(17)30181-X

Bai, S., Xiong, X., Tang, B., Ji, T., Li, X., Qu, X., & Li, W. (2021). Hsa-miR-199b-3p Prevents the Epithelial-Mesenchymal Transition and Dysfunction of the Renal Tubule by Regulating E-cadherin through Targeting KDM6A in Diabetic Nephropathy. *Oxidative Medicine and Cellular Longevity*, *2021*. https://doi.org/10.1155/2021/8814163

Banerjee, S., Lo, W. C., Majumder, P., Roy, D., Ghorai, M., Shaikh, N. K., Kant, N., Shekhawat, M. S., Gadekar, V. S., Ghosh, S., Bursal, E., Alrumaihi, F., Dubey, N. K., Kumar, S., Iqbal, D., Alturaiki, W., Upadhye, V. J., Jha, N. K., Dey, A., & Gundamaraju, R. (2022). Multiple roles for basement membrane proteins in cancer progression and EMT. *European Journal of Cell Biology*, *101*(2). https://doi.org/10.1016/j.ejcb.2022.151220

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4). https://doi.org/10.1016/j.cell.2007.05.009

Basharat, R., Rodenburg, K., Rodríguez-Hidalgo, M., Jarral, A., Ullah, E., Corominas, J., Gilissen, C., Zehra, S. T., Hameed, U., Ansar, M., & de Bruijn, S. E. (2023). Combined Single Gene Testing and Genome Sequencing as an Effective Diagnostic Approach for Anophthalmia and Microphthalmia Patients. *Genes*, *14*(8). https://doi.org/10.3390/genes14081573

Beckerman, P., & Susztak, K. (2018). APOL1: The Balance Imposed by Infection, Selection, and Kidney Disease. In *Trends in Molecular Medicine* (Vol. 24, Issue 8). https://doi.org/10.1016/j.molmed.2018.05.008

Bergamasco, M. I., Ranathunga, N., Abeysekera, W., Li-Wai-Suen, C. S. N., Garnham, A. L., Willis, S. N., McRae, H. M., Yang, Y., D'Amico, A., Di Rago, L., Wilcox, S., Nutt, S. L., Alexander, W. S., Smyth, G. K., Voss, A. K., & Thomas, T. (2024). The histone acetyltransferase KAT6B is required for hematopoietic stem cell development and function. *Stem Cell Reports*, *19*(4), 469–485. https://doi.org/10.1016/j.stemcr.2024.02.005

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., & Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2). https://doi.org/10.1073/pnas.231608898

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., Van Der Veen, L., Hombert, J. M., Quintana-Murci, L., & Comas, D. (2009). Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages. *Molecular Biology and Evolution*, *26*(7). https://doi.org/10.1093/molbev/msp069

Bhave, G., Colon, S., & Ferrell, N. (2017). The sulfilimine cross-link of collagen IV contributes to kidney tubular basement membrane stiffness. *American Journal of Physiology - Renal Physiology*, *313*(3). https://doi.org/10.1152/ajprenal.00096.2017

Bhave, G., Cummings, C. F., Vanacore, R. M., Kumagai-Cresse, C., Ero-Tolliver, I. A., Rafi, M., Kang, J. S., Pedchenko, V., Fessler, L. I., Fessler, J. H., & Hudson, B. G. (2012). Peroxidasin forms sulfilimine chemical bonds using hypohalous acids in tissue genesis. *Nature Chemical Biology*, *8*(9). https://doi.org/10.1038/nchembio.1038

Bikbov, B., Purcell, C. A., Levey, A. S., Smith, M., Abdoli, A., Abebe, M., Adebayo, O. M., Afarideh, M., Agarwal, S. K., Agudelo-Botero, M., Ahmadian, E., Al-Aly, Z., Alipour, V., Almasi-Hashiani, A., Al-Raddadi, R. M., Alvis-Guzman, N., Amini, S., Andrei, T., Andrei, C. L., … Murray, C. J. L. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *395*(10225). https://doi.org/10.1016/S0140-6736(20)30045-3

Bond, H. M., Scicchitano, S., Chiarella, E., Amodio, N., Lucchino, V., Aloisio, A., Montalcini, Y., Mesuraca, M., & Morrone, G. (2018). ZNF423: A new player in estrogen receptor-positive breast cancer. In *Frontiers in Endocrinology* (Vol. 9, Issue MAY). https://doi.org/10.3389/fendo.2018.00255

Bonnans, C., Chou, J., & Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. In *Nature Reviews Molecular Cell Biology* (Vol. 15, Issue 12). https://doi.org/10.1038/nrm3904

Brandenburg, J. T., Govender, M. A., Winkler, C. A., Romuald Boua, P., Agongo, G., Fabian, J., & Ramsay, M. (2022). Apolipoprotein L1 High-Risk Genotypes and Albuminuria in Sub-Saharan African Populations. *Clinical Journal of the American Society of Nephrology*, *17*(6). https://doi.org/10.2215/CJN.14321121

Brandenburg, J.-T., Chen, W. C., Boua, P. R., Govender, M. A., Agongo, G., Micklesfield, L. K., Sorgho, H., Tollman, S., Asiki, G., Mashinya, F., Hazelhurst, S., Morris, A. P., Fabian, J., & Ramsay, M. (2024). Genetic association and transferability for urinary albumin-creatinine ratio as a marker of kidney disease in four Sub-Saharan African populations and non-continental individuals of African ancestry. *Frontiers in Genetics*, *15*. https://doi.org/10.3389/fgene.2024.1372042

Brenner, B. M., Cooper, M. E., de Zeeuw, D., Keane, W. F., Mitch, W. E., Parving, H.-H., Remuzzi, G., Snapinn, S. M., Zhang, Z., & Shahinfar, S. (2001). Effects of Losartan on Renal and Cardiovascular Outcomes in Patients with Type 2 Diabetes and Nephropathy. *New England Journal of Medicine*, *345*(12). https://doi.org/10.1056/nejmoa011161

Briem, E., Budkova, Z., Sigurdardottir, A. K., Hilmarsdottir, B., Kricker, J., Timp, W., Magnusson, M. K., Traustadottir, G. A., & Gudjonsson, T. (2019). MiR-203a is differentially expressed during branching morphogenesis and EMT in breast progenitor cells and is a repressor of peroxidasin. *Mechanisms of Development*, *155*. https://doi.org/10.1016/j.mod.2018.11.002

Brodbeck, S., Besenbeck, B., & Englert, C. (2004). The transcription factor Six2 activates expression of the Gdnf gene as well as its own promoter. *Mechanisms of Development*, *121*(10). https://doi.org/10.1016/j.mod.2004.05.019

Brück, K., Stel, V. S., Gambaro, G., Hallan, S., Völzke, H., Ärnlöv, J., Kastarinen, M., Guessous, I., Vinhas, J., Stengel, B., Brenner, H., Chudek, J., Romundstad, S., Tomson, C., Gonzalez, A. O., Bello, A. K., Ferrieres, J., Palmieri, L., Browne, G., … Jager, K. J. (2016). CKD prevalence varies across the European general population. *Journal of the American Society of Nephrology*, *27*(7). https://doi.org/10.1681/ASN.2015050542

Buraczynska, M., Ksiazek, P., Drop, A., Zaluska, W., Spasiewicz, D., & Ksiazek, A. (2006). Genetic polymorphisms of the renin-angiotensin system in end-stage renal disease. *Nephrology Dialysis Transplantation*, *21*(4). https://doi.org/10.1093/ndt/gfk012

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Eichler, E. E., Korbel, J. O., Lee, C., Marschall, T., Devine, S. E., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18).

https://doi.org/10.1016/j.cell.2022.08.004

Carrero, J. J., Grams, M. E., Sang, Y., Ärnlöv, J., Gasparini, A., Matsushita, K., Qureshi, A. R., Evans, M., Barany, P., Lindholm, B., Ballew, S. H., Levey, A. S., Gansevoort, R. T., Elinder, C. G., & Coresh, J. (2017). Albuminuria changes are associated with subsequent risk of end-stage renal disease and mortality. *Kidney International*, *91*(1). https://doi.org/10.1016/j.kint.2016.09.037

Casoni, F., Croci, L., Bosone, C., D'Ambrosio, R., Badaloni, A., Gaudesi, D., Barili, V., Sarna, J. R., Tessarollo, L., Cremona, O., Hawkes, R., Warming, S., & Consalez, G. G. (2017). Zfp423/ZNF423 regulates cell cycle progression, the mode of cell division and the DNA-damage response in purkinje neuron progenitors. *Development (Cambridge)*, *144*(20). https://doi.org/10.1242/dev.155077

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., & Raschellà, G. (2017). Zinc-finger proteins in health and disease. In *Cell Death Discovery* (Vol. 3, Issue 1). https://doi.org/10.1038/cddiscovery.2017.71

Casson, R. J., & Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. In *Clinical and Experimental Ophthalmology* (Vol. 42, Issue 6). https://doi.org/10.1111/ceo.12358

Cha, R. H., Kang, S. W., Park, C. W., Cha, D. R., Na, K. Y., Kim, S. G., Yoon, S. A., Han, S. Y., Chang, J. H., Park, S. K., Lim, C. S., & Kim, Y. S. (2016). A randomized, controlled trial of oral intestinal sorbent AST-120 on renal function deterioration in patients with advanced renal dysfunction. *Clinical Journal of the American Society of Nephrology*, *11*(4). https://doi.org/10.2215/CJN.12011214

Chaki, M., Airik, R., Ghosh, A. K., Giles, R. H., Chen, R., Slaats, G. G., Wang, H., Hurd, T. W., Zhou, W., Cluckey, A., Gee, H. Y., Ramaswami, G., Hong, C. J., Hamilton, B. A., Červenka, I., Ganji, R. S., Bryja, V., Arts, H. H., Van Reeuwijk, J., … Hildebrandt, F. (2012). Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling. *Cell*, *150*(3). https://doi.org/10.1016/j.cell.2012.06.028

Chaki, M., Hoefele, J., Allen, S. J., Ramaswami, G., Janssen, S., Bergmann, C., Heckenlively, J. R., Otto, E. A., & Hildebrandt, F. (2011). Genotype-phenotype correlation in 440 patients with NPHP-related ciliopathies. *Kidney International*, *80*(11). https://doi.org/10.1038/ki.2011.284

Chan, Y., Lim, E. T., Sandholm, N., Wang, S. R., McKnight, A. J., Ripke, S., Daly, M. J., Neale, B. M., Salem, R. M., & Hirschhorn, J. N. (2014). An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *American Journal of Human Genetics*, *94*(3). https://doi.org/10.1016/j.ajhg.2014.02.006

Chang, J., & Chaudhuri, O. (2019). Beyond proteases: Basement membrane mechanics and cancer invasion. In *Journal of Cell Biology* (Vol. 218, Issue 8). https://doi.org/10.1083/JCB.201903066

Chen, H., Tian, Y., Shu, W., Bo, X., & Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041374

Chen, T. K., Knicely, D. H., & Grams, M. E. (2019). Chronic Kidney Disease Diagnosis and Management: A Review. In *JAMA - Journal of the American Medical Association* (Vol. 322, Issue 13). https://doi.org/10.1001/jama.2019.14745

Chen, W. C., Lin, H. H., & Tang, M. J. (2015). Matrix-Stiffness–Regulated Inverse Expression of Krüppel-Like Factor 5 and Krüppel-Like Factor 4 in the Pathogenesis of Renal Fibrosis. *The American Journal of Pathology*, *185*(9), 2468–2481. https://doi.org/10.1016/J.AJPATH.2015.05.019

Cheng, G., Li, H., Cao, Z., Qiu, X., McCormick, S., Thannickal, V. J., & Nauseef, W. M. (2011). Vascular peroxidase-1 is rapidly secreted, circulates in plasma, and supports dityrosine cross-linking reactions. *Free Radical Biology and Medicine*, *51*(7). https://doi.org/10.1016/j.freeradbiomed.2011.07.002

Cheng, G., Salerno, J. C., Cao, Z., Pagano, P. J., & Lambeth, J. D. (2008). Identification and characterization of VPO1, a new animal heme-containing peroxidase. *Free Radical Biology and Medicine*, *45*(12). https://doi.org/10.1016/j.freeradbiomed.2008.09.009

Cheng, G., & Shi, R. (2022). Mammalian peroxidasin (PXDN): From physiology to pathology. In *Free Radical Biology and Medicine* (Vol. 182). https://doi.org/10.1016/j.freeradbiomed.2022.02.026

Chevalier, R. L. (2016). The proximal tubule is the primary target of injury and progression of kidney disease: Role of the glomerulotubular junction. In *American Journal of Physiology - Renal Physiology* (Vol. 311, Issue 1). https://doi.org/10.1152/ajprenal.00164.2016

Chew, C., & Lennon, R. (2018). Basement membrane defects in genetic kidney diseases. In *Frontiers in Pediatrics* (Vol. 6). https://doi.org/10.3389/fped.2018.00011

Choi, A., Lao, R., Ling-Fung Tang, P., Wan, E., Mayer, W., Bardakjian, T., Shaw, G. M., Kwok, P. Y., Schneider, A., & Slavotinek, A. (2015). Novel mutations in PXDN cause microphthalmia and anterior segment dysgenesis. *European Journal of Human Genetics*, *23*(3). https://doi.org/10.1038/ejhg.2014.119

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., Ghoorah, A. W., Dareng, E., Odia, T., Falola, O., Adebiyi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O. A., Mbiyavanga, M., … Hanchard, N. A. (2020). High-depth African genomes inform human migration and health. *Nature*, *586*(7831). https://doi.org/10.1038/s41586-020-2859-7

Christov, M., Clark, A. R., Corbin, B., Hakroush, S., Rhee, E. P., Saito, H., Brooks, D., Hesse, E., Bouxsein, M., Galjart, N., Jung, J. Y., Mundel, P., Jüppner, H., Weins, A., & Greka, A. (2018). Inducible podocyte-specific deletion of CTCF drives progressive kidney disease and bone abnormalities. *JCI Insight*, *3*(4). https://doi.org/10.1172/jci.insight.95091

Cockerill, P. N. (2011). Structure and function of active chromatin and DNase i hypersensitive sites. In *FEBS Journal* (Vol. 278, Issue 13). https://doi.org/10.1111/j.1742-4658.2011.08128.x

Coleman, R. A., & Pugh, B. F. (1995). Evidence for functional binding and stable sliding of

the TATA binding protein on nonspecific DNA. *Journal of Biological Chemistry*, *270*(23). https://doi.org/10.1074/jbc.270.23.13850

Colon, S., Luan, H., Liu, Y., Meyer, C., Gewin, L., & Bhave, G. (2019). Peroxidasin and eosinophil peroxidase, but not myeloperoxidase, contribute to renal fibrosis in the murine unilateral ureteral obstruction model. *American Journal of Physiology-Renal Physiology*, *316*(2). https://doi.org/10.1152/ajprenal.00291.2018

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7). https://doi.org/10.1101/gr.3577405

Cui, S., Lv, X., Li, W., Li, Z., Liu, H., Gao, Y., & Huang, G. (2018). Folic acid modulates VPO1 DNA methylation levels and alleviates oxidative stress-induced apoptosis in vivo and in vitro. *Redox Biology*, *19*. https://doi.org/10.1016/j.redox.2018.08.005

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., & Gilad, Y. (2014). The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genetics*, *10*(3). https://doi.org/10.1371/journal.pgen.1004226

Davicioni, E., Finckenstein, F. G., Shahbazian, V., Buckley, J. D., Triche, T. J., & Anderson, M. J. (2006). Identification of a PAX-FKHR gene expression signature that defines molecular classes and determines the prognosis of alveolar rhabdomyosarcomas. *Cancer Research*, *66*(14). https://doi.org/10.1158/0008-5472.CAN-05-4578

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12). https://doi.org/10.1371/journal.pcbi.1001025

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology*, *11*(4). https://doi.org/10.1371/journal.pcbi.1004219

de Sainte Agathe, J. M., Filser, M., Isidor, B., Besnard, T., Gueguen, P., Perrin, A., Van Goethem, C., Verebi, C., Masingue, M., Rendu, J., Cossée, M., Bergougnoux, A., Frobert, L., Buratti, J., Lejeune, É., Le Guern, É., Pasquier, F., Clot, F., Kalatzis, V., … Baux, D. (2023). SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*, *17*(1). https://doi.org/10.1186/s40246-023-00451-1

De Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., Van Helden, P. D., Seoighe, C., & Hoal, E. G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, *128*(2). https://doi.org/10.1007/s00439-010-0836-1

Deshpande, O., Lara, R. Z., Zhang, O. R., Concepcion, D., & Hamilton, B. A. (2020). Znf423 patient variants, truncations, and in-frame deletions in mice define an allele-dependent range of midline brain abnormalities. *PLoS Genetics*, *16*(9 September). https://doi.org/10.1371/journal.pgen.1009017

Devuyst, O., Olinger, E., & Rampoldi, L. (2017). Uromodulin: From physiology to rare and complex kidney disorders. In *Nature Reviews Nephrology* (Vol. 13, Issue 9). https://doi.org/10.1038/nrneph.2017.101

Di Gregorio, J., Robuffo, I., Spalletta, S., Giambuzzi, G., De Iuliis, V., Toniato, E., Martinotti, S., Conti, P., & Flati, V. (2020). The Epithelial-to-Mesenchymal Transition as a Possible Therapeutic Target in Fibrotic Disorders. In *Frontiers in Cell and Developmental Biology* (Vol. 8). https://doi.org/10.3389/fcell.2020.607483

Di, Y., Chen, D., Yu, W., & Yan, L. (2019). Bladder cancer stage-associated hub genes revealed by WGCNA co-expression network analysis. *Hereditas*, *156*. https://doi.org/10.1186/s41065-019-0083-y

Dougan, J., Hawsawi, O., Burton, L. J., Edwards, G., Jones, K., Zou, J., Nagappan, P., Wang, G., Zhang, Q., Danaher, A., Bowen, N., Hinton, C., & Odero-Marah, V. A. (2019). Proteomics-metabolomics combined approach identifies peroxidasin as a protector against metabolic and oxidative stress in prostate cancer. *International Journal of Molecular Sciences*, *20*(12). https://doi.org/10.3390/ijms20123046

Dupont, S., & Wickström, S. A. (2022). Mechanical regulation of chromatin and transcription. In *Nature Reviews Genetics* (Vol. 23, Issue 10). https://doi.org/10.1038/s41576-022-00493-6

Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Beracochea, M., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., … Yates, A. D. (2025). Ensembl 2025. *Nucleic Acids Research*, *53*(D1), D948–D957. https://doi.org/10.1093/nar/gkae1071

Ezer, D., Zabet, N. R., & Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. In *Computational and Structural Biotechnology Journal* (Vol. 10, Issue 17). https://doi.org/10.1016/j.csbj.2014.07.005

Fabian, J., Gondwe, M., Mayindi, N., Chipungu, S., Khoza, B., Gaylard, P., Wade, A. N., Gómez-Olivé, F. X., Tomlinson, L. A., Ramsay, M., Tollman, S., Winkler, C., George, J. A., & Naicker, S. (2022a). Chronic kidney disease (CKD) and associated risk in rural South Africa: a population-based cohort study. *Wellcome Open Research*, *7*, 236. https://doi.org/10.12688/wellcomeopenres.18016.1

Fabian, J., Gondwe, M., Mayindi, N., Chipungu, S., Khoza, B., Gaylard, P., Wade, A. N., Gómez-Olivé, F. X., Tomlinson, L. A., Ramsay, M., Tollman, S., Winkler, C., George, J. A., & Naicker, S. (2022b). Chronic kidney disease (CKD) and associated risk in rural South Africa: A population-based cohort study. *Wellcome Open Research*, *7*. https://doi.org/10.12688/wellcomeopenres.18016.2

Fabian, J., Kalyesubula, R., Mkandawire, J., Hansen, C. H., Nitsch, D., Musenge, E., Nakanga, W. P., Prynn, J. E., Dreyer, G., Snyman, T., Ssebunnya, B., Ramsay, M., Smeeth, L., Tollman, S., Naicker, S., Crampin, A., Newton, R., George, J. A., Tomlinson, L., … Bird, N. (2022). Measurement of kidney function in Malawi, South Africa, and Uganda: a multicentre cohort study. *The Lancet Global Health*, *10*(8).

https://doi.org/10.1016/S2214-109X(22)00239-X

Falkson, S. R., & Bordoni, B. (2020). Anatomy, Abdomen and Pelvis, Bowman Capsule. In *StatPearls*

Fatumo, S., Chikowore, T., Kalyesubula, R., Nsubuga, R. N., Asiki, G., Nashiru, O., Seeley, J., Crampin, A. C., Nitsch, D., Smeeth, L., Kaleebu, P., Burgess, S., Nyirenda, M., Franceschini, N., Morris, A. P., Tomlinson, L., & Newton, R. (2021). Discovery and fine-mapping of kidney function loci in first genome-wide association study in Africans. *Human Molecular Genetics*, *30*(16). https://doi.org/10.1093/hmg/ddab088

Fidler, A. L., Vanacore, R. M., Chetyrkin, S. V., Pedchenko, V. K., Bhave, G., Yin, V. P., Stothers, C. L., Rose, K. L., McDonald, W. H., Clark, T. A., Borza, D. B., Steele, R. E., Ivy, M. T., Aspirnauts, T., Hudson, J. K., & Hudson, B. G. (2014). A unique covalent bond in basement membrane is a primordial innovation for tissue evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1). https://doi.org/10.1073/pnas.1318499111

Fine, D. M., Wasser, W. G., Estrella, M. M., Atta, M. G., Kuperman, M., Shemer, R., Rajasekaran, A., Tzur, S., Racusen, L. C., & Skorecki, K. (2012). APOL1 risk variants predict histopathology and progression to ESRD in HIV-related kidney disease. *Journal of the American Society of Nephrology*, *23*(2). https://doi.org/10.1681/ASN.2011060562

Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. A., Eisen, M. B., Bickel, P. J., Biggin, M. D., & Celniker, S. E. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(52). https://doi.org/10.1073/pnas.1209589110

Fogelgren, B., Yang, S., Sharp, I. C., Huckstep, O. J., Ma, W., Somponpun, S. J., Carlson, E. C., Uyehara, C. F. T., & Lozanoff, S. (2009). Deficiency in Six2 during prenatal development is associated with reduced nephron number, chronic renal failure, and hypertension in Br/+ adult mice. *American Journal of Physiology - Renal Physiology*, *296*(5). https://doi.org/10.1152/ajprenal.90550.2008

Freedman, B. I. (2013). APOL1 and nephropathy progression in populations of african ancestry. *Seminars in Nephrology*, *33*(5). https://doi.org/10.1016/j.semnephrol.2013.07.004

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8). https://doi.org/10.1101/gr.240663.118

Garcia-Garcia, G., & Jha, V. (2015). CKD in disadvantaged populations. In *Kidney International* (Vol. 87, Issue 2). https://doi.org/10.1038/ki.2014.369

Garrett, M. E., Soldano, K. L., Erwin, K. N., Zhang, Y., Gordeuk, V. R., Gladwin, M. T., Telen, M. J., & Ashley-Koch, A. E. (2023). Genome-wide meta-Analysis identifies new candidate genes for sickle cell disease nephropathy. *Blood Advances*, *7*(17). https://doi.org/10.1182/bloodadvances.2022007451

Gauderman, W. J. (2002). Sample size requirements for association studies of gene-gene interaction. *American Journal of Epidemiology*, *155*(5). https://doi.org/10.1093/aje/155.5.478

Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. In *Nature Reviews Genetics* (Vol. 21, Issue 5). https://doi.org/10.1038/s41576-019-0209-0

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., & Pollak, M. R. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, *329*(5993). https://doi.org/10.1126/science.1193032

George, C., Yako, Y. Y., Okpechi, I. G., Matsha, T. E., Kaze Folefack, F. J., & Kengne, A. P. (2018). An African perspective on the genetic risk of chronic kidney disease: A systematic review. *BMC Medical Genetics*, *19*(1). https://doi.org/10.1186/s12881-018-0702-x

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H. N., Jung, H., Nam, S., Chung, M., Kim, J. H., & Lee, I. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1). https://doi.org/10.1093/nar/gkx1013

Hanafusa, N., Nakai, S., Iseki, K., & Tsubakihara, Y. (2015). Japanese society for dialysis therapy renal data registry - A window through which we can view the details of Japanese dialysis population. *Kidney International Supplements*, *5*(1). https://doi.org/10.1038/kisup.2015.5

Hanmer, K. L., & Mavri-Damelin, D. (2018). Peroxidasin is a novel target of the redox-sensitive transcription factor Nrf2. *Gene*, *674*. https://doi.org/10.1016/j.gene.2018.06.076

Harb, O. A., Elfeky, M. A., El Shafaay, B. S., Taha, H. F., Osman, G., Harera, I. S., Gertallah, L. M., Abdelmonem, D. M., & Embaby, A. (2018). SPOP, ZEB-1 and E-cadherin expression in clear cell renal cell carcinoma (cc-RCC): Clinicopathological and prognostic significance. *Pathophysiology*, *25*(4). https://doi.org/10.1016/j.pathophys.2018.05.004

Hariparshad, S., Bhimma, R., Nandlal, L., Jembere, E., Naicker, S., & Assounga, A. (2023). The prevalence of chronic kidney disease in South Africa - limitations of studies comparing prevalence with sub-Saharan Africa, Africa, and globally. *BMC Nephrology*, *24*(1). https://doi.org/10.1186/s12882-023-03109-1

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-184

He, X., Riceberg, J., Soucy, T., Koenig, E., Minissale, J., Gallery, M., Bernard, H., Yang, X., Liao, H., Rabino, C., Shah, P., Xega, K., Yan, Z. H., Sintchak, M., Bradley, J., Xu, H., Duffey, M., England, D., Mizutani, H., … Pulukuri, S. M. (2017). Probing the roles of

SUMOylation in cancer cell biology by using a selective SAE inhibitor. *Nature Chemical Biology*, *13*(11). https://doi.org/10.1038/nchembio.2463

Heerspink, H. J. L., & Gansevoort, R. T. (2015). Albuminuria Is an appropriate therapeutic target in patients with CKD: The pro view. *Clinical Journal of the American Society of Nephrology*, *10*(6). https://doi.org/10.2215/CJN.11511114

Heinlein, C. A., & Chang, C. (2002). Androgen receptor (AR) coregulators: An overview. In *Endocrine Reviews* (Vol. 23, Issue 2). https://doi.org/10.1210/edrv.23.2.0460

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., & Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(4). https://doi.org/10.1073/pnas.1510805112

Hertel, K. J., Lynch, K. W., & Maniatis, T. (1997). Common themes in the function of transcription and splicing enhancers. *Current Opinion in Cell Biology*, *9*(3). https://doi.org/10.1016/S0955-0674(97)80007-5

Hildebrandt, F., Benzing, T., & Katsanis, N. (2011). Ciliopathies. *New England Journal of Medicine*, *364*(16). https://doi.org/10.1056/nejmra1010172

Hilgarth, R. S., Murphy, L. A., Skaggs, H. S., Wilkerson, D. C., Xing, H., & Sarge, K. D. (2004). Regulation and function of SUMO modification. In *Journal of Biological Chemistry* (Vol. 279, Issue 52). https://doi.org/10.1074/jbc.R400021200

Hinz, B. (2015). The extracellular matrix and transforming growth factor-β1: Tale of a strained relationship. In *Matrix Biology* (Vol. 47). https://doi.org/10.1016/j.matbio.2015.05.006

Hishikawa, A., Hayashi, K., & Itoh, H. (2018). Transcription factors as therapeutic targets in chronic kidney disease. In *Molecules* (Vol. 23, Issue 5). https://doi.org/10.3390/molecules23051123

Hoefield, R. A., Kalra, P. A., Baker, P., Lane, B., New, J. P., O'Donoghue, D. J., Foley, R. N., & Middleton, R. J. (2010). Factors associated with kidney disease progression and mortality in a referred CKD population. *American Journal of Kidney Diseases*, *56*(6). https://doi.org/10.1053/j.ajkd.2010.06.010

Huang, Y., Hong, W., & Wei, X. (2022). The molecular mechanisms and therapeutic strategies of EMT in tumor progression and metastasis. In *Journal of Hematology and Oncology* (Vol. 15, Issue 1). https://doi.org/10.1186/s13045-022-01347-8

Ingraham, L. L., & Meyer, D. L. (1985). Catalases and Peroxidases. In *Biochemistry of Dioxygen*. https://doi.org/10.1007/978-1-4613-2475-1_6

Inker, L. A., Astor, B. C., Fox, C. H., Isakova, T., Lash, J. P., Peralta, C. A., Kurella Tamura, M., & Feldman, H. I. (2014). KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *American Journal of Kidney Diseases*, *63*(5). https://doi.org/10.1053/j.ajkd.2014.01.416

Jankowski, J., Floege, J., Fliser, D., Böhm, M., & Marx, N. (2021). Cardiovascular Disease in

Chronic Kidney Disease Pathophysiological Insights and Therapeutic Options. In *Circulation* (Vol. 143, Issue 11). https://doi.org/10.1161/CIRCULATIONAHA.120.050686

Jayab, N. A., Abed, A., Talaat, I. M., & Hamoudi, R. (2024). The molecular mechanism of NF-κB dysregulation across different subtypes of renal cell carcinoma. *Journal of Advanced Research*. https://doi.org/10.1016/J.JARE.2024.07.030

Jefferis, J., Hudson, R., Lacaze, P., Bakshi, A., Hawley, C., Patel, C., & Mallett, A. (2024). Monogenic and polygenic concepts in chronic kidney disease (CKD). In *Journal of Nephrology* (Vol. 37, Issue 1). https://doi.org/10.1007/s40620-023-01804-8

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830). https://doi.org/10.1126/science.1141319

Joynt, A. T., Evans, T. A., Pellicore, M. J., Davis-Marcisak, E. F., Aksit, M. A., Eastman, A. C., Patel, S. U., Paul, K. C., Osorio, D. L., Bowling, A. D., Cotton, C. U., Raraigh, K. S., West, N. E., Merlo, C. A., Cutting, G. R., & Sharma, N. (2020). Evaluation of both exonic and intronic variants for effects on RNA splicing allows for accurate assessment of the effectiveness of precision therapies. *PLoS Genetics*, *16*(10). https://doi.org/10.1371/journal.pgen.1009100

Kalluri, R., & Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. In *Journal of Clinical Investigation* (Vol. 119, Issue 6). https://doi.org/10.1172/JCI39104

Kalyesubula, R., Wearne, N., Semitala, F. C., & Bowa, K. (2014). HIV-associated renal and genitourinary comorbidities in Africa. *Journal of Acquired Immune Deficiency Syndromes*, *67*(SUPPL.1). https://doi.org/10.1097/QAI.0000000000000259

Katz, I. (2005). Kidney and kidney related chronic diseases in South Africa and chronic disease intervention program experiences. *Advances in Chronic Kidney Disease*, *12*(1). https://doi.org/10.1053/j.ackd.2004.10.005

Kaze, A. D., Ilori, T., Jaar, B. G., & Echouffo-Tcheugui, J. B. (2018). Burden of chronic kidney disease on the African continent: A systematic review and meta-analysis. In *BMC Nephrology* (Vol. 19, Issue 1). https://doi.org/10.1186/s12882-018-0930-5

Khan, K., Rudkin, A., Parry, D. A., Burdon, K. P., McKibbin, M., Logan, C. V., Abdelhamed, Z. I. A., Muecke, J. S., Fernandez-Fuentes, N., Laurie, K. J., Shires, M., Fogarty, R., Carr, I. M., Poulter, J. A., Morgan, J. E., Mohamed, M. D., Jafri, H., Raashid, Y., Meng, N., … Ali, M. (2011). Homozygous mutations in PXDN cause congenital cataract, corneal opacity, and developmental glaucoma. *American Journal of Human Genetics*, *89*(3). https://doi.org/10.1016/j.ajhg.2011.08.005

Khoshnoodi, J., Pedchenko, V., & Hudson, B. G. (2008). Mammalian collagen IV. In *Microscopy Research and Technique* (Vol. 71, Issue 5). https://doi.org/10.1002/jemt.20564

Khoza, S., George, J. A., Naicker, P., Stoychev, S. H., Fabian, J., & Govender, I. S. (2024). Proteomic Analysis Identifies Dysregulated Proteins in Albuminuria: A South African

Pilot Study. *Biology*, *13*(9), 680. https://doi.org/10.3390/biology13090680

Kiberd, B. A., & Clase, C. M. (2002). Cumulative risk for developing end-stage renal disease in the US population. *Journal of the American Society of Nephrology*, *13*(6). https://doi.org/10.1097/01.ASN.0000014251.87778.01

Kido, T., Sikora-Wohlfeld, W., Kawashima, M., Kikuchi, S., Kamatani, N., Patwardhan, A., Chen, R., Sirota, M., Kodama, K., Hadley, D., & Butte, A. J. (2018). Are minor alleles more likely to be risk alleles? *BMC Medical Genomics*, *11*(1). https://doi.org/10.1186/s12920-018-0322-5

Kiffel, J., Rahimzada, Y., & Trachtman, H. (2011). Focal Segmental Glomerulosclerosis and Chronic Kidney Disease in Pediatric Patients. In *Advances in Chronic Kidney Disease* (Vol. 18, Issue 5). https://doi.org/10.1053/j.ackd.2011.03.005

Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. In *Experimental and Molecular Medicine* (Vol. 47, Issue 6). https://doi.org/10.1038/EMM.2015.33

Kintu, C., Soremekun, O., Machipisa, T., Mayanja, R., Kalyesubula, R., Bagaya, B. S., Jjingo, D., Chikowore, T., & Fatumo, S. (2023). Meta-analysis of African ancestry genome-wide association studies identified novel locus and validates multiple loci associated with kidney function. *BMC Genomics*, *24*(1). https://doi.org/10.1186/s12864-023-09601-0

Klahr, S., Levey, A. S., Beck, G. J., Caggiula, A. W., Hunsicker, L., Kusek, J. W., & Striker, G. (1994). The Effects of Dietary Protein Restriction and Blood-Pressure Control on the Progression of Chronic Renal Disease. *New England Journal of Medicine*, *330*(13). https://doi.org/10.1056/nejm199403313301301

Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., Neiman, P. E., & Lobanenkov, V. V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and Cellular Biology*, *13*(12). https://doi.org/10.1128/mcb.13.12.7612

Klinkhammer, B. M., Goldschmeding, R., Floege, J., & Boor, P. (2017). Treatment of Renal Fibrosis—Turning Challenges into Opportunities. In *Advances in Chronic Kidney Disease* (Vol. 24, Issue 2). https://doi.org/10.1053/j.ackd.2016.11.002

Kobayashi, A., Valerius, M. T., Mugford, J. W., Carroll, T. J., Self, M., Oliver, G., & McMahon, A. P. (2008). Six2 Defines and Regulates a Multipotent Self-Renewing Nephron Progenitor Population throughout Mammalian Kidney Development. *Cell Stem Cell*, *3*(2). https://doi.org/10.1016/j.stem.2008.05.020

Kopp, J. B., Smith, M. W., Nelson, G. W., Johnson, R. C., Freedman, B. I., Bowden, D. W., Oleksyk, T., McKenzie, L. M., Kajiyama, H., Ahuja, T. S., Berns, J. S., Briggs, W., Cho, M. E., Dart, R. A., Kimmel, P. L., Korbet, S. M., Michel, D. M., Mokrzycki, M. H., Schelling, J. R., … Winkler, C. A. (2008). MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nature Genetics*, *40*(10). https://doi.org/10.1038/ng.226

Kopp, J. B., Winkler, C. A., & Nelson, G. W. (2010). MYH9 genetic variants associated with glomerular disease: What is the role for genetic testing? *Seminars in Nephrology*, *30*(4). https://doi.org/10.1016/j.semnephrol.2010.06.007

Köttgen, A., Cornec-Le Gall, E., Halbritter, J., Kiryluk, K., Mallett, A. J., Parekh, R. S., Rasouly, H. M., Sampson, M. G., Tin, A., Antignac, C., Ars, E., Bergmann, C., Bleyer, A. J., Bockenhauer, D., Devuyst, O., Florez, J. C., Fowler, K. J., Franceschini, N., Fukagawa, M., … Gharavi, A. G. (2022). Genetics in chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney International*, *101*(6). https://doi.org/10.1016/j.kint.2022.03.019

Kulkarni, M. M., & Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, *130*(26). https://doi.org/10.1242/dev.00890

Lameire, N. H., Levin, A., Kellum, J. A., Cheung, M., Jadoul, M., Winkelmayer, W. C., Stevens, P. E., Caskey, F. J., Farmer, C. K. T., Ferreiro Fuentes, A., Fukagawa, M., Goldstein, S. L., Igiraneza, G., Kribben, A., Lerma, E. V., Levey, A. S., Liu, K. D., Małyszko, J., Ostermann, M., … Srisawat, N. (2021). Harmonizing acute and chronic kidney disease definition and classification: report of a Kidney Disease: Improving Global Outcomes (KDIGO) Consensus Conference. *Kidney International*, *100*(3). https://doi.org/10.1016/j.kint.2021.06.028

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. In *Cell* (Vol. 152, Issue 6). https://doi.org/10.1016/j.cell.2013.02.014

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J. M., Montgomery, S. B., & Griffith, O. L. (2016). ORegAnno 3.0: A community-driven resource for curated regulatory annotation. *Nucleic Acids Research*, *44*(D1). https://doi.org/10.1093/nar/gkv1203

Levenstien, M. A., & Klein, R. J. (2011). Predicting functionally important SNP classes based on negative selection. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-26

Levey, A. S., Becker, C., & Inker, L. A. (2015). Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: A systematic review. *JAMA - Journal of the American Medical Association*, *313*(8). https://doi.org/10.1001/jama.2015.0602

Levin, A., Hemmelgarn, B., Culleton, B., Tobe, S., McFarlane, P., Ruzicka, M., Burns, K., Manns, B., White, C., Madore, F., Moist, L., Klarenbach, S., Barrett, B., Foley, R., Jindal, K., Senior, P., Pannu, N., Shurraw, S., Akbari, A., … Tonelli, M. (2008). Guidelines for the management of chronic kidney disease. In *CMAJ. Canadian Medical Association Journal* (Vol. 179, Issue 11). https://doi.org/10.1503/cmaj.080351

Levitt, N. S., Steyn, K., Lambert, E. V., Reagon, G., Lombard, C. J., Fourie, J. M., Rossouw,

K., & Hoffman, M. (1999). Modifiable risk factors for Type 2 diabetes mellitus in a peri-urban community in South Africa. *Diabetic Medicine*, *16*(11). https://doi.org/10.1046/j.1464-5491.1999.00185.x

Lewis, E. J., Hunsicker, L. G., Bain, R. P., & Rohde, R. D. (1993). The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. The Collaborative Study Group. *New England Journal of Medicine*, *329*(20).

Li, C., Dong, X., Yuan, Q., Xu, G., Di, Z., Yang, Y., Hou, J., Zheng, L., Chen, W., & Wu, G. (2023). Identification of novel characteristic biomarkers and immune infiltration profile for the anaplastic thyroid cancer via machine learning algorithms. *Journal of Endocrinological Investigation*, *46*(8). https://doi.org/10.1007/s40618-023-02022-6

Li, H., Chen, D., & Zhang, J. (2012). Analysis of Intron Sequence Features Associated with Transcriptional Regulation in Human Genes. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0046784

Lin, Y. C., Boone, M., Meuris, L., Lemmens, I., van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R., Chen, J., Speleman, F., Lambrechts, D., van de Peer, Y., Tavernier, J., & Callewaert, N. (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms576

Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C. C., De Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., Carninci, P., Kasukawa, T., & Kawaji, H. (2019). Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1099

Lopez-Garcia, J., Periyasamy, M., Thomas, R. S., Christian, M., Leao, M., Jat, P., Kindle, K. B., Heery, D. M., Parker, M. G., Buluwela, L., Kamalati, T., & Ali, S. (2006). ZNF366 is an estrogen receptor corepressor that acts through CtBP and histone deacetylases. *Nucleic Acids Research*, *34*(21). https://doi.org/10.1093/nar/gkl875

Lu, H., Cassis, L. A., Kooi, C. W. V., & Daugherty, A. (2016). Structure and functions of angiotensinogen. In *Hypertension Research* (Vol. 39, Issue 7). https://doi.org/10.1038/hr.2016.17

Luscombe, N. M., Austin, S. E., Berman, H. M., & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biology*, *1*(1).

Luyckx, V. A., Tuttle, K. R., Garcia-Garcia, G., Gharbi, M. B., Heerspink, H. J. L., Johnson, D. W., Liu, Z. H., Massy, Z. A., Moe, O., Nelson, R. G., Sola, L., Wheeler, D. C., & White, S. L. (2017). Reducing major risk factors for chronic kidney disease. In *Kidney International Supplements* (Vol. 7, Issue 2). https://doi.org/10.1016/j.kisu.2017.07.003

Lv, D., Jia, F., Hou, Y., Sang, Y., Alvarez, A. A., Zhang, W., Gao, W. Q., Hu, B., Cheng, S. Y., Ge, J., Li, Y., & Feng, H. (2017). Histone acetyltransferase KAT6A upregulates PI3K/AKT signaling through TRIM24 binding. *Cancer Research*, *77*(22). https://doi.org/10.1158/0008-5472.CAN-17-1388

Lv, Z., Wang, M., Hou, H., Tang, G., Xu, H., Wang, X., Li, Y., Wang, J., & Liu, M. (2023). FOXMI-regulated ZIC2 promotes the malignant phenotype of renal clear cell carcinoma

by activating UBE2C/mTOR signaling pathway. *International Journal of Biological Sciences*, *19*(11). https://doi.org/10.7150/ijbs.84067

Machiela, M. J., & Chanock, S. J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, *31*(21). https://doi.org/10.1093/bioinformatics/btv402

Madhavan, S. M., & O'Toole, J. F. (2014). The biology of APOL1 with insights into the association between APOL1 variants and chronic kidney disease. In *Clinical and Experimental Nephrology* (Vol. 18, Issue 2). https://doi.org/10.1007/s10157-013-0907-4

Maestro, M. A., Cardalda, C., Boj, S. F., Luco, R. F., Servitja, J. M., & Ferrer, J. (2007). Distinct roles of HNF1beta, HNF1alpha, and HNF4alpha in regulating pancreas development, beta-cell function and growth. *Endocrine Development*, *12*. https://doi.org/10.1159/0000109603

Mahmoodi, B. K., Yatsuya, H., Matsushita, K., Sang, Y., Gottesman, R. F., Astor, B. C., Woodward, M., Longstreth, W. T., Psaty, B. M., Shlipak, M. G., Folsom, A. R., Gansevoort, R. T., & Coresh, J. (2014). Association of kidney disease measures with ischemic versus hemorrhagic strokes: Pooled analyses of 4 prospective community-based cohorts. *Stroke*, *45*(7). https://doi.org/10.1161/STROKEAHA.114.004900

Marable, S. S., Chung, E., Adam, M., Potter, S. S., & Park, J. S. (2018). Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome. *JCI Insight*, *3*(14). https://doi.org/10.1172/jci.insight.97497

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., & Califano, A. (2006). Reverse engineering cellular networks. *Nature Protocols*, *1*(2). https://doi.org/10.1038/nprot.2006.106

Martinez, M. F., Martini, A. G., Sequeira-Lopez, M. L. S., & Ariel Gomez, R. (2020). Ctcf is required for renin expression and maintenance of the structural integrity of the kidney. *Clinical Science*, *134*(13). https://doi.org/10.1042/CS20200184

Marutha, T., Williams, S., Novellie, M., Dillon, B., Carstens, N., & Mavri-Damelin, D. (2025). Exome sequencing identifies existing and novel variants in a South African cohort presenting with anterior segment dysgenesis. *Gene*, *943*, 149273. https://doi.org/10.1016/J.GENE.2025.149273

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. In *Annual Review of Genomics and Human Genetics* (Vol. 7). https://doi.org/10.1146/annurev.genom.7.080505.115623

Matsha, T. E., Kengne, A. P., Masconi, K. L., Yako, Y. Y., & Erasmus, R. T. (2015). APOL1 genetic variants, chronic kidney diseases and hypertension in mixed ancestry South Africans. *BMC Genetics*, *16*(1). https://doi.org/10.1186/s12863-015-0228-6

Matsha, T. E., Masconi, K., Yako, Y. Y., Hassan, M. S., Macharia, M., Erasmus, R. T., & Kengne, A. P. (2012). Polymorphisms in the Non-Muscle Myosin Heavy Chain Gene (MYH9) Are Associated with Lower Glomerular Filtration Rate in Mixed Ancestry Diabetic Subjects from South Africa. *PLoS ONE*, *7*(12).

https://doi.org/10.1371/journal.pone.0052529

Matsha, T. E., Yako, Y. Y., Rensburg, M. A., Hassan, M. S., Kengne, A. P., & Erasmus, R. T. (2013). Chronic kidney diseases in mixed ancestry south African populations: Prevalence, determinants and concordance between kidney function estimators. *BMC Nephrology*, *14*(1). https://doi.org/10.1186/1471-2369-14-75

Mauviel, A. (2005). Transforming growth factor-beta: a key mediator of fibrosis. In *Methods in molecular medicine* (Vol. 117).

Mayosi, B. M., Flisher, A. J., Lalloo, U. G., Sitas, F., Tollman, S. M., & Bradshaw, D. (2009). The burden of non-communicable diseases in South Africa. In *The Lancet* (Vol. 374, Issue 9693). https://doi.org/10.1016/S0140-6736(09)61087-4

McAnulty, R. J., Campa, J. S., Cambrey, A. D., & Laurent, G. J. (1991). The effect of transforming growth factor β on rates of procollagen synthesis and degradation in vitro. *BBA - Molecular Cell Research*, *1091*(2). https://doi.org/10.1016/0167-4889(91)90066-7

McCall, A. S., Cummings, C. F., Bhave, G., Vanacore, R., Page-Mccaw, A., & Hudson, B. G. (2014). Bromine is an essential trace element for assembly of collagen IV scaffolds in tissue development and architecture. *Cell*, *157*(6). https://doi.org/10.1016/j.cell.2014.05.009

Meng, X. M., Nikolic-Paterson, D. J., & Lan, H. Y. (2016). TGF-β: The master regulator of fibrosis. In *Nature Reviews Nephrology* (Vol. 12, Issue 6). https://doi.org/10.1038/nrneph.2016.48

Menon, M. C., Chuang, P. Y., Li, Z., Wei, C., Zhang, W., Luan, Y., Yi, Z., Xiong, H., Woytovich, C., Greene, I., Overbey, J., Rosales, I., Bagiella, E., Chen, R., Ma, M., Li, L., Ding, W., Djamali, A., Saminego, M., … Murphy, B. (2015). Intronic locus determines SHROOM3 expression and potentiates renal allograft fibrosis. *Journal of Clinical Investigation*, *125*(1). https://doi.org/10.1172/JCI76902

Mills, K. T., Bundy, J. D., Kelly, T. N., Reed, J., Kearney, P. M., Reynolds, K., Chen, J., & He, J. (2015). Abstract 16828: Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-based Studies From 90 Countries. *Circulation*, *132*(suppl_3). https://doi.org/10.1161/circ.132.suppl_3.16828

Miner, J. H. (2012). The glomerular basement membrane. *Experimental Cell Research*, *318*(9), 973–978. https://doi.org/10.1016/J.YEXCR.2012.02.031

Miyajima, A., Chen, J., Lawrence, C., Ledbetter, S., Soslow, R. A., Stern, J., Jha, S., Pigato, J., Lemer, M. L., Poppas, D. P., Vaughan, E. D., & Felsen, D. (2000). Antibody to transforming growth factor-β ameliorates tubular apoptosis in unilateral ureteral obstruction. *Kidney International*, *58*(6). https://doi.org/10.1046/j.1523-1755.2000.00414.x

Miyazaki, K., & Miyazaki, M. (2021). The Interplay Between Chromatin Architecture and Lineage-Specific Transcription Factors and the Regulation of Rag Gene Expression. In *Frontiers in Immunology* (Vol. 12). https://doi.org/10.3389/fimmu.2021.659761

Mohamed, S. A., Fernadez-Tajes, J., Franks, P. W., & Bennet, L. (2022). GWAS in people of

Middle Eastern descent reveals a locus protective of kidney function—a cross-sectional study. *BMC Medicine*, *20*(1). https://doi.org/10.1186/s12916-022-02267-7

Moman, R. N., Gupta, N., & Varacallo, M. (2023). Physiology, Albumin. In *StatPearls* (Vol. 8, Issue 8).

Moodley, N., Hariparshad, S., Peer, F., & Gounden, V. (2018). Evaluation of the CKD-EPI creatinine based glomerular filtration rate estimating equation in Black African and Indian adults in KwaZulu-Natal, South Africa. *Clinical Biochemistry*, *59*. https://doi.org/10.1016/j.clinbiochem.2018.06.014

Moosa, M. R., Van Der Walt, I., Naicker, S., & Meyers, A. M. (2015). Important causes of chronic kidney disease in South Africa. *South African Medical Journal*, *105*(4). https://doi.org/10.7196/SAMJ.9535

Moscona, R., Janssen, S. M., Elchebly, M., Papadakis, A. I., Rubin, E., & Spatz, A. (2023). BORIS/CTCFL-mediated chromatin accessibility alterations promote a pro-invasive transcriptional signature in melanoma cells. *Pigment Cell and Melanoma Research*, *36*(3–4). https://doi.org/10.1111/pcmr.13089

Mreich, E., Chen, X. M., Zaky, A., Pollock, C. A., & Saad, S. (2015). The role of Krüppel-like factor 4 in transforming growth factor-β-induced inflammatory and fibrotic responses in human proximal tubule cells. *Clinical and Experimental Pharmacology and Physiology*, *42*(6). https://doi.org/10.1111/1440-1681.12405

Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., & Lu, T. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. In *Cell Cycle* (Vol. 13, Issue 18). https://doi.org/10.4161/15384101.2014.949201

Murakami, K., Takemura, T., Hino, S., & Yoshioka, K. (1997). Urinary transforming growth factor-β in patients with glomerular diseases. *Pediatric Nephrology*, *11*(3). https://doi.org/10.1007/s004670050289

Navise, N. H., Mokwatsi, G. G., Gafane-Matemane, L. F., Fabian, J., & Lammertyn, L. (2023). Kidney dysfunction: prevalence and associated risk factors in a community-based study from the North West Province of South Africa. *BMC Nephrology*, *24*(1). https://doi.org/10.1186/s12882-023-03068-7

Niang, A., & Luyckx, V. A. (2022). Measuring kidney function: the voice of Africa. In *The Lancet Global Health* (Vol. 10, Issue 8). https://doi.org/10.1016/S2214-109X(22)00279-0

Nichols, B., Jog, P., Lee, J. H., Blackler, D., Wilmot, M., D'agati, V., Markowitz, G., Kopp, J. B., Alper, S. L., Pollak, M. R., & Friedman, D. J. (2015). Innate immunity pathways regulate the nephropathy gene Apolipoprotein L1. *Kidney International*, *87*(2). https://doi.org/10.1038/ki.2014.270

Nogueira, A., Pires, M. J., & Oliveira, P. A. (2017). Pathophysiological mechanisms of renal fibrosis: A review of animal models and therapeutic strategies. In *In Vivo* (Vol. 31, Issue 1). https://doi.org/10.21873/invivo.11019

Nørregaard, R., Mutsaers, H. A. M., Frøkiær, J., & Kwon, T. H. (2023). OBSTRUCTIVE

NEPHROPATHY AND MOLECULAR PATHOPHYSIOLOGY OF RENAL INTERSTITIAL FIBROSIS. In *Physiological Reviews* (Vol. 103, Issue 4). https://doi.org/10.1152/PHYSREV.00027.2022

Nqebelele, N. U., Dickens, C., Dix-Peek, T., Duarte, R., & Naicker, S. (2019). Low prevalence of apolipoprotein L1 gene variants in Black South Africans with hypertension-attributed chronic kidney disease. *Clinical Nephrology*, *91*(1). https://doi.org/10.5414/CN109432

Sanda, T., Lawton, L. N., Barrasa, M. I., Fan, Z. P., Kohlhammer, H., Gutierrez, A., Ma, W., Tatarek, J., Ahn, Y., Kelliher, M. A., Jamieson, C. H. M., Staudt, L. M., Young, R. A., & Look, A. T. (2012). Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T Cell Acute Lymphoblastic Leukemia. *Cancer Cell*, *22*(2). https://doi.org/10.1016/j.ccr.2012.06.007

Saran, R., Robinson, B., Abbott, K. C., Bragg-Gresham, J., Chen, X., Gipson, D., Gu, H., Hirth, R. A., Hutton, D., Jin, Y., Kapke, A., Kurtz, V., Li, Y., McCullough, K., Modi, Z., Morgenstern, H., Mukhopadhyay, P., Pearson, J., Pisoni, R., … Shahinian, V. (2020). US Renal Data System 2019 Annual Data Report: Epidemiology of Kidney Disease in the United States. In *American Journal of Kidney Diseases* (Vol. 75, Issue 1). https://doi.org/10.1053/j.ajkd.2019.09.003

Saritas, T., & Floege, J. (2020). Cardiovascular disease in patients with chronic kidney disease. *Herz*, *45*(2). https://doi.org/10.1007/s00059-019-04884-0

Schaefer, B., & Wühl, E. (2012). Educational paper: Progression in chronic kidney disease and prevention strategies. In *European Journal of Pediatrics* (Vol. 171, Issue 11). https://doi.org/10.1007/s00431-012-1814-5

Schiller, M., Javelaud, D., & Mauviel, A. (2004). TGF-β-induced SMAD signaling and gene regulation: Consequences for extracellular matrix remodeling and wound healing. In *Journal of Dermatological Science* (Vol. 35, Issue 2). https://doi.org/10.1016/j.jdermsci.2003.12.006

Schmidt, D., & Müller, S. (2003). PIAS/SUMO: New partners in transcriptional regulation. In *Cellular and Molecular Life Sciences* (Vol. 60, Issue 12). https://doi.org/10.1007/s00018-003-3129-1

Scholz, H., Boivin, F. J., Schmidt-Ott, K. M., Bachmann, S., Eckardt, K. U., Scholl, U. I., & Persson, P. B. (2021). Kidney physiology and susceptibility to acute kidney injury: implications for renoprotection. In *Nature Reviews Nephrology* (Vol. 17, Issue 5). https://doi.org/10.1038/s41581-021-00394-7

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S., Gómez-Olivé, F. X., Norris, S., Mashinya, F., Alberts, M., Hazelhurst, S., Schlebusch, C. M., & Ramsay, M. (2021). Genetic substructure and complex demographic history of South African Bantu speakers. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22207-y

Shaul, O. (2017). How introns enhance gene expression. *The International Journal of Biochemistry & Cell Biology*, *91*, 145–155. https://doi.org/10.1016/J.BIOCEL.2017.06.016

Shaw, G., Morse, S., Ararat, M., & Graham, F. L. (2002). Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *16*(8). https://doi.org/10.1096/fj.01-0995fje

Shi, R., Cao, Z., Li, H., Graw, J., Zhang, G., Thannickal, V. J., & Cheng, G. (2018). Peroxidasin contributes to lung host defense by direct binding and killing of gram-negative bacteria. *PLoS Pathogens*, *14*(5). https://doi.org/10.1371/journal.ppat.1007026

Shi, X., Young, C. D., Zhou, H., & Wang, X. (2020). Transforming growth factor-β signaling in fibrotic diseases and cancer-associated fibroblasts. In *Biomolecules* (Vol. 10, Issue 12). https://doi.org/10.3390/biom10121666

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. D. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8). https://doi.org/10.1101/gr.3715005

Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, *22*(14). https://doi.org/10.1093/bioinformatics/btl227

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. In *Cell* (Vol. 177, Issue 1). https://doi.org/10.1016/j.cell.2019.02.048

Sitole, B. N., & Mavri-Damelin, D. (2018). Peroxidasin is regulated by the epithelial-mesenchymal transition master transcription factor Snai1. *Gene*, *646*. https://doi.org/10.1016/j.gene.2018.01.011

Somma, M. P., Pisano, C., & Lavia, P. (1991). The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. *Nucleic Acids Research*, *19*(11). https://doi.org/10.1093/nar/19.11.2817

Song, A., Yan, R., Xiong, W., Xiang, H., Huang, J., Jiang, A., & Zhang, C. (2024). Early growth response protein 2 promotes partial epithelial-mesenchymal transition by phosphorylating Smad3 during renal fibrosis. *Translational Research*, *271*, 13–25. https://doi.org/10.1016/j.trsl.2024.04.005

Sönmezer, C., Kleinendorst, R., Imanci, D., Barzaghi, G., Villacorta, L., Schübeler, D., Benes, V., Molina, N., & Krebs, A. R. (2021). Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Molecular Cell*, *81*(2). https://doi.org/10.1016/j.molcel.2020.11.015

Soudi, M., Paumann-Page, M., Delporte, C., Pirker, K. F., Bellei, M., Edenhofer, E., Stadlmayr, G., Battistuzzi, G., Boudjeltia, K. Z., Furtmüller, P. G., Van Antwerpen, P., & Obinger, C. (2015). Multidomain human peroxidasin 1 is a highly glycosylated and stable homotrimeric high spin ferric peroxidase. *Journal of Biological Chemistry*,

290(17). https://doi.org/10.1074/jbc.M114.632273

Staden, R. (1994). Staden: searching for motifs in nucleic acid sequences. *Methods in Molecular Biology (Clifton, N.J.)*, *25*. https://doi.org/10.1385/0-89603-276-0:93

Stanifer, J. W., Jing, B., Tolan, S., Helmke, N., Mukerjee, R., Naicker, S., & Patel, U. (2014). The epidemiology of chronic kidney disease in sub-Saharan Africa: A systematic review and meta-analysis. *The Lancet Global Health*, *2*(3). https://doi.org/10.1016/S2214-109X(14)70002-6

Stevens, P. E., & Levin, A. (2013). Evaluation and management of chronic kidney disease: Synopsis of the kidney disease: Improving global outcomes 2012 clinical practice guideline. *Annals of Internal Medicine*, *158*(11). https://doi.org/10.7326/0003-4819-158-11-201306040-00007

Stewart, A. J., & Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, *192*(3). https://doi.org/10.1534/genetics.112.143370

Stojceski, F., Buetti-Dinh, A., Stoddart, M. J., Danani, A., Della Bella, E., & Grasso, G. (2023). Influence of dexamethasone on the interaction between glucocorticoid receptor and SOX9: A molecular dynamics study. *Journal of Molecular Graphics and Modelling*, *125*. https://doi.org/10.1016/j.jmgm.2023.108587

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *64*(3). https://doi.org/10.1111/1467-9868.00346

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, *31*(6). https://doi.org/10.1214/aos/1074290335

Sun, D., Xie, N., Wang, X., Wu, W., Li, X. Y., Chen, X., Qian, G., Li, C., Zhang, H., Jiang, Y., Ye, D., Liu, D., Hu, Y., Wang, J., Chen, W., Zhao, Q., Zeng, M., Zhang, J., Wang, L., & Zhang, X. (2021). Serum RelB is correlated with renal fibrosis and predicts chronic kidney disease progression. In *Clinical and Translational Medicine* (Vol. 11, Issue 5). https://doi.org/10.1002/ctm2.362

Sun, L., Liu, L., Jiang, J., Liu, K., Zhu, J., Wu, L., Lu, X., Huang, Z., Yuan, Y., Crowley, S. D., Mao, H., Xing, C., & Ren, J. (2024). Transcription factor Twist1 drives fibroblast activation to promote kidney fibrosis via signaling proteins Prrx1/TNC3. *Kidney International*. https://doi.org/10.1016/j.kint.2024.07.028

Sung, Y. J., Winkler, T. W., de las Fuentes, L., Bentley, A. R., Brown, M. R., Kraja, A. T., Schwander, K., Ntalla, I., Guo, X., Franceschini, N., Lu, Y., Cheng, C. Y., Sim, X., Vojinovic, D., Marten, J., Musani, S. K., Li, C., Feitosa, M. F., Kilpeläinen, T. O., … Chasman, D. I. (2018). A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. *American Journal of Human Genetics*, *102*(3). https://doi.org/10.1016/j.ajhg.2018.01.015

Taefehshokr, S., Key, Y. A., Khakpour, M., Dadebighlu, P., & Oveisi, A. (2017). Early growth response 2 and Egr3 are unique regulators in immune system. In *Central European Journal of Immunology* (Vol. 42, Issue 2). https://doi.org/10.5114/ceji.2017.69363

Takahashi, H., Lassmann, T., Murata, M., & Carninci, P. (2012). 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols*, *7*(3). https://doi.org/10.1038/nprot.2012.005

Tang, J., Yu, B., Li, Y., Zhang, W., Alvarez, A. A., Hu, B., Cheng, S., & Feng, H. (2019). TGF-β-activated lncRNA LINC00115 is a critical regulator of glioma stem-like cell tumorigenicity. *EMBO Reports*, *20*(12). https://doi.org/10.15252/embr.201948170

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*, *45*(W1). https://doi.org/10.1093/nar/gkx247

Tayo, B. O., Kramer, H., Salako, B. L., Gottesman, O., McKenzie, C. A., Ogunniyi, A., Bottinger, E. P., & Cooper, R. S. (2013). Genetic variation in APOL1 and MYH9 genes is associated with chronic kidney disease among Nigerians. *International Urology and Nephrology*, *45*(2). https://doi.org/10.1007/s11255-012-0263-4

Thomson, R., Genovese, G., Canon, C., Kovacsics, D., Higgins, M. K., Carrington, M., Winkler, C. A., Kopp, J., Rotimi, C., Adeyemo, A., Doumatey, A., Ayodo, G., Alper, S. L., Pollak, M. R., Friedman, D. J., & ARaper, J. (2014). Evolution of the primate trypanolytic factor APOL1. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20). https://doi.org/10.1073/pnas.1400699111

Tian, Y., Lv, Y., Wang, H., Che, J., Cui, F., Guo, J., Tian, W., Peng, J., Yang, B., Li, H., Zhou, B., Zhu, X., Cui, X., & Liu, L. (2024). Prenatal Phenotypic Analysis of Branchio-Oto-Renal Spectrum Disorder Attributable to EYA1 Gene Pathogenic Variants and Systematic Literature Review. *Prenatal Diagnosis*. https://doi.org/10.1002/pd.6673

Tindall, A. J., Pownall, M. E., Morris, I. D., & Isaacs, H. V. (2005). Xenopus tropicalis peroxidasin gene is expressed within the developing neural tube and pronephric kidney. *Developmental Dynamics*, *232*(2). https://doi.org/10.1002/dvdy.20226

Tishkoff, S. A., & Williams, S. M. (2002). Genetic analysis of African populations: Human evolution and complex disease. In *Nature Reviews Genetics* (Vol. 3, Issue 8). https://doi.org/10.1038/nrg865

Torrecilla, I., Oehler, J., & Ramadan, K. (2017). The role of ubiquitin-dependent segregase p97 (VCP or Cdc48) in chromatin dynamics after DNA double strand breaks. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 372, Issue 1731). https://doi.org/10.1098/rstb.2016.0282

Tran, N., Broun, A., & Ge, K. (2020). Lysine Demethylase KDM6A in Differentiation, Development, and Cancer. *Molecular and Cellular Biology*, *40*(20). https://doi.org/10.1128/mcb.00341-20

Trifonov, S., Yamashita, Y., Kase, M., Maruyama, M., & Sugimoto, T. (2016). Overview and assessment of the histochemical methods and reagents for the detection of β-galactosidase activity in transgenic animals. *Anatomical Science International*, *91*(1). https://doi.org/10.1007/s12565-015-0300-3

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., & Feolo, M. (2014). NCBI's database of genotypes and

phenotypes: DbGaP. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1211

Turner, J. M., Bauer, C., Abramowitz, M. K., Melamed, M. L., & Hostetter, T. H. (2012). Treatment of chronic kidney disease. In *Kidney International* (Vol. 81, Issue 4). https://doi.org/10.1038/ki.2011.380

Tzur, S., Rosset, S., Shemer, R., Yudkovsky, G., Selig, S., Tarekegn, A., Bekele, E., Bradman, N., Wasser, W. G., Behar, D. M., & Skorecki, K. (2010). Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene. *Human Genetics*, *128*(3). https://doi.org/10.1007/s00439-010-0861-0

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Ulasi, I. I., Tzur, S., Wasser, W. G., Shemer, R., Kruzel, E., Feigin, E., Ijoma, C. K., Onodugo, O. D., Okoye, J. U., Arodiwe, E. B., Ifebunandu, N. A., Chukwuka, C. J., Onyedum, C. C., Ijoma, U. N., Nna, E., Onuigbo, M., Rosset, S., & Skorecki, K. (2013). High population frequencies of apol1 risk variants are associated with increased prevalence of non-diabetic chronic kidney disease in the igbo people from south-eastern nigeria. *Nephron - Clinical Practice*, *123*(1–2). https://doi.org/10.1159000353223

Van Agtmael, T., Schlötzer-Schrehardt, U., McKie, L., Brownstein, D. G., Lee, A. W., Cross, S. H., Sado, Y., Mullins, J. J., Pöschl, E., & Jackson, I. J. (2005). Dominant mutations of Col4a1 result in basement membrane defects which lead to anterior segment dysgenesis and glomerulopathy. *Human Molecular Genetics*, *14*(21). https://doi.org/10.1093/hmg/ddi348

van der Valk, T., de Manuel, M., Marques-Bonet, T., & Guschanski, K. (2021). Estimates of genetic load in small populations suggest extensive purging of deleterious alleles. In *bioRxiv*.

Van Velde, M. Der, Halbesma, N., De Charro, F. T., Bakker, S. J. L., De Zeeuw, D., De Jong, P. E., & Gansevoort, R. T. (2009). Screening for albuminuria identifies individuals at increased renal risk. *Journal of the American Society of Nephrology*, *20*(4). https://doi.org/10.1681/ASN.2008060655

Vanacore, R., Ham, A. J. L., Voehler, M., Sanders, C. R., Conrads, T. P., Veenstra, T. D., Sharpless, B., Dawson, P. E., & Hudson, B. G. (2009). A Sulfilimine Bond Identified in Collagen IV. *Science*, *325*(5945). https://doi.org/10.1126/science.1176811

Vandewalle, C., Van Roy, F., & Berx, G. (2009). The role of the ZEB family of transcription factors in development and disease. In *Cellular and Molecular Life Sciences* (Vol. 66, Issue 5). https://doi.org/10.1007/s00018-008-8465-8

Vashee, S., Melchert, K., Ding, W. V., Johnstont, S. A., & Kodadek, T. (1998). Evidence for two modes of cooperative DNA binding in vivo that do not involve, direct protein-protein interactions. *Current Biology*, *8*(8). https://doi.org/10.1016/S0960-

9822(98)70179-4

Vicente-Manzanares, M., Ma, X., Adelstein, R. S., & Horwitz, A. R. (2009). Non-muscle myosin II takes centre stage in cell adhesion and migration. In *Nature Reviews Molecular Cell Biology* (Vol. 10, Issue 11). https://doi.org/10.1038/nrm2786

Vihinen, M. (2021). Functional effects of protein variants. *Biochimie*, *180*, 104–120. https://doi.org/10.1016/J.BIOCHI.2020.10.009

Vilar, J. M. G., & Saiz, L. (2005). DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. In *Current Opinion in Genetics and Development* (Vol. 15, Issue 2). https://doi.org/10.1016/j.gde.2005.02.005

Vivante, A. (2024). Genetics of Chronic Kidney Disease. *New England Journal of Medicine*, *391*(7), 627–639. https://doi.org/10.1056/nejmra2308577

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, *15*(10). https://doi.org/10.1093/bioinformatics/15.10.776

Wallace, M. A. (1998). Anatomy and physiology of the kidney. In *AORN journal* (Vol. 68, Issue 5). https://doi.org/10.1016/s0001-2092(06)62377-6

Wang, F., Zhang, R., Jian, J., Sun, Y., & Li, Q. (2024). Identification and Functional Study of Enhancers of EYA1: The Causative Gene of Branchio-Oto-Renal Syndrome. *Developmental Neuroscience*. https://doi.org/10.1159/000536260

Warming, S., Rachel, R. A., Jenkins, N. A., & Copeland, N. G. (2006). Zfp423 Is Required for Normal Cerebellar Development . *Molecular and Cellular Biology*, *26*(18). https://doi.org/10.1128/mcb.02255-05

Webster, A. C., Nagler, E. V, Morton, R. L., & Masson, P. (2017). Chronic Kidney Disease. *The Lancet*, *389*(10075), 1238–1252. https://doi.org/10.1016/S0140-6736(16)32064-5

Weckerle, A., Snipes, J. A., Cheng, D., Gebre, A. K., Reisz, J. A., Murea, M., Shelness, G. S., Hawkins, G. A., Furdui, C. M., Freedman, B. I., Parks, J. S., & Ma, L. (2016). Characterization of circulating APOL1 protein complexes in African Americans. *Journal of Lipid Research*, *57*(1). https://doi.org/10.1194/jlr.M063453

Wei, X., Zhu, X., Jiang, L., Huang, X., Zhang, Y., Zhao, D., & Du, Y. (2020). Recent advances in understanding the role of hypoxia-inducible factor 1α in renal fibrosis. In *International Urology and Nephrology* (Vol. 52, Issue 7). https://doi.org/10.1007/s11255-020-02474-2

Weidemüller, P., Kholmatov, M., Petsalaki, E., & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. In *Proteomics* (Vol. 21, Issues 23–24). https://doi.org/10.1002/pmic.202000034

Weiner, D. E., Tabatabai, S., Tighiouart, H., Elsayed, E., Bansal, N., Griffith, J., Salem, D. N., Levey, A. S., & Sarnak, M. J. (2006). Cardiovascular Outcomes and All-Cause Mortality: Exploring the Interaction Between CKD and Cardiovascular Disease. *American Journal of Kidney Diseases*, *48*(3). https://doi.org/10.1053/j.ajkd.2006.05.021

Weingarten-Gabbay, S., & Segal, E. (2014). The grammar of transcriptional regulation. In *Human Genetics* (Vol. 133, Issue 6). https://doi.org/10.1007/s00439-013-1413-1

Wells, J., & Farnham, P. J. (2002). Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. *Methods*, *26*(1). https://doi.org/10.1016/S1046-2023(02)00007-5

Wen, Y., Lu, X., Ren, J., Privratsky, J. R., Yang, B., Rudemiller, N. P., Zhang, J., Griffiths, R., Jain, M. K., Nedospasov, S. A., Liu, B. C., & Crowley, S. D. (2019). KLF4 in macrophages attenuates TNFa-mediated kidney injury and fibrosis. *Journal of the American Society of Nephrology*, *30*(10). https://doi.org/10.1681/ASN.2019020111

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., & Freund, C. (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. In *Frontiers in Immunology* (Vol. 8, Issue MAR). https://doi.org/10.3389/fimmu.2017.00292

Wiesel-Motiuk, N., & Assaraf, Y. G. (2020). The key roles of the lysine acetyltransferases KAT6A and KAT6B in physiology and pathology. In *Drug Resistance Updates* (Vol. 53). https://doi.org/10.1016/j.drup.2020.100729

World Health Organization. (2020). Meeting report of the WHO expert consultation on drug-resistant tuberculosis treatment outcome definitions, 17-19 November 2020. In *World Health Organization* (Issue October).

Wuttke, M., & Köttgen, A. (2016). Insights into kidney diseases from genome-wide association studies. In *Nature Reviews Nephrology* (Vol. 12, Issue 9). https://doi.org/10.1038/nrneph.2016.107

Wyllie, K., Panagopoulos, V., & Cox, T. R. (2023). The role of peroxidasin in solid cancer progression. In *Biochemical Society Transactions* (Vol. 51, Issue 5). https://doi.org/10.1042/BST20230018

Xiao, X., Tang, W., Yuan, Q., Peng, L., & Yu, P. (2015). Epigenetic repression of Krüppel-like factor 4 through Dnmt1 contributes to EMT in renal fibrosis. *International Journal of Molecular Medicine*, *35*(6). https://doi.org/10.3892/ijmm.2015.2189

Yamamoto, T., Noble, N. A., Cohen, A. H., Nast, C. C., Hishida, A., Gold, L. I., & Border, W. A. (1996). Expression of transforming growth factor-β isoforms in human glomerular diseases. *Kidney International*, *49*(2). https://doi.org/10.1038/ki.1996.65

Yamashita, N., & Kramann, R. (2024). Mechanisms of kidney fibrosis and routes towards therapy. In *Trends in Endocrinology and Metabolism* (Vol. 35, Issue 1). https://doi.org/10.1016/j.tem.2023.09.001

Yan, X., Sabrautzki, S., Horsch, M., Fuchs, H., Gailus-Durner, V., Beckers, J., de Angelis, M. H., & Graw, J. (2014). Peroxidasin is essential for eye development in the mouse. *Human Molecular Genetics*, *23*(21). https://doi.org/10.1093/hmg/ddu274

Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A., & Weinberg, R. A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell*, *117*(7).

https://doi.org/10.1016/j.cell.2004.06.006

Yang, L., Bai, Y., Li, N., Hu, C., Peng, J., Cheng, G., Zhang, G., & Shi, R. (2013). Vascular VPO1 expression is related to the endothelial dysfunction in spontaneously hypertensive rats. *Biochemical and Biophysical Research Communications*, *439*(4). https://doi.org/10.1016/j.bbrc.2013.09.012

Yang, Z., Wang, Y. X., Wen, J. K., Gao, H. T., Han, Z. W., Qi, J. C., Gu, J. F., Zhao, C. M., Zhang, H., Shi, B., Wang, D. D., Wang, X. L., & Qu, C. B. (2023). SF3B4 promotes Twist1 expression and clear cell renal cell carcinoma progression by facilitating the export of KLF 16 mRNA from the nucleus to the cytoplasm. *Cell Death and Disease*, *14*(1). https://doi.org/10.1038/s41419-022-05534-w

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, *31*(14). https://doi.org/10.1093/bioinformatics/btv145

Yuan, H., Kajiyama, H., Ito, S., Chen, D., Shibata, K., Hamaguchi, M., Kikkawa, F., & Senga, T. (2015). HOXB13 and ALX4 induce SLUG expression for the promotion of EMT and cell invasion in ovarian cancer cells. *Oncotarget*, *6*(15). https://doi.org/10.18632/oncotarget.3673

Yuan, Q., Tang, B., & Zhang, C. (2022). Signaling pathways of chronic kidney diseases, implications for therapeutics. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). https://doi.org/10.1038/s41392-022-01036-5

Zabidi, M. A., & Stark, A. (2016). Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. In *Trends in Genetics* (Vol. 32, Issue 12). https://doi.org/10.1016/j.tig.2016.10.003

Zanoli, L., Lentini, P., Briet, M., Castellino, P., House, A. A., London, G. M., Malatino, L., McCullough, P. A., Mikhailidis, D. P., & Boutouyrie, P. (2019). Arterial stiffness in the heart disease of CKD. *Journal of the American Society of Nephrology*, *30*(6). https://doi.org/10.1681/asn.2019020117

Zederbauer, M., Furtmüller, P. G., Brogioni, S., Jakopitsch, C., Smulevich, G., & Obinger, C. (2007). Heme to protein linkages in mammalian peroxidases: Impact on spectroscopic, redox and catalytic properties. In *Natural Product Reports* (Vol. 24, Issue 3). https://doi.org/10.1039/b604178g

Zhang, H., Gao, J., Wang, H., Liu, M., Lu, S., Xu, H., Tang, W., & Zheng, G. (2024). Novel likely pathogenic variant in the EYA1 gene causing Branchio oto renal syndrome and the exploration of pathogenic mechanisms. *BMC Medical Genomics*, *17*(1). https://doi.org/10.1186/s12920-024-01858-y

Zhang, H., & Sun, S. C. (2015). NF-KB in inflammation and renal diseases. In *Cell and Bioscience* (Vol. 5, Issue 1). https://doi.org/10.1186/s13578-015-0056-4

Zhang, L., Zhai, S. bo, Zhao, L. yue, Zhang, Y., Sun, B. chao, & Ma, Q. shan. (2018). New PAX2 heterozygous mutation in a child with chronic kidney disease: a case report and review of the literature. *BMC Nephrology*, *19*(1). https://doi.org/10.1186/s12882-018-1044-9

Zhang, P., Sun, Y., & Ma, L. (2015). ZEB1: At the crossroads of epithelial-mesenchymal transition, metastasis and therapy resistance. In *Cell Cycle* (Vol. 14, Issue 4). https://doi.org/10.1080/15384101.2015.1006048

Zhang, S., Hong, Z., Chai, Y., Liu, Z., Du, Y., Li, Q., & Liu, Q. (2017). CSN5 promotes renal cell carcinoma metastasis and EMT by inhibiting ZEB1 degradation. *Biochemical and Biophysical Research Communications*, *488*(1), 101–108. https://doi.org/10.1016/J.BBRC.2017.05.016

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9). https://doi.org/10.1186/gb-2008-9-9-r137

Zheng, Y. Z., & Liang, L. (2018). High expression of PXDN is associated with poor prognosis and promotes proliferation, invasion as well as migration in ovarian cancer. *Annals of Diagnostic Pathology*, *34*. https://doi.org/10.1016/j.anndiagpath.2018.03.002

Zhong, M., Zhu, E., Li, N., Gong, L., Xu, H., Zhong, Y., Gong, K., Jiang, S., Wang, X., Fei, L., Tang, C., Lei, Y., Wang, Z., & Zheng, Z. (2023). Identification of diagnostic markers related to oxidative stress and inflammatory response in diabetic kidney disease by machine learning algorithms: Evidence from human transcriptomic data and mouse experiments. *Frontiers in Endocrinology*, *14*. https://doi.org/10.3389/fendo.2023.1134325

Zhou, X., Sun, Q., Xu, C., Zhou, Z., Chen, X., Zhu, X., Huang, Z., Wang, W., & Shi, Y. (2022). A systematic pan-cancer analysis of PXDN as a potential target for clinical diagnosis and treatment. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.952849

Zhu, A. Y., Costain, G., Cytrynbaum, C., Weksberg, R., Cohn, R. D., & Ali, A. (2021). Novel heterozygous variants in PXDN cause different anterior segment dysgenesis phenotypes in monozygotic twins. *Ophthalmic Genetics*, *42*(5). https://doi.org/10.1080/13816810.2021.1925929

Zivotic, M., Dundjerovic, D., Naumovic, R., Kovacevic, S., Ivanov, M., Karanovic, D., Nikolic, G., Markovic-Lipkovski, J., Radojevic Skodric, S., & Nesovic Ostojic, J. (2022). Clinicopathological Relevance of PAX8 Expression Patterns in Acute Kidney Injury and Chronic Kidney Diseases. *Diagnostics*, *12*(9). https://doi.org/10.3390/diagnostics12092036

Zoccali, C., Vanholder, R., Massy, Z. A., Ortiz, A., Sarafidis, P., Dekker, F. W., Fliser, D., Fouque, D., Heine, G. H., Jager, K. J., Kanbay, M., Mallamaci, F., Parati, G., Rossignol, P., Wiecek, A., & London, G. (2017). The systemic nature of CKD. In *Nature Reviews Nephrology* (Vol. 13, Issue 6). https://doi.org/10.1038/nrneph.2017.52

Zou, Z., Ohta, T., Miura, F., & Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Research*, *50*(W1). https://doi.org/10.1093/nar/gkac199

APPENDICES

Appendix A

The GFR equations are as follows:

- MDRD: GFR = (ml/min/1.73 m^2) = 175 × {[plasma creatinine (μmol/l) / 88.4] $^{-1.154}$} × age (years)$^{-0.203}$ × 0.742 (if female) × 1.212 (if black))
- CKD-EPI: GFR = 141 × min(serum creatinine/k, 1)α× max(serum creatinine / k, 1) $^{-1.209}$ × 0.993^{Age} × 1.018 (if female) × 1.159 (if black), k is 0.7 for females and 0.9 for males, α is −0.329 for females and −0.411 for males
- Cockcroft-Gault: GFR = (140 – age) x bodyweight/plasma creatinine x 72 (x 0.85 if female)

Appendix B

Table 1 describes the TFs identified in ChIP-Atlas to be associated with PXDN and their potential implication in kidney function or disease pathogenesis with supporting literature is noted in the final column. Renal fibrosis affected kidney development and EMT are among the potential kidney disease processes mentioned. *ARNT*, *CREB1* and *TWIST1*, identified in Figures 1 and 2 of Appendix B, were added to the main analysis due to their association to disease pathways in the kidney.

Table 1: TFs identified using ChIP-Atlas and their implications in kidney disease.

TF	Potential implication in kidney disease pathogenesis	Source
ARNT	ARNT attenuates chronic kidney failure.	Tampe, B., Tampe, D., Nyamsuren, G., Klöpper, F., Rapp, G., Kauffels, A., Lorf, T., Zeisberg, E. M., Müller, G. A., Kalluri, R., Hakroush, S., & Zeisberg, M. (2018). Pharmacological induction of hypoxia-inducible transcription factor ARNT attenuates chronic kidney failure. *Journal of Clinical Investigation*, *128*(7). https://doi.org/10.1172/JCI89632
CREB1	Increased phosphorylated CREB1 protein correlates with poor prognosis in clear cell renal cell carcinoma	Zhang, Z., Guan, B., Li, Y., He, Q., Li, X., & Zhou, L. (2021). Increased phosphorylated CREB1 protein correlates with poor prognosis in clear cell renal cell carcinoma. *Translational Andrology and Urology*, *10*(8). https://doi.org/10.21037/tau-21-371
CTCF	CTCF is required for renin expression and maintenance of the structural integrity of the kidney and podocyte-specific deletion of CTCF drives progressive kidney disease.	Christov, M., Clark, A. R., Corbin, B., Hakroush, S., Rhee, E. P., Saito, H., Brooks, D., Hesse, E., Bouxsein, M., Galjart, N., Jung, J. Y., Mundel, P., Jüppner, H., Weins, A., & Greka, A. (2018). Inducible podocyte-specific deletion of CTCF drives progressive kidney disease and bone abnormalities. JCI Insight, 3(4). https://doi.org/10.1172/jci.insight.95091 Martinez, M. F., Martini, A. G., Sequeira-Lopez, M. L. S., & Ariel Gomez, R. (2020). Ctcf is required for renin expression and maintenance of the structural integrity of the kidney. Clinical Science, 134(13). https://doi.org/10.1042/CS20200184
EGR2	EGR2 promotes partial EMT in renal fibrosis.	Song, A., Yan, R., Xiong, W., Xiang, H., Huang, J., Jiang, A., & Zhang, C. (2024). Early growth response protein 2 promotes partial epithelial-mesenchymal transition by phosphorylating Smad3 during renal fibrosis. *Translational Research*, *271*, 13–25. https://doi.org/10.1016/j.trsl.2024.04.005
EP300	EP300 promotes renal tubular epithelial cell fibrosis in diabetic nephropathy.	Gong, Y., Dou, Y., Wang, L., Wang, X., & Zhao, Z. (2022). EP300 promotes renal tubular epithelial cell fibrosis by increasing HIF2α expression in diabetic nephropathy. *Cellular Signalling*, *98*. https://doi.org/10.1016/j.cellsig.2022.110407
KAT6A	KAT6A plays a beneficial role in kidney renal clear cell carcinoma.	Liang, F., Li, X., Shen, X., Yang, R., & Chen, C. (2023). Expression profiles and functional prediction of histone acetyltransferases of the MYST family in kidney renal clear cell carcinoma. *BMC Cancer*, *23*(1). https://doi.org/10.1186/s12885-023-11076-x
KAT6B	KAT6 plays a beneficial role in kidney renal clear cell carcinoma.	Liang, F., Li, X., Shen, X., Yang, R., & Chen, C. (2023). Expression profiles and functional prediction of histone acetyltransferases of the MYST family in kidney renal clear cell carcinoma. *BMC Cancer*, *23*(1). https://doi.org/10.1186/s12885-023-11076-x
KDM6A	KDM6A has been proposed as a tumour suppressor in kidney renal papillary cell carcinoma.	Chang, S., Yim, S., & Park, H. (2019). The cancer driver genes IDH1/2, JARID1C/ KDM5C, and UTX/ KDM6A: crosstalk between histone demethylation and

		hypoxic reprogramming in cancer metabolism. In *Experimental and Molecular Medicine* (Vol. 51, Issue 6). https://doi.org/10.1038/s12276-019-0230-6
MEF2B	MEF2B overexpression promotes HEK293A cell migration and EMT.	Pon, J. R., Wong, J., Saberi, S., Alder, O., Moksa, M., Grace Cheng, S. W., Morin, G. B., Hoodless, P. A., Hirst, M., & Marra, M. A. (2015). MEF2B mutations in non-Hodgkin lymphoma dysregulate cell migration by decreasing MEF2B target gene activation. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms8953
NR3C1	NR3C1 knockdown inhibits clear cell renal cell carcinoma proliferation and migration.	Yan, M., Wang, J., Wang, H., Zhou, J., Qi, H., Naji, Y., Zhao, L., Tang, Y., & Dai, Y. (2023). Knockdown of NR3C1 inhibits the proliferation and migration of clear cell renal cell carcinoma through activating endoplasmic reticulum stress–mitophagy. *Journal of Translational Medicine*, *21*(1). https://doi.org/10.1186/s12967-023-04560-2
OSR2	OSR2 is expressed at sites of EMT during kidney development.	Lan, Y., Kingsley, P. D., Cho, E. S., & Jiang, R. (2001). Osr2, a new mouse gene related to Drosophila odd-skipped, exhibits dynamic expression patterns during craniofacial, limb, and kidney development. *Mechanisms of Development*, *107*(1–2). https://doi.org/10.1016/S0925-4773(01)00457-9
RELA	In tubular epithelial cells activated NF-kappaB and pro-inflammatory chemokines are markers of progressive diabetic nephropathy.	Mezzano, S., Aros, C., Droguett, A., Burgos, M. E., Ardiles, L., Flores, C., Schneider, H., Ruiz-Ortega, M., & Egido, J. (2004). NF-κB activation and overexpression of regulated genes in human diabetic nephropathy. *Nephrology Dialysis Transplantation*, *19*(10). https://doi.org/10.1093/ndt/gfh207
SIX2	SIX2 mutations are linked to anomalous kidney development.	Weber, S., Taylor, J. C., Winyard, P., Baker, K. F., Sullivan-Brown, J., Schild, R., Knüppel, T., Zurowska, A. M., Caldas-Alfonso, A., Litwin, M., Emre, S., Ghiggeri, G. M., Bakkaloglu, A., Mehls, O., Antignac, C., Schaefer, F., & Burdine, R. D. (2008). SIX2 and BMP4 mutations associate with anomalous kidney development. *Journal of the American Society of Nephrology*, *19*(5). https://doi.org/10.1681/ASN.2006111282
SMAD4	Smad4 promotes diabetic nephropathy.	Li, J., Sun, Y. B. Y., Chen, W., Fan, J., Li, S., Qu, X., Chen, Q., Chen, R., Zhu, D., Zhang, J., Wu, Z., Chi, H., Crawford, S., Oorschot, V., Puelles, V. G., Kerr, P. G., Ren, Y., Nilsson, S. K., Christian, M., … Yu, X. (2020). Smad4 promotes diabetic nephropathy by modulating glycolysis and OXPHOS . *EMBO Reports*, *21*(2). https://doi.org/10.15252/embr.201948781
SMARCA4	SMARCA4 deficiency induces oncogene upregulation and hyperproliferation of tubular and interstitial cells during kidney development.	Xu, J., Zhou, X., Zhang, T., Zhang, B., & Xu, P. X. (2023). Smarca4 deficiency induces Pttg1 oncogene upregulation and hyperproliferation of tubular and interstitial cells during kidney development. *Frontiers in Cell and Developmental Biology*, *11*. https://doi.org/10.3389/fcell.2023.1233317
SOX17	SOX17 plays a role in kidney development with mutations observed in patients with congenital anomalies of the kidney and urinary tract.	Gimelli, S., Caridi, G., Beri, S., McCracken, K., Bocciardi, R., Zordan, P., Dagnino, M., Fiorio, P., Murer, L., Benetti, E., Zuffardi, O., Giorda, R., Wells, J. M., Gimelli, G., & Ghiggeri, G. M. (2010). Mutations in SOX17 are associated with congenital anomalies of the kidney and the urinary tract. *Human Mutation*, *31*(12). https://doi.org/10.1002/humu.21378
TWIST1	TWIST1 promotes fibroblast activation inducing kidney fibrosis.	Sun, L., Liu, L., Jiang, J., Liu, K., Zhu, J., Wu, L., Lu, X., Huang, Z., Yuan, Y., Crowley, S. D., Mao, H., Xing, C., & Ren, J. (2024). Transcription factor Twist1 drives fibroblast activation to promote kidney fibrosis via signaling proteins Prrx1/TNC3. *Kidney International*. https://doi.org/10.1016/j.kint.2024.07.028
ZEB1	ZEB1 represses E-cadherin and induces EMT.	Sánchez-Tilló, E., Lázaro, A., Torrent, R., Cuatrecasas, M., Vaquero, E. C., Castells, A., Engel, P., & Postigo, A. (2010). ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene*, *29*(24). https://doi.org/10.1038/onc.2010.102
ZEB2	ZEB2 knockout leads to kidney fibrosis.	Kumar, S., Fan, X., Rasouly, H. M., Sharma, R., Salant, D. J., & Lu, W. (2023). ZEB2 controls kidney stromal

		progenitor differentiation and inhibits abnormal myofibroblast expansion and kidney fibrosis. *JCI Insight*, *8*(1). https://doi.org/10.1172/jci.insight.158418
ZHX2	ZHX2 plays an oncogenic driver role in clear cell renal cell carcinoma.	Zhang, J., Wu, T., Simon, J., Takada, M., Saito, R., Fan, C., Liu, X. de, Jonasch, E., Xie, L., Chen, X., Yao, X., Teh, B. T., Tan, P., Zheng, X., Li, M., Lawrence, C., Fan, J., Geng, J., Liu, X., … Zhang, Q. (2018). VHL substrate transcription factor ZHX2 as an oncogenic driver in clear cell renal cell carcinoma. *Science*, *361*(6399). https://doi.org/10.1126/science.aap8411
ZIC2	ZIC2 promotes malignant phenotype of renal clear cell carcinoma.	Lv, Z., Wang, M., Hou, H., Tang, G., Xu, H., Wang, X., Li, Y., Wang, J., & Liu, M. (2023). FOXMI-regulated ZIC2 promotes the malignant phenotype of renal clear cell carcinoma by activating UBE2C/mTOR signaling pathway. *International Journal of Biological Sciences*, *19*(11). https://doi.org/10.7150/ijbs.84067

Appendix C

Figures 1 and 2 illustrate the number of studies in which each TF was identified in using a MACS score threshold of 50. Additionally, whether MACS scores were above 500 were shown for the promoter and intron 1 region, respectively. For both the promoter and first intronic regions, the majority of MACS scores were less than 500. For the promoter region, *CTCF* was identified in the greatest number of studies (Figure 1). Conversely, eight TFs were identified in only one experiment in the promoter region. For the first intronic region, *AR, NR3C1, RELA* and *SMARCA4* were identified in the greatest number of experiments, with 12 TFs identified in only one experiment (Figure 2).

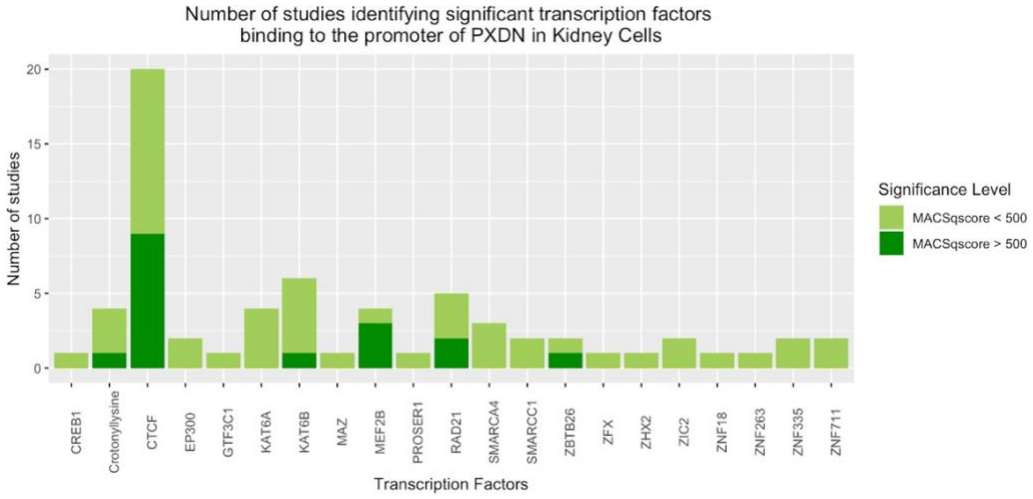


Figure 1: Bar graph showing TFs binding to the promoter of the *PXDN* gene in kidney cells, as identified from ChIP-Atlas, using a threshold of 50. The graph displays the number of studies supporting each TF, with MACS scores indicated. Dark green bars represent TFs with a MACS score greater than 500, while lighter green bars indicate TFs with a MACS score less than 500.

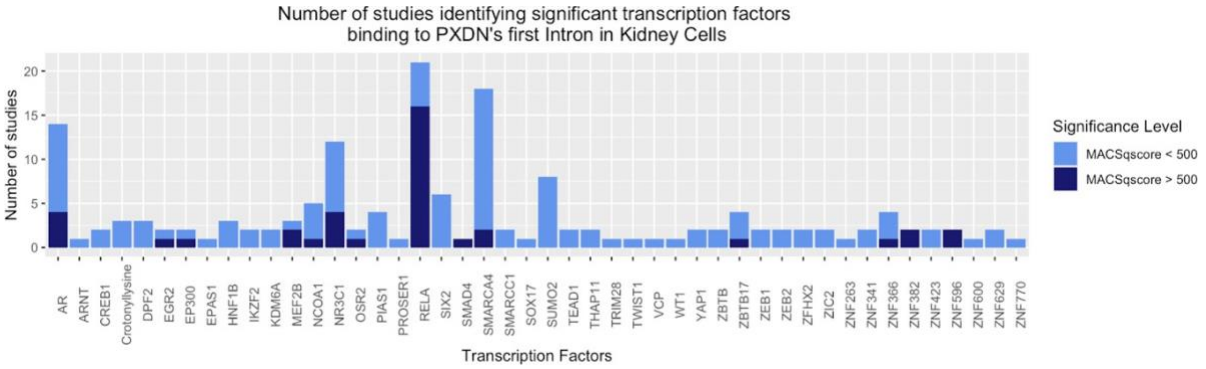


Figure 2: Bar graph showing TFs binding to the first intronic region of *PXDN* in kidney cells, as identified from ChIP-Atlas, using a threshold of 50. The graph displays the number of supporting each TF binding, with MACS scores indicated. Dark blue bars represent TFs with a MACS score greater than 500, while lighter blue bars indicate TFs with a MACS score less than 500.

Appendix D

For cell treatments mentioned in the Tables 2 and 3, untreated refers to cells which have not been treated. Normoxia refers to normal levels of oxygen. B-Galactosidase, often used as a reporter gene, has been expressed using a plasmid that has been engineered to not contain CpG dinucleotides (Trifonov et al., 2016). BORIS is a TF which is involved in chromatin organisation and gene regulation, often associated with cancer (Moscona et al., 2023). R1881 refers to metribolone, a synthetic androgen, which strongly activates the Androgen receptor (AR) TF (Heinlein & Chang, 2002). Dexamethasone treatment has roles in regulating transcription of target genes through binding to the glucocorticoid receptor, a nuclear receptor which mediates the effects of metabolism, immune response and stress response (Stojceski et al., 2023). ML-792 is an inhibitor of the SUMO-activating enzyme, affecting transcription, DNA repair and signal transduction (He et al., 2017). DMSO is used either as a control or as a solvent to dissolve compounds (AlOkda & Van Raamsdonk, 2022). Bortezomib is a proteasome inhibitor put forward as an anticancer therapy (Richardson et al., 2006). PBRM1 knockout refers to cells where the *PBRM1* gene has been deleted. CHIR refers to cells being exposed to CHIR99021, a selective inhibitor of glycogen synthase kinase 3 (GSK-3) (Roca & Campillo, 2020).

Appendix E

Table 1 describes the TFs and genes selected to be included in the association analysis. The source of the TF/gene is listed in the second column. Promoter and Intron 1 TFs are TFs identified in Aim 1 (Chapter 2), while remaining TFs and genes were found in literature. Literature evidencing their implication in kidney dysfunction is noted in the second column.

Table 1. Table listing the TF/gene regions included in the association analysis with their source.

TF/gene name	Reference
AR	Intron 1 TF
CREB1	Promoter TF
CTCF	Promoter TF
EGR1	Ai, K., Li, X., Zhang, P., Pan, J., Li, H., He, Z., Zhang, H., Yi, L., Kang, Y., Wang, Y., Chen, J., Li, Y., Xiang, X., Chai, X., & Zhang, D. (2022). Genetic or siRNA inhibition of MBD2 attenuates the UUO- and I/R-induced renal fibrosis via downregulation of EGR1. *Molecular Therapy Nucleic Acids*, *28*. https://doi.org/10.1016/j.omtn.2022.02.015
EGR2	Intron 1 TF
EP300	Promoter TF
Eya1	Wang, F., Zhang, R., Jian, J., Sun, Y., & Li, Q. (2024). Identification and Functional Study of Enhancers of EYA1: The Causative Gene of Branchio-Oto-Renal Syndrome. *Developmental Neuroscience*. https://doi.org/10.1159/000536260
FOS	Pengrattanachot, N., Thongnak, L., & Lungkaphin, A. (2022). The impact of prebiotic fructooligosaccharides on gut dysbiosis and inflammation in obesity and diabetes related kidney disease. In *Food and Function* (Vol. 13, Issue 11). https://doi.org/10.1039/d1fo04428a
FoxD1	Gu, X., Mallipattu, S. K., Guo, Y., Revelo, M. P., Pace, J., Miller, T., Gao, X., Jain, M. K., Bialkowska, A. B., Yang, V. W., He, J. C., & Mei, C. (2017). The loss of Krüppel-like factor 15 in Foxd1+ stromal cells exacerbates kidney fibrosis. *Kidney International*, *92*(5), 1178–1193. https://doi.org/10.1016/J.KINT.2017.03.037
HNF1B	Bantounas, I., Rooney, K. M., Lopes, F. M., Tengku, F., Woods, S., Zeef, L. A. H., Lin, I. H., Kuba, S. Y., Bates, N., Hummelgaard, S., Hillman, K. A., Cereghini, S., Woolf, A. S., & Kimber, S. J. (2024). Human pluripotent stem cell-derived kidney organoids reveal tubular epithelial pathobiology of heterozygous HNF1B-associated

	dysplastic kidney malformations. *Stem Cell Reports*, *19*(6). https://doi.org/10.1016/j.stemcr.2024.04.011
HNF4A	Marable, S. S., Chung, E., Adam, M., Potter, S. S., & Park, J. S. (2018). Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome. *JCI Insight*, *3*(14). https://doi.org/10.1172/jci.insight.97497
JUN	Pillebout, E., Weitzman, J. B., Burtin, M., Martino, C., Federici, P., Yaniv, M., Friedlander, G., & Terzi, F. (2003). JunD protects against chronic kidney disease by regulating paracrine mitogens. *Journal of Clinical Investigation*, *112*(6). https://doi.org/10.1172/JCI200317647
KDM6A	Intron 1 TF
KLF4	Wen, Y., Lu, X., Ren, J., Privratsky, J. R., Yang, B., Rudemiller, N. P., Zhang, J., Griffiths, R., Jain, M. K., Nedospasov, S. A., Liu, B. C., & Crowley, S. D. (2019). KLF4 in macrophages attenuates TNFa-mediated kidney injury and fibrosis. *Journal of the American Society of Nephrology*, *30*(10). https://doi.org/10.1681/ASN.2019020111
MEF2B	Promoter and Intron 1 TF
NCOA1	Intron 1 TF
NFKB1	Coto, E., Díaz-Corte, C., Tranche, S., Gómez, J., Alonso, B., Iglesias, S., Reguero, J. R., López-Larrea, C., & Coto-Segura, P. (2018). Gene variants in the NF-KB pathway (NFKB1, NFKBIA, NFKBIZ) and their association with type 2 diabetes and impaired renal function. *Human Immunology*, *79*(6). https://doi.org/10.1016/j.humimm.2018.03.008
NFKB2	Song, N., Thaiss, F., & Guo, L. (2019). NFκB and kidney injury. *Frontiers in Immunology*, *10*(MAR). https://doi.org/10.3389/fimmu.2019.00815
NR3C1	Intron 1 TF
NFE2L2	Gómez-Garciá, E. F., Cortés-Sanabria, L., Cueto-Manzano, A. M., Vidal-Martínez, M. A., Medina-Zavala, R. S., López-Leal, J., Renteriá-Padilla, J., & Mendoza-Carrera, F. (2022). Association of Variants of the NFE2L2 Gene with Metabolic and Kidney Function Parameters in Patients with Diabetes and/or Hypertension. *Genetic Testing and Molecular Biomarkers*, *26*(7–8). https://doi.org/10.1089/gtmb.2022.0041
OSR2	Intron 1 TF
Pax2	Zhang, L., Zhai, S. bo, Zhao, L. yue, Zhang, Y., Sun, B. chao, & Ma, Q. shan. (2018). New PAX2 heterozygous mutation in a child with chronic kidney disease: a case report and review of the literature. *BMC Nephrology*, *19*(1). https://doi.org/10.1186/s12882-018-1044-9
Pax8	Zivotic, M., Dundjerovic, D., Naumovic, R., Kovacevic, S., Ivanov, M., Karanovic, D., Nikolic, G., Markovic-Lipkovski, J., Radojevic Skodric, S., & Nesovic Ostojic, J. (2022). Clinicopathological Relevance of PAX8 Expression Patterns in Acute Kidney Injury and Chronic Kidney Diseases. *Diagnostics*, *12*(9). https://doi.org/10.3390/diagnostics12092036
RAD21	Promoter TF

RBPJ	Lin, E. E., Sequeira-Lopez, M. L. S., & Gomez, R. A. (2014). RBP-J in FOXD1+ renal stromal progenitors is crucial for the proper development and assembly of the kidney vasculature and glomerular mesangial cells. *American Journal of Physiology - Renal Physiology*, *306*(2). https://doi.org/10.1152/ajprenal.00313.2013
REL	Ta, M. H., Schwensen, K. G., Liuwantara, D., Huso, D. L., Watnick, T., & Rangan, G. K. (2016). Constitutive renal Rel/nuclear factor-κB expression in Lewis polycystic kidney disease rats. *World Journal of Nephrology*, *5*(4). https://doi.org/10.5527/wjn.v5.i4.339
RELA	Intron 1 TF
RELB	Sun, D., Xie, N., Wang, X., Wu, W., Li, X. Y., Chen, X., Qian, G., Li, C., Zhang, H., Jiang, Y., Ye, D., Liu, D., Hu, Y., Wang, J., Chen, W., Zhao, Q., Zeng, M., Zhang, J., Wang, L., & Zhang, X. (2021). Serum RelB is correlated with renal fibrosis and predicts chronic kidney disease progression. In *Clinical and Translational Medicine* (Vol. 11, Issue 5). https://doi.org/10.1002/ctm2.362
SIX2	Intron 1 TF
SMAD4	Intron 1 TF
SMARCA4	Intron 1 TF
SNAI1	Grande, M. T., Sánchez-Laorden, B., López-Blau, C., de Frutos, C. A., Boutet, A., Arévalo, M., Rowe, R. G., Weiss, S. J., López-Novoa, J. M., & Nieto, M. A. (2015). Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nature Medicine*, *21*(9). https://doi.org/10.1038/nm.3901
SOX17	Intron 1 TF
Sp1	Chen, S. J., Wu, P., Sun, L. J., Zhou, B., Niu, W., Liu, S., Lin, F. J., & Jiang, G. R. (2017). MiR-204 regulates epithelial-mesenchymal transition by targeting SP1 in the tubular epithelial cells after acute kidney injury induced by ischemia-reperfusion. *Oncology Reports*, *37*(2). https://doi.org/10.3892/or.2016.5294
SRF	Zhao, L., Li, C., Guan, C., Song, N., Luan, H., Luo, C., Jiang, W., Bu, Q., Wang, Y., Che, L., & Xu, Y. (2021). Serum response factor, a novel early diagnostic biomarker of acute kidney injury. *Aging*, *13*(2). https://doi.org/10.18632/aging.202381
TWIST1	Intron 1 TF
WT1	Dong, L., Pietsch, S., & Englert, C. (2015). Towards an understanding of kidney diseases associated with WT1 mutations. In *Kidney International* (Vol. 88, Issue 4). https://doi.org/10.1038/ki.2015.198
ZBTB17	Intron 1 TF
ZBTB26	Promoter TF
ZEB1	Intron 1 TF
ZEB2	Intron 1 TF
ZFHX2	Intron 1 TF
ZHX2	Promoter TF

ZIC2	Intron 1 TF
ZNF366	Intron 1 TF
ZNF382	Intron 1 TF
ZNF423	Intron 1 TF
ZNF596	Intron 1 TF
ZNF629	Intron 1 TF
ZNF711	Promoter TF

Appendix F

Table 1 shows the list of genes and regions included in the association study, including a 4 000 bp flanking regions.

Table 1. Genomic coordinates of gene regions included in the association study.

Gene	Chromosome	Start position	End position
AR	X	67544020	67730619
CREB1	2	207529961	207605988
CTCF	16	67570772	67639185
EGR1	5	138465478	138469303
EGR2	10	62811995	62819167
EP300	22	41092591	41180077
EYA1	8	71197432	71548094
FOS/AP1	14	75278827	75282230
FOXD1	5	73446265	73448777
HNF1B	17	37686430	37745059
HNF4A	20	44355698	44434596
JUN	1	58780790	58784047
KDM6A	X	44873187	45112779
KLF4	9	107484851	107489769
MEF2B	19	19145566	19170263
NCOA1	2	24491253	24770702
NFE2L2	2	177230302	177264727
NFKB1	4	102501358	102617302
NFKB2	10	102394109	102402529
NR3C1	5	143277930	143403686
OSR2	8	98944441	98952100
PAX2	10	100735395	100829944
PAX8	2	113215996	113278921
PXDN	2	1627887	1748852
RAD21	8	116845933	116874776
RBPJ	4	26105448	26435131
RELA	2	60881573	65663857
SIX2	2	45005181	45009452
SMAD4	18	51030212	51085042
SMARCA4	19	10961029	11062273

SNAI1	20	49982979	49988886
SOX17	8	54457934	54460892
SP1	12	53380175	53416446
SRF	6	43171268	43181506
TWIST1	7	19113046	19117636
WT1	11	32387774	32435539
ZBTB17	1	15941868	15976101
ZBTB26	9	122915565	122931512
ZEB1	10	31319215	31529804
ZEB2	2	144384080	144520119
ZFHX2	8	23520856	23529664
ZHX2	8	122780378	122974510
ZNF366	5	72439902	72507410
ZNF382	19	36605312	36634114
ZNF423	16	49487523	49640500
ZNF596	8	232199	247340
ZNF629	16	30778455	30787205
ZNF711	X	85243990	85273357

Appendix G

Table 1 shows the phenotype data of the 100 black South Africans included in the study. Sex (0 = female, 1 = male), age at collection, eGFR, measured in mL/min per 1.73m^{2,} and UACR, in mg/g, is shown.

Table 1. List of phenotype data for AWI-Gen samples

Study ID	Sex	Age	eGFR	UACR
AB1313	1	76	90.87781	
DFX0S	0	41	105.04822	
RFI0W	1	48	87.81825	
AB2502	1	57	100.838	0.93571428571429
AB2305	1	67	96.47098	
AB0009	0	42	119.22764	1.4583333333333
CPW0I	0	49	97.475845	
AB2513	0	51	99.40469	
CFJ0C	0	48	114.51954	
AZZ0A	1	41	108.25657	
CMM0V	0	47	64.09592	
ACH0L	1	56	97.95183	
TIH0I	0	69	100.47371	0.65178571428571
CXV0X	0	51	106.89126	
CCL0X	0	46	64.4289	
CBM0W	0	55	82.19978	
ANY0C	1	53	103.536385	12.674846625767
CYE0I	0	58	96.84076	
RWB0W	0	45	78.51265	
RRT0G	0	48	104.70734	0.5
BST0J	1	58	92.2028	
VLD0N	1	51	107.02111	
VGD0D	0	57	71.427956	
AB2484	1	51	74.419426	0.33727810650888
DBB0M	0	53	64.780945	
VKS0C	0	48	106.676476	0.47712418300654
AB2163	1	68	57.94811	0.90710382513661
ADC0I	1	52	104.6392	
AVR0I	1	50	103.517815	
RMP0S	1	40	81.8843	0.21794871794872
REN0A	0	50	106.05581	
RVB0T	0	40	102.44521	
AB2867	0	51	102.13103	4.6923076923077
AB2870	0	41	109.318474	0.39306358381503
RDH0Q	1	52	81.95871	
AB2533	1	45	105.785576	
ATM0C	1	53	74.57377	
AB2967	0	43	109.950775	
DGN0L	0	59	68.41012	0
VVJ0M	1	57	88.00755	11.551724137931
DEP0I	0	54	100.3062	
RJN0K	0	39	69.687836	
BSH0X	1	53	93.014244	0.38376383763838
AIZ0R	1	52	104.42372	0.29042904290429
DHY0Y	0	57	80.88842	
CFK0D	0	60	85.69958	
AB2092	1	73	105.32554	2.761316872428

AGM0B	1	55	53.820705	
RFA0M	0	59	95.50561	
AB0330	1	43	93.61855	0.18627450980392
RZD0F	1	56	73.79233	
RXV0S	0	50	102.86831	
AXA0V	1	42	126.195	
RFQ0E	1	61	75.40184	
AB2098	0	65	89.26289	0.3457249070632
AFY0J	1	58	94.40819	
BFH0W	1	41	116.12451	
AB1298	0	76	94.89965	
AB0867	0	57		0.36046511627907
REV0G	1	54	98.59314	0.32258064516129
BLP0R	1	53	112.18867	16.453608247423
AB2496	1	55	102.61321	
AB0874	0	59		
AB1276	0	72	54.32325	
VSM0L	0	52	100.8366	
REW0H	1	66	81.22929	
RDJ0S	0	58	77.455215	
CWF0F	0	48	105.56563	
DDB0R	0	52	113.89191	
VNB0Q	1	61	89.40921	
AB1089	0	79	85.20177	
CVN0L	0	56	76.943245	
DEC0V	0	48	104.62423	
AB2543	0	46	101.05494	0.25
CFF0X	0	46		
THT0S	0	48	103.84204	
AKM0J	1	41	110.76636	1.4964028776978
TDF0V	0	40	108.71895	
BKX0X	1	60	75.138695	0.64150943396226
CGB0V	0	46	93.995346	
AIR0J	1	49	111.76581	
CGI0D	0	54	97.2752	
AB2721	1	68	97.29198	16.673913043478
AJX0R	1	57	101.21254	
BFE0S	1	41	110.8402	2.3923076923077
AST0G	1	45	113.57306	
AQV0D	1	59	82.26272	0.41284403669725
VBC0Q	1	51	100.660065	
AJK0F	1	56	103.504036	17.956310679612
AB0528	0	52	79.98747	1.8
AGQ0E	1	46	103.16174	
RCZ0G	0	50	71.508026	0.58823529411765
ASP0C	1	48	100.9091	
CBB0J	0	50	92.487236	
AGN0C	1	40	120.23755	
BBX0D	1	47	119.021355	
ANG0J	1	55	109.52433	0.97058823529412
ACS0X	1	53	74.06735	
AB1223	0	80	84.755196	
DLB0J	0	41	102.39586	

Appendix H

Figure 1 includes a PCA of PC1 vs PC2 for the AWI-Gen samples with three 1000 Genomes reference populations.

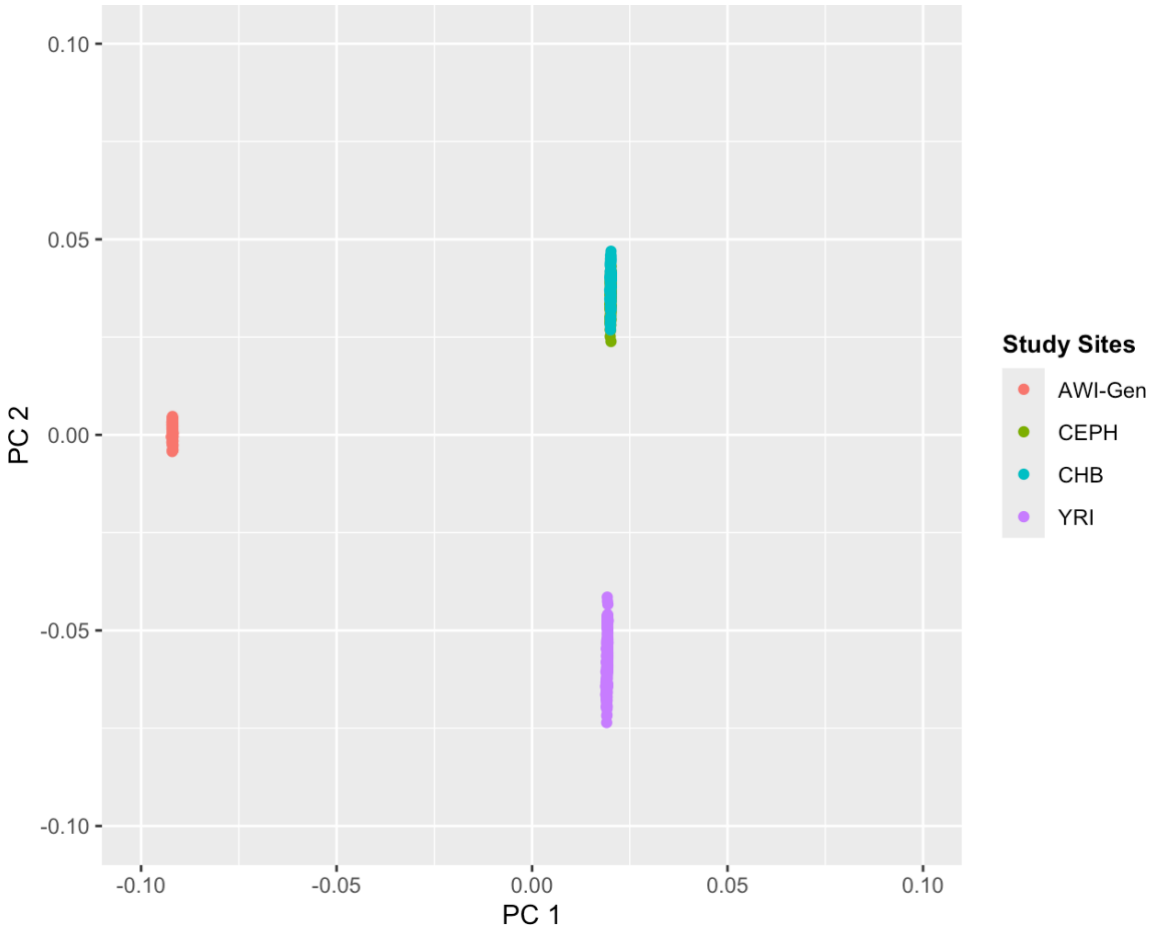


Figure 1. Principal component analysis (PCA) of AWI-Gen samples with 1000 Genomes reference populations. The included populations are South African AWI-Gen samples, Northern and Western Europeans (CEPH), Han Chinese (CHB) and Yoruba Nigerians (YRI).