Elucidating the genetic architecture of cystic kidney disease using whole genome sequencing

Omid Sadeghi-Alavijeh

Division of Medicine

University College London

PhD Supervisor: Professor Daniel Gale

A thesis submitted for the degree of

Doctor of Philosophy

September 2023

Declaration

I, Omid Sadeghi-Alavijeh confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Cystic kidney disease (CyKD) is the commonest life-threatening monogenic disorder, causing great morbidity and mortality. Whilst there is believed to be a strongly monogenic architecture, an unbiased whole genome sequencing approach to understanding the underlying genetic architecture has never previously been attempted.

In this thesis I used statistical genetics and bioinformatics methodology to investigate the genetic architecture of CyKD as well as two other rare disorders, urinary stone disease (USD) and extreme early onset hypertension (EEHTN), using whole genome sequencing data from the 100,000 Genomes Project. I used population-based tools to assess the rare and common variant associations in diverse ancestry matched cohorts seeking enrichment of single nucleotide/indel and structural variants on a genome-wide and per-gene basis.

In all three disorders this improved our understanding of the underlying architecture. CyKD is shown to be strongly monogenic as expected but low-frequency and common variants are shown to play an important role in pathogenesis and causation of this disease, revealing a role for polygenic factors. The heritability of USD is shown to be heavily influenced by low-frequency variants in the sodium-phosphate transporter gene *SLC34A3*, which explains much of the missing heritability not detected by previous large-scale common variant association studies. This finding bridges the gap between the traditional thinking that USD is either monogenic or polygenic/environmental. Finally, EEHTN is shown to likely be an extreme manifestation of primary hypertension, with a strong polygenic basis.

These results support the idea that with better sequencing and larger biobanks, an omnigenic model of disease will become more demonstrable for a broader range of phenotypes, consistent with genotype-phenotype heterogeneity, variable expressivity and incomplete penetrance observed in all three diseases. Finally, I demonstrate that population level approaches traditionally used to study common disease are applicable and useful in rare disease research.

Impact Statement

The findings from this study will have implications across multiple disciplines within and external to nephrology. From a patient perspective many of the findings are in patients who are unsolved by the clinical arm of the 100,000 genomes project. These are in the process of being fed back to the relevant patient's clinicians with a view to potentially offering them a molecular diagnosis.

Secondly, these results will be of interest to researchers in nephrogenetics as well as clinicians involved in rare renal disease. I hope that these results will be hypothesis forming for both *in silico* and functional analyses. For the wider genomics community, I have used a mixed ancestry in nearly all of my analyses without major genomic confounding. This demonstrates the scientific advantages of including a wider cohort for genomic analysis and normalises the representation of individuals from diverse ancestral backgrounds. At the start of my thesis, I was only using Europeans for my analyses, which as a researcher not of European ancestry, was odd to me. I am pleased that our group has developed methods to improve representation in rare disease analyses.

Finally, attempting to marry the common and rare variants domains via the analysis of low frequency variants in rare disease has great implications for the future of rare disease genomics. As rare disease cohorts become larger and sequencing improves, we really are at an exciting time to tease out the "missed heritability" of diseases. This will help guide understanding of biology and more importantly offer new avenues for therapeutics for a series of diseases that really lack personalised approaches.

The impact of this work will be disseminated primarily through publication in peer reviewed journals with lay summaries to increase public and patient engagement. My work on USD has already been published and I have created a tutorial on Twitter to increase the visibility and approachability of the work for patients, academics, and clinicians. Promotion through social media such as Twitter has increased the visibility of our work and has led to fruitful engagements with relevant stakeholders. My work

has been selected for presentation at a number of international and national conferences (oral presentations at UK Kidney Week 2021-2023, Wellcome Genomics of Rare Disease 2023 and Association of Physicians of Great Britain and Ireland Annual Meeting 2023; poster presentations at the American Society of Nephrology Kidney Week 2022) highlighting its broad appeal across genetics, nephrology, and medicine.

Acknowledgements

Professor Daniel Gale (Dr when I first met him!) has been an inspirational supervisor throughout my career to date. Working with such a talented and dedicated clinical academic has opened my eyes to the possibilities in our field and beyond and I would not be where I am today if not for him. In at a close second has to be Melanie Chan, my fellow PhD researcher and unofficial post-doc throughout my time with the group, her mentorship throughout my PhD enabled me to ask any question I had in an infinitely better way, she has an exceptionally bright future ahead of her. I am also grateful to the other PIs in the group, Horia Stanescu for his excellent seminars on the fundamentals of genetics and science in general as well as general guidance, Detlef Bockenhauer for his guidance on all things tubular, particularly on the urinary stone work, Adam Levine for dragging me through my ACF and helping to teach me to code, Patricia Wilson for providing a biological take on my bioinformatic findings and Robert Kleta for his support and mentorship throughout. Professor Richard Sandford, my secondary supervisor, has always been willing to listen whenever I have had findings of note and I am grateful.

To my fellow lab members and recent alumni, I would like to thank you for giving the lab such a convivial and hospitable atmosphere: Mallory Downie, Sanjana Gupta, Matthew Stubbs, Catalin Voinescu, Joshua Carmichael, Anna Ferlin, Anne Kesselheim, Joanna Smith, Gabriel Doctor, Katie Wong and Vaksha Patel.

I am grateful to the Medical Research Council for funding me and this project and Genomics England for generating the data for analysis, and to the patients and families involved in the 100,000 Genomes Project, without whom this research would not have been possible. The Bioinformatics team at Genomics England require a special mention for their ongoing technical support.

My parents (Professors Mo Alavijeh and Zoe Aslanpour) have been particularly supportive, especially helping with childcare during the writing process! Their ongoing support throughout my life has enabled me to get this far.

Finally, to my partner Anna and children Kassra and Albie, thank you for being patient with me, especially during the write-up. Your love, support and encouragement has been the base from which I have been able to complete this thesis.

Publications

During this PhD, I have authored or contributed to the following publications:

Rare variants in the sodium-dependent phosphate transporter gene *SLC34A3* explain missing heritability of urinary stone disease.
Sadeghi-Alavijeh, Omid; Chan, Melanie My; Moochhala, Shabbir H; Genomics England Research, Consortium; Howles, Sarah; Gale, Daniel P; Böckenhauer, Detlef; (2023) Rare variants in in the sodium-dependent phosphate transporter gene SLC34A3 explain missing heritability of urinary stone disease. *Kidney International* 10.1016/j.kint.2023.06.019

Diverse ancestry whole-genome sequencing association study identifies TBX5 and PTK7 as susceptibility genes for posterior urethral valves.
Melanie MY Chan, **Omid Sadeghi-Alavijeh**, Filipa M Lopes, Alina C Hilger, Horia C Stanescu, Catalin D Voinescu, Glenda M Beaman, William G Newman, Marcin Zaniew, Stefanie Weber, Yee Mang Ho, John O Connolly, Dan Wood, Alexander Stuckey, Athanasios Kousathanas, Genomics England Research Consortium, Robert Kleta, Adrian S Woolf, Detlef Bockenhauer, Adam P Levine and Daniel P Gale. *eLife*. 2022 Sep 20;11:e74777. doi.org/10.7554/eLife.74777. PMID: 36124557.

Shared genetic risk across different presentations of gene test–negative idiopathic nephrotic syndrome.
Mallory L Downie, Sanjana Gupta, Melanie MY Chan, **Omid Sadeghi-Alavijeh**, Jingjing Cao, Rulan S Parekh, Carmen Bugarin Diz, Agnieszka Bierzynska, Adam P Levine, Ruth J Pepper, Horia Stanescu, Moin A Saleem, Robert Kleta, Detlef Bockenhauer, Ania B Koziell, Daniel P Gale *Pediatric Nephrology* 38.6 (2023): 1793-1800.

Common Risk Variants in AHI1 Are Associated With Childhood Steroid Sensitive Nephrotic Syndrome
Mallory L Downie, Sanjana Gupta, Catalin Voinescu, Adam P Levine, **Omid Sadeghi-Alavijeh**, Stephanie Dufek-Kamperis, Jingjing Cao, Martin Christian, Jameela A Kari, Shenal Thalgahagoda, Randula Ranawaka, Asiri Abeyagunawardena, Rasheed Gbadegesin, Rulan Parekh, Robert Kleta, Detlef Bockenhauer, Horia C Stanescu, Daniel P Gale *Kidney International Reports*. 2023 May 27.

Pathogenicity of missense variants affecting the collagen IV α5 carboxy non-collagenous domain in X-linked Alport syndrome
Gibson JT, **Sadeghi-Alavijeh O**, Gale DP, Rothe H, Savige J. *Scientific Reports*. 2022 Jul 4;12(1):11257.

Genotype-phenotype correlations for COL4A3-COL4A5 variants resulting in Gly substitutions in Alport syndrome.
Joel T Gibson, Mary Huang, Marina Shenelli Croos Dabrera, Krushnam Shukla, Hansjorg Rothe, Pascale Hilbert, Constantinos Deltas,
Helen Storey, Beata S Lipska-Ziętkiewicz, Melanie MY Chan, **Omid Sadeghi-**

Alavijeh, Daniel P Gale, Genomics England Research Consortium, Agne Cerkauskaite, Judy Savige. Sci Rep. 2022 Feb;12(1):2722. doi: 10.1038/s41598-022-06525-9. PMID: 35177655.

Prevalence Estimates of Predicted Pathogenic COL4A3 - COL4A5 Variants in a Population Sequencing Database and Their Implications for Alport Syndrome.
Joel Gibson, Rachel Fieldhouse, Melanie Chan, **Omid Sadeghi-Alavijeh**, Leslie Burnett, Valerio Izzi, Anton Persikov, Daniel Gale, Helen Storey, and Judy Savige. J Am Soc Nephrol. 2021 Sep;32(9):2273-2290. doi: 10.1681/ASN.2020071065. PMID: 34400539.

Large-Scale Whole-Genome Sequencing Reveals the Genetic Architecture of Primary Membranoproliferative GN and C3 Glomerulopathy.
Levine AP, Chan MMY, **Sadeghi-Alavijeh O**, Wong EKS, Cook HT, Ashford S, Carss K, Christian MT, Hall
M, Harris CL, McAlinden P, Marchbank KJ, Marks SD, Maxwell H, Megy K, Penkett CJ, Mozere M, Stirrups KE, Tuna S, Wessels J, Whitehorn D; MPGN/DDD/C3 Glomerulopathy Rare Disease Group; NIHR BioResource, Johnson SA, Gale DP. J Am Soc Nephrol. 2020 Feb;31(2):365-373. doi: 10.1681/ASN.2019040433. PMID: 31919107.

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

Rare variants in the sodium-dependent phosphate transporter gene SLC34A3 explain missing heritability of urinary stone disease

b) **Please include a link to or doi for the work**

10.1016/j.kint.2023.06.019

c) **Where was the work published?**

Kidney International

d) **Who published the work?** (e.g. OUP)

Elsevier

e) **When was the work published?**

July 4th 2023

f) **List the manuscript's authors in the order they appear on the publication**

Omid Sadeghi-Alavijeh, Melanie M Y Chan, Shabbir H Moochhala; Genomics England Research Consortium; Sarah Howles, Daniel P Gale, Detlef Böckenhauer

g) **Was the work peer reviewed?**

Yes

h) **Have you retained the copyright?**

No

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

Yes - https://www.medrxiv.org/content/10.1101/2022.12.02.22283024v1

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

Table of Contents

Contents

Table of figures

List of tables

Abbreviations

100KGP	100,000 Genomes Project
AD	Autosomal Dominant
ADPKD	Autosomal Dominant Polycystic Kidney Disease
ADTKD	Autosomal Dominant Tubulo-interstitial kidney disease
AF	Allele Frequency
AG	(Splice site) Acceptor Gain
AL	(Splice site) Acceptor Loss
ALT	Alternate Allele
AR	Autosomal Recessive
ARPKD	Autosomal Recessive Polycystic Kidney Disease
bp	Base-pair
CADD	Combined Annotation Dependent Depletion
CAKUT	Congenital Anomalies of the Kidneys and Urinary Tract
CGH	Comparative Genomic Hybridization
CHR	Chromosome
CI	Confidence Interval
CKD	Chronic Kidney Disease
CNS	Central Nervous System
CNV	Copy Number Variant
CyKD	Cystic Kidney Disease
DEL	Deletion
DG	(Splice site) Donor Gain
DL	(Splice site) Donor Loss
DNA	Deoxyribonucleic Acid
DUP	Duplication
EEHTN	Extreme Early onset Hypertension
eGFR	Estimated Glomerular Filtration Rare
eQTL	Expression Quantitative Trait Loci
ER	Endoplasmic reticulum
ES	Exome Sequencing
ESRF	End Stage Renal Failure

FDR	False Discovery Rate
GLMM	Generalised Logistic Mixed Model
gnomAD	Genome Aggregation Database
GQ	Genotype Quality
GRM	Genomic Relationship Matrix
GT	Genotype
gVCF	Genomic Variant Call Format
GWAS	Genome-Wide Association Study
HC	High Confidence (loss-of-function calls)
HES	Hospital Episode Statistics
HES-EEHTN	Hospital Episode Statistics (derived) Extreme Early Onset Hypertension
HPO	Human Phenotype Ontology
HWE	Hardy-Weinberg Equilibrium
IBD	Identical by Descent
ICH	Intracerebral haemorrhage
Indel	Insertion/Deletion
INV	Inversion
IQR	Interquartile Range
kb	Kilobase
LMM	Linear Mixed Model
LD	Linkage Disequilibrium
LoF	Loss-of-function
MAC	Minor allele count
MAF	Minor allele frequency
Mb	Megabase
MODY	Mature Onset Diabetes of the Young
NGS	Next Generations Sequencing
NHS	National Health Service
OR	Odds Ratio
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PKD	Polycystic Kidney Disease

PLD	Polycystic Liver Disease
pLOF	Predicted Loss-of-Function
POS	(Chromosomal) Position
PP	Posterior Probability
QC	Quality Control
Q-Q	Quantile-Quantile
RCAD	Renal Cysts and Diabetes Syndrome
REF	Reference Allele
RNA	Ribonucleic Acid
RR-EEHTN	Renal Removed Extreme Early onset Hypertension
SD	Standard Deviation
SE	Standard Error
SKAT-O	Sequence Kernel Association Test – Optimal
SNV	Single Nucleotide Variant
SPA	Saddlepoint Approximation
SV	Structural Variants
TAD	Topologically Associated Domain
USD	Urinary Stone Disease
UTR	Untranslated Region
VCF	Variant Call Format
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
XLD	X-linked Dominant
XLR	X-linked Recessive

Chapter 1. Introduction

In this thesis I examined the genetic architecture of cystic kidney diseases (CyKD), urinary stone disease (USD) and extreme early onset hypertension (EEHTN) using whole genome sequencing (WGS) in an unbiased genome-wide manner. In the introduction I will define and elaborate on some of the key concepts to be analysed as well as a brief discussion of the current landscape of genomic analysis.

1.1 The current landscape of genomic analysis

1.1.1 From linkage to GWAS

The elucidation of the intricate relationship between genetic variations and complex traits has been a foundational pursuit in the field of genetics. Over the course of the last century, this pursuit has undergone a remarkable transformation driven by technological advancements and methodological innovations. This evolution of genetic association analysis has propelled the field from rudimentary observations of familial inheritance to sophisticated investigations at the level of the entire genome. Central to this progression is the advent of whole-genome sequencing (WGS), a revolutionary technique that has modernized our ability to comprehensively examine the genetic landscape underlying various phenotypes.

The history of genetic association analysis can be traced back to the pioneering work of early geneticists, who sought to decipher the patterns of inheritance of observable traits. Gregor Mendel's experiments with pea plants in the mid-19th century laid the groundwork for understanding the basic principles of inheritance, and the subsequent discovery of the DNA double helix by Watson and Crick in the 20th century unveiled the molecular basis of genetics. These foundational discoveries set the stage for the exploration of genetic variations' influence on phenotypes.

In the mid-20th century, the concept of genetic linkage emerged as researchers began to observe that certain traits co-segregated more often than expected by chance due to their physical proximity on chromosomes. This led to the development of linkage analysis, a

method used to identify genetic loci associated with traits through the study of familial inheritance patterns. Despite its success in pinpointing genes responsible for Mendelian disorders, linkage analysis faced limitations when applied to complex traits influenced by multiple genetic and environmental factors.

The late 20th century witnessed a shift in focus towards molecular markers and their application in genetic association studies. The introduction of restriction fragment length polymorphisms (RFLPs) and microsatellites enabled more accurate mapping of genetic loci. Family-based linkage studies paved the way for the identification of genes linked to inherited diseases, yet these methods remained inadequate for unravelling the genetic basis of complex traits affecting broader populations as well as in *de novo* disorders. Its efficacy also diminishes in cases of incomplete penetrance or when locus heterogeneity is at play. Population-based candidate gene studies, which have employed positional cloning methods and more recently targeted next-generation sequencing (NGS) approaches, adopt a hypothesis-driven selection of genes based on biological plausibility. These endeavours have contributed to gene discovery in rare diseases, but they remain constrained by elevated false-positive rates and often encounter challenges in terms of reproducibility. Today, both of these methodologies have largely ceded ground to whole-exome sequencing techniques in the domain of rare diseases. Meanwhile, genome-wide association studies (GWAS) conventionally serve as the method of choice for gene discovery in common, complex traits.

1.1.2 The Genomic Era

The shift away from linkage analysis (although it still has a vital role to play in elucidating mechanisms in rare disease) has been facilitated by the increasing number of patients and participants having their DNA sequenced. Now two decades after the groundbreaking publication of the draft human genome sequence (Lander et al. 2001) there has been a revolution in not just analytical techniques but the datasets that provide them. The International HapMap Project (International HapMap 3 Consortium et al. 2010) and the 1000 Genome Project (Sudmant et al. 2015) were harbingers to increasingly large international consortia creating biobanks of large-scale sequencing data from hundreds of thousands of individuals from across the globe (H3Africa

Consortium et al. 2014; Kurki et al. 2022; Halldorsson et al. 2022; Nagai et al. 2017; Turro et al. 2020; GenomeAsia100K Consortium 2019; Bycroft et al. 2018; Taliun et al. 2019; 100000 Genomes Project Pilot Investigators et al. 2021).

However, initial excitement that the draft human genome would lead to a revolution in clinical care has been cooled by the discovery of the vast complexity of the human genome both in sequencing and interpretation. The whole genome from telomere to telomere was only sequenced in 2022 (Nurk et al. 2022) and the vast amount of data generated since 2001 has required rapid advancement in computational and statistical techniques to draw meaningful inferences that inform biology. Treatments informed by genomics are now beginning to reach clinical pipelines and patients such as ribonucleic acid interference (RNAi) molecules for primary hyperoxaluria (Garrelfs et al. 2021), *PCSK9* inhibitors for primary hypercholesterolemia (Abifadel et al. 2003), the presence of a truncating variant in *PKD1* for access to tolvaptan (Müller et al. 2022) and most recently Inaxaplin for *APOL1* associated proteinuric renal disease (Egbuna et al. 2023). Equally impactfully, patients with cancer and rare diseases are now getting rapid and accurate molecular diagnoses which can personalise treatment, risk stratify by genotype and aid in screening programmes. The recent introduction of whole genome sequencing (WGS) into routine clinical care within the NHS means this is truly an exciting time for genomic medicine.

1.1.3 Sitting between rare and common

The common disease common variant (CDCV) debate has roots in the early 20th century conflict between the “Mendelians” led by William Bateson and Hugo de Vries and the “Biometricians” led by Karl Pearson. The latter camp rejected the Mendelian idea that discrete units of heredity could explain the continuous range of phenotypic variation. This was largely unified by RA Fisher and colleagues who showed that Mendel’s genes and laws could work additively to influence the expression of a phenotype both in a discrete and continuous capacity. This influenced discourse in more contemporary debates such as the CDCV vs common disease rare variant (CDRV) over the genetic architecture of hypertension where arguments centred on whether hypertension was rooted in low effect polygenic variants or high effect rare variants. The answer has been

shown to be both with GWAS and candidate gene studies highlighting how these two causes sit together in not just hypertension but many chronic diseases.

In their seminal paper discussing the common variant/common disease paradigm Eric Lander and David Reich gave rationale to this theory positing that it holds true for most diseases and therefore GWAS was a reasonable approach to the study of complex diseases (Reich and Lander 2001). However Pritchard argued that population dynamics are more likely to favour the contribution of multiple rare variants to disease (Pritchard 2001) . He contends that common variants, due to their lengthy presence in the human population, are more likely to have undergone potential selective pressures over time, diminishing the impact of negative selection. In contrast, rare variations, often newly arisen within only a few generations, tend to escape the influence of negative selection or are rare because they are being actively selected against, owing to their inherently deleterious nature. However, whilst conceptually disease can be seen as caused by a spectrum of variants across the allelic frequency spectrum the available tools have continued to silo researchers into "common variant/common disease" or "rare variant/rare disease" methods.

1.1.3.1 GWAS for common and complex diseases

Thousands of GWA studies have now been conducted looking at the relationship between common variants (initially taken to be those with a minor allele frequency[MAF] greater than 5% but now greater than 1% is accepted) and various diseases with great success (Abdellaoui et al. 2023). The rationale from the Lander paper above that the power to detect association in case-control studies is a function of the effect size of an allele and its frequency in the study population means it have been limited mainly to complex traits and disorders such as diabetes or schizophrenia.

Typically, GWAS deploy genome-wide single-nucleotide variant (SNV) microarrays, encompassing hundreds of thousands or millions of variants, often characterized by a minor allele frequency (MAF) greater than 1%. These microarrays enable the genotyping of cohorts under investigation, allowing subsequent comparison with

appropriate control populations. These genotype results are then typically imputed which involves leveraging ancestry-specific reference panels comprised of haplotypes reconstructed from sequencing data, exemplified by initiatives like the Haplotype Reference Consortium. Imputation serves to bridge gaps in data, utilizing the knowledge of linkage disequilibrium (LD), which captures non-random co-inheritance of alleles, to infer missing variants. However, it's crucial to note that the imputation accuracy diminishes when dealing with variants not in LD with those genotyped, particularly rare variants (present in less than 1% of the general population) and those manifesting in non-European populations. These imputed variants are then used for the association test of the trait of interest, with the variants severing markers or indirect proxies rather than direct indicators of the causal variants in the underlying genetic regions.

GWAS has now identified thousands of associations that have informed gene discovery, the generation of predictive risk score (Khera et al. 2018), estimations of heritability (Zhu and Zhou 2020) and prioritization of targets for drug development (Kiryluk et al. 2023). However, as successful as GWAS has been the results to date only explain a small fraction of the burden of any disease in the population at large. This “missing heritability” (Manolio et al. 2009) has been attributed to a) GWAS not capturing common variants with low effect sizes, b) the contribution of variants not detected by imputation of panel data, namely rare variants and structural variants (SVs), c) epistasis, where gene-gene interactions occur and d) genomic imprinting or parent of origin effects. For those variants that have been detected, ~90% of risk alleles are found in non-coding regions of the genome, making functional annotation difficult; although efforts to generate cell and context specific multi-omics data via such projects as ENCODE (Dunham et al. 2012), the RoadMap Epigenomics Consortium (Roadmap Epigenomics Consortium et al. 2015) and GTEx (The GTEx Consortium et al. 2020) have aided hugely with prioritization of causal variants for functional follow-up. Finally, >95% GWAS to date have been done in individuals of European ancestry (as of August 2023 https://gwasdiversitymonitor.com). Increasing ancestral diversity in genetic studies improves the power to detect associations (Ishigaki et al. 2022; Z. Lu et

al. 2022; Mahajan et al. 2022) and is ethically crucial (Peterson et al. 2019; Fatumo et al. 2022).

With these recognised limitations in GWAS, high-coverage WGS data is now being explored for investigating diseases. Its ability to give whole genome coverage to excellent depth has demonstrated improved power and sensitivity over conventional techniques. This coupled with larger multi-ancestry biobanks have identified novel associations in variants that are either rare or ancestry-specific (Hu et al. 2021). The major limiting factor has been the cost of WGS, but with falling costs (a whole genome can now be sequenced for <£500) this is set to become the standard method of analysis.

1.1.3.2 Sequencing in rare diseases

With GWAS requiring large case numbers and being unable to accurately impute rare variants, sequencing both targeted gene sequencing and whole exome/genome sequencing became the focus in Mendelian disease analysis. When Ng *et al* used whole exome sequencing to discover rare variants in *DHODH* as causative for the Miller syndrome in 2010 (Ng et al. 2010) it was hoped that a new era of precision medicine in rare disease would be enabled. It was cost effective and had the potential to overcome the issues with linkage studies such as the requirement for large pedigrees, often poor resolution of linked regions, inability to call *de novo* variants and locus heterogeneity. In the two years post the initial WES proof of concept experiment (Ng et al. 2009), 180 novel genes were described in Mendelian disorders alone (Boycott et al. 2013) and it soon found its way into clinical genetics pipelines and diagnostic labs (Y. Yang et al. 2013).

However, WES has methodological and conceptual issues. From a methods perspective WES gives heterogenous coverage of the exons due to the issues with the hybridisation/capture and PCR-amplification steps during library preparation (Kebschull and Zador 2015), WES also has lower per base coverage than WGS leading to it missing may variants in exons (Belkadi et al. 2015) and it is not a reliable approach for detecting copy number and structural variants (CNV/SV) due to most CNV/SVs extending beyond the boundaries of captured exons. The choice of preparation library is

particularly important, ~50% of pathogenic variants associated with hereditary nephrotic syndrome and Deny-Drash syndrome were poorly covered using leading WES capture kits (Park et al. 2015). Conceptually, variants causing Mendelian disorders are within coding regions ~95% of the time (Botstein and Risch 2003), a finding that spurned the initial drive towards WES, however, non-coding variants have been implicated in multiple diseases (Spielmann and Mundlos 2016; French and Edwards 2020) including kidney diseases such as atypical haemolytic uraemic syndrome (Mele et al. 2015), Alport syndrome (King et al. 2002) and Gitelman syndrome (Lo et al. 2011). WES ignores such variants, and it also ignores a large proportion of SV/CNVs as it is unable to reliably define their breakpoints (R. Tan et al. 2014). It also limits our ability to integrate findings with other lines of multi-omics evidence such as epigenetics or chromatin conformation where the interactions lie outside of the coding genome.

Whole genome sequencing (WGS) has remained in the shadows of WES for some time given its historically higher cost and the vast amounts of data created leading to issues with data storage, security, and downstream analysis. It undoubtedly has benefits over WES, allowing for full capture of non-coding variants, better and more uniform coverage of coding regions (Belkadi et al. 2015), more accurate capture of SV/CNVs (Hehir-Kwa, Pfundt, and Veltman 2015) and better phasing and thus assessment of compound heterozygosity (Hofmeister et al. 2023). The cost of WGS is now falling to that comparable to WES (Dewey et al. 2014), especially when WES may require multiple runs to increase read depth to a level to match the variant detection of WGS (Lelieveld et al. 2015). Alongside the falling cost, the establishment of large scale WGS biobanks such as deCODE (Gudbjartsson et al. 2015), TOPMed (Taliun et al. 2021) and gnomAD (Karczewski et al. 2020) has made the data generated by WGS integral to human genetics research. The metrics these have directly affected variant interpretation at a clinical level and helped inform the establishment of UK biobanks that serve a dual function of research and clinical utility such as the 100,000 genome project (100KGP) (100000 Genomes Project Pilot Investigators et al. 2021) and the NIHR Rare Disease Bioresource (Turro et al. 2020).

1.1.3.3 Rare variants – intermediate effect sizes

The large WGS biobanks have led to several insights that are pertinent to the CDCV/CDRV debate. As sequencing projects get larger in both number as well as read lengths (long read sequencing) the frequency and volume of genetic variation become apparent. There is an abundance of rare and private (seen in one individual) variation within the ~3 million SNVs and ~0.5 million indels in the average genome (Karczewski et al. 2020), loss-of-function variants that are predicted to truncate protein function are more (Lek et al. 2016) frequent than thought and SV/CNVs may in fact account for 25-29% of all such protein truncating events per genome (R. L. Collins et al. 2020).

Such insights give evidence to the theoretical models discussed in the past decade. Figure 1-1 references a now seminal paper by Manolio et al from 2009 whereby the "missing heritability" of diseases is theorized to originate from rare variants of small and intermediate effects size as well as structural variants (Manolio et al. 2009). Figure 1-2 highlights variants on this spectrum in relation to renal diseases from a 2020 review by Groopman et al. In the intervening 11 years there has been little exploration of the rare variant low effect size space. Like GWAS studies, analysis of this would require large cohorts, well sequenced, using the latest methodology to overcome issues around lack of statistical power.

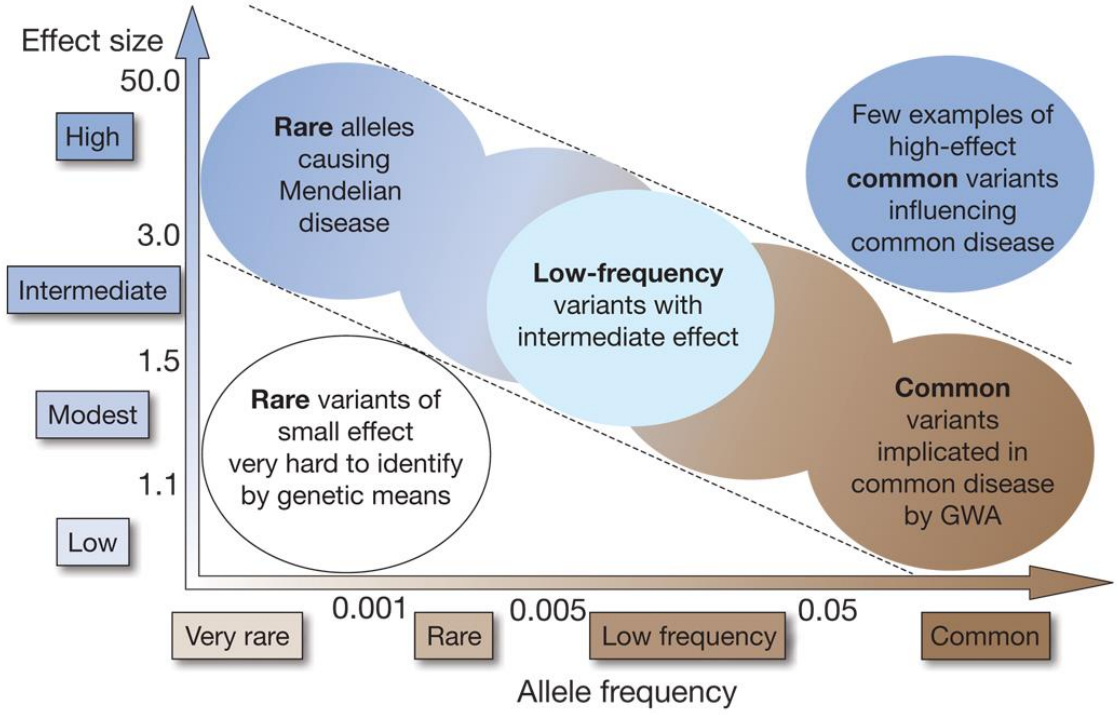

Figure 1-1 Feasibility of identifying genetic variant by risk allele frequency and strength of genetic effect.

The dotted lines represent the areas the author's thought variants were most likely to be found in 2009. The genetic research landscape has expanded greatly since then. Taken from Manolio et al 2009.

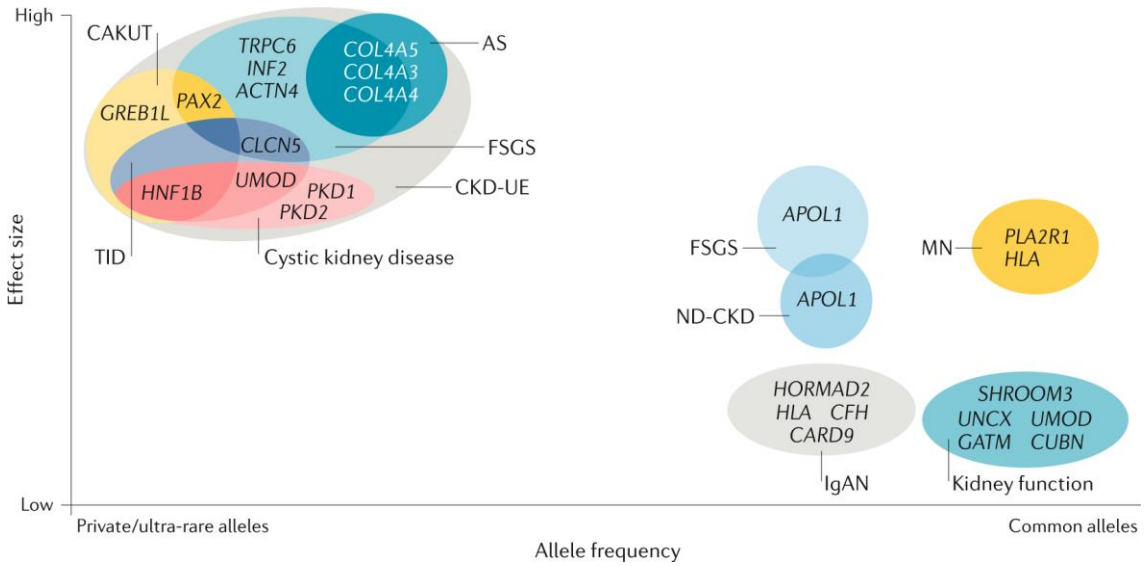

Figure 1-2 The genetic contribution of variants to renal diseases

Disease causing variants for renal diseases can be large effect, rare variants which tend to follow a Mendelian patten of inheritance and monogenic. More common alleles have smaller individual effect sizes although exceptions such as *APOL1* exist. Other genes such as *UMOD* have both common and rare

variant roles in renal diseases, or multiple phenotypes such as in *HNF1β*. Crucially the area of rare, low effect size variants remains relatively unexplored in the renal disease.

We now live in the era of large biobanks of WGS data. There are also techniques such as region-based testing that collapse information across genomic regions e.g., a gene, before testing for association with a phenotype. Such methods have seen success in describing a number of novel gene-disease relationships (Q. Wang et al. 2021; Deaton et al. 2021; Akbari et al. 2021) and more importantly begin to fill out the bottom left hand quadrant of both figure 1.1-1.2. One of the largest analysis of rare and low frequency variants to date, across 643,219 individuals and 744 phenotypes from the UKBB and Finngen, found 975 associations of which 145 were driven by unique variants in the allelic frequency between 0.1-2% with an average odds ratio of 2.8 (Benjamin B. Sun et al. 2022). Clearly these variants are not acting in purely Mendelian ways, their penetrance is likely to be low and they should be seen as risk factors or modifiers that transact with other genetic and environmental factors for a particular disease. This is an exciting era for genomic medicine and in this thesis, I will use similar methods to explore this genome space for a number of disorders to further elucidate their genetic architecture.

1.2 Genetic architecture

The term “genetic architecture” needs defining prior to its use in this thesis. It refers to the types of genetic variation and their respective effects on the observed variation in a phenotype. This is driven by both our knowledge of the types of variation that exist and in turn the technologies and methods available to detect them. This encompasses the arrangements and distribution of genetic variants, such as single nucleotide variants (SNVs), insertions/deletions (indels) and structural variants (SNVs) across a cohort’s genomes, their allele frequencies, and their effect sizes. In human population-based analyses genetic architecture describes the genetic variation that is responsible for broad-sense phenotypic heritability (Mackay 2001). This is compared to narrow-sense heritability which applies to additive genetic effect only (Visscher, Hill, and Wray 2008).

Genetic architecture also includes the interaction and combined effects of multiple genes (epistasis) and their interactions with environmental factors in determining the expression of traits. Defining this architecture for a trait or disease is a fundamental goal of human genetics both scientifically and clinically.

Genetic architecture is defined as much by the technologies available to the researchers as the underlying genomic variants. Historically, limited by both pedigree size and genetic mapping technology, linkage analysis and fine mapping where the technologies of the day during the 1980s and 1990s (Lipner and Greenberg 2018). Localization of genetics signals was typically followed up using Sanger sequencing and then functional studies in cellular and animal models (Heather and Chain 2016). This was a difficult and laborious process but by 2000 ~1000 of the ~7000 single gene inherited disease had been described such as Huntington's disease and cystic fibrosis (Kremer et al. 1994; F. S. Collins 1990).

The first draft of the human genome sequence reduced many of the barriers to disease-gene mapping (Schmutz et al. 2004). Microarray-based technologies allowed for structural variation to be analysed and exome and genome-wide sequencing have been instrumental in further elucidation of genetic architecture aided by the parallel development of *in silico* analysis of genetic variants (Heather and Chain 2016). Complex diseases and traits with polygenic architecture can now be described and biobanks of increasing size allow for the examination of low frequency, low effect rare variants in all forms of disease (Q. Wang et al. 2021). In fact for certain traits such as height, the "missing heritability" has been solved in European ancestry paving the way for further phenotypes to be elucidated in a similar fashion as biobanks increase in size (Yengo et al. 2022).

This increasing ability to sequence more of an individual's DNA more reliably and at scale had led to an evolution of the models of genetic architecture. Traditionally genetic architecture has been described as monogenic, oligogenic or polygenic implying differing levels of genetic variant contribution to the variability in a phenotype (Badano and Katsanis 2002). However, the "omnigenic" model describes a gene regulatory

network wired as such to allow all genes expressed to contribute to a trait, with peripheral gene networks having a no-zero effect and interaction with a core set of genes in a given tissue or cell type (Boyle, Li, and Pritchard 2017). Like Fisher's "infinitesimal model' whereby all variants have non-zero effects on trait variation, the omnigenic model effectively describes all traits as quantitative as variation throughout the genome affects the process as much as more closely related variants (Fisher 1919). Clearly it still holds true that some phenotypes are much more monogenic e.g. Cystic Fibrosis whilst others are more polygenic e.g. Type 2 diabetes but the increasing influence of "peripheral" genetic variants on "core" genes and subsequent phenotype is increasingly being appreciated. From a practical perspective, these influences have been best described at the intersection of monogenic disease and polygenic risk scores, such as in altering the penetrance of monogenic tier 1 genomic conditions (Fahed et al. 2020), describing the modification of chronic kidney disease (CKD) risk in monogenic causes of renal disease (Khan et al. 2023) and our work on the polygenic interaction with rare, low effect size variants in nephrolithiasis described later in this thesis (Sadeghi-Alavijeh et al. 2023).

A comprehensive understanding of genetic architecture allows for better screening, diagnosis, prognosis, and therapeutics for a given disease. In this thesis I describe the use of short-read whole genome sequencing (WGS) in a national cohort of cystic kidney disease (CyKD), urinary stone disease (USD) and extreme early onset hypertension (EEHTN) in order to describe the genetic architecture of these disorders.

1.3 Summary

In summary, our ability to plumb the full range of genetic variation and assign these findings a role in a trait or disease has exploded in the last decade. While previous studies of genetic causation in disease have been siloed by the technologies and methods available into "common – GWAS" or "rare- sequencing" our ability to overcome these challenges has improved to the point of being able to integrate a broad spectrum of variation into our models.

1.4 This study

In this study I use WGS data from the 100,000 Genomes Project to understand the genetic architecture of CyKD, USD and EEHTN. Population based rare and common variant association testing was performed in diverse ancestry case control cohorts looking for enrichment of single nucleotide/indel and structural variants on a genome-wide level. Polygenic risk scoring was utilised as a method to ascertain heritability and understand common variant contribution to these diseases. This study represents one of the largest WGS analyses of all three conditions using unbiased genome wide methods.

Chapter 2. Materials & Methods

In this chapter I discuss the methods used for generating genomic and cohort data used throughout the thesis. I then discuss the overarching theories behind the statistical methodology used for controlling for sources of bias and the subsequent association testing. I will go into more detailed practical methodology within each results chapter. All code used in this thesis can be found on my GitHub: https://github.com/oalavijeh/phd_scripts/tree/main, all workflows generated by Genomics England's bioinformatics team can be found here: https://re-docs.genomicsengland.co.uk/workflows/. All summary statistics have been uploaded to a shared drive at: https://liveuclac-my.sharepoint.com/:f:/r/personal/zchaf43_ucl_ac_uk/Documents/thesis_summary_stats?csf=1&web=1&e=oTKY8q and will be referenced to as "summary statistics" in the text, this will be hyperlinked to this location.

2.1 The 100,000 Genomes Project

In 2012, the UK launched the 100,000 Genomes Project (100KGP), an initiative to sequence 100,000 genomes from patients with cancer, rare disease and their unaffected relatives (100000 Genomes Project Pilot Investigators et al. 2021) . 13 National Health Service (NHS) Genomic Medicine Centres across the UK recruited participants which was completed in December 2018. In total 132,760 genomes had been sequenced by March 2023. The Genomics England dataset (version 15) consists of WGS data, clinical phenotypes encoded using a standardized vocabulary of phenotypic abnormalities called Human Phenotype Ontology (HPO) codes (Groza et al. 2015), and retrospective and prospectively ascertained NHS hospital records for 90,189 individuals. Ethical approval for the 100KGP was granted by the Research Ethics Committee for East of England Cambridge South (REC Ref 14/EE/1112). Written informed consent was obtained from all participants or their guardians.

The 100,000 Genomes Project (100KG) is one of the largest sequencing initiatives in the world offering a unique opportunity to combine high-quality, high-coverage

genomic data with rich clinical and phenotypic information from a national health system. Furthermore, a key strength of this dataset is the availability of sequence data from large numbers of people without the phenotype under study, drawn from the same population recruited and their samples processed and sequenced within a shared pipeline. This allows control for allele frequency and variant burden in the population. This is an advantage compared with previous sequencing studies in these disorders that have typically lacked such robust control.

2.2 Data Generation and Processing

DNA extraction, processing, whole genome sequencing, WGS alignment, variant calling, variant quality control and aggregation were all performed centrally by the Genomics England central bioinformatics team and will be detailed below. This resulted in an aggregated genomic variant calling file (gVCF) incorporating a majority of the 100KGP participants split into chunks by genomic position.

2.2.1 DNA extraction and preparation

Nearly all the DNA (99%) was harvested from blood and prepared using EDTA with the remaining coming from saliva or tissue. Samples underwent quality control assessment based on volume, concentration, purity, and degradation. Libraries were prepared using the Illumina TruSeq DNA PCR-Free High Throughput Sample Preparation kit to minimize PCR-induced sequencing bias. Where limited DNA was available (<1% samples) the Illumina TruSeq Nano High Throughput Sample Preparation kit was used.

2.2.2 Whole-genome sequencing and alignment

Illumina HiSeq X instruments were used to perform WGS, generating 150bp paired end reads which were processed on the Illumina North Star Version 4 Whole Genome Sequencing Workflow (version 2.6.53.23). Read were mapped to the Homo Sapiens NCBI GRCh38 reference assembly and decoys (partially assembled DNA sequences missing from the reference genome) using the Illumina Isaac Aligner (version

03.16.02.19). A quality threshold of ≥ 95% genome alignment at ≥15X with mapping quality > 10 for samples to be retained was set.

For the pilot arm of the 100KGP samples were aligned to NCBI GRCh37 reference, however, those patients in the pilot were mostly moved to build 38. For downstream analyses I only took those samples aligned to GRCh38, however, the clinical arm of 100KGP where diagnostic yields for the project are generated are calculated using probands aligned to either GRCh37 or GRCh38. In terms of coverage (the number of times a single base is read during sequencing) the 100KGP samples achieved 97.4% mean coverage at 15X with median genome-wide coverage of 39X. Samples with heterozygosity rates >2% (implying cross-contamination of samples) were removed (as determined by the VerifyBamID tool). Males and females were subset and analysed separately for sex chromosome quality control.

2.2.3 Variant calling

Variant calling was performed using Illumina's Starling software (version 2.4.7) for small SNVs and short insertions/deletions (INDELs). These were output to a genomic variant calling format file (gVCF). Starling uses a combination of read quality scores, allele counts to predict a genotype per locus before comparing it to a reference genome.

2.2.4 gVCF aggregation and variant-level quality control

Genomic variant call format files (gVCFs) were aggregated using gvcfgenotyper (Illumina, version: 2019.02.26) with variants normalized and multi-allelic variants decomposed using vt (version 0.57721). Variants were retained if they passed the following filters:

- missingness ≤ 5%
- median depth ≥ 10
- median GQ ≥ 15
- percentage of heterozygous calls not showing significant allele imbalance for reads supporting the reference and alternate alleles (ABratio) ≥ 25%
- percentage of complete sites (completeGTRatio) ≥ 50% and

- P value for deviations from Hardy-Weinberg equilibrium (HWE) in unrelated samples of inferred European ancestry ≥ 1×10-5.

HWE is the principal by which allele and genotype frequencies remain static between generations as long as mating is random and migration, mutation or selection do not occur. Variants that differ vastly by HWE normally represent genotyping or sequencing errors. However, HWE deviances can also represent population stratification or true associations. HWE is normally assessed separately in cases and controls to avoid removing true associations.

2.2.5 gVCF annotation

Annotation was performed using Variant Effect Predictor (VEP, version 98.2) (McLaren et al. 2016). Allele frequencies were annotated using gnomAD and TOPMed databases using both total population and ancestry specific values. Variants were further annotated with the Combined Annotation Dependent Depletion (CADD) scores (version 1.5) (Rentzsch et al. 2019), the loss-of-function transcript effect estimator (LOFTEE) tool (Karczewski et al. 2020) and SpliceAI splice site predictor tool (Jaganathan et al. 2019).

CADD incorporates more than 60 different annotations (including evolutionary constraint, epigenetic modifications, and functional predictions) into a machine learning model, generating a deleteriousness score for all ~9 billion potential coding and non-coding SNVs in the human genome (Rentzsch et al. 2019). A CADD PHRED adjusted score >20 for a variant means it is predicted to be in the top 1% damaging variants in the human genome. CADD scoring is a very popular method for variant deleteriousness calling and remains one of the top-performing and flexible tools (D. Wang et al. 2022) despite many other callers now being incorporated into Ensembl.

LOFTEE assesses variants that are stop-gained, splice site disrupting and frameshift variant only. It filters out variants based on sequence and transcript context (such as removing terminal truncation variants or well rescued splice variants) and flags exonic features such as conservation. It has been shown to effectively remove predicted loss of

function variants (pLoF) that are common in the population while retaining correctly ascertained pLoF variants (Karczewski et al. 2020). For these variants LOFTEE gives a flag indicating whether there is a “high confidence” (HC) or a “low confidence” that they cause pLoF.

SpliceAI is deep neural network that predicts cryptic splice mutations from genomic sequence data using an unsupervised deep learning model (Jaganathan et al. 2019). The output for each variant is a delta score ranging from 0-1 for each type of splice variant (donor loss, donor gain, acceptor loss, acceptor gain) with higher scores indicating a higher probability of the variant affecting splicing; a score >0.8 is used by the authors as a high precision cut-off.

2.2.6 Bioinformatics tools

The gVCF files were filtered using bcftools (version 1.11) (Danecek et al. 2021) and BEDtools (Quinlan and Hall 2010) in the command. Phenotype data including hospital episode statistics (HES) and human phenotype ontology data (HPO) was extracted from LabKey tables using the LabKey R package (Nelson et al. 2011). The outputs of the association analyses were manipulated, analysed and plotted in R (Version 4.0.3) using the data.table, tidyverse, qqman (D. Turner 2018) and ggplot2 packages (Wickham. 2016). Survival analysis was performed and plotted with the survival package in R (Therneau 2023).

2.3 Relatedness Estimation

Case-control analyses in genomics looks for shared areas of the genome, pre-defined at the point of testing e.g. SNV, gene, structural variant etc that are more common in either cases or control outputs a statistical probability as to the confidence of the association as well as an effect size as to the magnitude of the association. Related individuals share more common tracts of genomic information and if grouped together in such analyses lead to spurious associations and biased estimated of effect sizes if

unaccounted for. Common practice entails using “unrelated” individuals, usually defined as more distant than second-degree relatives.

Genetic relatedness can be ascertained using identify-by-descent (IBD), a concept that refers to the sharing of genetic material between two individuals inherited from a common ancestor. IBD assumes that individuals who are closely related are more likely to share longer segments of their DNA. The proportion of loci where a pair of individuals share 0,1 or 2 alleles from a common ancestor is calculated, with these estimated used to create a pair-wise kinship coefficient (Φ). The Φ is defined as the probability that a randomly selected allele from two individuals is IBD. A coefficient of 0.5 is equivalent to monozygotic twins, 0.25 to first-degree relatives and 0.125 to second-degree relatives.

Genomics England had generated a set of 127,747 high quality autosomal biallelic SNVs with a minor allele frequency (MAF) > 1% using PLINK (version 1.9) (Purcell et al. 2007). SNVs were included if they met the following criteria:

- missingness < 1%
- median GQ ≥ 30
- median depth ≥ 30
- AB Ratio ≥ 0.9
- completeness ≥ 0.9

SNVS that were ambiguous due to strand uncertainty were excluded. To prevent further confounding linkage disequilibrium (LD) pruning was performed using a squared correlation coefficient (r2) threshold of 0.1 and window of 500kb to remove correlated variants. Variants in regions of long-range high LD (https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD) were also removed. SNVs out of HWE in any of the African (AFR), East Asian (EAS), European (EUR) or South Asian (SAS) 1000 Genomes populations were also removed (pHWE <1×10-5).

With this pruned set of variants, I employed the KING-Robust algorithm (Manichaikul et al. 2010) to infer relationships in the presence of population substructure. KING generates pairwise kinship matrices, which I generated for cases and controls separately first. I then subset my sampled into unrelated individuals with a kinship coefficient threshold of 0.0884 (second degree relative). I then combined the case/control subsets and re-ran KING with the same threshold, removing controls that were related to the causes using a custom Python script (Mr. Catalin Voinescu, UCL).

2.4 Population Stratification

Removing close relatives from genetic analyses removes one source of bias however, population structure presents another similar challenge. Population stratification refers to the presence of systematic genetic differences between subpopulations within a larger population i.e., the distribution of genetic variants differs between subpopulations. These differences can arise due to various factors such as geographical isolation (with non-random mating), migration patterns, genetic drift (random fluctuations in the frequency of genetic variants or alleles within a population over successive generations), genetic admixture, gene flow (the transfer of genetic material from one population to another) and evolutionary processes. The result is false positive associations and inflated test statistics. Various statistical approaches have been developed to minimise confounding by population structure.

2.4.1 Genomic Control

A genome-wide inflation factor, often denoted as λ (lambda), is a statistical measure used in genome-wide association studies (GWAS) to assess and correct for potential inflation of test statistics due to population stratification or other sources of systematic bias. The inflation factor is a measure of the inflation of test statistics compared to what is expected under the null hypothesis of no association. If the test statistics are inflated due to population stratification or other sources of bias, the inflation factor will be greater than 1. A value of $\lambda = 1$ indicates no inflation, meaning that the test statistics follow the expected null distribution. The inflation factor can then be used to correct the

test statistics in the GWAS (Devlin and Roeder 1999). Genomic inflation under a collapsing rare variant model is less straightforward and further discussed in the collapsing rare variant section (3.3.2.4).

2.4.2 Principal component analysis (PCA)

Principal component analysis (PCA) helps to identify patterns, structure, and relationships within high-dimensional genetic data by reducing the dimensionality and visualizing the data in a more manageable form. In the context of genomics, PCA is often applied to genotype or gene expression data, where each individual or sample is represented by many variables (e.g., genetic variants or gene expression levels). By employing PCA, these high-dimensional datasets can be transformed into a lower-dimensional space while retaining the most important patterns of variation. The steps of PCA analysis are:

- Covariance Matrix: PCA calculates the covariance matrix from the data, in this case PLINK files containing sample and genomic variant data which quantifies the relationships and dependencies between the genetic. The covariance matrix captures the variance and co-variance of the variables in the dataset.

- Eigendecomposition: The covariance matrix is then eigendecomposed to obtain the eigenvectors (principal components) and eigenvalues. Each eigenvector represents a principal component, and the corresponding eigenvalue indicates the amount of variance explained by that component (Patterson, Price, and Reich 2006).

- Dimension Reduction: The eigenvectors are ranked based on their associated eigenvalues, and the top-ranked eigenvectors capture the most significant patterns of variation in the data. By selecting a subset of the top principal components, the dimensionality of the data is reduced.

To aid in interpretation this data is visualised in a scatter plot, where each individual is represented by its scores on the selected principal components. This visualization allows for the identification of clusters, outliers, and patterns of genetic similarity or dissimilarity among individuals. This is particularly useful for detecting population stratification or genetic ancestry differences in genomic datasets. It can reveal underlying genetic substructure or relationships between populations, which is crucial for controlling population stratification in genetic association studies.

Usually, the top ten PCs are included as fixed (non-random) effects in the regression model of an association analysis to control for population stratification. However, it should be noted that PCA is less reliable in small sample sizes or when estimating population substructure (Stoltzfus 2011; Johnstone and Lu 2009).

2.4.3 Linear mixed models (LMM)

LMMs (also known as a mixed effect model) now play an integral role in accounting for population stratification in genetic association studies and can be used on both continuous and binary traits (if using binary input it is known as logistic mixed model) (Z. Zhang et al. 2009; Dandine-Roulland and Perdry 2015; G. Li and Zhu 2013).

LMM is a statistical modelling approach that incorporates both fixed effects and random effects into the analysis. The response variable is modelled as a linear combination of fixed effects and random effects, along with an error term. Fixed effects (covariates) represent the systematic or non-random factors that influence the response variable. Fixed effects can be categorical (e.g., treatment groups, sex) or continuous (e.g., principal components). The coefficients associated with the fixed effects estimate the relationship between the covariates and the response variable. Random effects capture the variability due to factors that are not of primary interest but are still important to account for. Random effects account for correlation or clustering within the data and are typically used to model the hierarchical or nested structure of the data. In genetic studies, random effects can account for the genetic relatedness between individuals or clustering within families and are calculated via a genomic relationship

matrix (GRM). The random effects are assumed to follow a specific probability distribution, often a multivariate normal distribution. Finally, the error term in an LMM accounts for the residual variation that cannot be explained by the fixed and random effects. It represents the within-group or within-subject variability that is not accounted for by the model. The error term is assumed to follow a normal distribution with mean zero and constant variance.

LMMs are flexible and can handle unbalanced or missing data, accommodate different data structures (e.g., repeated measures, nested designs), and provide estimates of both fixed and random effects, along with associated uncertainty measures (e.g., standard errors, confidence intervals). LMMs are commonly estimated using maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML). It should be noted that LMMs are computationally intensive but allow for the accounting of inter-and intra-population structure and cryptic relatedness.

2.4.4 Control of population structure

Published methodology from our group has shown that incorporating two of the above approaches controls confounding from population structure in a mixed ancestry case/control population (Chan et al. 2022). The first method is using a matching algorithm that matches cases to controls within a distance threshold as calculated using the first ten principal components (generated with PLINK using the 127,747 high quality autosomal biallelic SNVs with MAF > 1%) weighted by the percentage of genetic variation explained by each component (Figure 2-1 for an example from CyKD). Only controls within a specified distance of a case were included, with each case having to match a minimum of two controls to be included in the final cohort.

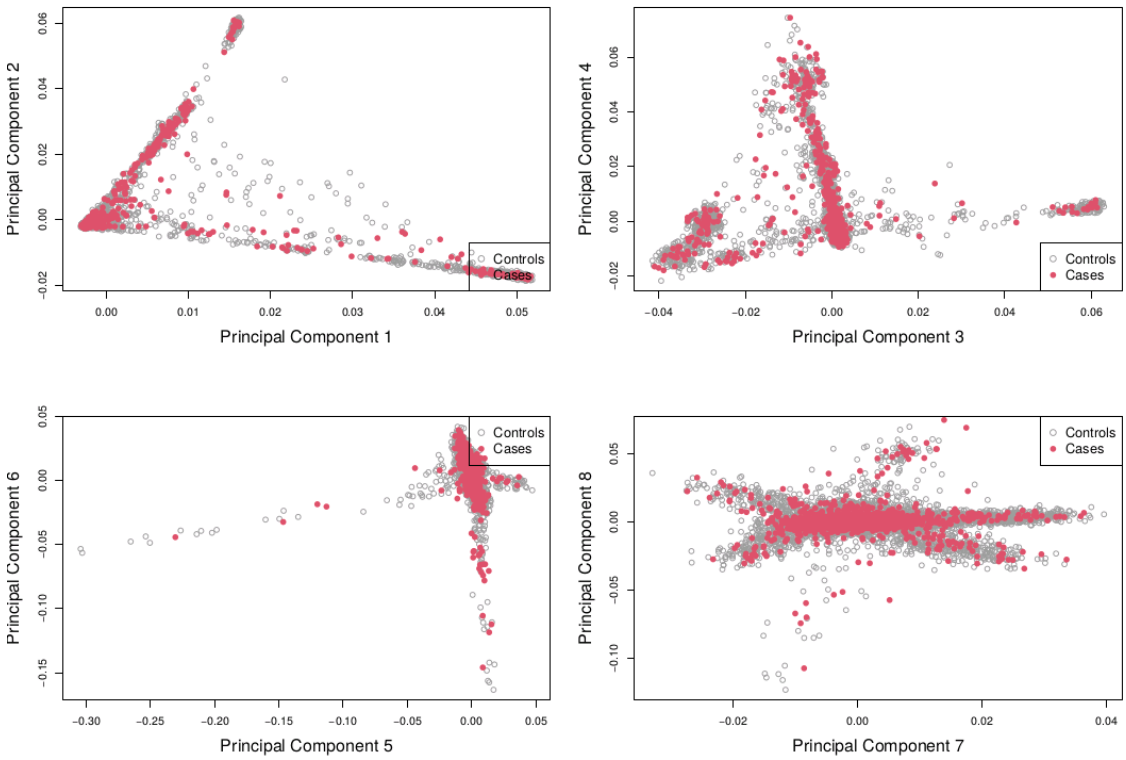

Figure 2-1 Principal component matching

Principal component analysis showing the first eight principal components for CyKD cases (red) and controls (white) prior to ancestry matching (1294 cases and. 27660 controls).

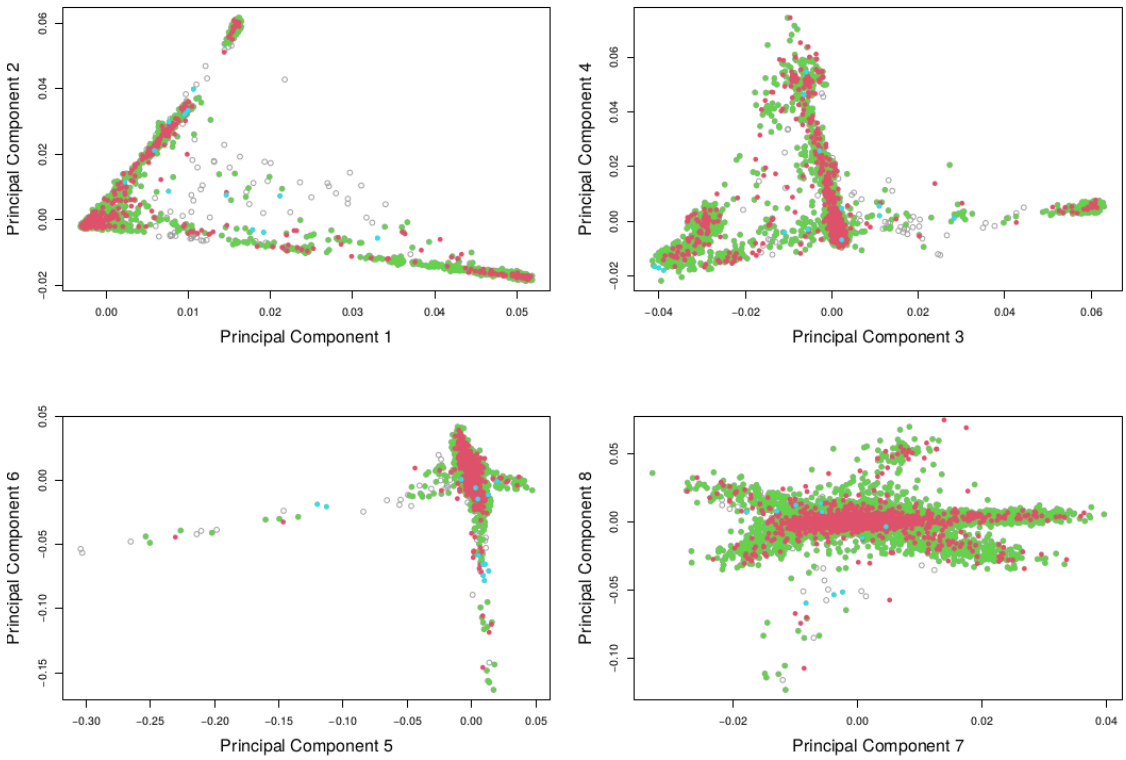

Figure 2-2 Ancestry Matching

Principal component analysis showing the first eight principal components for matched cases (red) and controls (green) and unmatched controls (grey) in a CyKD cohort. This highlights that cases are taken from multiple different ancestries with the appropriate matched controls. After ancestry matching there were 1209 cases to 29096 controls.

Secondly, a logistic mixed model was implemented using SAIGE (W. Zhou et al. 2018) and SAIGE-GENE (W. Zhou et al. 2022). This in addition to ancestry matching allowed for further control of population structure and cryptic relatedness.

2.5 SAIGE

The Scalable and Accurate Implementation of Generalized mixed model or SAIGE and its extension SAIGE-GENE have been developed to deal with the increasing challenges in running association testing in large biobank scale genomic datasets. It is computationally intensive running mixed models on such datasets and controlling type 1 error is challenging in unbalanced case control ratios (roughly greater than 100 controls per case). These tools have now become the standard for genome-wide single variant and exome-wide region-based association testing in large cohorts (W. Zhou et al. 2018, 2022). Given the importance of SAIGE in this thesis I have broken down its key features below:

2.5.1 Generalised logistic mixed model

SAIGE uses a logistic mixed model written as:

$$logit(\mu_i) = X_i\alpha + G_i\beta + b_i$$

where μ_i is the probability of individual i being affected by the disease or trait in question based on their covariates, genotype, and random effects. X_i is a vector of covariates (e.g., sex and top ten principal components), α is a vector of fixed covariate effects including the intercept, G_i is the matrix of allele counts (0,1,2) for each qualifying variant and β is the fixed genotype effect. b_i is a vector of random effects that incorporates relatedness (and consequently population structure) between individuals estimated using an N x N GRM. SAIGE wraps this methodology around optimised computational strategies to reduce the cost of fitting null logistic mixed models, making it ideal for large scale biobanks.

2.5.2 Saddlepoint approximation

In unbalance case-control cohorts (roughly greater than 100 controls to 1 case) there is not a normal distribution of test statistics for single variants leading to inflated type 1 error rates. In order to control for this SAIGE utilises saddlepoint approximation (SPA) (Dey et al. 2017). SPA is a mathematical technique used to approximate the distribution of a random variable when its exact distribution is difficult to determine analytically. It is particularly useful when dealing with complex distributions, such as those encountered in genomic association tests. In some scenarios, the null distribution of the test statistic is not readily available in closed form. This can be the case when the sample size is small or when the distribution of the test statistic is complicated. SPA provides an efficient and accurate method to estimate the null distribution and compute p-values in such situations.

SPA involves finding the saddlepoint of a Laplace-type integral equation, which is a point in the domain of the characteristic function of the random variable where the integral equation is satisfied. The saddlepoint approximation constructs an asymptotic expansion around this saddlepoint, allowing for the estimation of the tail probabilities of

the distribution. This approach is particularly effective for approximating the tails of distributions, which is essential for calculating p-values.

In genomic association tests, saddlepoint approximation can be applied to compute accurate p-values for test statistics under various null distributions, such as the chi-square distribution or the logistic distribution. By accurately estimating the null distribution, one can determine the statistical significance of genetic associations and make more reliable inferences about the relationship between genetic variants and traits or diseases.

However, when variants have a minor allele count (MAC) < 10, considered to be “rare”, then SPA loses accuracy. SAIGE-GENE tunes this signal by employing efficient resampling methods to further control for type 1 error rates (Seunggeun Lee et al. 2016). Efficient resampling refers to methods such as bootstrapping or permutation tests that involve generating multiple resamples or permutations from the observed data to assess the sampling variability and make statistical inferences. In the case of SAIGE-GENE, permutation testing is performed only in those individuals carrying the minor allele to estimate the sampling distribution and generate an empirical *P* value.

2.5.3 Workflow

There are two main steps behind SAIGE and SAIGE-GENE:

1. Variance component estimation using a generalized linear mixed model (GLMM): The first step involves fitting a null GLMM using sex and the first ten principal components without the genetic variants (fixed effects). Next a GRM is constructed using variants with a MAF>1% with the variance components used as random effects. This account for both genetic relatedness and population structure.
2. Score test for association analysis: After estimating the variance components, the second step involves performing association tests to assess the significance of genetic variants. SAIGE and SAIGE-GENE use a score test, which is a

variant of the standard likelihood ratio test (LRT). The score test compares the likelihood of the model under the null hypothesis (no association) to the likelihood under the alternative hypothesis (presence of association). The test statistic is derived from the score vector, which represents the derivative of the log-likelihood function with respect to the variant effect size. The saddlepoint approximation is used to account for case-control imbalance.

Whilst SAIGE has been widely adopted, there exist several relevant limitations. With any use of logistic regression, if the event rate is low then estimations of effect size (β) can be inaccurate; this holds true for rare variants and the authors of SAIGE now recommend Firth logistic regression be used instead in such scenarios. Secondly, SAIGE has been shown to be slightly conservative when case-control ratios are very unbalanced.

2.6 Power

For single-variant association analysis statistical power was calculated using the R package genpwr (Moore, Jacobson, and Fingerlin 2019) assuming an additive model and a $P<5x10^{-08}$, the standard genome-wide significance threshold. Figure 2-3 illustrates the power for the GWAS at different allele frequencies and odds ratios (OR) for the CyKD cohort. At an allele frequency of 1% single variant association testing is well powered (>80%) to detect alleles with an OR >3. USD and EEHTN are discussed in more detail in their respective chapters.

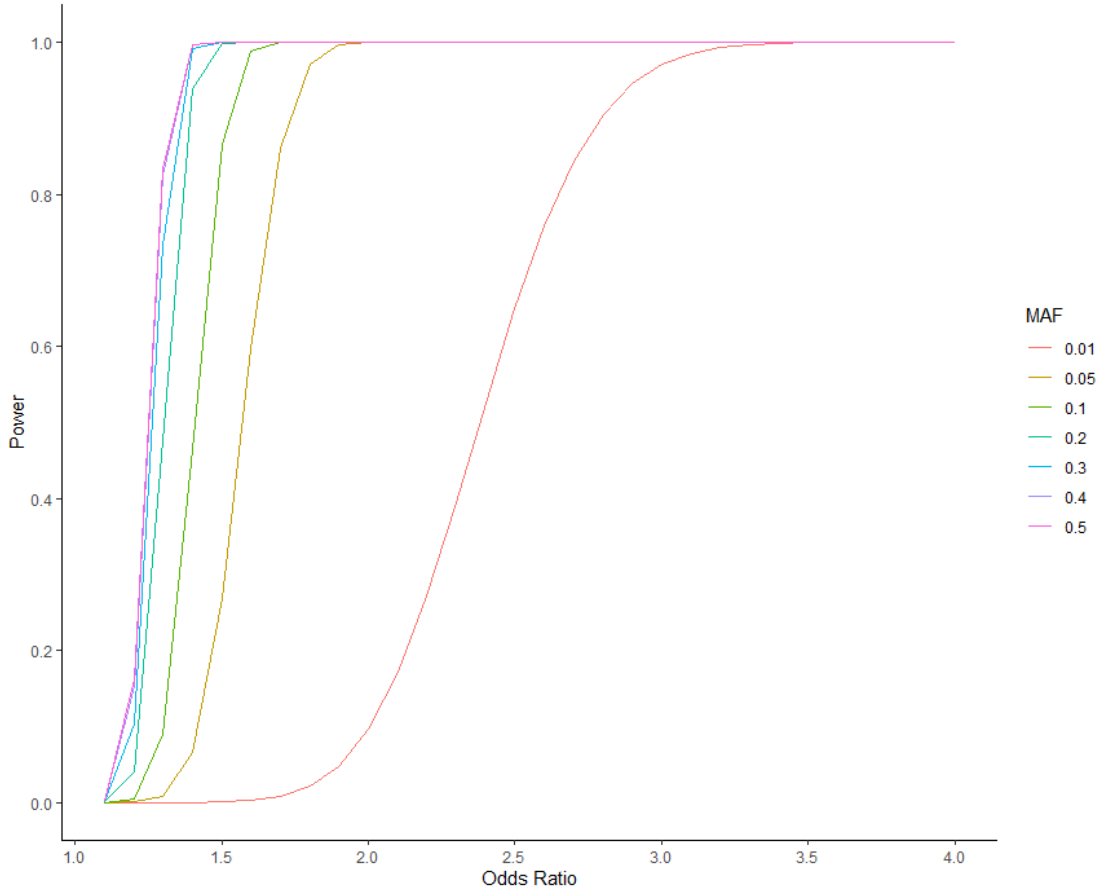

Figure 2-3 Statistical power of CyKD GWAS

Power to detect single variant association under an additive model for 1209 CyKD cases and 26096 controls at a genome-wide significance threshold of $5x10^{-8}$. MAF, minor allele frequency.

For region-based association testing, establishing power is more challenging due to the myriad parameters that need to be accounted for such as allele frequency and effect sizes of individual variants. PAGEANT (Derkach, Zhang, and Chatterjee 2018) was developed to aid in calculating power for gene-based collapsing tests by using distributions derived from the precursor to the gnomAD database, ExAC (Lek et al. 2016). PAGEANT was used to calculate the minimum proportion of cases explained by a single gene detected with 80% power in the rare variant analyses (discussed further in chapter 3), assuming 80% of the qualifying variants used for the collapsing test were causal. The genome-wide threshold used was $P<2.5x10^{-06}$; the less stringent p-value reflecting a Bonferroni correction per gene rather than per SNV.

2.7 Statistical Significance

The more tests one carries out, the more likely one is to see a statistically significant result by chance. This means that the chances of rejecting the null hypothesis when it is true also increases. This is Type 1 error and require careful consideration in the context of genome-wide association testing where millions of independent tests are carried out. I will discuss some of the current thinking behind controlling for multiple testing below:

2.7.1 Bonferroni correction

If α is the desired significance level, usually 0.05, and *n* is the number of independent tests, then a Bonferroni correction can be represented as α/*n*. This is widely used in the genomics community and represents the most stringent of type 1 error control methods. The underlying assumption that every variant tested is independent does not always hold true and this method is weighted in favour of minimising false positives (type 2 error) at the expense of potentially missing real signals (Devlin and Roeder 1999).

2.7.2 False discovery rate

False discovery rate (FDR) is recently gained more favour in the genomics community as an alternative to Bonferroni correction. FDR determines a proportion of false positive that are acceptable within the significant results, typically 1% or 5% (Benjamini and Hochberg 1995). FDR has an increased type 1 error rate but greater power for signal detection. They are often used for hypothesis-generating where the results will not directly impact patient care.

2.7.3 Permutation testing

Rather than assuming an underlying distribution, permutation testing calculates a distribution for the test statistic under the null hypothesis in order to give an empirical *P* value. By randomly rearranging the data and recalculating the test statistic of interest, a null distribution of the test statistic is calculated (Salmaso et al. 2011). This is then compared to the observed test statistic to determine the p-value. Permutation testing is

computationally intensive but particularly useful when the underlying assumptions required by traditional parametric tests are not met or when the sample size is small.

2.7.4 Bayesian approaches

Bayesian methods provide an alternative approach to determining statistical significance compared to classical frequentist methods. In Bayesian statistics, statistical significance is typically expressed in terms of posterior probabilities or credible intervals rather than p-values. Bayesian methods use prior probabilities to fine tune and generate a conditional probability using the observed data. The Bayes factor quantifies the relative strength of evidence for one hypothesis compared to another. It represents the ratio of the likelihood of the data under one hypothesis to the likelihood under an alternative hypothesis, after considering prior beliefs. A Bayes factor greater than 1 indicates evidence in favour of the hypothesis in the numerator, while a value less than 1 favours the hypothesis in the denominator. The strength of evidence can be interpreted using widely accepted guidelines but as of yet has not been widely adopted by the genomics community (Fernando and Garrick 2013).

Bayesian methods have found more favour in generating “credible sets” of variants that make up a significant GWAS signal, rather than the traditional p-value thresholding as it allows for a more nuanced interpretation of the results by providing a range of plausible effect sizes or variants rather than a binary significant/not significant determination. It also enables researchers to quantify and compare the evidence for different variants or effect sizes, aiding in prioritizing follow-up investigations.

2.7.5 Significance thresholds

Bonferroni correction was selected as the p-value adjustment methods throughout this thesis as my aim was to replicate any significant findings in an independent cohort such as the UK Biobank. In order to do so I wanted to make sure any significant findings were as robust as possible and thus I wanted to minimise noise.

Single variant association analysis uses a genome-wide significance threshold of $5x10^{-08}$, a figure derived from the International HapMap Consortium based on estimates of the number of common independent variants ($r2 < 0.8$) with MAF > 5% in a European ancestry population (~1 million). Given I used a lower MAF of 1% and have access to WGS data and therefore test more variants than the HapMap consortium there is an argument to be made that a lower p-value threshold is applicable. However, this has yet to be implemented in the genomics community.

For gene-level rare variant association analyses $P<2.5x10^{-06}$ is the exome wide significance: 0.05 Bonferroni corrected for the number of protein coding genes in the human genome ($\alpha = 0.05/\sim19,000$).

2.8 Summary

In this chapter I have discussed how the central data used for this thesis was constructed as well as the underlying theories behind the statistical genetic approaches. In the subsequent chapters I will go into more detail specific to the results presented per phenotype.

Chapter 3. Cystic kidney disease

Cystic kidney disease (CyKD) is a catch all term encompassing a wide group of diseases with differing causes that all involve the formation of fluid filled cysts in one or both kidneys. CyKD can present at any point in life and the many causes are usually distinguished by their respective clinical feature, imaging characteristics, cyst distribution and whether extra-renal features are present. However, in the age of genomic testing becoming more widely available, a molecular approach as the first step in diagnosis is increasingly popular and has led to a better understanding of CyKD pathogenesis and improved diagnostic accuracy. To this end the causes of CyKD can be divided into hereditary and non-hereditary causes and are detailed in the Table 3-1 below. I will give more in-depth analysis of the causes of CyKD that are directly relevant to this thesis, namely dominant and recessive causes of CyKD as well as *HNF1β* -related CyKD; other relevant and important genes and pathways involved in CyKD will be referred to in the "pathophysiology of cyst formation" section.

Table 3-1 Causes of cystic kidney disease.

Gene	Disease	Renal Phenotype	Extra-renal phenotype	Mode	OMIM#	Reference
ADPKD						
PKD1 (truncating)	ADPKD-*PKD1*	Bilateral kidney cysts, renal enlargement, median age ESRF ~55 years	PLD, ICH, heart valve abnormalities, aortic root dilatation, hernias, diverticular disease, cysts in other organs	AD	173900	Harris *et al.* 1994
PKD1 (non-truncating)		Bilateral kidney cysts, renal enlargement, median age ESRF ~67 years			173900	Harris *et al.* 1994
PKD2	ADPKD-*PKD2*	Bilateral kidney cysts, renal enlargement, median age ESRF ~79 years			613095	Mochizuki *et al.* 1996
GANAB	ADPKD-GANAB	Bilateral cysts, preserved renal function	PLD		600666	Porath *et al.* 2016
DNAJB11	ADPKD-DNAJB11	Multiple small cysts with normal/small kidneys, possible ESRF after 60 years	PLD		618061	Cornec-Le Gall *et al.* 2018
ALG5	ADPKD-ALG5	Interstitial fibrosis with non-enlarging cystic kidneys, possible ESRF after 60 years	Rarely mild PLD		620056	Lemoine *et al.* 2022
ALG8	ADPKD-ALG8	Bilateral kidney cysts, nephrolithiasis	None to date		Pending	Apple *et al.* 2023
ALG9	ADPKD-ALG9	Moderate bilateral kidney cysts, rarely progressing to ESRF	Rarely mild PLD		Pending	Besse *et al.* 2019
IFT140	ADPKD-IFT140	Bilateral enlarging kidney cysts with ESRF comparable to *PKD2*	Rarely mild PLD		Pending	Senum *et al.* 2022
ADTKD						
HNF1B	ADTKD-*HNF1B*	Bilateral kidney cysts in ~45% of affected individuals, ESRF highly variable	Diabetes, gout, hyperuricaemia, hypomagnesaemia, elevated liver enzymes, bicornate	AD	137920	Bingham *et al.* 2001

			uterus, solitary kidney			
MUC1	ADTKD-*MUC1*	Normal to small-sized kidneys, ~50% small renal cysts; variable progression to ESRF in adulthood	Gout		174000	Kirby *et al.* 2013
SEC61A1	ADTKD-*SEC61A1*	Normal or small-sized kidneys, ~50% small bilateral renal cysts	Intrauterine growth retardation, neutropenia, anaemia (congenital)		617056	Bolar *et al*. 2016
UMOD	ADTKD-*UMOD*	Normal to small-sized kidneys, 1/3 small kidney cysts (uni/bilateral), variable ESRF in adulthood	Gout		162000	Dahan *et al.* 2003
ADPLD						
PRKCSH	ADPLD	Occasional kidney cysts	PLD	AD	174050	Li *et al*. 2003
SEC63		Occasional kidney cysts	PLD		617004	Davila *et al.* 2004
ALG8		Occasional kidney cysts	PLD		617874	Besse *et al.* 2017
LRP5		Occasional kidney cysts	PLD		617875	Cnossen *et al.* 2014
ARPKD						
PKHD1	ARPKD	Antenatally enlarged hyperechogenic kidneys; multiple bilateral small cysts; 50% ESRF within first 10 years, milder presentation associated with increased age of diagnosis	Congenital hepatic fibrosis with associated portal HTN, Caroli syndrome, small liver cysts in heterozygous patients	AR	263200	Onuchic *et al. 2002*
DZIP1L	ARPKD	Antenatally enlarged hyperechogenic kidneys; multiple bilateral small cysts; variable ESRF in second and third decade of life	None to date		617610	Lu *et al.* 2017
PMM2	Hyperinsulinaemic hypoglycaemia with PKD	Antenatally enlarged hyperechogenic kidneys; multiple bilateral small cysts; variable ESRF	Hyperinsulinaemic hypoglycaemia; occasional PLD		Pending	Cabezas *et al.* 2017

Collagenopathies						
COL4A3	Alport Syndrome	Occasional kidney cysts. Thinning of basement membrane with microscopic haematuria and progressive ESRF.	Sensorineural deafness, anterior lenticonus, perimacular flecks	AR	203780/104200	Mochizuki et al. 1994
COL4A4		Occasional kidney cysts. Thinning of basement membrane with microscopic haematuria and progressive ESRF.	Sensorineural deafness, anterior lenticonus, perimacular flecks	AR	203780/141200	Mochizuki et al. 1994
COL4A5	X-linked Alport Syndrome	Occasional kidney cysts. Thinning of basement membrane with microscopic haematuria and progressive ESRF.	Sensorineural deafness, anterior lenticonus, perimacular flecks	XLD	301050	M'Rad et al. 1992
Syndromic forms of CyKD						
TSC1 or *TSC2*	Tuberous sclerosis	Multiple and bilateral angiomyolipomas and kidney cysts; kidney function usually preserved; possible evolution to ESRF, contiguous gene deletion of *TSC2* and *PKD1* leads to severe CyKD with ESRF <30 years	CNS (cortical tubers, astrocytomas, epilepsy, and intellectual disabilities); skin lesions (facial angiofibromas and hypopigmented spots); pulmonary lymphangioleiomyomatosis; cardiac rhabdomyoma and retinal hamartoma; PLD in contiguous deletion	AD	191100/613254	Kandt *et al.* 1992
VHL	Von Hippel-Lindau disease	Bilateral kidney cysts, renal cell carcinoma	Hemangioblastomas of the retina, spine, or brain; pheochromocytoma; neuroendocrine	AD	193300	Carsillo *et al.* 2000

			tumour of the pancreas			
COL4A1	HANAC syndrome or *COL4A1*-related disease	Bilateral kidney cysts, ESRF in later adulthood	Microscopic haematuria, aneurysms, muscle cramps, elevated creatine phosphokinase, tortuosity of the retinal arteries	AD	611773	Plaisier *et al.* 2007
OFD1	Oro-facial-digital syndrome type 1	X-linked, embryonically lethal in boys, CyKD in women	Cleft palate, facial dysmorphia; syndactyly, clinodactyly, or polydactyly; PLD	XLD	311200	Ferrante *et al.* 2001
FLCN	Birt-Hogg-Dubé syndrome	Kidney cysts and kidney tumours	Hair follicle hamartomas, lung cysts with spontaneous pneumothorax	AD	135150	Nickerson *et al.* 2002
NPHP1-6	Nephronophthisis/Joubert/Senior-Løken syndrome	Bilateral kidney cysts	Retinal degeneration, polydactyly, liver disease, severe CNS disease	AR	PS256100	Review: Wolf *et al.* 2011
BBS1-12	Bardet-Biedl syndrome	Broad range of structural kidney issues including unilateral or bilateral cysts	Cone-rod dystrophy, obesity, polydactyly, cognitive impairment, hypogonadism, neurological issues, olfactory dysfunction, diabetes	AR	PS209900	Review: Florea *et al.* 2021

MGS1-6	Meckel-Gruber syndrome	Bilateral cysts	Encephalocele with CNS involvement, hepatic fibrosis, polydactyly, genitourinary malformation	AR	PS249000	Review: Hartill *et al.* 2017
Acquired cystic renal disease						
N/A	Multiple benign simple cysts	Multiple benign simple cysts - more common with increasing age	N/A	N/A	N/A	N/A
N/A	Acquired kidney cystic disease	CKD associated especially with patient on renal replacement therapy. Usually small and bilateral.	N/A	N/A	N/A	N/A
N/A	Lithium induced kidney cysts	Normal/small kidneys with small bilateral cysts - history of lithium exposure, interstitial fibrosis	N/A	N/A	N/A	N/A

AD – autosomal dominant, AR – Autosomal Recessive, XLD – X-linked Dominant, PLD – Polycystic Liver Disease, ESRF – End Stage Liver Failure, CNS – Central Nervous System, CKD – Chronic Kidney Disease, ICH – Intracerebral haemorrhage

3.1 Introduction to CyKD

3.1.1 Autosomal dominant polycystic kidney disease

Autosomal dominant polycystic kidney disease (ADPKD) is the most common monogenic cause of renal failure worldwide and one of the commonest single-gene disorders generally (Bergmann et al. 2018). Present in roughly 1:400-1:1000 live births, it is present in equal distribution worldwide and is a huge health burden, representing roughly 10% of all patients receiving renal replacement therapy (RRT) (K. Evans et al., 2018). It is typically late onset and is multi-system, characterised by bilateral real cysts, liver cysts, and an increased risk of intracranial aneurysm and haemorrhage. Extra-renal manifestations include cysts in other organs such as the pancreas and seminal vesicles, aortic root dilatation, mitral valve prolapse and abdominal wall hernias. Clinically this can present as early onset hypertension, flank pain and eventually renal failure. 50% of ADPKD patients reach end stage renal failure by 60 years old (Cornec-Le Gall, Alam, and Perrone 2019). There is, however, substantial phenotype variability between patients even within families suggesting either secondary genetic effects and/or environmental factors play an important role in disease modulation (Harris and Rossetti 2010).

Diagnosis is made based on imaging criteria or genetic testing confirming the presence of a heterozygous variant in one of the known pathogenic genes, predominately *PKD1* or *PKD2* or the less common and more recently discovered genes *GANAB, ALG5,ALG8 ALG9, DNAJB11* and *IFT140* (Hughes et al. 1995; The European Polycystic Kidney Disease Consortium. 1994; T. Mochizuki et al. 1996.; Apple et al. 2023; Lemoine et al. 2022; Porath et al. 2016; Cornec-Le Gall et al. 2018; Senum et al. 2022) with roughly 5% of cases remaining unsolved (Bergmann et al. 2018). Genotype-phenotype correlations are described in more detail in Table 3-2

Table 3-2 Genotype-phenotype correlation of the causes of ADPKD

Gene	ADPKD attributable to gene	Protein	Renal phenotype	Liver phenotype
PKD1	78%	Polycystin-1	Truncating: Innumerable bilateral kidney cysts leading to progressive kidney enlargement, median age of ESRF about 55 years	Polycystic liver disease, mild to severe
			Non-truncating: Innumerable bilateral kidney cysts leading to progressive kidney enlargement, median age of ESRF about 67 years	
PKD2	15%	Polycystin-2	Innumerable bilateral kidney cysts leading to progressive kidney enlargement, median age of ESRF about 79 years	Polycystic liver disease, mild to severe
ALG5	<0.5%	Dolichyl-phosphate beta-glucosyltranseferase	Non-enlarging cystic kidneys with some interstitial fibrosis. ESRF in those greater than 65 potentially	Polycystic liver disease, absent or mild
ALG8	<0.5%	Alpha-1,3-glucosyltransferase	Bilateral cysts with normal kidney size, nephrolithiasis	Some liver cysts but not unique

ALG9	<0.5%	Alpha-1,2-mannosyltransferase	Moderate number of bilateral cysts. Rarely progresses to ESRF	Polycystic liver disease, absent to mild
DNAJB11	<0.5%	DnaJ homolog subfamily B member 11	Normal to small kidneys, small cysts, potential evolution to ESRF after 60 years	Polycystic liver disease, absent to moderate
GANAB	<0.5%	Neutral alpha-glucosidase AB	Bilateral cysts, normal renal function	Polycystic liver disease, mild to severe
IFT140	1-2%	Intraflagellar transport protein 140 homolog	Bilateral renal cysts, mild effect on renal function akin to non-truncating *PKD2* variants	Occasional polycystic liver disease

ADPKD – Autosomal dominant polycystic kidney disease

Until recently, treatment of ADPKD centred on the management of symptoms secondary to renal cyst formation and chronic kidney disease. However, with the approval for the use of Tolvaptan (Torres et al. 2012), there is now a treatment designed to retard disease progression, with many novel compounds currently going through clinical trials (J. X. Zhou and Torres 2023). Treatment initiation is now focused on those patients with rapidly progressive disease as evidenced by several factors including rate of kidney decline, rapidity of cyst growth, family history of ESRF and crucially genotype, with truncating *PKD1* variants requiring treatment initiation early (Cornec-Le Gall et al. 2016). This has given added impetus to the need to molecularly screen individuals.

3.1.2 Autosomal recessive polycystic kidney disease

Autosomal recessive polycystic kidney (ARPKD) is a severe disorder occurring in 1 in 20,000 births. It causes severe dilatation of the kidney collecting ducts and malformation of the portobiliary system. Often diagnosed in utero or at birth, the patients suffer from large echogenic kidneys leading to poorly functioning kidneys and consequent oligohydramnios. Perinatal mortality is roughly 30%, with children that survive mostly reaching ESRF by adulthood (42% renal survival by 20 years old) (Bergmann et al. 2018). Nearly all patients suffer a gamut of issues related to renal failure, portal hypertension and biliary failure. Patients who are diagnosed later tend to have a better renal prognosis. Later presentations of the disease have phenotypic overlap with ADPKD and can lead to diagnostic misclassification (Sekine et al. 2022).

ARPKD is predominately caused by variants in the polycystic and hepatic disease gene 1 (*PKHD1)* and codes for the fibrocystin-polyductin complex (FPC) (L. F. Onuchic et al. 2002). Most affected patients are compound heterozygotes. Management is largely supportive with no dedicated treatments at present.

Other recessive cystic diseases include *PMM2* associated hyperinsulinaemic hypoglycaemia with PKD and *DZIP1L* associated ARPKD (Cabezas et al. 2017; H. Lu et al. 2017). Antenatal enlarge hyperechogenic kidneys and bilateral small cysts are present in both. In *PMM2* associated disease there are small liver cysts and hyperinsulinaemic hypoglycaemia whilst *DZ1P1L* disease has no associated extra-renal manifestations.

Treatment is supportive, with genetic testing allowing for a molecular diagnosis and genetic counselling.

3.1.3 *HNF1β* associated cystic renal disease.

Broad renal involvement is now seen as one of the earliest manifestations of *HNF1β* associated disease. Various phenotypes have been attributed to *HNF1β*, classically

starting with maturity onset diabetes of the young (MODY) coupled with renal cysts leading to the term “renal cysts and diabetes syndrome” (Horikawa et al. 1997). However, there are now over 10 different renal pathologies associated with *HNF1β*-nephropathy (Izzi et al. 2020). This is mirrored in the ever-expanding list of extra-renal side phenotypes such as exocrine pancreatic failure, liver function abnormalities, gout, and genital tract malformations.

HNF1β-codes hepatocyte nuclear factor homeobox B found on chromosome 17q12 and plays an integral role in early embryonic development. Its protein product, transcription factor-2 (TCF2) is a necessary component in tissue specific gene expression in many epithelial tissues including kidney, pancreas, liver and genitourinary tract (Kolatsi-Joannou et al. 2001; Ferrè and Igarashi 2019).

Structural variants involving *HNF1β* are of note. In ~45% of cases with *HNF1β* variants, a whole gene deletion of *HNF1β* occurs as part of the 17q12 deletion syndrome, causing a multi-system disorder with renal involvement (OIM #614527) (Mitchel et al. 2016). The other cases are mainly heterozygous SNVs (Fokkema et al. 2011). There is little correlation between phenotype and genotype but large cohorts studying this condition are lacking (Nagano et al. 2019; Dubois-Laforgue et al. 2017) with none assessing the burden of *HNF1β* at a genome wide level using WGS.

There are no specific treatments for this condition bar supportive care and surveillance for multi-organ involvement in patients with 17q12 deletions.

3.1.4 Pathophysiology of cyst formation from genetic insights

3.1.4.1 ADPKD

PKD1 is a large gene with 46 exons, of which the first 34 are homologous (and therefore very similar in sequence) to several nearby pseudogenes as well as being GC rich making sequencing challenging. *PKD2* is much smaller (15 exons) and is therefore easier to sequence. Their respective discoveries in 1995 and 1996 have led to the development of a polycystin model of cyst formation.

The polycystins 1 and 2 coded (PC1 and PC2) by *PKD1/2* respectively are found predominantly in the primary cilium although are expressed in epithelial cells, vascular smooth muscle, cardiac myocytes as well as other locations (A. C. Ong 2000). Within the cells PC1 is found throughout lateral membrane junctions, focal adhesions, apical vesicles and primary cilia whereas PC2 is mainly found at the endoplasmic reticulum although the two proteins do also co-express (A. C. M. Ong and Harris 2005) . PC1 is a 4303 amino acid membrane bound protein with 11 transmembrane domains, a large extracellular domain and a ~200 amino acid intracellular carboxy-terminal tail thought to be integral in the regulation in multiple signalling cascades (Harris and Torres 2014). The cleavage of PC1 at its G protein couple receptor regulates biogenesis and trafficking of PC1 (Kurbegovic et al. 2014) as well as modulating signalling pathways via the release of the intracellular C-terminal tail, freeing PC1 fragments into the cytoplasm and nucleus (Y. Xu et al. 2016). PC2 is less than 968 amino acids and has six transmembrane spanning domains acting as a calcium permeable channel, it sits within the transient receptor potential (TRP) family and in isolation forms a tetrameric channel structure with a pore loop and voltage sensing domain (Shen et al. 2016).

PC2 co-localises with PC1 within the primary cilia shaft and basal body in renal epithelia (Geng et al. 1997), and their correct localisations and function are dependent on both elements functioning correctly (H. Kim et al. 2014; Cai et al. 2014). The C-terminal tail of PC1 facilitates the interaction between PC1 and PC2 (Tsiokas et al. 1997), which together act as an ion transporter involved in calcium signalling but many functions of both PC1 and PC2 remain unclear. The structure of a modified PC1-PC2 complex was solved by cryo-electron microscopy in 2018, revealing a structure (1:3 PC1:PC2) resembling a TRP with a novel pore like structure in which the C-terminal domain of PC1 contributes one side of the tetrameric channel (Su et al. 2018). This asymmetric pore loop structure makes it very different to TRP channels as it potentially ameliorates the cation selectivity of the polycystin channel, explaining why electrophysiology to date have found it difficult to reach a consensus on the cation selectivity of the channel (Delling et al. 2016). A ligand for this polycystin complex is yet to be elucidated with a recent study suggesting the cleaved N-terminus of PC1(a mutational hotspot) as a candidate (Ha et al. 2020). This last point is of note as Su et al

were unable to include the entire extracellular N terminus of PC1 in their structure due to the protein being too large and unstable to analyse.

Given the similarities in phenotype between ADPKD phenotypes caused by *PKD* variants and the physical proximity of PC1 and PC2 in cells, common signalling pathways have been sought. Cyclic AMP (cAMP), mammalian target of rapamycin complex 1 (mTORC1), extracellular signal-regulated kinases (ERK), 5' AMP-activated protein kinase (AMPK) and JAK-STAT have all been shown to be affected by aberrant polycystin functioning (Harris and Torres 2014). cAMP in particular has been targeted for downregulation via vasopressin receptor 2 antagonism using Tolvaptan, successfully retarding cyst growth and disease progression. These disrupted signalling pathways have then been postulated to cause cyst formation and growth vial clonal expansion of epithelial cells, alterations in apical-basal polarity, planar cell polarity, increased extracellular matrix production and cellular metabolism creating a snowball effect in which the secondary events take on an increasing role in cyst formation and growth (Figure 3-1).

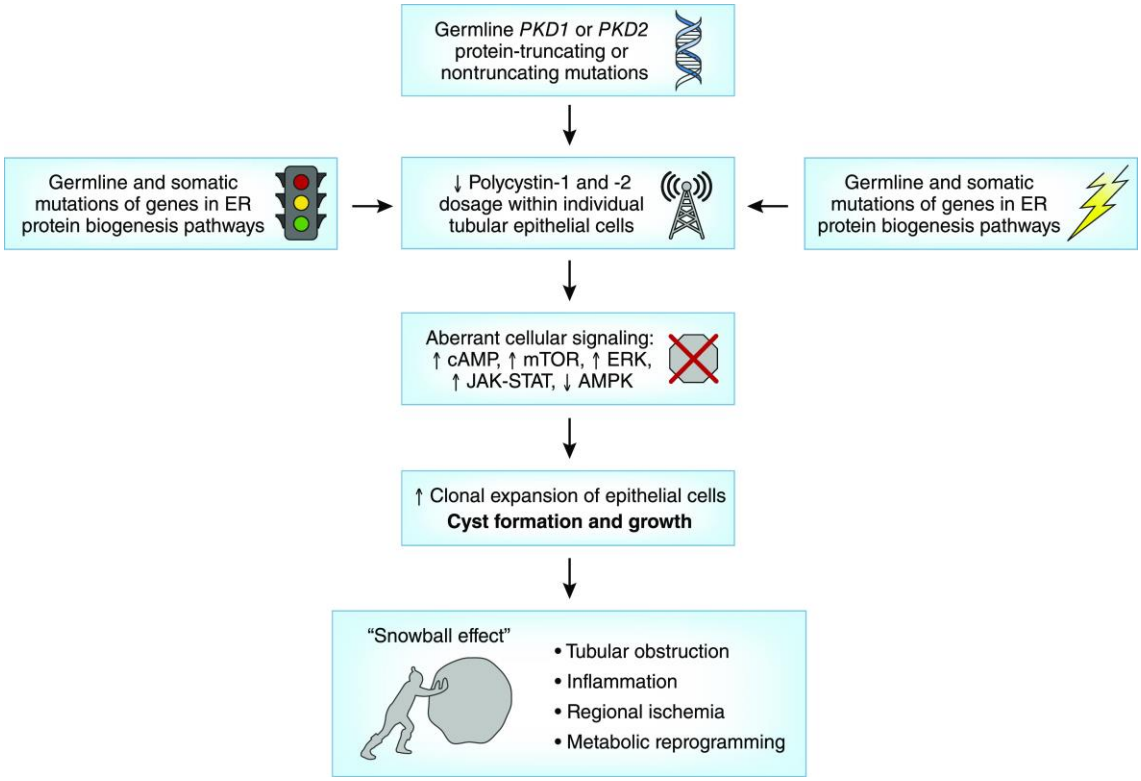

Figure 3-1 Proposed model of ADPKD pathology

Proposed mechanisms of the pathobiology of ADPKD. Taken from (Lanktree, Haghighi, et al. 2021). Even though the germline variants in ADPKD are present in all cells, cysts form in less than 10% of tubules focally (Grantham 1996). This led to a "second hit" hypothesis of cystogenesis. Under this model a somatic second hit is required to alter the remaining normal copy of PKD1 or PKD2. This has been shown to be the case in studies of kidney cysts in patients with PKD1, showing loss of heterozygosity due to a somatic variant, supporting a cellularly recessive mechanism for cyst formation in ADPKD (A. Y. Tan et al. 2018; Brasier and Henske 1997; Watnick et al. 1998).

However, there are numerous examples of patient with hypomorphic variants, which reduced the level of the gene product in the polycystin genes suffering CyKD suggesting a "threshold" mechanism of cystogenesis (Gallagher, Germino, and Somlo 2010; Rossetti et al. 2009; Harris 2010). It has been suggested that a dose of functional polycystin below ~10-30% of normal within tubular epithelial cells is enough to start cyst formation (Hopp et al. 2012; Lantinga-van Leeuwen et al. 2004) leading to various downstream aberrant cellular processes via multiple signalling pathways (S.-T. Jiang et al. 2006; Boca et al. 2006; Song et al. 2009; Lanktree, Haghighi, et al. 2021). Timing of gene inactivation is also vital. *PKD1* inactivation up to 13 days prior to birth in a mouse model led to severe CyKD compared to *PKD1* inactivation after day 14 of age in the same model which results in a far milder form of CyKD (Piontek et al. 2007).

More recently, genes linked to ADPKD have been discovered that affect protein creation, modification, and trafficking within the endoplasmic reticulum (ER). These genes effect the entry of unfolded protein into the ER (*SEC63* and *SEC61B* (Besse et al. 2017), the control of protein through the ER (Cornec-Le Gall et al. 2018), *N*-glycosylation of nascent proteins (a vital step in the trafficking of glycoproteins) (Apple et al. 2023; Besse et al. 2019; Lemoine et al. 2022; Cabezas et al. 2017) and the removal of glucose molecules to allow export from the ER to the Golgi complex (Porath et al. 2016). These variants have all been shown to lower the "dose" of polycystins, particularly PC1, within the cell, helping to a) confirm their role in pathogenesis and b) further elucidate the pathway the polycystin complex takes from transcription to final destination.

The latest discovery of monoallelic variants in *IFT140* causing a mild form of CyKD is of particular interest as it the first description of a protein involved in ciliary structure and function being described as causing ADPKD (Senum et al. 2022). ADPKD has

been seen by many as a “ciliopathy” with many of the experimental assays for the various causative genes showing clear ciliary disruption but never in monoallelic human disease. As part of the IFT-A complex that is responsible for retrograde transport in cilia, *IFT140* has a clearly defined ciliary role. However, *IFT140* is not required for the assembly of the IFT-A complex but does account for roughly half of the *TULP3* binding surface of the complex (M. Jiang et al. 2023). Combined with work by Legue et al showing *TULP3*’s involvement in ciliary trafficking in CyKD (Legué and Liem 2019), it as has been proposed that truncating *IFT140* variants disrupt *TULP3*-mediated cargo transport. *IFT140* disruption may lead to disruption in the trafficking of the polycystins to the cilia but this requires further experimental work.

Figure 3-2 details a schematic of the journey the polycystins take to their target with listed genes (modified from Lanktree et al. 2021). We can see that genes along the entire route of PC-1/2 journey to the cilium have been discovered to affect gene formation. As shown in the results chapter, these more recent genes have lower effect sizes than *PKD1/PKD2* and their recent discoveries is down to larger cohorts being sequenced with the latest technologies. This has enabled a higher diagnostic yield and an excellent elucidation of ciliary biology.

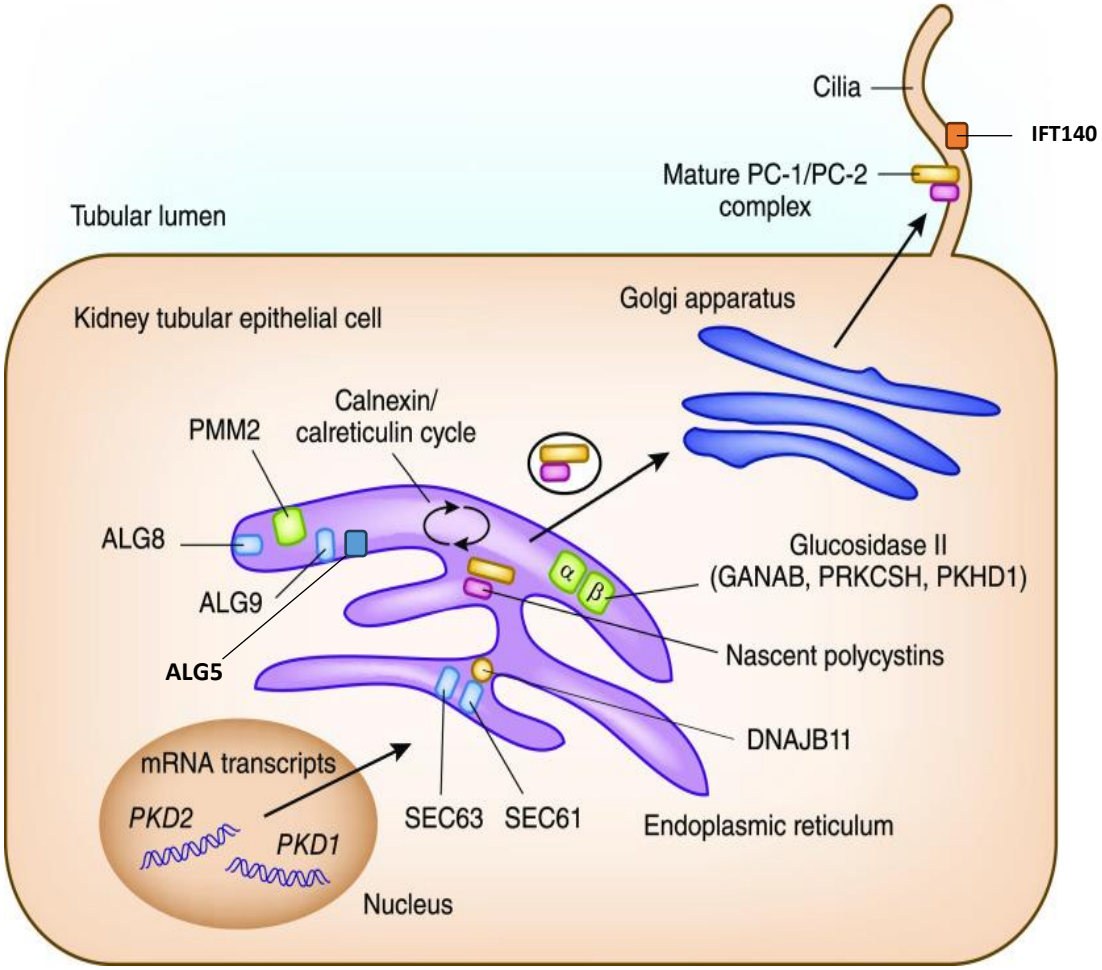

Figure 3-2 Genes in implicated in ADPKD and their effect on PC1/2 maturation.

Schematic representing the genes indicating in ADPKD and their effects on the maturation of PC1/PC2 (Modified from Lanktree et al 2021)

3.1.4.2 ARPKD

Most ARPKD cases are caused by variants in *PKHD1*, found on chromosome 6p21 which encodes fibrocystin. *PKHD1* has multiple transcripts, with a 4074 amino acid made up of a single transmembrane domain, an extensive extracellular N-terminal domain, and a short C-terminal cytoplasmic tail, comprising the largest one. The function of fibrocystin is still debated, but it is found throughout the kidney and epithelial cells of hepatic bile ducts and localises to the primary cilia membrane (Ward et al. 2003, 2002; L. F. Onuchic et al. 2002). The proteolytic cleavage of fibrocystin releases its C terminus, and this cleaved product has been the focus of much of the

speculated pathophysiology of ARPKD as its motifs are associated with ciliary targeting and interactions with the polycystin complex (Follit et al. 2010).

Many of the signalling pathways affected in ADPKD are also disrupted in ARPKD including cAMP (X. Wang et al. 2005) and mTOR (Fischer et al. 2009) but the general pathophysiology remains poorly understood. This has been compounded by mouse models of *PKHD1* including knockouts having minimal renal disease before adulthood (Moser et al. 2005; S. S. Williams et al. 2008) making functional characterisations of fibrocystin challenging. Recent work has highlighted the role the cleaved C-terminus of fibrocystin may have in preventing cytogenesis via its interaction with mitochondrial pathways (R. Walker et al. 2022), work which carriers homology with that of Caplan et al which showed the C-terminal of PC-1 suppresses cystic disease via a mitochondrial pathway (L. Onuchic et al. 2023). The potential for a mechanism other than direct interaction was confirmed by work on a digenic system combining *PKHD1* knock out mice with a hypomorphic *PKD1* mutant showing no interaction between the fibrocystin protein and polycystins (Olson et al. 2019) directly. It is likely that a shared ciliary mechanism or mitochondrial process is the missing link.

Variants in the gene *DZIPI1L*, coding for the ciliary transition zone protein DAZ-interacting protein 1-like protein has been described as a moderate cause of ARPKD. Working at the barrier between the cell and cilium, variants in this gene have been shown to disrupt the transport of PC21 and PC2 into the cilium (H. Lu et al. 2017). Phosphomannomutase2 (*PMM2)* variants have also been described as causing ARPKD, whilst the pathophysiology has not been fully delineated *PMM2* is an enzyme critical to *N*-linked glycosylation, potentially causing a trafficking issue with the polycystins similar to the other ER linked monogenic causes of CyKD (Cabezas et al. 2017).

3.1.4.3 HNF1β and ADPKD

HNF1β is thought to be an autosomal dominant condition with haploinsufficiency as the molecular mechanism, as patients with whole-gene deletions have a similar phenotype to those with coding or splice variants. Up to 50% of *HNF1β* cases are thought to be *de novo* (Ulinski et al. 2006; Edghill et al. 2007).

Mice with renal specific depletion of *HNF1β* develop CyKD and renal dysfunction (Gresh et al. 2004; Hiesberger et al. 3 2004). Further molecular analysis reveals HNF1β's role in regulating the transcription of *PKHD1, PKD2, UMOD* and gene implicated in nephronophthisis (Attanasio et al. 2007; Song et al. 2009; Gong et al. 2009) which explains the variable nature of the phenotype as well as the mechanism of disease. Bar the effects on known monogenic causes of cystogenesis, *HNF1β* also directly increases cAMP levels via the regulation of the expression of phosphodiesterase 4C which catabolises cAMP in the primary cilium (Y.-H. Choi et al. 2011), inhibition of cAMP being a primary mechanism of Tolvaptan's function. Further functional work is required to map the full molecular pathway of *HNF1β* associated renal disease.

3.2 Cystic kidney disease as a monogenic disorder

3.2.1 Introduction

Rare variants that cause classical Mendelian disease are kept rare through the process of natural selection whereby rare deleterious variants are prevented from becoming common in the general population by negatively affecting reproductive fitness. Rare diseases are typically caused by rare variants and CyKD is no exception. Rare, highly penetrant alleles that are very damaging make up the bulk of causative variants in CyKD and offer the most clues as to the underlying biology of the disease. This is reflected in the vast swathes of published genetics research on CyKD being focused on monogenic causes in patients and families (Bergmann et al. 2018). This chapter focuses on rare variants, defined as those with a MAF<0.1%, as the primary driver of CyKD.

3.2.2 Aims

1. To determine the prevalence of known monogenic disease in a large cohort of patients with cystic kidney disease.
2. To discover novel candidate genes using an unbiased exome-wide rare variant association testing approach.

3.2.3 Methods

3.2.3.1 Case selection

Cases were recruited under the "Cystic kidney disease" 100KGP cohort by clinicians across the country using the following inclusion criteria:

- >5 cysts affecting one or both kidneys with one of the following features:
    - cysts not clinically characteristic of ADPKD
    - onset before the age of 10
    - syndromic features
    - where a genetic diagnosis would influence management

- Unaffected individuals had to have undergone appropriate screening for cryptic disease.

A second tranche of recruitment widened the inclusion criteria to include participants with features suggestive of classical ADPKD who had not undergone prior genetic testing of *PKD1* and *PKD2*. Participants were excluded if they suffered from end-stage kidney failure due to identified (non-cystic) disease, if they had multicystic dysplastic kidney(s) or if they had a prior genetic diagnosis for their condition. This recruitment strategy led the total cohort being a mixture of unsolved cystic cases and those more obviously ADPKD-*PKD1* and ADPKD-*PKD2*.

3.2.3.2 Control selection

Controls were made up of unaffected relatives of non-renal rare disease participants in the 100KGP. We refined this further by excluding those with HPO and/or hospital episode statistics (HES) terms related to kidney disease or failure. Within the >20,000 controls there was the possibility that some harboured undetected cystic kidney disease as we did not have access to imaging data, however, it was felt their contribution to statistical signal would not significantly affect the outcome.

3.2.3.3 Identification of pathogenic variants

All cases recruited for had been assessed in the clinical interpretation arm of the 100KGP (100000 Genomes Project Pilot Investigators et al. 2021). For this, patients' WGS data is extracted variants that are rare (MAF < 1% for autosomal recessive and MAF < .1% or autosomal dominant inheritance), protein-truncating or missense. These are then cross referenced with an expertly curated (Antonio Rueda Martin et al. 2019) panel of 28 CyKD associated genes (https://nhsgms-panelapp.genomicsengland.co.uk/panels/283/v4.0) CNVs losses with a 60% overlap with the 2q13 loci associated with *NPHP1* loss and the 17q12 loci associated with *HNF1β* loss were also ascertained for.

These results then underwent multi-disciplinary (MDT) review with the recruiting clinical team, local genomic medicine centre with support from Genomics England. Candidate variants were assessed against the Association for Clinical Genomic Science (ACGS) Best Practice Guidelines for Variant Classification in Rare Disease (https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf). These criteria are based on recommendations from the American College of Molecular Genetics (ACMG) criteria to determine pathogenicity (Richards et al. 2015) using a host of metrics including population frequency of alleles, *in silico* predictions of deleteriousness, functional localization, putative mechanism of disease and known associations with phenotypes in validated disease databases to assign one of the following classifications: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign or benign.

3.2.3.4 Aggregate rare coding variant analysis

3.2.3.4.1 Overview of rare variant association tests

Rare variant analysis is considered more challenging than common variant analysis due to several factors. Firstly, rare variants, by definition, occur at a low frequency in the population. As a result, large sample sizes are often required to have sufficient statistical power to detect associations. Secondly, when rare variants are called single variant association testing is underpowered due to the scarcity of variants in a given population (Seunggeung Lee et al. 2014).

In order to overcome some of these issues, collapsing analyses have been employed to boost power. In this statistical method, variants are “collapsed” by some kind of region, most commonly per gene, and their effect sizes cumulated to test for an association with the disease or trait of interest. This is particularly helpful for allelic heterogeneity where multiple different alleles account for a disease, with no single allele explaining a large fraction of risk; a situation found in ADPKD (Paul et al. 2014). I will discuss below the broad categories of rare variant collapsing tests as well as the rationale for selecting my chosen method.

3.2.3.4.1.1 Burden tests

Collapsing burden tests combined multiple genetic variants into a single genetic score per region, testing for an association between this region and the disease of interest. A simple example would involve counting the number of minor alleles in all variants in each region. The score statistic would be:

$$Q_{burden} = \left(\sum_{j=1}^{m} w_j S_j \right)^2$$

where m = the number of variants in the region, w_j = the weight for variant j (e.g., using MAF or beta effect size), and S_j = the score statistic for variant j generated from the sum of allele counts (0,1, or 2) for each individual at variant j, accounting for phenotype. S_j is positive when variant j increases disease risk, and negative when associated with decreased disease risk. A P value is then obtained by comparing the burden test statistic to chi-squared distribution with 1 degree of freedom (Seunggeung Lee et al. 2014)

Multiple different implementations of the burden test exist which make different assumptions about disease mechanism and architecture. In the main, a dominant mode of inheritance is assumed to maximise power, with the various methods differing on how they define the weight (w_j) of variants to be collapsed. The MZ test for example counts individuals with at least one minor allele in the region (Morris and Zeggini 2010) whilst the cohort allelic sums test (CAST) assumes any rare variant in a region can cause loss of function (Morgenthaler and Thilly 2007) both of which give a binary weight to w_j. The Madsen and Browning method weights by the MAF as a proportion to give a beta of densities (Madsen and Browning 2009) . Finally, the Combined Multivariate and Collapsing (CMC) test uses the CAST approach but by collapsing groups based on their MAF per region and then using a non-regression technique (Hoteling's t test) to combine the effects (B. Li and Leal 2008). Burden tests assume

most the variants collapsed are causative and have the same direction of effect with violations of these assumptions resulting in a significant loss of power.

Burden testing has been further adapted with "adaptive burden" tests to help account for the null variants and variants affecting disease risk in either direction. Whilst such methods such as the kernel-based adaptive cluster (KBAC) method (D. J. Liu and Leal 2010) have overcome these issues most adaptive methods require *P* value permutation and are therefore computationally intensive, making them unsuitable for large scale biobank studies.

3.2.3.4.1.2 Variance Component tests

Variance component tests use a random effects model to overcome some of the issues of unknown underlying genetic architecture and variant effects. Instead of aggregating variants and then generating a combined test statistic on the whole region, variance-component tests look at the distribution of individual test statistics per variant and then aggregate these to compute an overall *P* value. The most used variance-component test is the sequence kernel association test (SKAT) (Wu et al. 2011) which can be represented as:

$$Q_{SKAT} = \sum_{j=1}^{m} w_j^2 \, S_j^2$$

The SKAT test uses the weighted sum of squares of single variant score statistics S_j. By collapsing S_j^2 instead of S_j as per the burden test, SKAT is robust to both non-causal and variants acting in direction of effect. The addition of covariates to this analysis allows for adjustment for population stratification. SKAT testing has two major issues: firstly, contrary to the burden test, if a large proportion of variants are causal, variance-component tests lose power; secondly, for binary traits calculating many *P* values on a per variant basis and then combining them per region can lead to high type 1 error rates especially when the minor allele count is low or the sample size small.

3.2.3.4.1.3 Combined tests

An understanding of the genetic architecture of a disease is often lacking at the time of analysis effecting the power of both approaches to collapsing analysis discussed. Even for conditions such as CyKD, the architecture of disease may differ on a gene-by-gene basis as evidenced by both recessive and dominant Mendelian conditions causing kidney cysts. This has led to the development of a combined method which uses a linear combination of burden and SKAT testing based on the underlying data to maximise power, SKAT-O (Seunggeun Lee et al. 2012). SKAT-O can be represented as:

$$Q_\rho = (1-\rho)\,Q_{\text{SKAT}} + \rho Q_{\text{burden}}\,,\ 0 \le \rho \le 1$$

The key term here is the parameter ρ which represents the pairwise correlation between genetic-effect coefficients (β). $\rho = 1$ when all variants act in the same direction, meaning the test statistic resolves as the Q_{burden}, whereas if the variants are uncorrelated in their direction and magnitude of effects then the $\rho = 0$ and the test statistic approximates to the Q_{SKAT}. In reality the ρ is unknown so SKAT-O uses an adaptive procedure to approximate the value and calculate *P* values analytically, allowing for a combined method that uses that uses the best of both methods and allows for uncertainty in the underlying genetic architecture. SKAT-O has been shown to perform well across a wide range of disease models and is widely used in association tests (Seunggeun Lee et al. 2012).

3.2.3.4.2 Selection of qualifying variants

As powerful as the collapsing methods discussed above are, if the majority of variants selected to be collapsed per region have little or no effect then the power gained by collapsing variants is limited. Including qualifying variants that are more likely to be damaging and therefore disease causing will increase the power to detect association. In general, “damaging” or “deleterious” variants include those that are rare, loss-of-function e.g. protein truncating, or predicted *in silico* to be damaging. For my analyses I collapsed variants across genes using a number of parameters applied as a “mask” detailed below. Thanks to having access to WGS data I was also able to include masks that included intronic variants also. I applied the “missense+” and “LoF” mask to the

total CyKD cohort and then removed cases that had qualifying variants in statistically significant genes until we had a cohort of patients with “no variants detected” (NVD). To this cohort we applied all the masks listed:

1. Likely damaging (“missense+”):
    - MAF < 0.1% or absent from gnomAD (version 3.1.1)
    - Annotated as missense, in-frame insertion, in-frame deletion, start loss, stop gain, frameshift, splice donor or splice acceptor.
    - CADD (version 1.5) score ≥ 20 corresponding to the top 1% of all predicted deleterious variants in the genome. Indels without CADD scores were also kept as most frameshift variants do not have assigned CADD scores.

2. Loss-of-function (“LoF”):
    - MAF < 0.01% or absent from gnomAD (version 3.1.1)
    - ‘High confidence’ loss-of-function variants (stop gain, splice site, or frameshift) as determined by LOFTEE (Karczewski et al. 2020).

3. Intronic:
    - MAF < 0.01% or absent from gnomAD (version 3.1.1)
    - Variants labelled as intronic
    - CADD score ≥ 20

4. 5’ untranslated region (“5’ UTR”):
    - MAF < 0.01% or absent from gnomAD (version 3.1.1)
    - Variants labelled as 5’UTR
    - CADD score ≥ 10

5. 3’ untranslated region (“3’ UTR”):
    - MAF < 0.01% or absent from gnomAD (version 3.1.1)
    - Variants labelled as 3’UTR
    - CADD score ≥ 10

6. Splicing (“donor loss,” “donor gain,” “acceptor loss,” “acceptor gain”):
    - SpliceAI score ≥ 0.8 (discussed in Methods 2.2.5)

Variants meeting the following quality control filters were retained: MAC ≤ 20, median site-wide sequencing depth in non-missing samples > 20 and median GQ ≥ 30. Sample-level QC metrics for each site were set to minimum depth per sample of 10, minimum GQ per sample of 20 and ABratio P value > 0.001. Variants with significantly different missingness between cases and controls (P<10-5) or >5% missingness overall were excluded.

3.2.3.4.3 SAIGE-GENE

I employed SAIGE-GENE (W. Zhou et al. 2020) to ascertain whether rare coding variation was enriched in cases on a per-gene basis exome-wide. SAIGE-GENE uses a generalized mixed-model to correct for population stratification and cryptic relatedness as well as a saddle point approximation and efficient resampling adjustment to account for the inflated type 1 error rates seen with unbalanced case-control ratios (see chapter 2.5 for further details). It combines single-variant score statistics and their covariance estimate to perform SKAT-O gene-based association testing, upweighting rarer variants using the beta (1,25) weights option. Sex and the top ten principal components were included as fixed effects when fitting the null model. A Bonferroni adjusted P value of 2.58×10^{-6} (0.05/19,364 genes) was used to determine the exome-wide significance threshold. Binary odds ratios and 95% confidence intervals were calculated for exome-wide significance genes by extracting the number of cases and controls carrying qualifying variants per gene in the collapsing analysis and applying a Fisher’s test in R.

3.2.3.4.4 Genomic inflation in rare variant collapsing tests

Genomic inflation estimates using rare variants is unreliable as the variant distribution under the null model is unknown when allele counts are low, making inferences about population stratification difficult. Equally, different set or gene based association tests have different numbers of variants per set meaning inflation statistics are incomparable

to each other (Q. Liu, Nicolae, and Chen 2013). For reference most of my genomic inflation values for rare variant analyses fell below 1 with figures between 0.5-0.9 however, these are unreliable. I have provided quantile-quantile plots (QQ-plots) as they continue to provide a good visual method of assessing inflation and have taken the well-controlled inflation values from the common variant seqGWAS analysis as evidence of a lack of population stratification in my rare variant analyses (please see chapter 5 for further details). Furthermore, there is good evidence that the addition of controls allows for appropriate stratification correction in rare variant analyses, even in situations of large case: control imbalances when a GLMM and PC approach is used for correction (Bouaziz et al. 2021); an approach I have adhered to in my rare variant analyses.

3.2.3.5 Stratification by primary variant and depleting analysis

The type of variant driving ADPKD is known to affect the renal prognosis with truncating *PKD1* variants carrying the worst prognosis (Cornec-Le Gall et al. 2016). Within families it is also known that those with the same variant can have vastly different phenotypes (Harris and Rossetti 2010) with the heritability of time to ESRF ranging from 45-50% (Paterson et al. 2005; Fain et al. 2005). Whilst there are known environmental factors affecting disease progression such as caffeine and smoking (Tanner and Tanner 2001; Orth et al. 1998) it is clear that there are genetic modifiers of ADPKD. This will likely hold true for other causes of CyKD but has yet to be studied in detail.

Until now, candidate gene studies have been unsuccessful in identifying modifier genes due to small study sizes, lack of clinical characterisation and problematic endpoints (Baboolal et al. 1997; A. Persu et al. 2002; D. Walker et al. 2003). In the biobank era with access to WGS we are now able to stratify cohorts based on the primary driving disease causing variant and conduct genetic association studies to look for secondary genetic markers causing disease. As will be discussed in the time to event analysis chapter, biobanks also contain renal function endpoints, allowing for association studies to look for markers of disease progression within each molecular cohort.

CyKD patients who have their phenotype "solved" by the clinical multi-disciplinary team (MDT) had a report issued with the details of the molecular diagnosis. These were available to researchers in the 100KGP and could be manipulated in R using the LabKey tool (Nelson et al. 2011). Depending on the molecular diagnosis CyKD patients were placed into different cohorts: *PKD1*-truncating (*PKD1*-T), *PKD2*-truncating (*PKD2*-T), *PKD1*-non truncating (*PKD1*-NT), *PKD2* non-truncating (*PKD2*-NT), "other gene" (encompassing other green genes in the PanelApp list of approved genes thought to cause CyKD) and no variant detected (NVD). In the patients with NVD I bioinformatically reanalysed them looking for variants that met the "missense+" or "loss-of-function mask" (detailed below), in the approved cystic kidney disease panel of genes in PanelApp (Antonio Rueda Martin et al. 2019) and placing them in the relevant cohort. The filtering was performed using BCFtools and filter-VEP (McLaren et al. 2016). For each subsequent round of analysis if a gene or structural variant was found to be significantly enriched in cases, I identified the cases that contained qualifying variants and removed them from the NVD cohort and re-analysed the cohort, eventually leaving 266 cases with no clear genetic cause of disease.

I performed all single-variant, gene-burden, and structural variant analysis in each molecular subgroup (bar the "other genes" group as this was a heterogenous group of disorder). I used the same controls for each subgroup without repeating ancestry matching as there was no evidence of genomic inflation within each subgroup and the controls (lambda between 0.99-1.02 in all common variant analyses).

3.2.3.6 Pathway analysis using collapsing rare variant summary statistics

Gene set analysis (GSA), similar to collapsing tests via genes, aims to increase the power to detect signal by collapsing variant signals across sets of genes associated with a molecular pathway. GSA aggregates signals from genes into sets sharing biological or functional characteristics. This reduces the number of tests performed and can provide insight into the pathways of cellular mechanisms involved in a trait or phenotype (Pers 2016).

For the cohort of patients that had no molecular diagnosis the summary statistics from their rare variant SKAT-O analysis with SAIGE-GENE was analysed using the Gene set analysis Association using Sparse Signals method (GAUSS) with default settings (Dutta et al. 2021). The summary statistics were analysed using the canonical curated gene set pathways from the Gene Set Enrichment Analysis (GSEA) group (Subramanian et al. 2005). GAUSS was selected as it has been shown to be more powerful than existing methods, whilst controlling for type I error and scaling to biobank level datasets.

3.2.4 Results

All variants contributing to significant associations in the collapsing tests can be found in the summary statistics available in the supplementary data.

3.2.4.1 Diagnostic yield of WGS in CyKD

3.2.4.1.1 Cohort description

1558 participants were recruited to the 100KGP under cystic kidney disease. 1294 were probands. 921 were recruited as singletons (59.11%), 187 (12%) as a duo with their mother or father, 147(9.44%) as a trio with their mother and father, 124 (7.96%) as a duo, 81 (5.2%) as a family with more than three participants, 66 (4.24%) as a trio with one of their mothers or fathers and another biological relation, 32 (2.05%) as a trio with other biological relatives. The median age of the cohort was 50 with a family history in 58% of the cohort. 25% of the cohort had reached ESRF with a median age of 52. The demographic information of probands and the ancestry matched controls is set out in table 3-3. The top five most frequent human phenotype ontology codes are set out in table 3-4.

Table 3-3 Demographic breakdown of the recruited cystic kidney disease probands and controls

Demographics	Case	Control
Female	669(51.75%)	14557(55.78%)
Median age	50 (IQR 37-61)	47.89(IQR 39-54)
Affected 1st degree relative	752(58.03%)	NA
Consanguinity in parents	41(3.17%)	NA
End-stage kidney disease	398(25.55%)	NA
Median age ESRF	52(IQR 44-60)	NA
Self-reported ethnicity		
European	924(71.41%)	18445 (70.68%)
African	58(4.42%)	564 (2.16%)
Other Asian	12(0.93%)	461 (17.67%)
South Asian	54(4.17%)	2308 (8.84%)
East Asian	6(0.46%)	73 (0.28%)
Mixed	25(1.93%)	357 (1.37%)
Not stated/unknown	215(16.62%)	3888 (14.90%)

IQR – Interquartile Range, ESRF – End Stage Renal Failure

Table 3-4 Top 5 most frequent HPO terms in the CyKD cohort

HPO code	Count(percentage)
Multiple renal cysts	1085(83.85%)
Hypertension	697(53.86%)
Enlarged Kidney	513(39.64%)
Hepatic cysts	383(29.60%)
Haematuria	162(12.52%)

HPO – Human phenotype ontology

3.2.4.1.2 Prevalence of monogenic disease in the clinical arm of 100KGP

Of these probands 1290 had outcome data from the 100KGP clinical pipeline: 640 (52.93%) were solved, 34 (2.81%) partially solved, 79 (6.54%) unaccounted for and 537 (44.42%) unsolved. The full breakdown of solved cases and their types of variants can be found in table 3-5 (3 patients were solved for primary conditions unrelated to their cystic kidney disease e.g. intellectual disability and were not included in this table and 12 cases did not have listed genes despite being listed as solved). Of the 1290 cases 578 had data regarding kidney function in the form of HPO or HES codes with 398 having reached ESRF.

Table 3-5 Molecular diagnosis in cystic kidney disease cases that were solved by the 100,000-genome project clinical pipeline.

Gene (Condition)	Consequence	Count
PKD1 (ADPKD)	Protein truncating	340
	Non protein truncating	118
PKD2 (ADPKD)	Protein truncating	122
	Non protein truncating	13
PKHD1 (ARPKD)	Compound heterozygous	7
	Homozygous	5
DNAJB11 (ADPKD)	Protein truncating	5
	Non protein truncating	1
BBS1 (Bardet-Biedl syndrome 1; biallelic)	Non protein truncating	2
HNF1B (Renal cysts and diabetes syndrome)	Protein truncating	2
SALL1 (Townes-Brocks syndrome)	Protein truncating	1
	Non protein truncating	1
COL4A4 (Alport syndrome)	Truncating	1
FAN1 (Interstitial nephritis)	Protein truncating	1
GANAB (ADPKD)	Protein truncating	1
OFD1 (Joubert syndrome 10)	Protein truncating	1
SDCCAG8 (Bardet-Biedl syndrome 16; biallelic)	Protein truncating	1
TMEM67 (Joubert syndrome 6)	Protein truncating	1
UMOD (Tubulointerstitial kidney disease)	Non protein truncating	1
WT1 (Denys-Drash syndrome)	Non protein truncating	1

ADPKD – Autosomal dominant polycystic kidney disease, ARPKD – Autosomal recessive polycystic kidney disease

3.2.4.1.3 Survival analysis

Grouping the solved cases into their respective primary driving variants and performing survival analysis led to the graph in figure 3-3. Age of reaching ESRF was the endpoint and in keeping with the known literature, patients with truncating *PKD1* variants carried the worse prognosis with a median age of ESRF of 58 years. There were not enough events in the *PKD2* non-truncating group to be included in the Kaplan-Meier plot (two

events with median age of ESF 71 years). The no-variant detected group had a survival profile (median age of ESRF 76 years) between that of *PKD1* non-truncating variants (median age 63 years) and *PKD2* truncating variants (median age 89 years) highlighting their unmet clinical need.

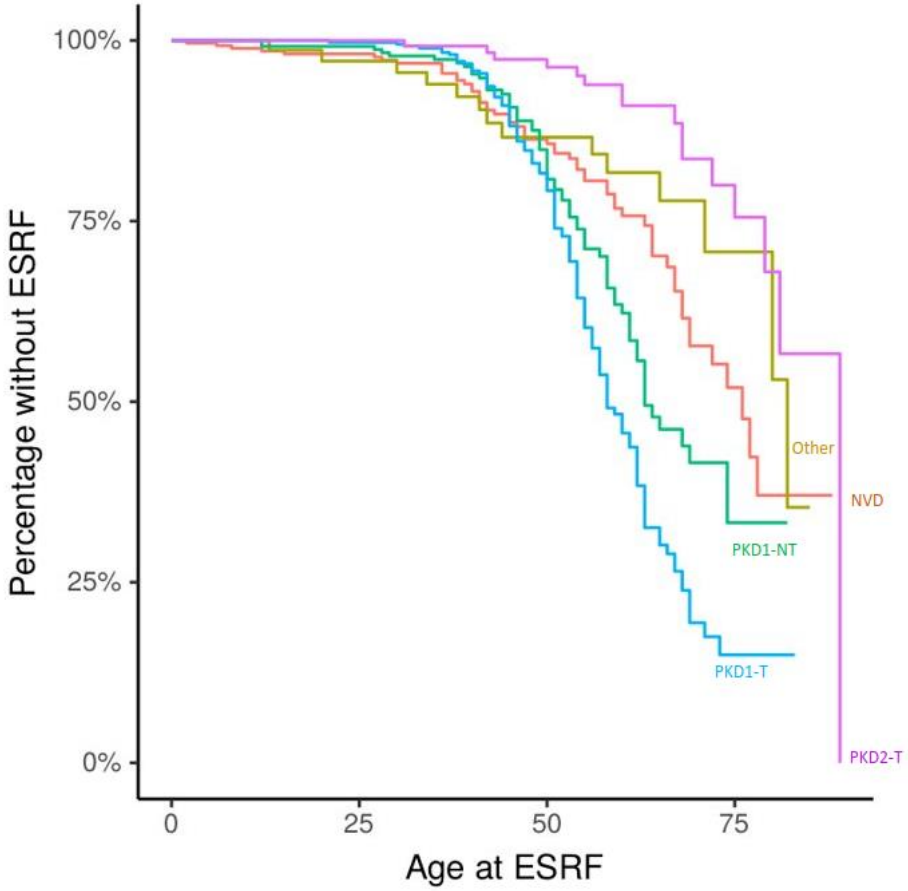

Figure 3-3 Kaplan-Meier plot of renal survival plotted by primary driving variant

PKD1-T *PKD1*-truncating variant, *PKD1*-NT *PKD1*-nontruncating variant, *PKD2*-T *PKD2*-truncating variant, Other-another variant in the PanelApp cystic kidney disease gene panel, NVD – no variant detected. Note *PKD2*-NT is not plotted due to the low number of events.

3.2.4.2 Rare variant association testing

3.2.4.2.1 Depleting analysis of cases

I performed SKAT-O testing as implemented via SAGIE-GENE in 1209 CyKD cases ancestry matched to 26096 unrelated controls in all coding genes collapsed by the "missense+" (likely damaging [CADD >20, MAF <0.01, at least a missense

annotation]) and “LoF” tags (MAF<0.01, high confidence of causing loss-of-function). I then consequently depleted the cases for those solved by the 100KGP project or those who carried variants that made up the significantly associated gene signals under the “missense+” or “LoF” masks. At each step of analysis, I removed those cases making up significant associations until any positive signal was ameliorated. Unless otherwise stated, all individuals with qualifying variants for the results presented here were heterozygous for their variants.

3.2.4.2.2 Likely damaging variants (“missense+”)

Rare variant analysis of the total ancestry matched cohort of 1209 cases and 26096 controls under the “missense+” mask showed a significant enrichment of cases for *PKD1* ($P=1.17x10^{-309}$, OR=10.60, 95% CI = 9.35-12.01), *PKD2* ($P=1.96x10^{-150}$, OR=19.07, 95% CI 15.13-23.99), *DNAJB11*($P=3.52x10^{-07}$, OR 1.07, 95% CI 0.95-1.21), and *COL4A3*($P=1.26x10^{-06}$, OR=3.02, 95% CI 2.10-4.22). Notable genes just below genome wide significance included *IFT140* ($P=1.02x10^{-05}$, OR=2.04, 95% CI 1.53-2.75)) and *PKHD1* ($P=8.17x10^{-06}$, OR=1.60, 95%CI=1.27-2.00) (Figure 3-4). There was no evidence of genomic inflation (lambda<1 and Figure 3-5).

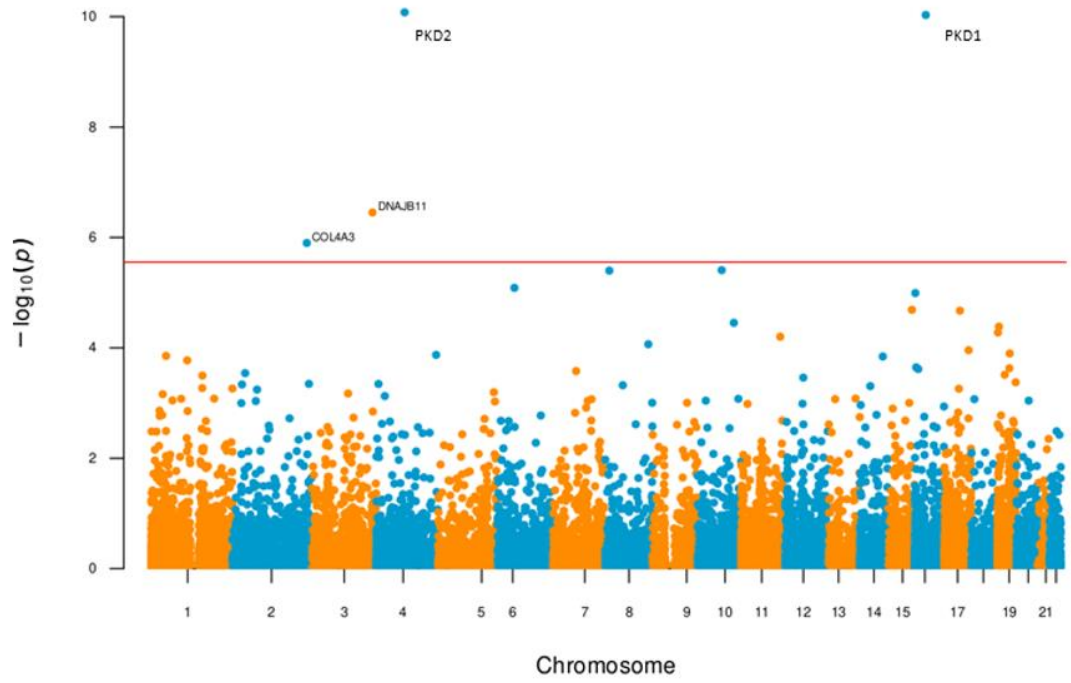

Figure 3-4 Gene based Manhattan for the association of likely damaging variants between all CyKD cases and control.

Manhattan plot of exome-wide gene-based rare, likely damaging variant association testing for 1209 CyKD cases and 26096 ancestry matched controls. SAIGE-GENE was performed for 19,168 genes with loss-of-function and likely damaging missense variants with MAF < 0.1%. Each dot represents a gene. The red line indicates the exome-wide significance threshold of $P=2.58\times10^{-6}$. *PKD1* and *PKD2* are listed at the top of the graph to highlight they fall far out of bounds of the scale due to the strength of their association.

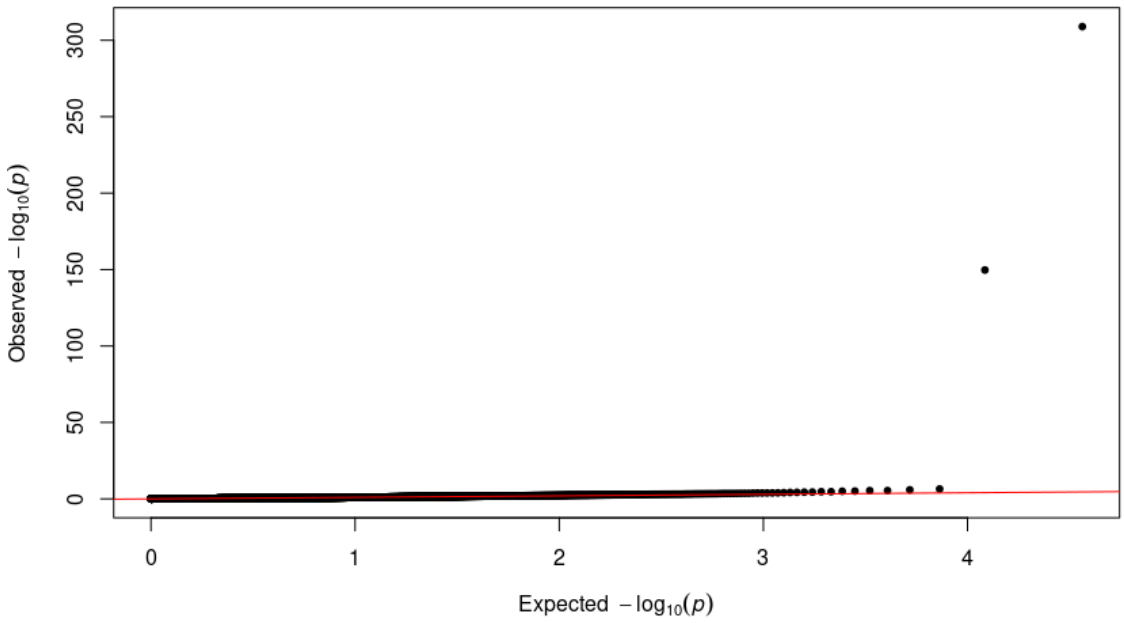

Figure 3-5 Q-Q plot for the association of likely damaging variants between all CyKD cases and control

Q-Q plot of exome-wide gene-based association testing for 1209 CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line signifies the observed versus the expected – $\log_{10}(P)$ for each gene tested.

Removing cases solved by 100KGP and patients that had a bioinformatically ascertained pathogenic variant in a known cystic gene left 308 cases. Performing rare variant analysis under the "missense+" tag showed a significant enrichment of cases with variants in *IFT140* ($P=1.26\times10^{-16}$, OR=5.57, 95%CI 3.63-8.21) and *COL4A3* ($P=6.83\times10^{-07}$, OR=4.93 95%CI 2.77-8.11) compared with 26096 controls (Figure 3-6, QQ-plot 3-7).

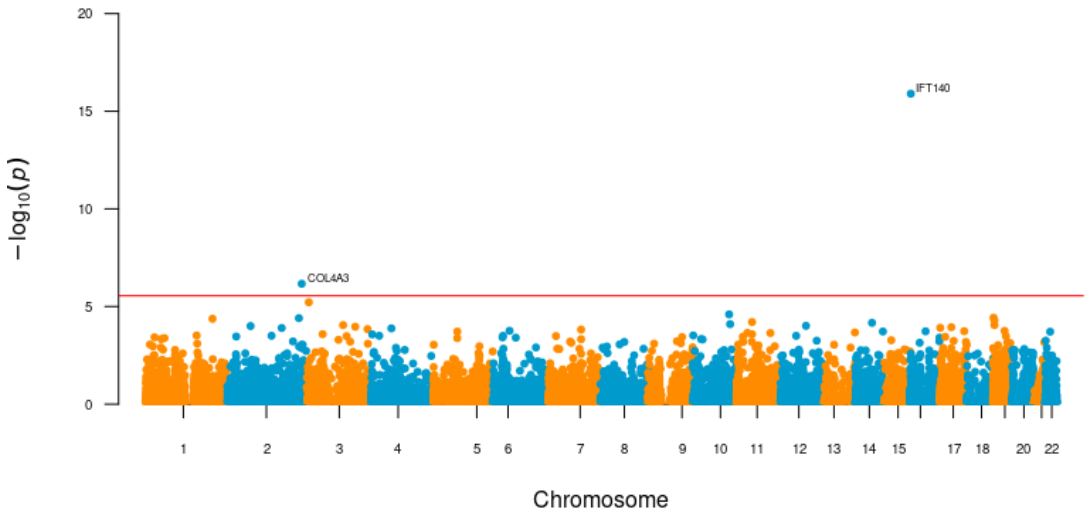

Figure 3-6 Gene based Manhattan for the association of likely damaging variants between unsolved CyKD case and controls.

Manhattan plot of exome-wide gene-based rare, likely damaging variant association testing for 308 unsolved CyKD cases and 26096 ancestry matched controls. SAIGE-GENE was performed for 19,168 genes with loss-of-function and likely damaging missense variants with MAF < 0.1%. Each dot represents a gene. The red line indicates the exome-wide significance threshold of P=2.58×10-6.

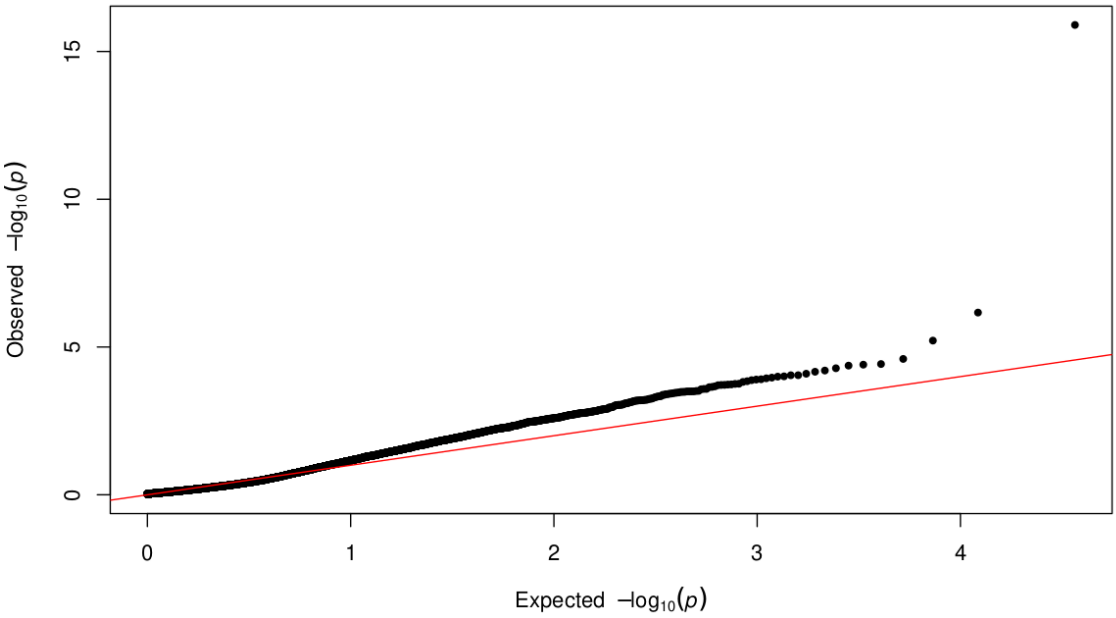

Figure 3-7 Q-Q plot for the association of likely damaging variants between unsolved CyKD cases and controls

Q-Q plot of exome-wide gene-based association testing for 308 unsolved CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line signifies the observed versus the expected $-\log_{10}(P)$ for each gene tested.

3.2.4.2.3 Loss of function variants

Collapsing rare variants that had a high confidence call for loss-of-function under the "LoF" mask revealed significant enrichment for *PKD2* ($P=3.05\times10^{-147}$, OR=130.85, 95% CI = 83.66-215.37), *PKD1*($P=1.29\times10^{-117}$,OR=36.01, 95% CI 30.52-42.23), *IFT140* ($P=3.00\times10^{-25}$, OR=14.03, 95%CI 7.91-24.45), *DNAJB11*($P=1.84\times10^{-12}$, OR 1.07, 95% CI 0.95-1.21) and *PKHD1* ($P=2.98\times10^{-08,}$ OR=4.07 95%CI 2.24-6.88) (Figure 3-8 and QQ Figure 3-9).

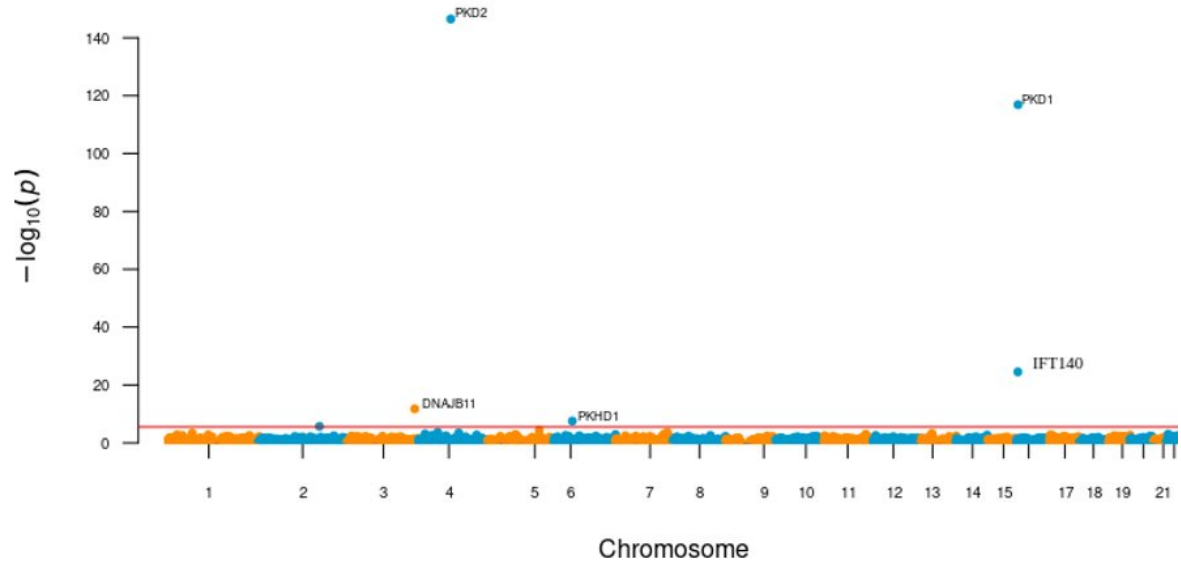

Figure 3-8 Gene based Manhattan for the association of loss-of-function variants between all CyKD cases and controls

Manhattan plot of exome-wide gene-based loss-of-function variant association testing for 1209 CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line indicates the exome-wide significance threshold of P=2.58×10-6.

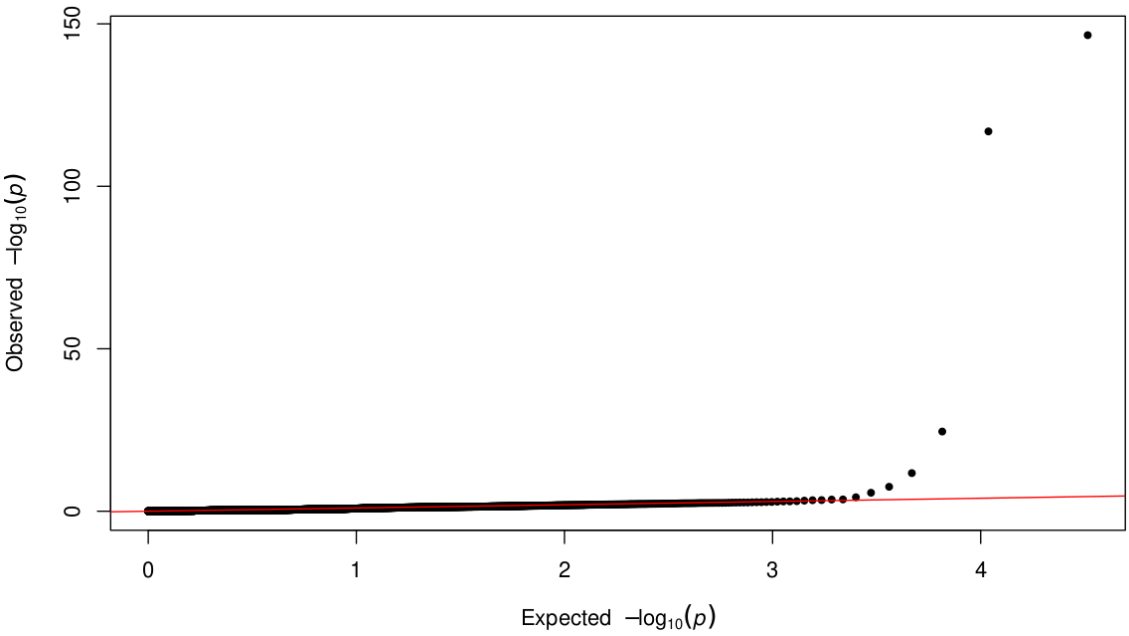

Figure 3-9 Q-Q plot for the association of loss-of-function variants between unsolved CyKD cases and controls

Q-Q plot of exome-wide gene-based association testing for 1209 CyKD cases and 26096 ancestry matched controls under the loss-of-function mask. Each dot represents a gene. The red line signifies the observed versus the expected $-\log_{10}(P)$ for each gene tested.

Analysing the 308 unsolved cases against controls revealed enrichment for *IFT140* *(P=1.35x10^{-17}, OR=5.11, 95% CI 0.77-16.82)* (Figure 3-10 and QQ Figure 3-11). Further depletion of cases by those with qualifying variants that made up the *IFT140* and *COL4A3* signals led to 266 cases remaining which did not reveal any further significant associations on rare variant testing for either mask (Figure 3-12).

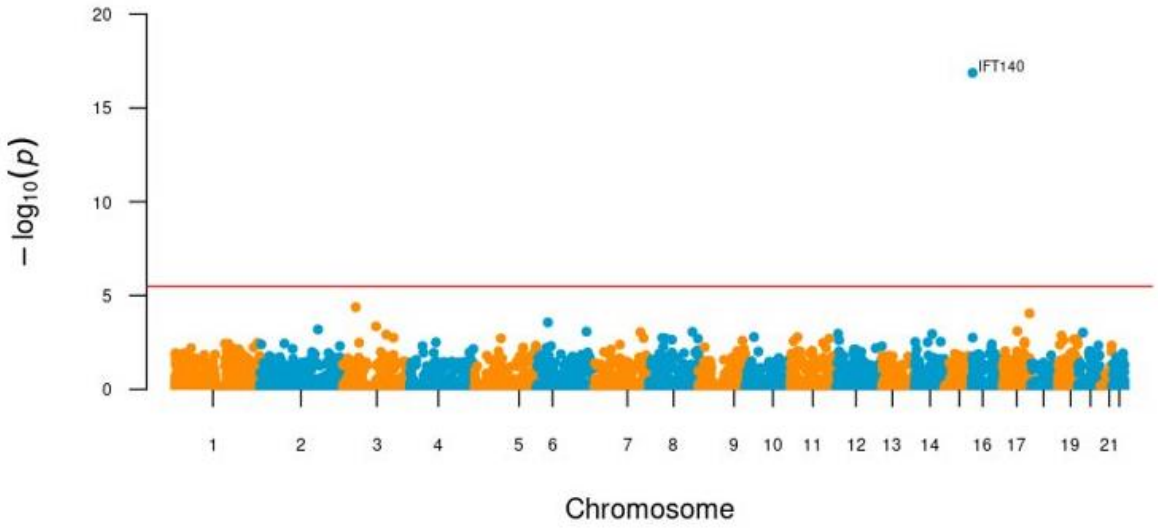

Figure 3-10 Gene based Manhattan for the association of loss-of-function variants between the unsolved CyKD cases and controls

Manhattan plot of exome-wide gene-based loss-of-function variant association testing for 308 unsolved CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line indicates the exome-wide significance threshold of P=2.58×10-6.

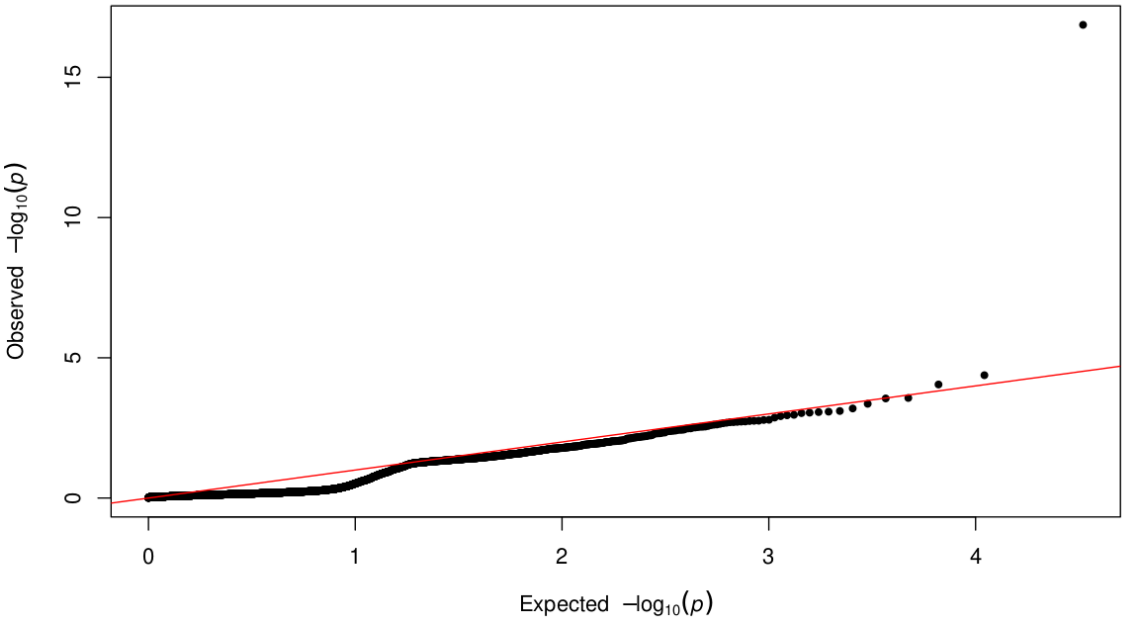

Figure 3-11 Q-Q plot for the association of loss-of-function variants between unsolved CyKD cases and controls

Q-Q plot of exome-wide gene-based association testing for 308 unsolved CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line signifies the observed versus the expected $-\log_{10}(P)$ for each gene tested.

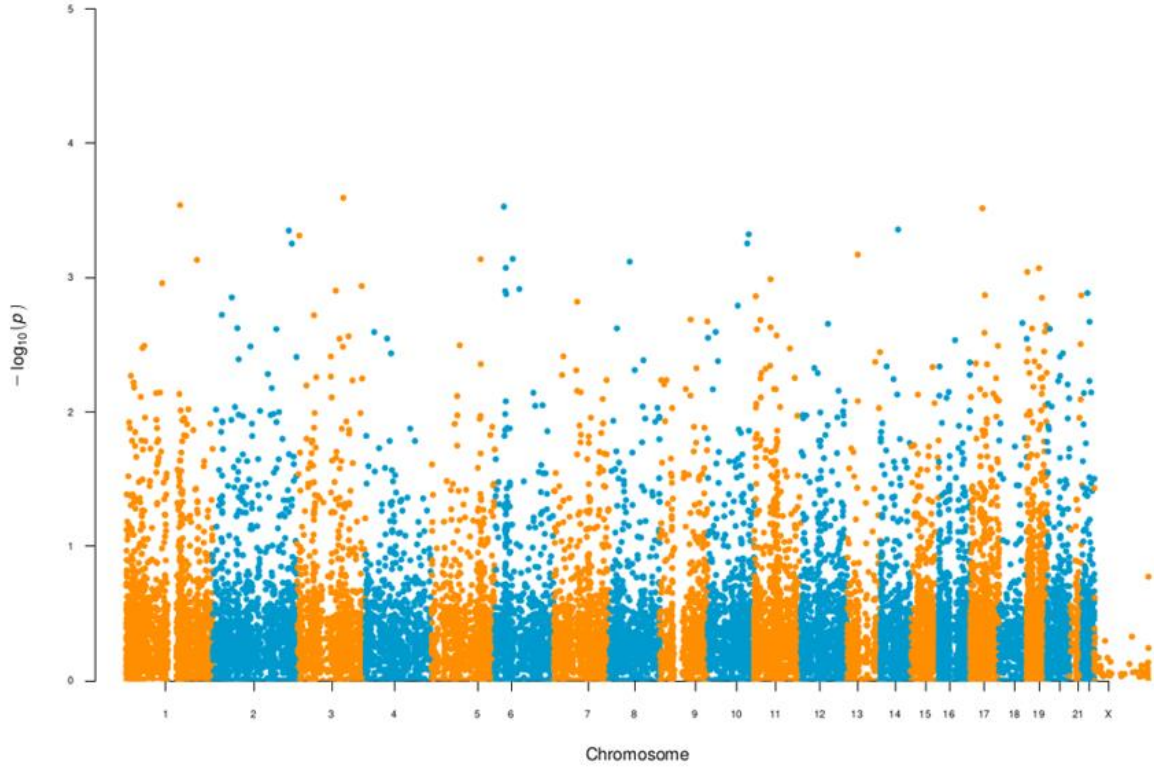

Figure 3-12 Gene based Manhattan for the association of likely damaging variants between the depleted unsolved CyKD cases and controls

Manhattan plot of exome-wide gene-based likely damaging variant association testing for 266 unsolved CyKD cases and 26096 ancestry matched controls. Each dot represents a gene. The red line indicating the exome-wide significance threshold of $P=2.58\times10^{-6}$ is not visible due to the lack of association and adjusted y-axis.

In all presented analysis the patients were heterozygous for their qualifying variants bar in *DNAJB11* where 59 of the 369 cases that had qualifying variants within the "missense+" tag were homozygous.

3.2.4.2.4 Non-coding collapsing analysis

Removing the cases with qualifying *IFT140* and *COL4A3* variants led to no further enrichment in the no variant detected cohort under the "missense+" or "LoF" gene collapsing tests. However, in the remaining 266 cases versus 26096 controls there was significant enrichment in acceptor gain (AG), acceptor loss (AL) and donor loss (DL) splice variants for *PKD1 (AG P= $6.70x10^{-11}$ OR=150.57 95% CI 35.39-730-24, AL P= $4.22x10^{-08}$ OR=398.51 95% CI 39.10-16384, DL P=$6.32x10^{-06}$ OR=no variants in controls)* and for DL in *PKD2 (P=$5.97x10^{-10}$ OR=no variants in controls) (*Figure 3-13*).*

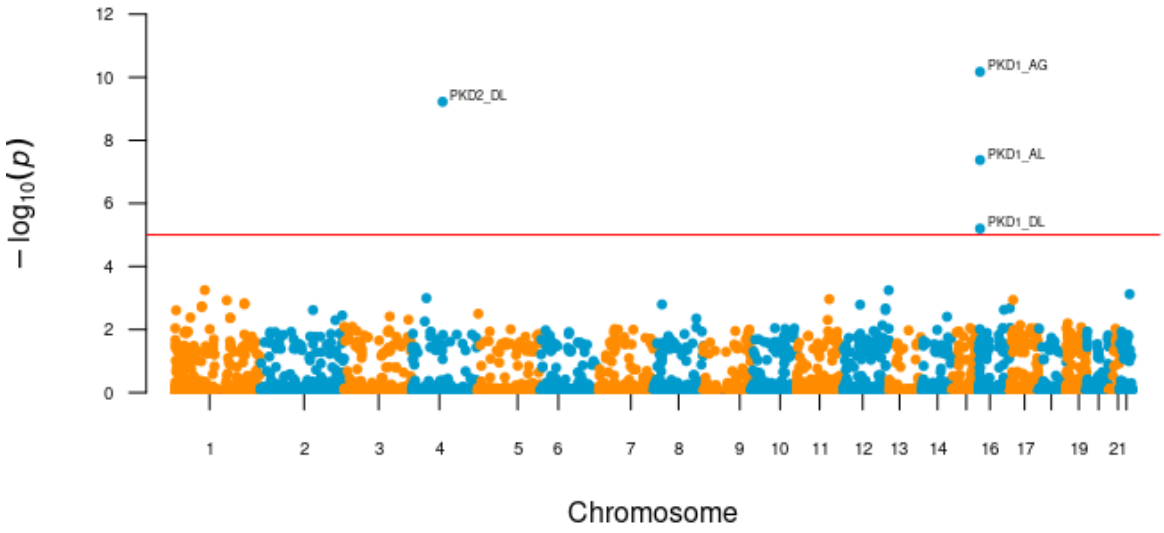

Figure 3-13 Gene based Manhattan for the association of splice variants between the depleted unsolved CyKD cases and controls

Gene based Manhattan plot of the SAIGE-GENE analysis with the splice mask. Each point is a gene representing the significance of the association with cystic kidney disease in 266 cases versus 26096 controls, made up of variants that are highly likely (SpliceAI score >0.8) to impact on splicing. The horizontal line indicates the threshold for genome wide significance.

There was no further enrichment in the 3' or 5'-UTR regions, intronic regions with a CADD score>20 or donor gain splice sites on a genome wide basis (all results available in the supplementary information).

3.2.4.2.5 *PKHD1* analysis

PKHD1 variants are responsible for ARPKD. The majority of ARPKD patients are compound heterozygous, carrying two variants at two different alleles, with 20% of all cases carrying a missense variant on exon 3 (c.107>T). 61 predicted LoF variants in *PKHD1* made up the association signal in the LoF mask analysis of the whole cystic disease cohort.

These were seen in 50 cases of which 22 were solved, 2 were partially solved, 2 were unable to ascertain their solved status and 24 were unsolved. All 50 cases were heterozygous for the variant that made up the signal.

Of the 22 solved cases 3 patients were solved with a diagnosis of ARPKD secondary to *PKHD1* variants, and 19 had variants secondary to *PKD1* or *PKD2*. In the 2 partially solved cases both patients had a second *PKHD1* variant deemed to be a variant of unknown significance (VUS).

Of the 24 unsolved cases with a single LoF *PKHD1* variant, four had a computationally predicted high impact non-truncating variant in *PKD1*, and 1 had a predicted high impact non truncating *PKD2* variant.

In the remaining 18 cases there were no single nucleotide variants, structural variants or copy number variants that would account for compound heterozygous diagnosis of ARPKD. Two patients had high CADD scoring variants in *PKHD1*, but both had been deemed "likely benign" by Clinvar (Clinvar ID: 1187104 and 102305).

In total 634 controls of the 26096 carried qualifying *PKHD1* LoF variants. When compared to the 18 out of 266 unsolved cases with no clear molecular cause of disease there is a significant enrichment of *PKHD1* variants in the unexplained cystic disease cohort ($P=5.85x10^{-06}$, OR=2.92, 95% CI 1.69-4.76).

3 of the 18 (16.67%) monoallelic *PKHD1* cases had reached ESRF with a median age of ESRF of 42 years. There was no statistical difference between the rates of liver cysts between the monoallelic *PKHD1* cohort and the general CyKD cohort (P=0.31). The full demographics table can be found below in table 3-6.

Table 3-6 Demographics of the *PKHD1* cohort

#	*Age*	*ESRF*	*CON*	*HGVSc*	*HGVSp*	*rsID*	*HPO*	*FH*	*Consanguinity*
1	15-20	No	Frameshift	c.525del	p.Asp175fs	rs1810924520	Hepatic fibrosis, cryptorchidism, hypertension, hypogonadism, CKD, polycystic kidney dysplasia, renal cysts	No	No
2	60-65	No	Frameshift	c.5411del	p.Arg1804fs	rs1554194574	Hepatic cysts, enlarged kidneys, renal cysts, gout, nephrolithiasis	No	Yes
3	60-65	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts	Yes	No
4	50-55	Yes	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts, hepatic cysts, CKD	No	No
5	70-75	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Enlarged kidneys, multiple renal cysts, hypertension	No	No
6	15-20	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts, hyperechogenic kidneys, hypertension, medullary cysts	No	No
7	75-80	No	Stop gain	c.1830T>A	p.Tyr610Ter	rs749293235	Enlarged kidneys, multiple renal cysts, cortical cysts	Yes	No
8	55-60	No	Stop gain	c.474G>A	p.Trp158Ter	rs1350620976	Hepatic cysts, multiple renal cysts,	No	No
9	65-70	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Hepatic cysts, multiple renal cysts, hypertension	No	No
10	60-65	Yes	Frameshift	c.3106dup	p.Trp1036fs	.	Multiple renal cysts, hypertension	Yes	No
11	30-35	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts, hypertension	Yes	No

12	5-10	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts, multiple cortical cysts	No	No
13	40-45	No	Frameshift	c.9689del	p.Asp3230fs	rs398124502	Hepatic cysts, multiple renal cysts, nephrolithiasis	No	No
14	60-65	No	Stop gain	c.1690C>T	p.Arg564Ter	rs765251347	Multiple renal cysts, cortical cysts, medullary cysts	Yes	No
15	60-65	No	Frameshift	c.3761_3762delinsG	p.Ala1254fs	rs398124484	Multiple renal cysts, hypertension, microscopic haematuria	No	No
16	35-40	No	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Clear cell renal cell carcinoma	No	No
17	20-25	No	Stop gain	c.10565C>A	p.Ser3522Ter	.	Hepatic fibrosis, multiple renal cysts, medullary cysts	No	No
18	60-65	Yes	Frameshift	c.5895dup	p.Leu1966fs	rs746838237	Multiple renal cysts, proteinuria, anemia, CKD, hypertension	No	No

CKD – Chronic Kidney Disease, ESRF – End stage renal failure, CON – Consequence, HPO – Human phenotype ontology, FH – Family history

3.2.4.2.6 *IFT140* analysis

27 cases within the 308 unsolved cases had a qualifying variant in IFT140 under the "missense+" tag. Of the 27 cases, all were heterozygous for the qualifying variants. None of the variants individually reached genome wide significance. There were no plausible second variants within *IFT140* that were candidates for a compound heterozygous mechanism of disease.

3.2.4.2.7 *COL4A3* analysis

Amongst the 15 unsolved cystic kidney disease patents with qualifying variants in COL4A3 under the "missense+" tag all were heterozygous for their respective variants and did not overlap with the unsolved *IFT140* cohort listed above. None of the variants individually reached genome wide significance. 5 of the 15 (33%) patients had reached ESRF (median age 58). 7/15 (46.76%) had hypertension and 4/15 (26.76%) had liver cysts.

3.2.4.2.8 Analysis of cohorts divided by primary driving variant.

Using the primary variant, the cystic cohort was divided into those cases with *PKD1* and *PKD2* truncating and non-truncating variants, respectively. Bar the primary gene in each cohort there was no further enrichment of any other gene genome wide (full results at supplementary data).

3.2.4.2.9 Pathway analysis

Using the NVD rare variant disease summary statistics from the SAIGE-GENE analysis as the input for the pathway analysis with GAUSS, did not reveal any significant associations in either the canonical or hallmark gene set pathways from the molecular signatures database (supplementary data).

3.2.5 Summary

- 936 of the 1209 probands had a likely monogenic SNV causing for their CyKD (77.42%)
- *PKD1* and *PKD2* were the major drivers of this with other known and recently described genes making up the remaining signal.
- *IFT140* despite being only newly described is the third most prevalent monogenic cause for CyKD in this cohort.
- The first population level evidence that *COL4A3* variants are associated with CyKD.
- Monoallelic *PKHD1* variants represent a risk factor for developing CyKD.
- Splice variants in *PKD1* and *PKD2* are a key cause of CyKD in an unsolved population.

3.2.6 Discussion

3.2.6.1 Prevalence of known monogenic disease and the use of WGS

WGS allows the capture of nearly all genomic variation in an unbiased way, allowing for analysis of SNV and SVs as well as systematic reanalysis of variants and genes (Costain et al. 2018). There is now compelling evidence that WGS gives superior diagnostic yield in rare diseases when compared to microarray, gene panels and WES; in particular the detection of deep intronic variants, splice variants, mitochondrial DNA and small SV/CNVs which have poor coverage on WES (Gilissen et al. 2014; Taylor et al. 2015; Stavropoulos et al. 2016; Lionel et al. 2018; Turro et al. 2020; 100000 Genomes Project Pilot Investigators et al. 2021). WES is also inappropriate for CyKD due to the pseudogene homology to *PKD1* (Ali et al. 2019).

995 of the 1209 (77.42%) tested cystic kidney disease cases had a monogenic cause for their disease identified in an unbiased way using statistically validated methods of analysing biobank-scale WGS data. Our diagnostic yield is comparable to the only other WGS study of CyKD where a clinically reportable results was found in 70% of an unselected diagnostic cohort (Mallawaarachchi et al. 2021). In work by Mallawaarachchi *et al* they used stricter criteria for inclusion of a variant as causative

and only looked at established genes due to the study being designed to match the standards of a clinical genetics pipeline. *IFT140* was also not yet described as a cause of CyKD and represents the third most common cause of CyKD at a population level (Senum et al. 2022). This confirms the high diagnostic yield of WGS in practice, and this technology is now available to all suspected CyKD patients in the UK via the National Health Service's Genomics Medicine Service.

3.2.6.2 CyKD as a monogenic disorder

Whilst a proportion of these variants would not necessarily meet the specificity requirements for issuing a clinically actionable molecular diagnosis, they give a firm basis for understanding the underlying genetic architecture of cystic kidney disease, namely that it is extensively driven by monogenic mechanisms.

The arguments for this position are compelling. Firstly, this unbiased method has confirmed the importance of established and newly described genes in the pathogenesis of cystic kidney disease (*PKD1, PKD2, IFT140, DNAJB11*), acting as a positive control. Contrast this to a recently published WGS analysis of posterior urethral valves (PUV), a disease thought to follow a non-Mendelian complex pattern of genetic aetiology, using the same methodology and sequencing platform published by our group (Chan et al. 2022) that showed no enrichment of rare monogenic causes of disease and confirmed a complex genetic architecture.

Secondly, we give robust statistical evidence that *COL4A3* is associated with cystic kidney disease. Smaller studies have hinted at this association(Gulati et al. 2020) and in sequencing of unexplained renal failure patients in an American cohort, a significant proportion of unexplained cystic cases were attributed to the *COL4A* family of genes (Groopman et al. 2019).

Finally, our findings are replicated in the UK Biobank with the top gene associations with cystic kidney disease using SAIGE-GENE being *PKD1* ($P=9.83x10^{-63}$), *PKD2*($P=1.64x10^{-60}$) and *IFT140* ($P=4.52x10^{-15}$) in a cohort of 531 patients versus

239,516 controls (Q. Wang et al. 2021). The inclusion of *COL4A3* in our cohort reflects the recruitment strategy of 100KGP which initially was biased towards those with atypical or molecularly unaccounted for CyKD.

3.2.6.3 Splice-sites and non-coding analysis

Using WGS we have also undertaken the first systematic assessment of non-coding variant contribution to CyKD. These contribute to unsolved cases highlighting the power WGS has in identifying sites previously untested by traditional sequencing techniques. Whilst splice site variants have been implicated in individual families with unexplained CyKD (Claverie-Martin, Gonzalez-Paredes, and Ramos-Trujillo 2015; K. Wang et al. 2009) and more recently in modestly sized cohorts (Hort et al. 2023) this analyses give robust statistical evidence at a population level that these sites should be scrutinized in *PKD1* and *PKD2*. These findings should help inform decisions about the sensitivity of other potential sequencing approaches in the clinical setting such as RNA-sequencing or long read DNA (Borràs et al. 2017) sequencing. Given the lack of RNA sequencing we are unable to functionally characterize the discovered splice variants however, so our conclusions rest on the enrichment of such variants in cases compared with controls and clinical actionability for participants in 100KG would be subject to cDNA confirmation on a case-by-case basis.

Whilst the analysis of the intronic and UTR space did not yield any results, by using robust case-controlled methodology I can be confident that the lack of signal is due to either lack of probands to detect what will likely be mild effect sizes or a true lack of signal.

3.2.6.4 Speculated mechanisms for COL4A3 and monoallelic PKHD1 variants

3.2.6.4.1 *COL4A3*

Diagnostic variants in the collagen genes are found not just in Alport syndrome but in a host of nephropathies including cystic kidney disease when unexplained CKD patients

are ascertained for genetic causes of their disease. (Groopman et al. 2019; Lata et al. 2018; Gulati et al. 2020). These genes encode collagen IV α3, α4 and α5 proteins, which combine to form the collagen IV α3α4α5 trimer — a key constituent of basement membranes in the glomerulus, eye and inner ear (Kamiyoshi et al. 2016).

Traditionally it was thought that cysts were more common in carriers of heterozygous pathogenic variants in *COL4A4* compared to *COL4A5* (Sevillano et al. 2014). However, there are many case reports of *COL4A3* variants being associated with renal cysts. Prior to this study renal cysts were more likely to be found in patients with proteinuria and decreased renal function, appearing before the age of 50 and not being associated with hypertension or liver cysts. There is no correlation between variant type and the likelihood of cysts developing (Savige and Harraka 2021).

However, in our analysis we find only *COL4A3* is associated with renal cysts and that a significant minority of the patients have both hypertension and liver cysts. Pathogenic collagen variants affect the collagen IV α3α4α5 network leading to basement membrane weakening which could in theory distend causing cysts. Given this collagen network is only expressed in the glomeruli and distal tubules one would expect the cysts to originate from there. Mouse models for *COL4A4* and *COL4A5* have dilated tubules but not *COL4A3* (Blake et al. 2003) and pathogenic variants in *COL4A1* which cause HANAC (hereditary angiopathy, nephropathy, aneurysms, and muscle cramps) have kidney cysts (Plaisier et al. 2007). Given a third of the *COL4A3* patients have ESRF in the 100KGP cohort, it seems plausible that cyst formation in this cohort could represent an accelerated form of acquired cystic kidney disease found in patients with ESRF, particularly those on dialysis. Without further imaging or more granular longitudinal data about renal function, further inferences are hard to make.

It is known that the penetrance of *COL4A* variants is highly variable. Whilst heterozygotes generally have a milder disease course than X-linked or autosomal recessive carriers, some heterozygous carriers exhibit phenotypes as severe (Fallerini et al. 2014; Rosado et al. 2015). Further studies using multi-omics data as well as saturation editing to assess function consequences of a given variant will aid in

facilitating a further understanding of the role of collagen variants in the pathogenesis of cyst formation (Findlay et al. 2018).

3.2.6.4.2 PKHD1

Historically, monoallelic variants in *PKHD1* were thought to be of no consequence, with the parents of ARPKD children being healthy, moreover, there is no kidney phenotype in heterozygous rodent models (Moser et al. 2005). However, an increasing number of case reports have associated monoallelic *PKHD1* variants with mild CyKD and PLD (J. Wang et al. 2021; Van Buren, Neuman, and Sidlow 2023). Work by Besse et al found an enrichment of monoallelic *PKHD1* variants ($P = 1.55 \times 10^{-9}$) in patients with PLD using a case-control analysis of gene burden in a European cohort of unexplained PLD patients who did not meet the criteria for diagnosis with ADPKD. 4 out of the 10 (40%) patients with qualifying variants also had atypical CyKD (Besse et al. 2017); this made up 10% of asymptomatic carrier adults studied.

In terms of mechanism of pathogenesis, the two most likely causes are either haploinsufficiency or somatic inactivation of the second *PKHD1* allele. Mice heterozygous for *PKHD1* will develop a cystic phenotype in keeping with medullary sponge kidney when aged ~1.5 years but this is in keeping with proximal tubule ectasia rather than a collecting duct pathology seen in ADPKD (Shan et al. 2019). An human ultrasound study of monoallelic carriers of pathogenic *PKHD1* variants found increased medullary echogenicity 6/110 (5.5%) of the cohort (Gunay-Aygun et al. 2011). These findings seem to suggest a somatic second hit due to the later presentation of the phenotype. Crucially, in both the mouse model and human studies mentioned above, liver fibrosis, a hallmark of ARPKD, seems to be missing, hinting that the timing of *PKHD1* expression is vitally important. This may be similar to *IFT140* where protein truncating variants are associated with a mild CyKD phenotype, whilst recessive variants cause a devastating multi-system ciliopathy with early paediatric presentation and high morbidity (Senum et al. 2022).

There is mixed data that the product of *PKHD1*, Fibrocystin/Polyductin (FPC) interacts with the polycystin complex. FPC may bind to PC2 and regulate calcium permeability and it interacts with PC1 in a dose dependant manner to affect cystogenesis in PLD (I. Kim et al. 2008; Fedeles et al. 2011). FPC itself has been implicated in ciliary targeting (Follit et al. 2010), nuclear translocation (Hiesberger et al. 2006) as well as a host of downstream molecular signalling cascades making unpicking it's precise role in the pathogenesis of cyst formation difficult. *PKHD1* also has many different expressed isoforms, making translation of bioinformatic variant analysis challenging without tissue specific expression data (Kaimori et al. 2007) (Menezes et al. 2004). This combined with the animal models for *PKHD1* and ARPKD not mimicking the human phenotype well (Wilson 2008) make further deduction of mechanism challenging. However, as discussed in chapter 3.1.2 there is increasing evidence of a shared mitochondrial pathway between *PKHD1* and *PKD1*.

3.2.6.5 Strengths and Limitations

This represents the largest sample of WGS CyKD disease patients analysed with this methodology. Having the cases and controls sequenced on the same platform removes confounding signals from sequencing artefacts and allows for quality control and variant processing to be done on the same pipeline. The use of a generalised logistic mixed model and case-control ancestry matching minimised confounding by population structure.

The main limitations lie in the power to detect signal in the remaining unsolved cases. Given the "positive control" findings in the total CyKD cohort (i.e. that *PKD1* and *PKD2* are by far the strongest monogenic signals) it is unlikely there was a methodological flaw when analysing the unsolved CyKD cohort. It is much more likely that I lacked the power to detect additional monogenic signals in this group – either because they have reduced effect size or are individually extremely rare. Alternatively, it may be that a proportion of this group exhibited cystic kidney disease because of non-monogenic developmental disorders or undocumented environmental exposures, such as to Lithium (Grünfeld and Rossier 2009). Irrespective this work gives an estimate of

the cohort size needed to power future studies to discover additional monogenic causes of CyKD using unbiased genome-wide approaches. Future work will involve rare-variant meta-analysis with other unsolved cohorts to improve power. Equally as more patients with CyKD are sequenced as part of their routine healthcare in the UK further monogenic causes will be discovered using this methodology.

I also have not combined variant types for this analysis. This chapter has focused on rare, damaging variants collapsed per gene. As the SV work will show in the next chapter, *HNF1β* SVs play in important role in CyKD, and I have not incorporated that with rare variant collapsing gene analysis from this chapter; such an approach leveraging multiple variant types may reveal novel genetic associations.

Equally, whilst I have attempted to incorporate some non-coding variants, I have only collapsed per gene. It is well documented that the non-coding genome is enriched for regulator elements that affect gene expression. Extending rare variant association analysis into this space is challenging (Bocher and Génin 2020) for two main reasons. Firstly, predicting pathogenicity in non-coding variants is challenging making it difficult to select variants to form part of a set. Secondly, deciding how to collapse variants isn't clear. Recent advances in the field have led to the development of a noncoding RV association detection framework, the STARRpipeline, that uses dynamic windows to annotate across the genome (Z. Li et al. 2022). This approach has led to non-coding variants being associated with lipid levels, eGFR and a number of other phenotypes (Selvaraj et al. 2022). Other approaches include using cell-specific experimentally predicted regulator regions to guide variant selection, an approach that led to the discovery of *TET2* as a casual gene in neurodegenerative disease (Cochran et al. 2020) however, this requires a wealth of experimentally generated data to work from. However, until bioinformatics tools are better validated for functional prediction, rare non-coding association studies remain challenging.

Finally, it can be argued that the very high "solve" rate from the research arm of my analysis is simply because I have lowered the threshold for a variant to be thought of as causative. It is true that missense variants in *PKD1* and *PKD2* have an incomplete and

variable penetrance (Rossetti et al. 2009). A recent WES analysis of an American ADPKD population only found 31.2% of patient with *PKD1* missense variants reported as “likely pathogenic” by the Mayo PKD database had ADPKD, although using different thresholds for classifying these variants as pathogenic (A. R. Chang et al. 2022). However, as an exercise in defining genetic architecture the missense variants included in this analysis are at the least “hot” variants of uncertain significance by ACMG criteria and serve their purpose in highlighting statistically robust signals. These signals have also been replicated in the UK Biobank under many different models that include missense variation. I am confident that these findings are reflective of the underlying genetic architecture of CyKD

3.2.7 Conclusion

In this chapter, I have presented evidence that CyKD is predominantly a monogenic disease. There was a significant and likely causative gene found in the majority of probands, representing one of the largest cohorts every studied. However, other forms of genetic variation play an important role and I will spend the next chapter discussing structural variation.

3.3 Structural variants in cystic kidney disease

3.3.1 Introduction

Structural variation (SV) represents a broad range of variants ≥ 50bps that can either be unbalanced or balanced (Figure 3-14). Unbalanced SVs include gains or losses of DNA, including copy-number variants (CNVs), whilst balanced rearrangements do not alter the dosage of the variant such as in inversions or translocations. Balanced variants are notable for not being detectable using conventional microarray-based methods. SV are increasingly recognised as having an important influence on genome structure and function (R. L. Collins et al. 2020).

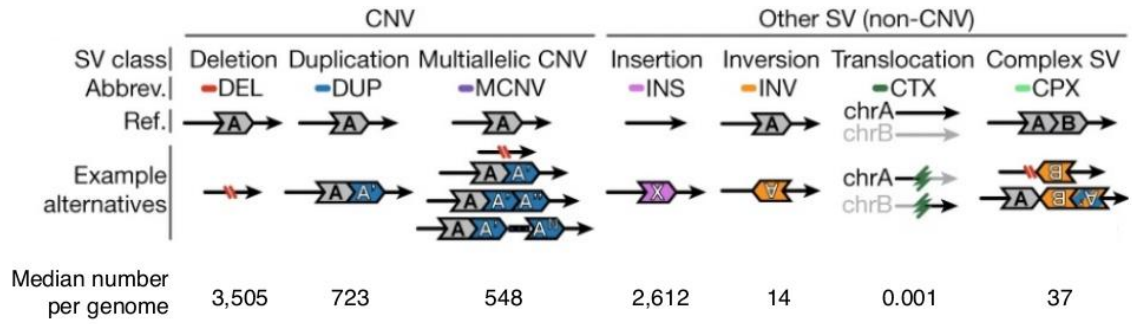

Figure 3-14 Types of structural variation

Median number of SVs per genome based on short-read WGS detection. CNV, copy number variant; SV, structural variant. Adapted from Collins et al. 2020.

SVs account for roughly 0.1% of all variants but due to their large size, they make a large contribution to the diversity between two human genomes compared to any other form of variation (Sudmant et al. 2015). As sequencing technology has improved the number of SVs predicted to exist per genome as also risen from 4500 SVs per genome using short-read WGS (Abel et al. 2020; R. L. Collins et al. 2020) to >25,000 SVs per genome with long-read WGS (Audano et al. 2019; Chaisson et al. 2019).

SVs are more damaging than SVs by either directly affecting gene function, altering the gene dosage or disrupting regulatory elements (Lappalainen et al. 2019). This is reflected in the negative selection seen against all SV types that overlap with genes or *cis*-regulatory elements (Abel et al. 2020; R. L. Collins et al. 2020). Copy-gain

duplications, however, have no selection pressure against them, in keeping with the role this type of event has in deriving new functions (Dennis and Eichler 2016).

Comparing detection rate of SVs in CyKD is made difficult by the shifting technologies by which SVs are called. SVs often occur in repeat-rich segmentally duplicated regions, making it problematic to detect and resolve breakpoints. Historically, microarray-based approached such as array CGH (aCGH) have been used to detect CNVs to a resolution of roughly 25kb. Array CGH cannot resolve small SVs, balanced SVs such as inversions or SVs not found in a reference genome. aCGH in the clinical lab has largely been superseded by multiplex ligation dependent probe amplification (MLPA), which has better resolution but is limited by the need to develop specific probes per gene. Next generation sequencing techniques and developments in the *in silico* tools used to SVs has enabled breakpoint resolution down to a single base pair (Ho, Urban, and Mills 2020). WGS for SV detection is still seen as a research tool but will likely replace arrays for clinical grade analysis in the future, allowing for hypothesis free genome wide testing.

Microarray and MLPA based SV/CNV analysis for CyKD has focused on targeted testing of known causative genes namely *PKD1/2*. Diagnostic yields range between 1.6%-7.1% in various ADPKD cohorts (D.-Y. Hwang et al. 2014; Schönauer et al. 2020; Kinoshita et al. 2016; D. Xu et al. 2018; Fujimaru et al. 2018; M. Zhang et al. 2019; Toshio Mochizuki et al. 2019), with most findings being in *PKD1/2* or occasionally *HNF1β*.

A single WGS paper published by Mallawaarachchi *et al.* in 2021 did detect SVs in 5.8% of the 40 cases tested using ClinSV a bioinformatic tool to call SVs genome wide (Mallawaarachchi et al. 2021; Minoche et al. 2021), with 3 exon crossing SVs in *HNF1β* and 1 in *PKD2*. However, this study was limited by small numbers.

At present, very little is known about the contribution of SVs to the CyKD phenotype at a population level. In this chapter I use the superior detection capabilities of WGS to

investigate the contribution rare, gene disrupting SVs make to the pathogenesis of CyKD.

3.3.2 Aims

1. To ascertain whether rare gene-disrupting structural variation is associated with CyKD when compared with an ancestry matched control population using an unbiased collapsing exome-wide approach.

3.3.3 Methods

3.3.3.1 Variant calling

Structural variants were called from WGS using the Genomics England pipeline that incorporates CANVAS (Roller et al., 2016) to detect copy number (>10kb) and MANTA (Xiaoyu Chen et al. 2016) to identify SVs greater than 50bp, but less than 10kb. CANVAS uses read depth to assign CNV losses and gain. MANTA uses both discordant read-pair and split-read data to identify SV regions. While MANTA can detect deletions and tandem duplications < 10kb, inversions, and interchromosomal translocations it cannot reliably identify dispersed duplications, small inversions (< 200bp), fully assembled large insertions (> 2x150bp) or breakends where repeat lengths approach the read size (150 bp). Very few insertions were identified in this cohort using MANTA and in view of this they were excluded from downstream analysis. In addition, variants classified as translocations, single breakends or complex SVs which are more difficult to accurately resolve were filtered out. Both tools are widely used in the bioinformatics community and perform consistently well when compared with other SV callers (Cameron, Di Stefano, and Papenfuss 2019).

3.3.3.2 Quality control

The following quality control filters were applied to the variants:

- CNV length > 10kb and Q-score ≥ Q10 indicating 90% confidence there is a variant present.

- A quality score ≥ 20 indicating 99% confidence that there is a variant at the site, GQ ≥ 15 indicating 95% confidence that the genotype assigned to a sample is correct, and MaxMQ0Frac < 0.4 which indicates the proportion of uniquely mapped reads around either breakend.

Variants without paired read support, inconsistent ploidy, or depth > 3x the mean chromosome depth near one or both breakends were excluded.

3.3.3.3 Extraction of exon crossing SVs and filtering by allele frequency

For each sample BEDTools (Quinlan and Hall 2010) was used to extract SVs that intersected at least one exon by a minimum of 1 base pair.

Variants were then separated by type into CNV, deletion (DEL), duplication (DUP), and inversion (INV) sets before being filtered using BEDTools to remove common SVs of the same type. SVs were removed if they had a minimum 70% reciprocal overlap with:

a) the gnomAD SVs (R. L. Collins et al. 2020) with allele frequency >1%

and/or

b) dataset of common (AF>0.1%) SVs generated from 12,243 cancer patients recruited to 100KGP. SVs were then merged using SURVIVOR (Jeffares et al. 2017) allowing a maximum distance of 300bp between pairwise breakpoints and allele frequencies calculated using BCFtools.

Overlapping common variants were removed and then a custom Perl script (Dr Helen Griffin, Newcastle University) was used to calculate allele frequencies for each type of SV across the combined case-control cohort using bins of 10kb across the entire genome. SVs with an AF < 0.1% were retained for further analysis.

3.3.3.4 Burden analysis

Exome-wide gene-based burden testing was carried out using custom R scripts stratified by SV type. SVs were aggregated across 19,005 autosomal protein-coding genes. A two-sided Fisher's exact test was used to compare the burden of rare (MAF<0.1%) SVs in cases and controls under a dominant inheritance model globally and per type of SV. The length of SVs was compared in cases versus controls using a Wilcox-Mann-Whitney test. The Bonferroni correction for the number of genes ($P=0.05/19{,}005=2.6x10^{-6}$) tested was applied, although with the knowledge that this is likely to be too stringent given the tests are not truly independent (one SV can affect multiple genes).

3.3.4 Results

Summary statistics used to generate all the Manhattan plots below is available in the supplementary information.

3.3.4.1 Burden of gene-disrupting structural variation

Across each type of SV in rare (AD<0.1%), exon crossing variants of at least 50bps there was no enrichment in cases versus controls globally. The median size of inversions in cases was nearly double that of controls (table 3-7).

Table 3-7 Burden of rare, autosomal, exonic structural variants in CyKD probands versus controls

SV type		Cases (n=1209)	Controls (n=26096)
CNV	n(%)	976(81%)	20607(79%)
	OR (CI)	1.0(0.92-1.08)	-
	Fisher's P	1.0	-
	Median size (kb) (IQR)	90 (170)	89 (170)
	P (Wilcoxon)	0.86	-

DEL	n(%)	1059(88%)	22449(86%)
	OR (CI)	1.0(0.92-1.07)	-
	Fisher's P	0.84	-
	Median size (kb) (IQR)	1.6(4.3)	1.7(4.4)
	P (Wilcoxon)	0.06	-
DUP	n(%)	453 (38%)	9746 (37%)
	OR (CI)	1.0(0.88-1.08)	-
	Fisher's P	0.66	-
	Median size (kb) (IQR)	3.1 (5.4)	3.1 (5.4)
	P (Wilcoxon)	0.78	-
INV	n(%)	169 (14%)	3215(12%)
	OR (CI)	1.12 (0.95-1.31)	-
	Fisher's P	0.19	-
	Median size (kb) (IQR)	202 (2141)	440(3395)
	P (Wilcoxon)	0.02	-

CNV – copy number variation, DEL – deletion, DUP – duplication, INV – inversion, OR – odds ratio, IQR – interquartile range, CI – Confidence interval (95%)

Analysing CNVs by size did not show any significant difference between cases and controls (Figure 3-15)

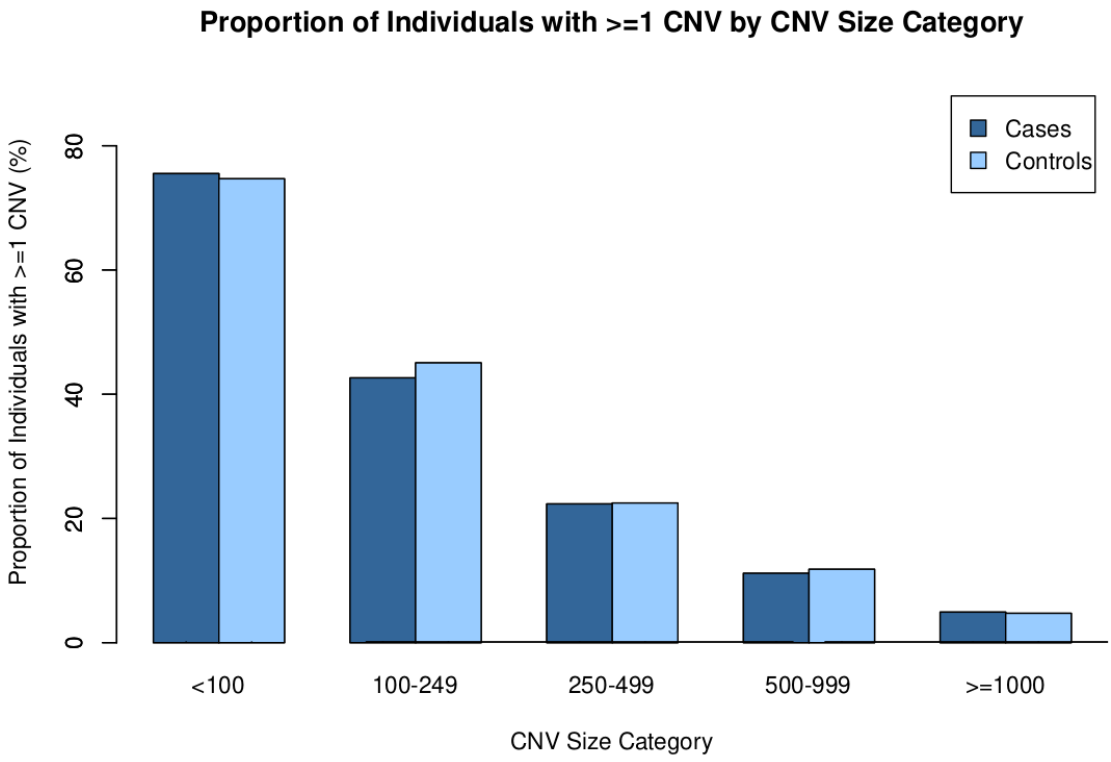

Figure 3-15 CNV sizes in cases vs controls (kb)

Individuals with at least one CNV were included and the proportion of their CNV burden by size is illustrated here with CNV sizes in kilobases (kb).

3.3.4.2 Exome-wide gene-based burden testing

Exome wide gene-based SV/CNV analysis was performed in CyKD cases and ancestry matched controls. Across all combined types of SV and CNV there was significant enrichment in *PKD1*($P=2.02\times10^{-14}$, OR=2.52 95%CI 1.69-3.63) and *PKD2*($P=7.48\times10^{-12}$, OR=3.51, 95% CI 1.74-6.37). The next set of genes reached genome wide significance on a per gene basis but combined represented the genes found within the 17q12 locus including *HNF1B* ($P=8.81\times10^{-09}$, OR=7.11, 95% CI 3.41-13.66). Of note two genes within proximity of *PKD2* also reached genome wide significance *SPARCL1* ($P=5.76\times10^{-07}$) and *HSD17B11* ($P=8.69\times10^{-06}$) but these were made up of large deletions that encompassed *PKD2* also (Figure 3-16).

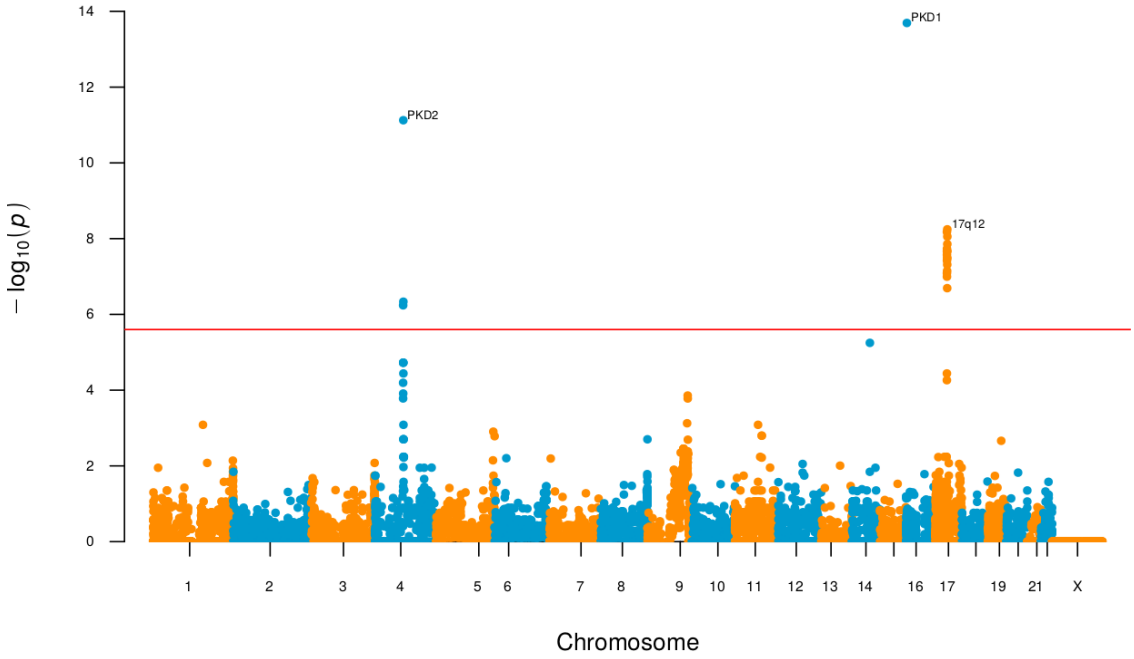

Figure 3-16 Gene based Manhattan of the association of structural variants between all CyKD cases and controls.

Exome wide gene-based SV/CNV analysis was performed in 1209 CyKD cases and 26096 ancestry matched controls. SVs had to be exon crossing and rare to be included in the analysis (MAF<0.1%) with no overlap with a selection of common SVs taken from the GEL cancer cohort and the gnomAD SV cohort. The red line represent genome-wide significance.

The *PKD1* signal was predominately driven by small deletions, with no other genes reaching genome wide significance for deletions <10 kb ($P=2.17x10^{-22}$, OR=8.11 95% CI 4.58-13.83). For *PKD2* ($P=7.48x10^{-12}$, OR=13.03 95% CI 5.02-31.87) and the 17q12 locus ($P=4.12x10^{-08}$, OR=8.70, 95% CI 3.72-18.80) this signal was driven by large deletions with no other loci reaching genome wide significance. No genes reached genome wide significance for duplications or insertions (Table 3-8).

Table 3-8 Comparison of SV sizes in gene enriched in the CyKD cohort

Gene	Driving SV type	Median Case Size(kb) (IQR)	Median Control Size (IQR)	Wilcoxon (P-value)
PKD1	<10Mb deletions	1.14(0.46-3.06)	N/A	N/A
PKD2	>10Mb deletions	405.52 (97.89-1371)	1107 (1107-1107)	0.8
HNF1β	>10Mb deletions	1550 (1545-1639)	1549 (1546-1550)	0.8

IGQ – interquartile range, kb – kilobase, Mb - Megabase

3.3.4.3 Genotype/phenotype correlation

Of the 46 patients with rare exon crossing SV/CNVs in *PKD1* or *PKD2*, 13 also harboured predicted LoF variants in *PKD1* or *PKD2*, thus leaving 33 patients with cystic kidney disease attributable to SV/CNVs in *PKD1* or *PKD2*.

Of the 10 patients with 17q12 loci CNVs in the cystic disease cohort, 1 patient had a *PKD1* non truncating single nucleotide variant and 2 had *PKD1* truncating single nucleotide variants that met the criteria for being likely causative and 1 patient was solved for a known *HNF1β* CNV detected by a separate diagnostic lab prior to the return of 100KGP results.

Analysing the subgroup of patients without an identified molecular diagnosis (n=266), there was significant enrichment for CNVs at the 17q12 loci (lowest significant P=9.21x10-09, OR=24.04 94% CI 8.00-60.71).

Of the 7 17q12 patients none had reached ESRD, and the median age of the cohort was 13.5 years, significantly lower than the total cystic disease cohort median age ($P<0.05$). None of the patients had HPO or HES codes pertaining to diabetes; a full breakdown of phenotypic profile can be found in table 3-9).

Table 3-9 Phenotype breakdown of patients with *HNF1β* CNVs as their likely causative variant for CyKD

ID	*Renal HPO terms*	*Extra-Renal HPO terms*	*HES codes*
1	Renal Cyst	Decreased liver function, unilateral cryptorchidism	Cardiomegaly, elevated transaminases
2	Cystic renal dysplasia, multicystic kidney	Patent ductus arteriosus, coarctation of aorta, splenic cyst	Cardiomegaly, essential hypertension
3	Enlarged kidney, multiple renal cysts, renal cortical cysts, multiple glomerular cysts, multiple small medullary renal cysts	None	Unobstructed inguinal hernia
4	Renal cortical cysts, calculus of kidney	None	Congenital hypotonia, dermatitis
5	Renal cortical cysts	None	None
6	Renal cortical cysts	None	None
7	Multiple renal cysts, unilateral renal atrophy	Endometriosis	Adenocarcinoma in situ, costal chondritis; urinary tract disease, angina, cervical dysplasia, essential hypertension, family history of diabetes, gout

HES – hospital episode statistics, HPO – human phenotype ontology

3.3.5 Summary

- As a whole there was no significant enrichment of rare, exon spanning structural variants in CyKD versus controls.
- Structural variants play a large role in the variance landscape of CyKD in known monogenic drivers of CyKD namely *PKD1, PKD2* and the 17q12 loci.
- 3.35% (40/1209) of the CyKD burden is attributable to SV/CNV disease.
- Causative *PKD1* SVs tend to be small <10kb.
- Causative *PKD2* and 17q12 SVs tend to be large and >10kb.

3.3.6 Discussion

3.3.6.1 WGS and SV calling in CyKD

To date, small studies in ADPKD cohort have yielded a diagnostic CNV burden between 1.6-7.1% for SV/CNVs in ADPKD cohorts using a range of methods from Sanger sequencing to long range PCR (Claus et al. 2022). Only one study used WGS as mentioned in the introduction, and none used control datasets to understand the population significance of the data.

This study represents the first case-control genome wide analysis of SV/CNVs in CyKD. Our diagnostic yield of 3.4% is comparable to other studies but is far larger in power and scope. By analysing our data with a control cohort, we are able to quantify the risk of SVs in CyKD revealing a significant risk (odds ratios are between 8-24 which are broadly comparable to SNVs in *PKD1/2*) . These findings are consistent with known pathogenic mechanisms of CyKD, namely that *PKD1* and *PKD2* are disease genes and that 17q12 CNVs can manifest with renal cysts. However, whilst 17q12 has always been known to cause renal cysts, this study highlights its importance as a cause of renal cysts alone with none of the 7 patients diagnosed with diabetes. This confirms the need to screen the 17q12 loci for CNVs as part of a CyKD genomic workup.

3.3.6.2 Understanding the biology of the findings

Why are the structural variants in *PKD2* nearly 400 times longer on average than those found in *PKD1*? The answer may lie in the underlying genetic architecture of these two genes.

PKD1 is a 47.2kb gene made up of 46 exons with the first 33 exons sharing 97.7% sequence homology to six pseudogenes that lie proximally to *PKD1*. These pseudogenes have arisen due to successive segmental duplication events since primate evolution (Kirsch et al. 2008). Pseudogenes are under less selection pressure and thus tend to have a high mutation rate (Claes and De Leeneer 2014) and the region as a whole has a very high GC content. Segmental duplications (SD) are defined as areas of DNA that contain >90% sequence identity and >1kb in length in the reference haploid genome. Genome-wide analysis of SD and low copy repeats (LCR – an umbrella term encompassing SDs as well as other repeat sequences) highlight how they overlap with regions with high rates of genomic rearrangements that are often associated with disease (Bailey et al. 2002). LCRs affect genomic stability making the flanked regions more likely to undergo nonallelic homologous recombination (NAHR) or replication based mechanisms (RBMs) (Stankiewicz and Lupski 2002). Moreover, the formation of structural variants can both accompanied or caused by SNVs, highlighting a complex interplay between the two types of variants (Carvalho et al. 2013). These facts coupled with the fact that *PKD1* has a high GC rich content making it more prone to spontaneous deamination and gene conversion (Coulondre et al. 1978; Alexandrov et al. 2020) mean *PKD1* has multiple sites for potential SV breakpoint formation within itself and its surrounding regions (Figure 3-17).

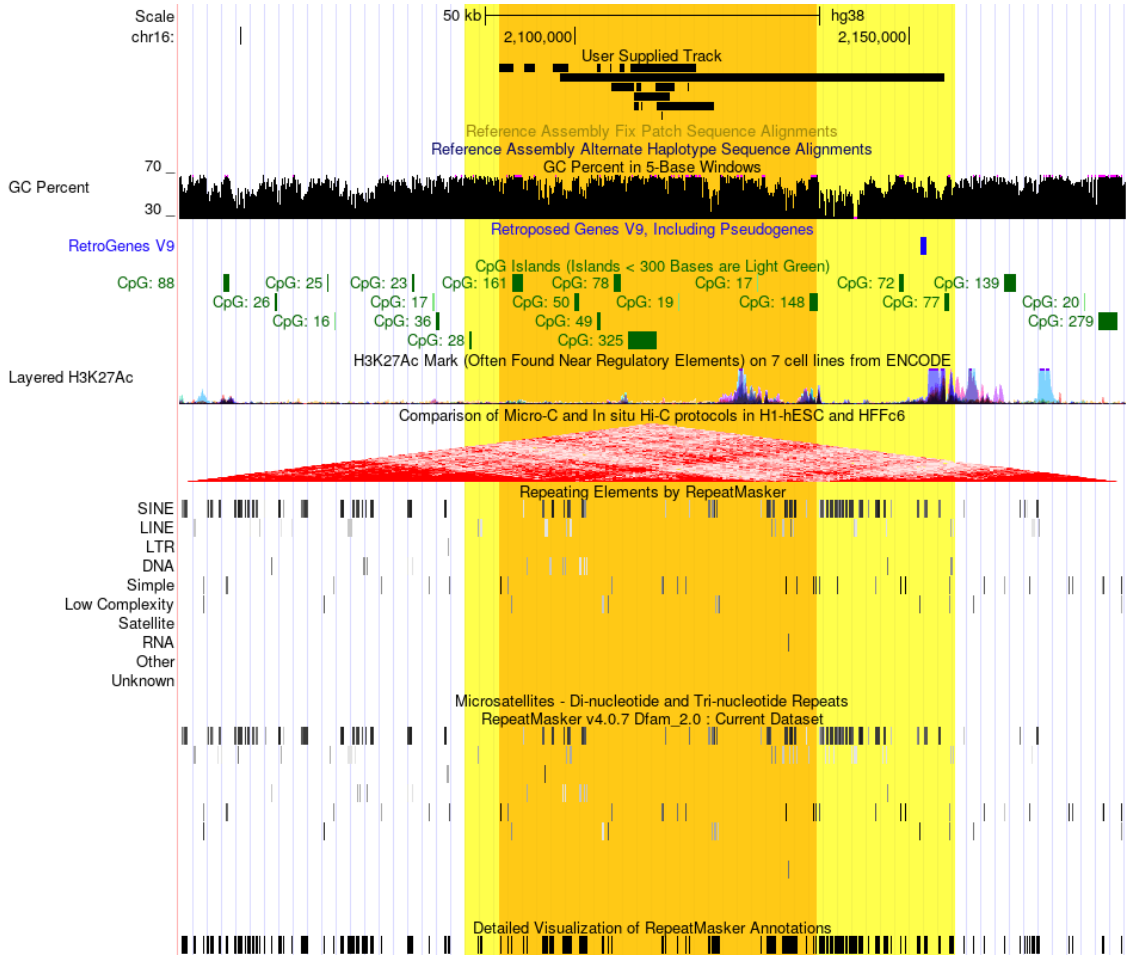

Figure 3-17 PKD1 deletions mapped to different genomic features

An image from the UCSC genome browser with the *PKD1* deletions at the top in yellow, the orange is the *PKD1* gene. The annotated features are the GC percentage (high), the presence of CpG islands, the layered chromatin bindings sites, the TAD and then areas of repeats sequence (SINE/LINE/LTR etc). Compared to Figure 3-18, there is far more GC content and far fewer repeating elements

PKD2 is a 68kb gene on chromosome 4 with 15 exons and no associated pseudogenes. Whilst not an LCR hotspot, it does have a significantly higher coverage by *Alu* elements, a transposable elements accounting for up to 10% of the human genome, when compared to *PKD1* (Deininger 2011). These elements form hotspots for nonrecurrent rearrangements via mechanisms other than NAHR such as microhomology-mediated break-induced replication (MMBIR) (Mayle et al. 2015). There is speculation that *Alu* elements may provide the microhomology islands that can act as potential breakpoints for larger structural rearrangements (Carvalho and Lupski 2016). However, the 17q12 loci pathology is mediated by microdeletions of this region caused by flanking segmental duplications via NAHR and *PKD2* is equally flanked by areas of low complexity. Added to this complexity is that fact that the rate of meiotic

NAHR correlates in monozygotic twins and is independent of age insinuating unknown genetic and/or environmental factors that affect the control of NAHR (J. A. L. MacArthur et al. 2014) . Figure 3-18 maps these features to the gene and deletions.

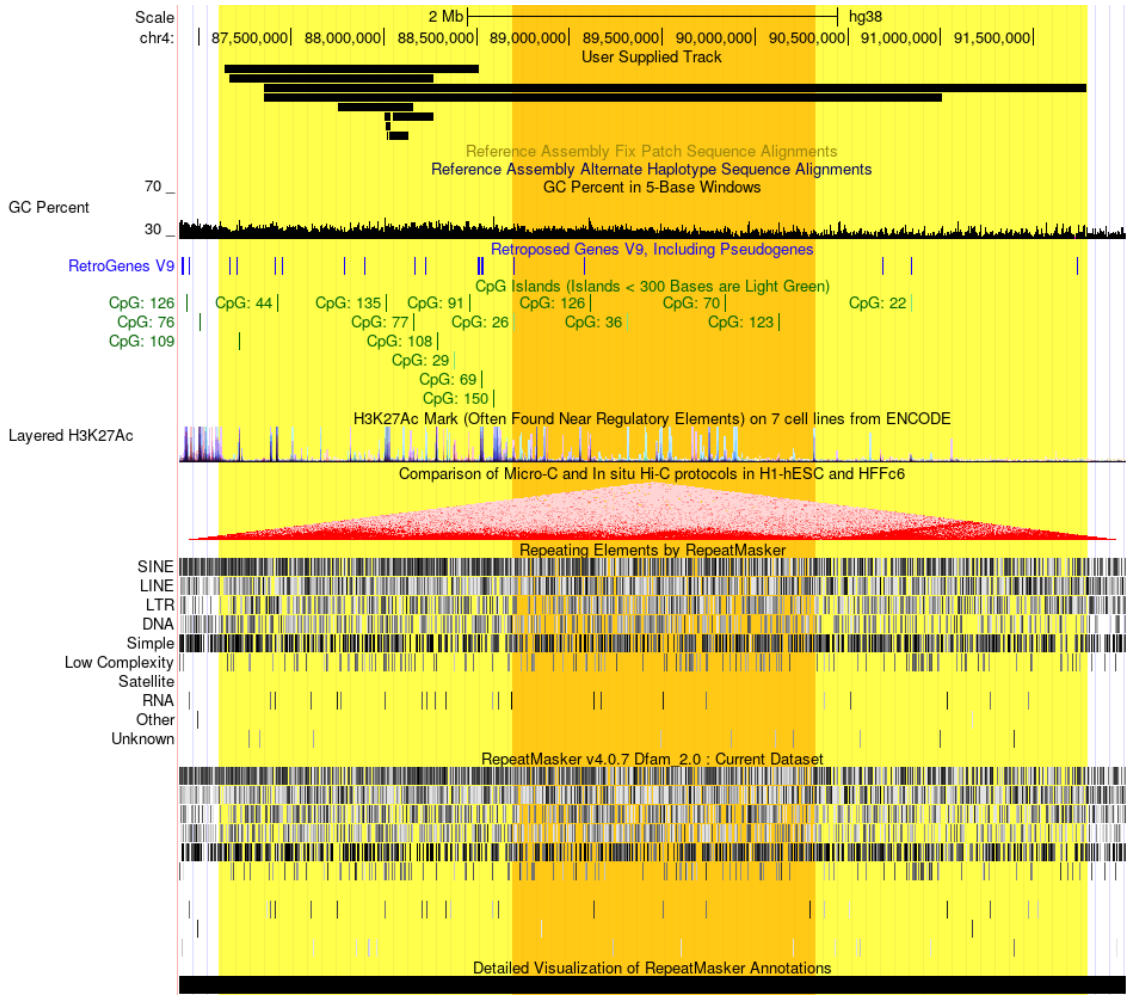

Figure 3-18 PKD2 deletions mapped to different genomic features

An image from the UCSC genome browser with the *PKD2* deletions at the top in yellow, the orange is the *PKD2* gene. The annotated features are the GC percentage (low), the presence of CpG islands, the layered chromatin bindings sites, the TAD and then areas of repeats sequence (SINE/LINE/LTR etc). Compared to Figure 3-17, there is far less GC content and far more repeating elements.

SV interpretation is a complex and nascent field both at the mechanistic and bioinformatic level. At this stage it is simply speculation as to the differences between the rare, exonic SV calls in the two top candidate genes.

3.3.7 Strengths and Limitations

This is the first comprehensive assessment of structural variation in CyKD using WGS data. By using a case control cohort, I was able to quantify the risk from the associated genes in the CyKD cohort, allowing for population inferences about the role SVs play.

Limitations wise, unlike my analyses using logistic mixed models, I was unable to control for population structure using principal components. The burden analysis as based on a dominant inheritance model and any additive or recessive effects may have been unascertained. I was also unable to confirm any of my findings using orthogonal approaches such as long read sequencing or PCR-based methods or independently verify my findings in another independent cohort. Finally, by calling SVs in 10kb bins and calling the MAF from these I may have incorrectly grouped some rare SVs, excluding them from the analysis.

From a technological standpoint, whilst WGS performs well at resolving SVs there are some limitations. WGS suffers from read mapping ambiguity when large variants are found in complex, repetitive or GC-rich regions and is easily outperformed by long read platforms such as PacBio (Chaisson et al. 2019). Finally, it has now become gold standard to use multiple SV callers on WGS data and then merge the outputs to create a consensus SV set to reduce the known problem of false positive calls. CANVAS and MANTA were used by Genomics England prior to such a consensus being reached in the SV calling community, in the future using a tool that calls SVs with multiple callers such as Parliament (Zarate et al. 2020) will allow for reduction in false positive calls.

3.3.8 Conclusion

Whilst the ability to sequence and call SVs has improved exponentially, the interpretation and analysis of them remains nascent. In this chapter I identified an enrichment of rare, exonic SVs in the CyKD population. Future work will be necessary to validate and replicate these findings as well as to tease apart the biological mechanisms involved.

3.4 Common variants in cystic kidney disease

3.4.1 Introduction

Thus far I have focused on rare variation in CyKD. In this chapter I focus on the role common variants have in CyKD. Whilst CyKD is seen as a monogenic disorder, there has been great interest in the role common variants have to play in CyKD. This is driven by observations around the intrafamilial variability of the phenotype in the presence of a known monogenic driver, hinting at modulating factors such as common variants or environmental factors.

GWAS traditionally was constrained by the cost of sequencing, meaning populations were sequenced using array-based platforms and imputation panels then applied. This limited the analysis in terms of genome coverage and populations studies, with most GWAS analyses occurring in European populations. With the advent of WGS and large scale biobanks, mixed ancestry WGS GWAS that are not reliant on imputation (using predominantly European panels) have been performed with success (Taliun et al. 2021; DiCorpo et al. 2022).

3.4.2 Aims

1. To assess the contribution of common and low-frequency SNVs and indels (MAF >0.1%) to the genetic architecture of CyKD as a whole and when stratified by primary driving variant and time to ESRF using a sequencing-based GWAS.
2. To perform a meta-analysis of CyKD GWAS studies across population biobanks to boost power to detect SNV associations.
3. To use this data to generate an estimate of the contribution to heritability of common and low-frequency variants.

3.4.3 Methods

3.4.3.1 Sequencing-based GWAS

Whole-genome single-variant association analysis was carried out using the R package SAIGE (version 0.42.1) (W. Zhou et al. 2018) which uses a GLMM to account for population stratification. High-quality, autosomal, bi-allelic, LD-pruned SNVs with MAF >5% were used to generate a genetic relationship matrix and fit the null GLMM. Sex and the top 10 principal components were used as covariates (fixed effects). SNVs and indels with MAF ≥0.1% that passed the following quality control filters were retained:

- MAC ≥20
- missingness <1%
- HWE $p>10^{-6}$
- differential missingness $p>10^{-5}$.

One limitation of SAIGE is that the betas estimated from score tests can be biased at low MACs and therefore odds ratios for variants with MAF <1% were calculated separately using allele counts in R.

The R packages qqman (D. Turner 2018) was used to create Manhattan and Q-Q (quantile-quantile) plots. A Bonferroni corrected P-value threshold of $5x10^{-08}$ for genome wide statistical significance was used to account for the number of independent haplotypes tested. The genomic inflation factor (lambda), calculated based on the 50th percentile, was between 0.99-1.02 in all analyses indicating no significant population stratification.

3.4.3.2 Conditional analysis and epistasis

For any genome-wide significant loci, SAIGE was used to perform conditional analysis on the lead variant to look for secondary associations as well as fine mapping of the locus by analysing variants with a MAC ≥ 3 to ascertain if rare variants were driving the observed signal.

PLINK (version 1.9) was used to assess epistasis between lead variants.

3.4.3.3 Heritability analysis

The proportion of phenotypic variance due to genetic factors is known as the heritability of a condition. Broad sense heritability (H^2) is the variance explained by all genetic factors under all models (dominant, additive and epistatic) but is hard to estimate without a number of assumptions (Visscher, Hill, and Wray 2008). Narrow sense heritability (h^2) is the contribution to phenotypic variance from additive genetic factors only and is hence easier to estimate, normally coming from GWAS outputs which assume an additive model.

To estimate the phenotypic variance from common and low-frequency variants in CyKD, heritability analysis was performed with GCTA (version 1.93.1beta) (J. Yang et al. 2011) on a European subset of total ancestry-matched CyKD cohort. Variants with MAF ≥ 0.1% were included. Using the GREML-LDMS approach (J. Yang et al. 2016) variants were stratified into seven different bins based on MAF (0.001-0.01, 0.01-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3- 0.4, 0.4-0.5) and for each bin of variants, SNP-based LD scores were calculated over a 200kb region (with 100kb overlap between two adjacent segments). For a given bin of variants defined by MAF, variants were further stratified into quartiles using LD scores. For each of the 28 bins subset by MAF and LD, GCTA was used to produce a GRM from the raw genotype files. The REML (restricted maximum likelihood) function was then used to conduct a GREML-LDMS analysis using the 28 GRMs, including the top four principal components as covariates. GREML has not been validated for mixed ancestry cohorts, so an estimation was made using a European subset of the CyKD cohort as defined by principle component analysis (903 cases and 20255 controls). The observed heritability was then liability adjusted to account for the population prevalence of CyKD relative to its representation in the 100KGP (S. H. Lee et al. 2011). In this analysis a CyKD prevalence of 0.001 was used to transform the observed heritability to a liability threshold model.

3.4.3.4 Polygenic association between monogenic and unsolved CyKD

GWAS studies have uncovered thousands of SNVs associated with traits throughout the human genome. The same studies also highlight how the majority of genetic variants have a small effect on the phenotype and are not predictive of disease risk (Manolio et al. 2009; Visscher et al. 2017). This has led to the development of polygenic risk scoring (PRS), a method to aggregate the germline genetic risk for a trait. It is a number that summarises the effect of many different genetic variants on an individual's phenotype. This is typically calculated as a weighted sum of trait-associated alleles as illustrated by the equation where $\hat{S}$ is the sum across the *m* numbers of SNVs with risk increasing alleles weighted by their $\hat{\beta}_j$ weights:

$$\hat{S} = \sum_{j=1}^{m} x_j \hat{\beta}_j$$

However, whilst the primary focus of PRS has been on risk prediction for diseases it can also be used as a method to assess heritability (S. W. Choi, Mak, and O'Reilly 2020). It has been shown that using a linear mixed model one can judge heritability of trait by evaluating the effects of all SNVs simultaneously (J. Yang et al. 2010). This led to other statistical methods being explored that simultaneously assess SNV contribution to a phenotype including LD score regression (Bulik-Sullivan et al. 2015) and PRS (Palla and Dudbridge 2015; Dudbridge 2013). PRS has been shown to compare favourably to these other methods in assessing heritability (L. M. Evans et al. 2018).

PRS can be derived through multiple methods with the focus being on how the number of SNVs to be included is derived, methods include: pruning and thresholding were a subset of genetic markers are selected through LD and P-value filtering of GWAS summary statistics (International Schizophrenia Consortium et al. 2009) , Bayesian polygenic prediction methods such as LDpred (Vilhjálmsson et al. 2015) that incorporate genome-wide markers and penalized regression (Mak et al. 2017). Generally, as the methods have developed more variants have been included with the methods to handle the effect sizes becoming more sophisticated.

Once the PRS has been generated it is fitted as a logistic model against the binary phenotype and any co-variates with the metrics of goodness of fit being used to give an estimate of heritability. This includes the variance explained or R^2, which is a well-defined concept in continuous traits. For binary outcomes several proxies for this have been developed (pseudo-R^2) such as Nagelkerke R^2 (Heinzl, Waldhör, and Mittlböck 2005). However, when estimating the heritability of a binary trait using PRS in a case control study the proportion of cases included is overall higher than the prevalence in the general population. This makes any estimation of heritability liable to ascertainment bias (S. H. Lee et al. 2011). This observed heritability can be transformed to an underlying continuous liability threshold model that uses the ratio of cases to controls and the population prevalence of the disease to give a less biased heritability. Of note PRS have limited portability across different ancestral groups (Alicia R. Martin et al. 2020), irrespective of the method used to generate the score (Alicia R. Martin et al. 2019; Duncan et al. 2019). Given I conducted the heritability analysis on a European cohort, and the only published estimate of heritability of ADPKD used a European population (Blair, Hoffmann, and Shieh 2022) I opted to conduct PRS fitting in a European cohort of CyKD also to allow for easier comparisons.

To this end the summary statistics from the GWAS of unsolved CyKD was used to generate a PRS that best explained the variance in the phenotype of a European subset of the molecularly solved cases to ascertain if there is shared polygenic architecture between the two cohorts. Given the GWAS studies used the same controls, I generated a new "monogenic" CyKD cohort taking those patients who had a clear monogenic cause for their disease and then ancestry matched them to patients recruited to the 100KGP with cancer who were of European ancestry, excluding those with urogenital cancers or any HPO code pertaining to renal disease (685 cases and 9856 controls). This was done using the PRSice2 tool (version 2.3.3) to generate multiple sets of polygenic risk scores at different P-value thresholds relative to the GWAS summary statistics (S. W. Choi and O'Reilly 2019). The best PRS is then selected (in this case at a P-value threshold of 0.027 made up of 55557 variants) and fitted as a logistic model (with sex and 4 PCs as covariates) to the data with a liability adjusted heritability calculated using the pseudo-

R^2 (calculated as the difference between the full model R^2 and the null model R^2 using a population prevalence of CyKD of 0.001) in R with the pscl tool (Zeileis, Kleiber, and Jackman 2008). A permutated P-Value is then generated to ascertain the model fit (10,000 permutations). The model AUC was calculated using DescTools (Signorell 2023) with the 95% confidence interval calculated using bootstrapping with 10,000 permutations using the boot tool ("Bootstrap Functions (Originally by Angelo Canty for S) [R Package Boot Version 1.3-28.1]" 2022).

3.4.3.5 Meta-analysis

Meta-analysis aims to combine the evidence for association from individual studies using appropriate weights. There are multiple statistical methods to meta-analyse GWAS summary statistics ranging from P value meta-analyses to Bayesian approaches (Evangelou and Ioannidis 2013). METAL was selected due to its ubiquity, ease of use and reliability.

A metanalysis of cystic kidney disease was performed using GWAS summary statistics from my analysis (10,377,276 markers), a combined UK/Japanese Biobank (UK/JBB) analysis of 220 phenotypes including polycystic kidney disease (19,093,042 markers) (Sakaue et al. 2021) and the Finnish Biobank (version 8) analysis of cystic kidney disease (19,441,692 markers) (Kurki et al. 2022). This was performed using METAL (Willer, Li, and Abecasis 2010). The summary statistics from the UK/JBB and Finngen were lifted over from build 37 to 38 using the UCSC liftover tool (Hinrichs et al. 2006). Between the three data sets 8,217,458 markers were shared with matching alleles. Meta-analysis was performed weighting the effect size estimates using the inverse of the standard errors. Variants showing heterogeneity of effect between the two datasets ($P < 1 \times 10^{-5}$) and those in which the minimum/maximum allele frequencies differed by >0.05 were excluded leaving 6,641,352 variants across 2923 cases and 900,824 controls. The genomic inflation factor (lambda), calculated based on the 50th percentile, was 1.01 indicating no significant population stratification.

3.4.3.6 Time to event analysis

There is large variability of the phenotype in CyKD, particularly ADPKD where this has been extensively studied (Harris and Rossetti 2010). Genetic differences within families with ADPKD have been estimated to account for 18%-59% of the phenotypic variance before ESRF, 45 to50% for time to ESRF (Paterson et al. 2005; Fain et al. 2005). Attempts to find potential modifier genes have been disappointing, all being candidate gene led and suffering from small sample sizes and a number of methodological issues (Baboolal et al. 1997; Pereira et al. 2006; A. Persu et al. 2002).

Whilst seqGWAS has opened an avenue to look for common variants that may modify the course of disease, the case-control label does not leverage any further information about the cohort that may add power to detect associations. Survival models, particularly the Cox proportional hazard model has been used in other fields of biomedical research to analyse time to event (TTE) outcomes (Cox 1972). Previous work has shown that using hazard models increases the power to detect SNV associated with age-of-onset TTE phenotypes compared to logistic regression models traditionally used in GWAS, although at a much greater computational cost (Staley et al. 2017). In the era of large genetic biobanks married to granular longitudinal phenotypic data there has been renewed interest in using genome wide TTE studies to identify genetic variants associated with disease. This represents a particularly appealing approach for CyKD given the variable presentations both between and within families.

Large scale genomic studies require controlling for population structure and relatedness, discussed in the methods section. Similar to GLMM models that include mixed effects to account for the above, frailty models, which are mixed effect survival models have been suggested (Hougaard 1995). The “frailties” are unobserved random effects.

To date developments in utilising frailty models with large scale biobank data have centred around optimising the model to account for complicated dependency structures (Therneau, Grambsch, and Pankratz 2003) but these do not scale to GWAS at biobank level. The Genetic Analysis of Time to Event (GATE) analysis has been developed to overcome these issues which accounts for population stratification, relatedness, type I

error in the face of heavy censoring (as often seen in rare disease phenotypes) and is scalable to biobank scale data (Dey et al. 2022). It effectively converts multivariate Gaussian frailty models into modified Poisson generalised linear mixed models similar to SAIGE (GATE is devised by the same group as SAIGE).

The 100KGP project participants consented to give access to their Hospital Episode Statistics (HES) which is a database containing details of all admission, emergency attendances and outpatient appointments at NHS hospitals in England. The database was searched for codes that would highlight whether a patient had reached end stage renal failure (ESRF). The age of ESRF was used as the end point in the TTE analysis and those who were yet to reach ESRF were censored. The same genomic and phenotype data as per the single variant seqGWAS was used to conduct the TTE GWAS.

3.4.4 Results

All GWAS summary statistics are available in the supplementary information.

3.4.4.1 seqGWAS and meta-analysis

A seqGWAS of 1209 CyKD cases compared with 26096 controls using 10377275 markers with a MAF>1% revealed only a single variant reaching genome wide statistical significance on chromosome 8, chr8:92259567:A:C ($P=1.38x10^{-08}$, OR 0.72) (Figure 3-19). There was no evidence of genomic inflation ($\lambda=0.99$, Figure 3-20).

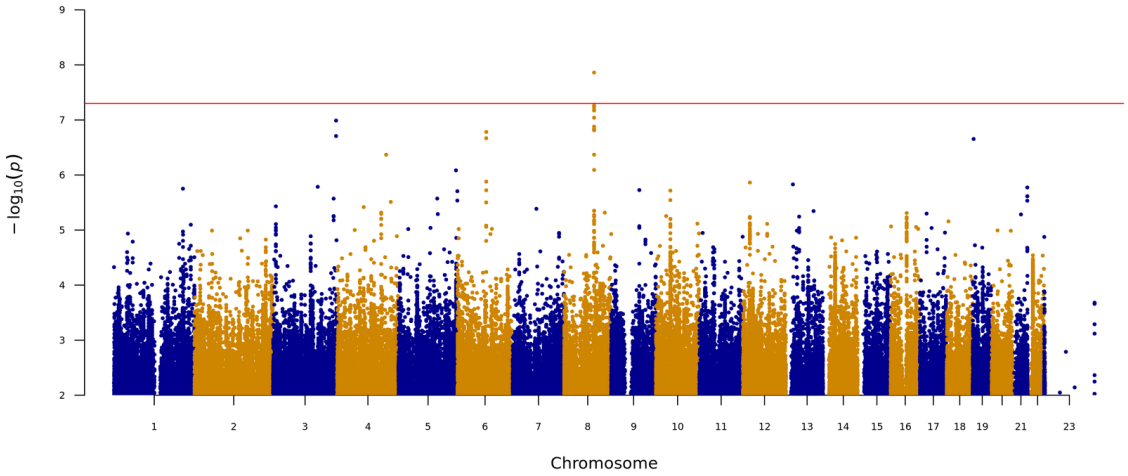

Figure 3-19 Variant Manhattan plot all CyKD GWAS

A sequencing based GWAS was carried out in 1209 unrelated CyKD cases and 26,096 controls for 10377275 variants with MAF ≤ 0.1%. Chromosomal position (GRCh38) is denoted along the x axis and strength of association using a –log10(P) scale on the y axis. Each dot represents a variant. The red line indicates the conventional threshold for genome-wide significance (P < 5x10-8).

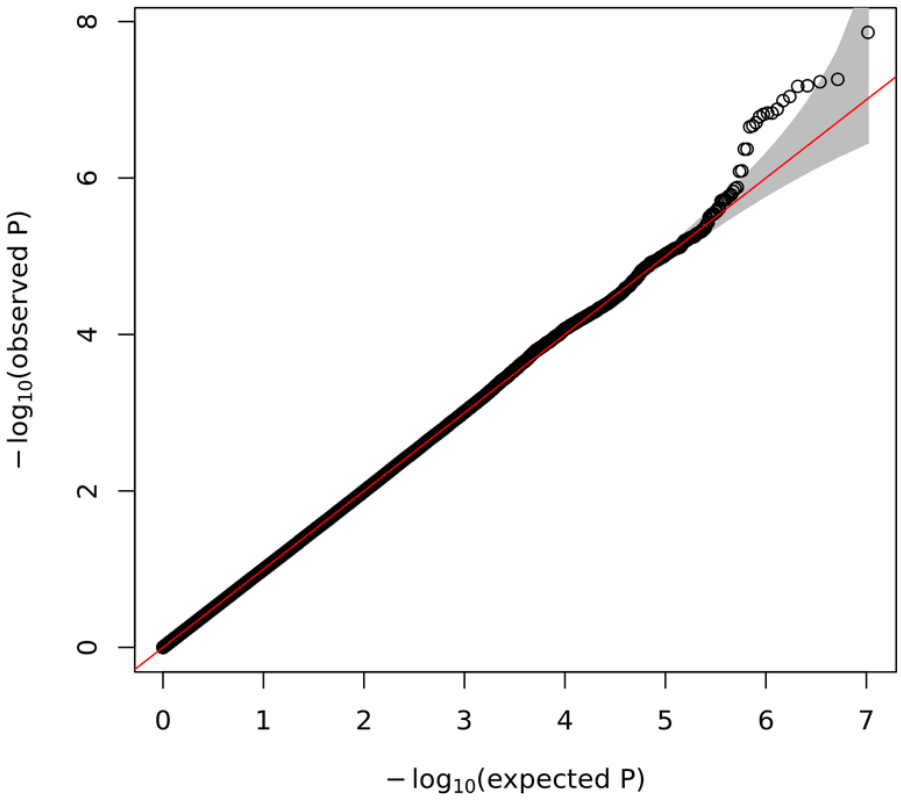

Figure 3-20 Q-Q plot for CyKD mixed-ancestry GWAS

Q-Q plot displaying the observed versus the expected –log10(P) for each variant tested. The grey shaded area represents the 95% confidence interval of the null distribution. The λ=0.99

Fine mapping of this locus and conditional analysis confirmed it to be the lead variant. This variant sits downstream from *RUNX1T1*, a transcriptional corepressor, and upstream from a long noncoding RNA (LOC102724710) (Figure 3-21). However, this variant sits in within its own topologically associating domain (TAD) separate to both gene, with no overlapping features of note that would at hint at the mechanism by which it would confer protection against CyKD (Figure 3-22).

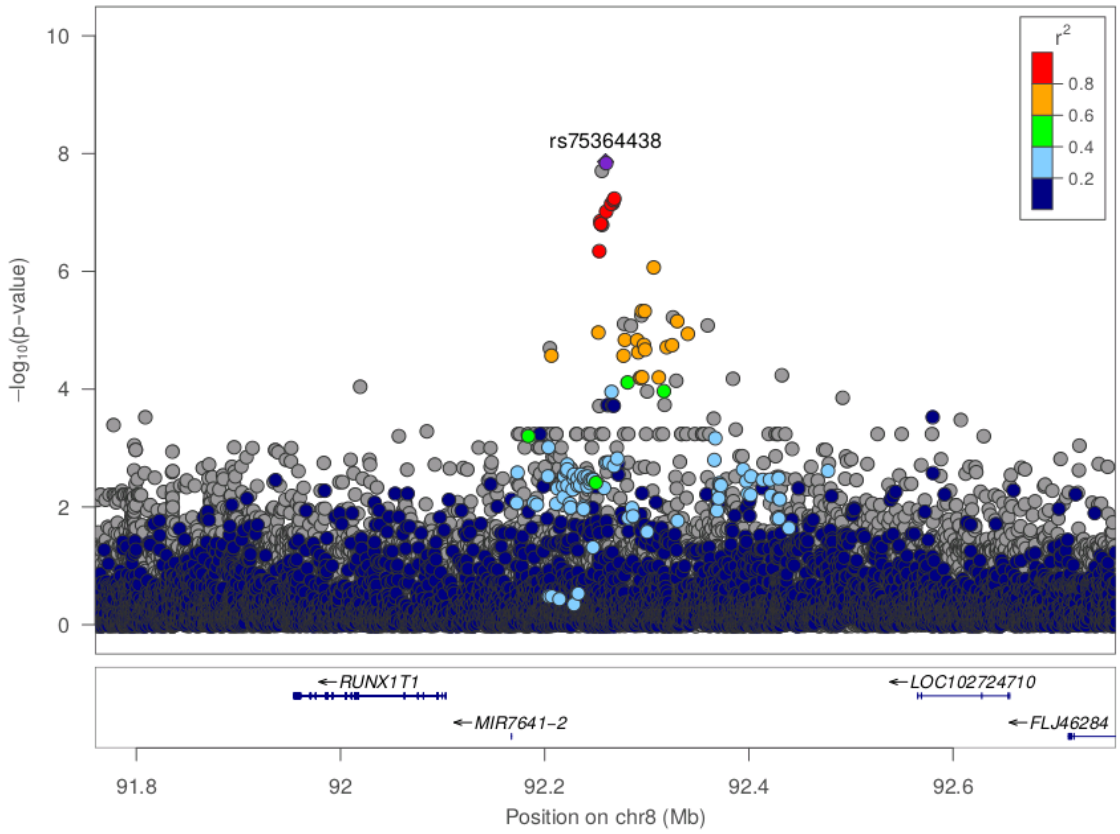

Figure 3-21 Regional association plot for lead SNV from CyKD GWAS

Each variant is represented by a dot. The lead variant is labelled (rs75364438). The remaining variants are colored in relation to the strength of their linkage with the lead SNP. Gene names in the region are listed against their chromosomal position (GRCh38).

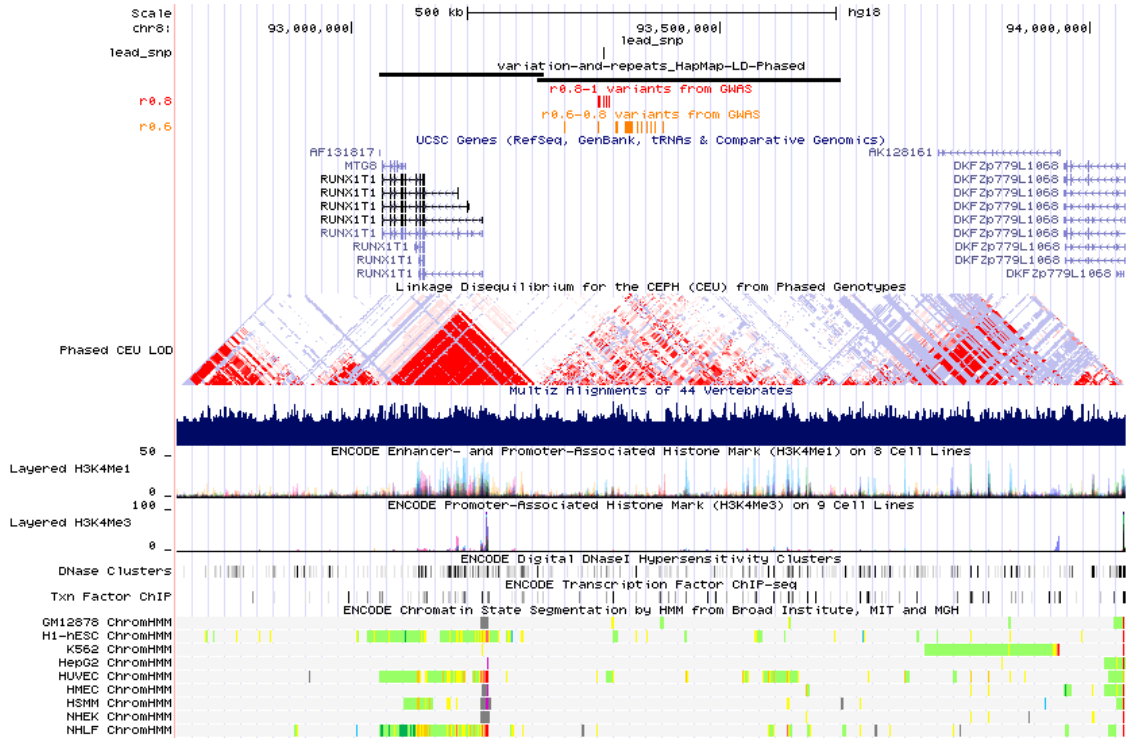

Figure 3-22 Functional annotation for lead SNV from CyKD GWAS

A plot taken from the UCSC genome browser showing the lead variant (black bar at top) and top variants in linkage disequilibrium with the lead variant (red r0.8 and r0.6 variants). Below are the nearest genes, the topologically associating domains and the enhancer/promoted assorted histone marks. The lead variant was within its own linkage domain (the long black bars below the lead_snp mark) with no overlying annotations of note that would hint at further mechanistic function.

Given the lack of further functional information, to confirm/refute this finding, I meta-analysed this dataset with those from the UK, Japanese and Finnish biobanks. In the Finnish cohort there was evidence of association in several loci, most notably a stop gain in *PKHD1* but the chr8:92259567 signal was not replicated and overall (Figure 3-23 and 3-24), the combined analysis of 2923 cases and 900824 controls across 6641351 markers did not reveal any genome wide significant markers (Figure 3-25).

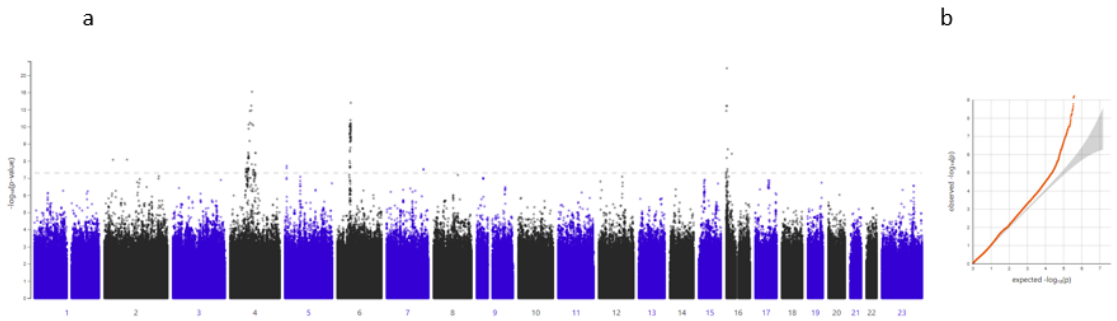

Figure 3-23 Manhattan plot of Finngen CyKD GWAS

3a GWAS Manhattan from Finngen plot of 780 cystic kidney disease cases against 375708 controls. 3b – Quantile-Quantile plot of the association test in 3a. The genomic inflation is 1.04.

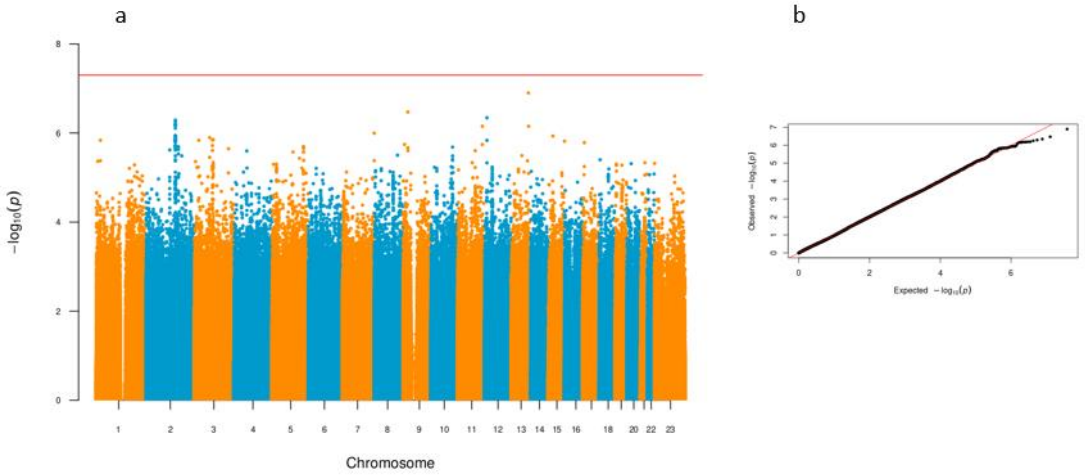

Figure 3-24 Manhattan plot of UKBB/JBB CyKD GWAS

4a GWAS Manhattan from UKBB/JBB analysis - plot of 932 cystic kidney disease cases against 534581 controls. 4b – Quantile-Quantile plot of the association test in 4a. The genomic inflation is 1.02.

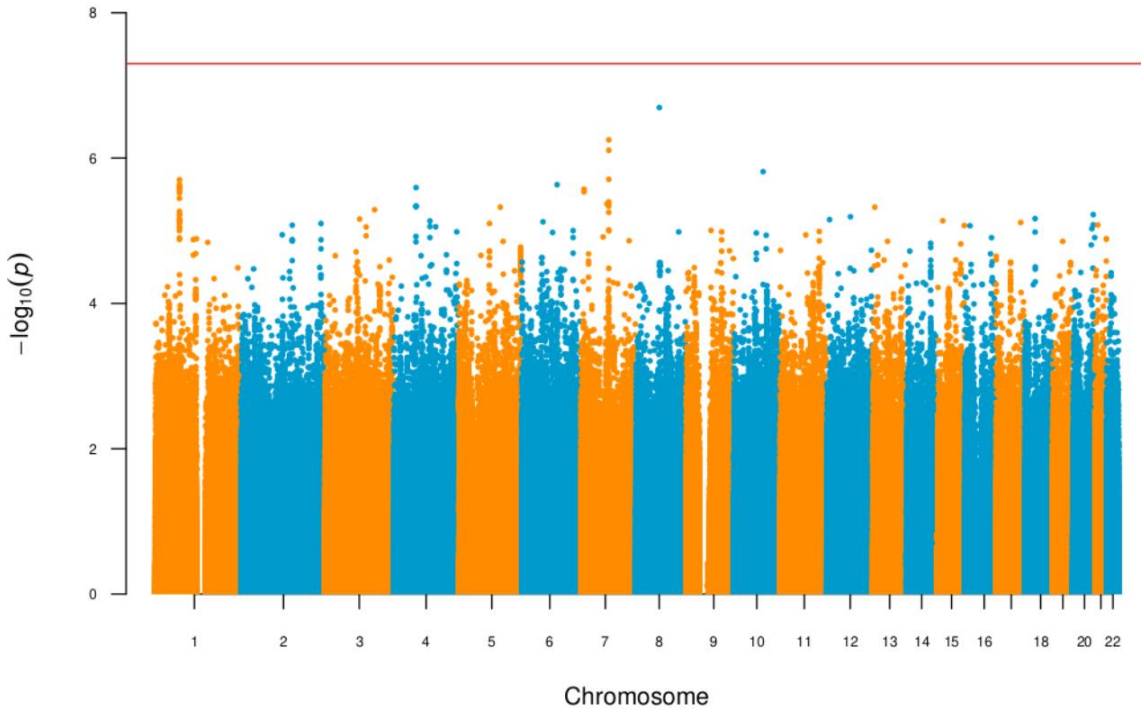

Figure 3-25 Manhattan plot of CyKD meta-analysis GWAS

A meta-analysis of my GWAS study and those from the UK, Japanese and Finnish biobanks was carried out across 2923 cases and 900824 controls across 6641351 markers. Chromosomal position (GRCh38) is denoted along the x axis and strength of association using a –log10(P) scale on the y axis. Each dot represents a variant. The dotted line indicates the conventional threshold for genome-wide significance ($P < 5x10-8$). $\lambda = 1.01$

Subgroup analysis by primary variant type did not reveal any genome wide significant loci (Figure 3-26, QQ plots Figure 3-27).

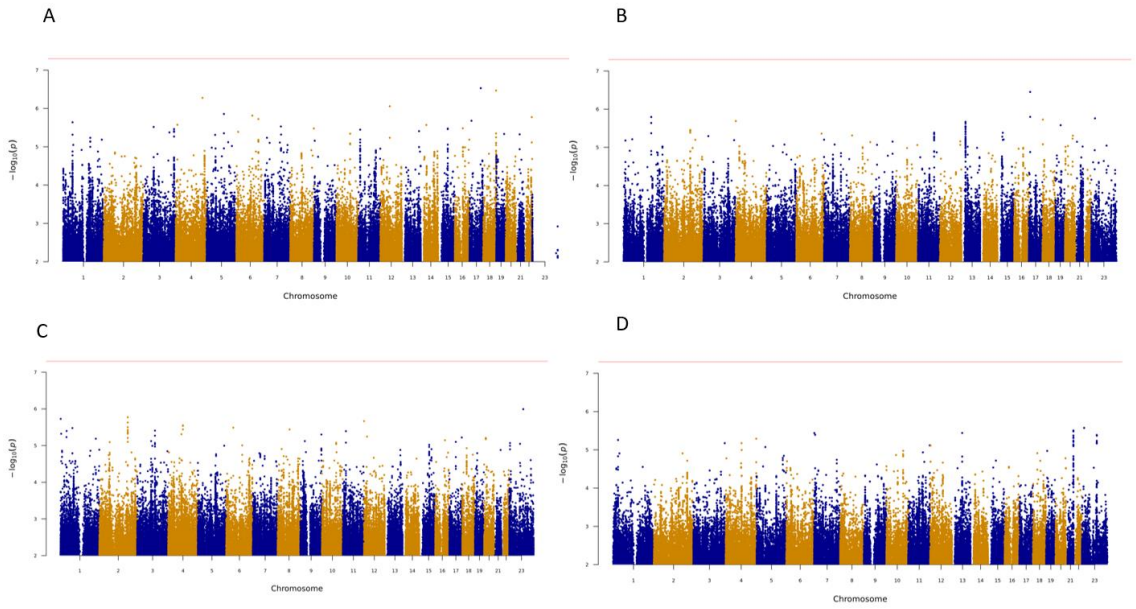

Figure 3-26 Manhattans of GWAS by primary variant type

Sequencing based GWAS was performed in case cohorts stratified by primary driving variant. No significant association were detected in each cohort with no evidence of genomic inflation. A – *PKD1*-truncating (10370320 markers), B- *PKD1* non-truncating (10409799 markers), C- *PKD2*-truncating (10408817 markers), D – *PKD2* non-truncating (10407729 markers). The red line represents the genome wide significance line.

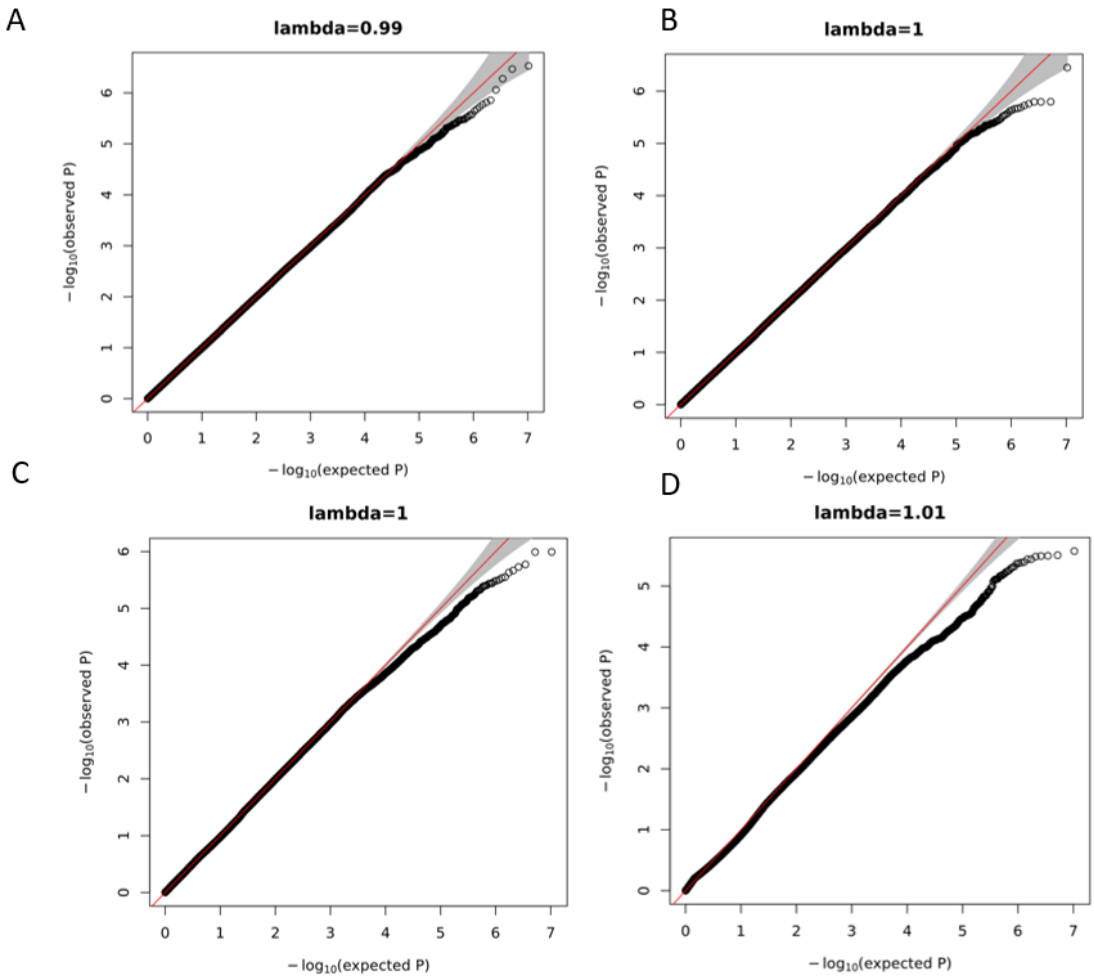

Figure 3-27 QQ-plots of the per driving variant GWAS

A – *PKD1*-truncating, B- *PKD1* non-truncating, C- *PKD2*-truncating, D – *PKD2* non-truncating. There was no evidence of genomic inflation. The grey area represents the 95% confidence interval.

3.4.4.2 Time To Event (TTE) analysis

Within the cohort 398 of the 1209 probands had reached ESRF with a median age of 52 years (IQR 44-60). Time to event analysis using GATE did not reveal any genome wide significant associations – either in the total cohort or stratified by primary gene or variant type. There was no evidence of genomic inflation. Of note analysis of *PKD2* was not possible due to the low number of events in the group causing a failure of model fitting (Figure 3-28, QQ plots Figure 3-29).

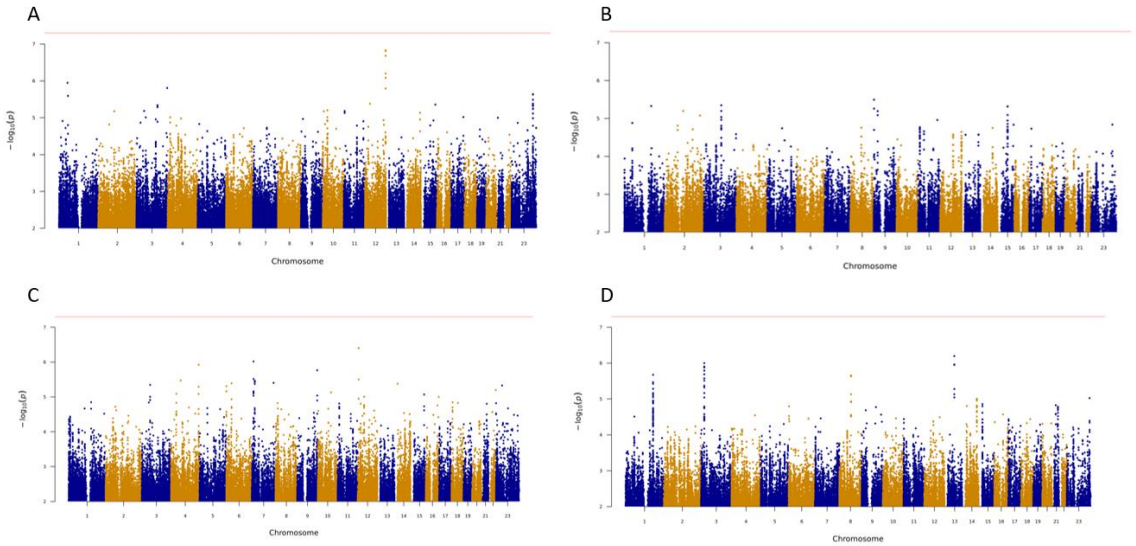

Figure 3-28 Manhattan of TTE GWAS

Manhattan plot of the time-to-event GWAS across the total CyKD cohort and divided by primary variant and no variant detected cohorts. A – Total CyKD cohort (11485299 markers), B – No variant detected cohort (7718522 markers), C – *PKD1*-truncating cohort (8543817 markers), D – *PKD1*-nontruncating cohort (7465234 markers). The red line represent genome wide significance.

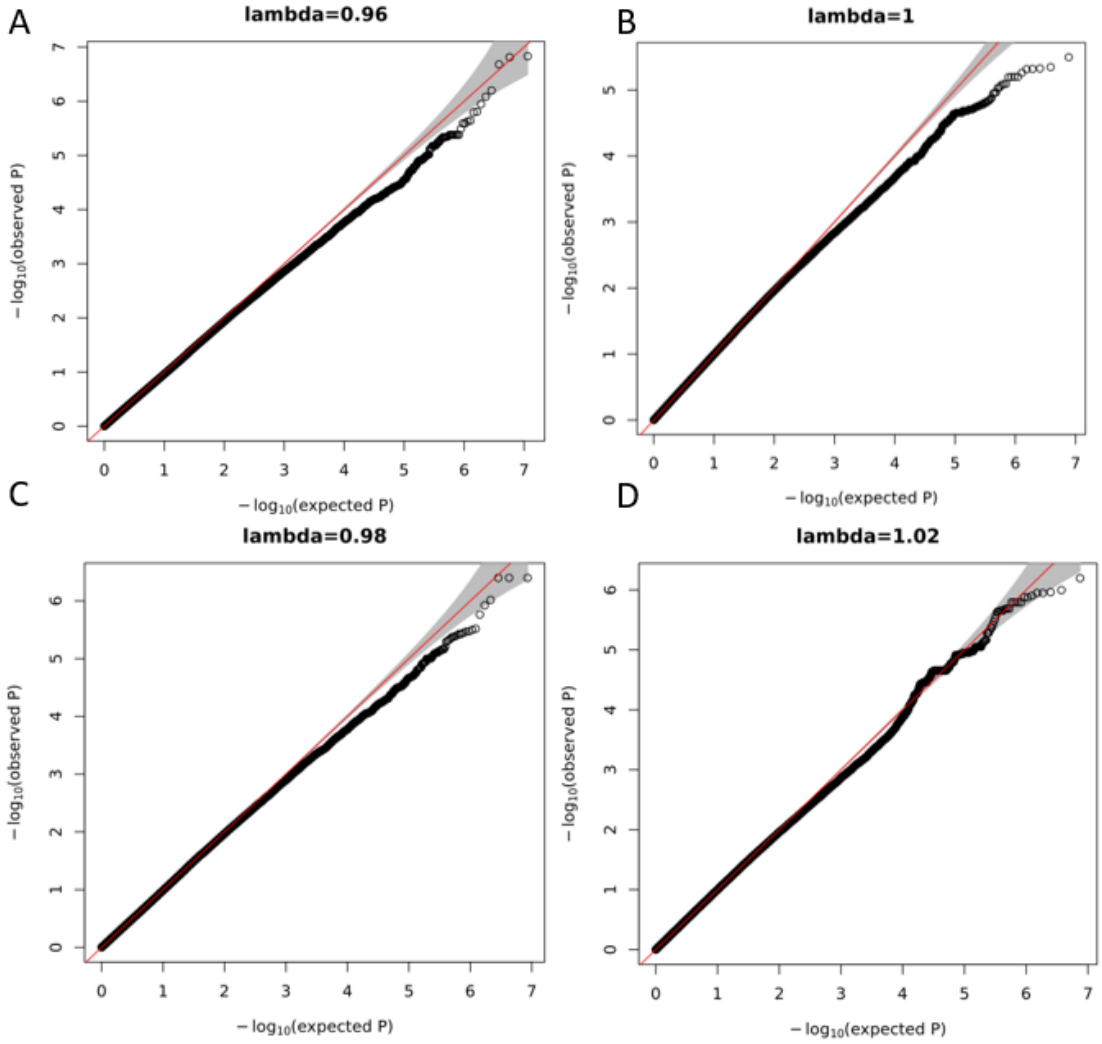

Figure 3-29 QQ-plots of the TTE GWAS results

A – Total CyKD cohort B – No variant detected cohort, C – *PKD1*-truncating cohort, D – *PKD1*-nontruncating cohort. There was no evidence of genomic inflation. The grey area represents the 95% confidence interval.

3.4.4.3 Heritability

Using the tools described to approximate narrow-sense heritability (h^2) I estimated that the proportion of phenotypic variance explained by additive common and low-frequency variation in a European CyKD cohort was 9.0% (SE 7.6%). Low frequency variants with MAF between 1% and 5% accounted for nearly all of the estimated heritability (Figure 3-30). This suggests that there are likely to be a significant number of contributory low-frequency variants with effect sizes too small to be detected in this cohort.

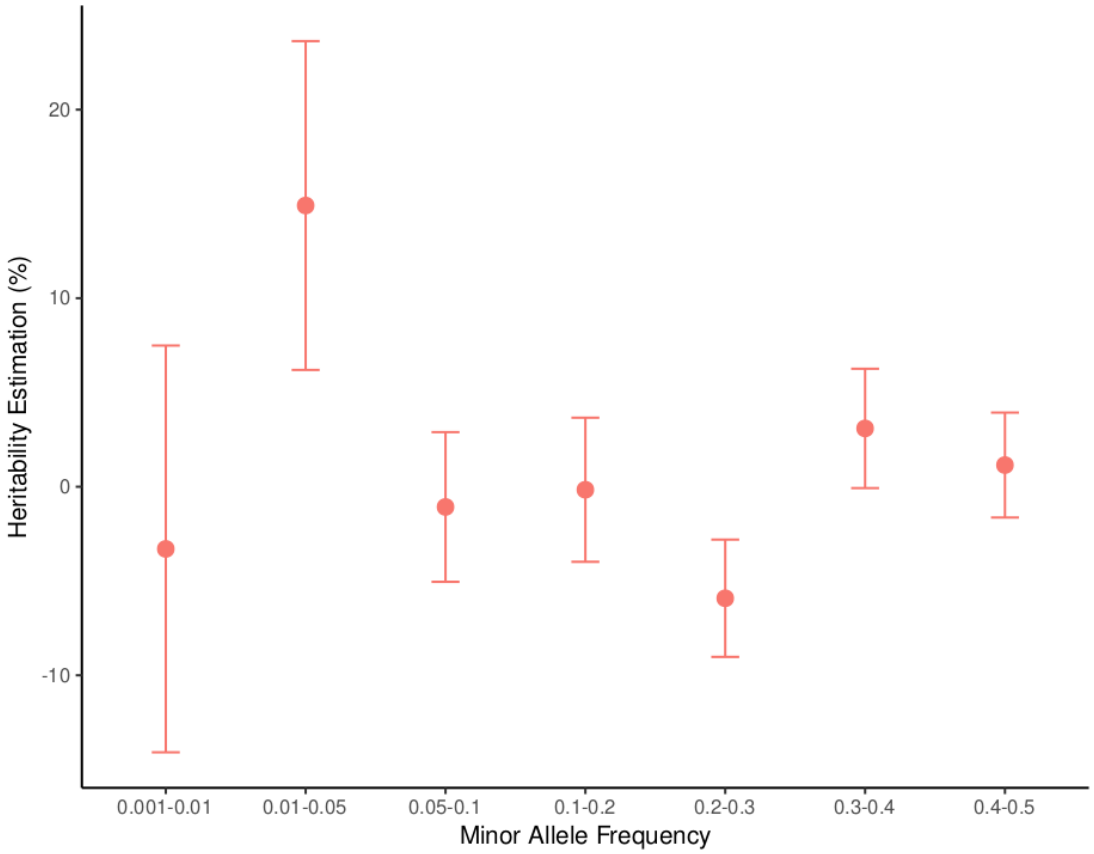

Figure 3-30 Partitioning of heritability by MAF in a European cohort of 903 CyKD cases and 20255 controls

Narrow-sense heritability (h^2) is represented using the liability threshold model based on a population disease prevalence of 1 in 1000. Bins indicate MAF of variants tested. Error bars indicate standard error (SE). h^2 follows a normal distribution and therefore unbiased estimates may be negative, as seen here, particularly if the sample size is small (and the variance large).

3.4.4.4 Polygenic association

There was a significant difference between a cohort of European monogenic CyKD cases and cancer controls when scored with a PRS generated from the summary statistics from the unsolved CyKD GWAS ($P=2.8x10^{-05}$) (Figure 3-31). Fitting a logistic model in the monogenic/cancer cohort of phenotype against the PRS, sex and 4 PCs revealed significant protective association between phenotype and PRS with an adjusted increase in one standard deviation of the PRS leading to the odds ratio of developing CyKD decreasing by 0.70 (95%CI 0.60-0.80, $P=1.14x10^{-06}$). Sex and PC1-PC6 were also significantly associated with the phenotype ($P<0.001$). The model AUC was 0.81 (95%CI 0.79-0.83).

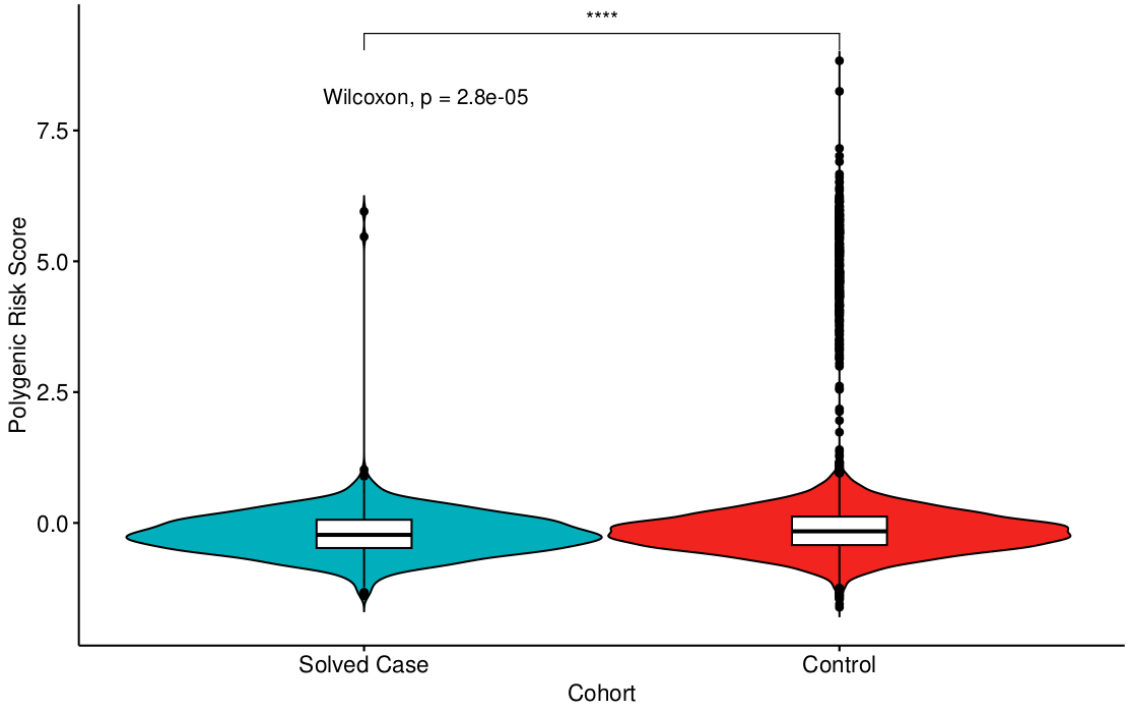

Figure 3-31 Violin and boxplot comparing polygenic risk score distribution in monogenic cases and cancer controls

Violin and boxplot showing the polygenic risk score (PRS) distributions between controls (ancestry matched cancer patients without renal disease or urogenital malignancy) and cases with monogenic CyKD. The means of the two PRS were significantly different *** = *statistical significance*

The liability adjusted pseudo-R^2, used as a surrogate for the heritability of the PRS, accounted for 8.68% (SE 7.63%), permutated P-value for model fit = $1.14x10^{-06}$). This represents the heritability of monogenic cystic kidney disease that is accounted for by common polygenic variants behind unsolved CyKD and in this instance represents a protective effect.

There was no statistical difference between each molecular subgroup of CyKD PRS, nor within each subgroup when stratified by end-stage renal failure (Figure 3-32).

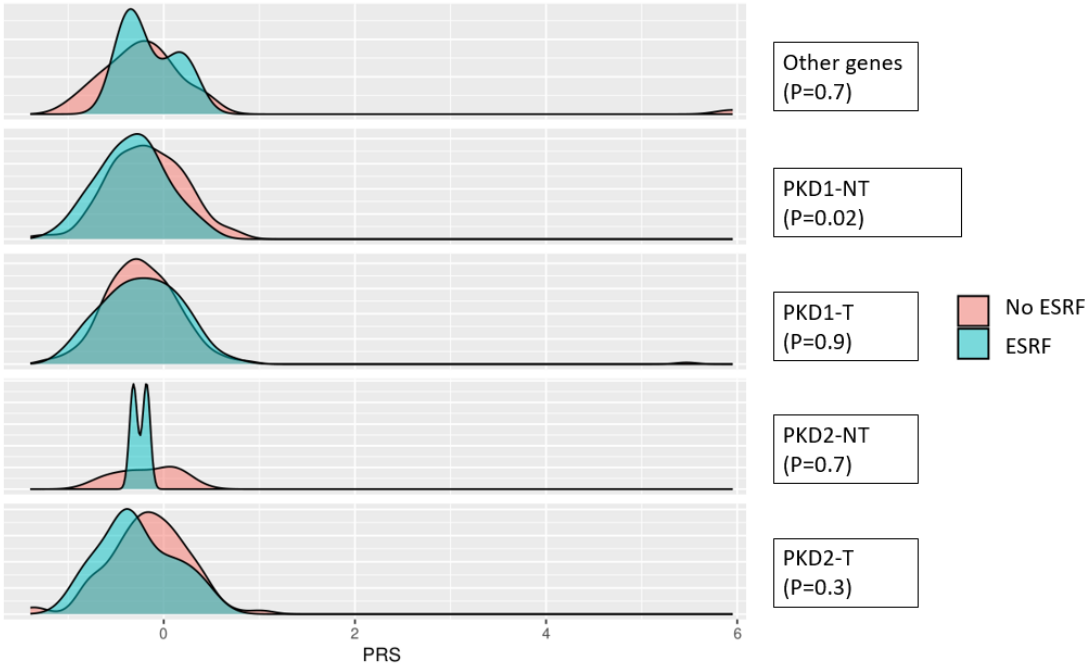

Figure 3-32 PRS distribution in CyKD cases

PRS distribution in European monogenic CyKD cases divided by primary driving variant and the presence of absence of ESRF. There was no significant association between the PRS scores in those with or without ESRF (bar *PKD2*-NT) nor between each molecular cohort.

These results suggest a protective effect of common variants in those with unsolved CyKD in the monogenic CyKD cohort.

3.4.5 Summary

- GWAS of the cystic disease cohort revealed a single SNV reaching genome wide significance in the 100KGP cohort.
- This was lost on metanalysis with multiple biobanks and no further signals were found.
- GWAS by primary driving variant and with time to event analysis did not reveal any genome wide significant associations.
- Heritability assessment estimates a 9% contribution to heritability from common variants, almost entirely coming from variants between a MAF of 1-5%.
- Common variants associated with non-monogenic CyKD make a ~9% contribution to the heritability of monogenic CyKD in a protective direction.

3.4.6 Discussion

3.4.6.1 CyKD GWAS and subgroup analysis

In this chapter I have used mixed-ancestry sequencing-based GWAS to examine the common variant contribution to CyKD and time to ESRF in CyKD in the total cohort and stratified by primary driving variant. This is the largest GWAS conducted in CyKD to date. The lack of genomic inflation highlights how well controlled the mixed ancestry WGS study is and allows for an increase in power by including individuals from more ancestries. Further to this, I then conduct a metanalysis of my results with other large biobanks to conduct by far the largest GWAS of CyKD to date.

This represents the largest systematic analysis of whether oligogenic or polygenic mechanisms are important in the aetiology of CyKD, highlighting how strong the monogenic drive for CyKD is. Despite a large number of cases and controls, with enough power to detect signals, we do not find any common variant signals. In fact, in the Finnish population that has undergone significant genetic bottlenecks causing increased frequency of certain recessive variants there is an enrichment in a known pathogenic *PKHD1* variant (rs137852949, OR=4.69) at an allelic frequency that borders the "rare" variant tag (ClinVar ID: 4114, MAF in Finnish population = $7.48x10^{-03,}$ MAF in non-Finnish European population =$3.24x10^{-04}$) but meets inclusion for seqGWAS in Finngen. This variant has been implicated as a heterozygous cause of polycystic liver disease (Besse et al. 2017) and the enrichment of rare monoallelic *PKHD1* variants in our cohort give further strong evidence for its role as a monoallelic cause of cystic kidney disease, potentially in a similar model to *IFT140*.

There are no other published GWAS studies of CyKD to compare to bar the summary statistics from large scale biobank projects where multiple phenotypes are analysed and results are made publicly available. The UK Biobank and Japanese Biobank CyKD GWAS summary statistics were taken from a joint Japanese/Biobank analysis of imputed data which did not reveal any genome wide significant associations (Sakaue et al. 2021), whilst the Finnish analysis did contain a number of significant associations.

Looking at the credible set of variants in the Finnish analysis in detail (https://r9.finngen.fi/pheno/Q17_CYSTIC_KIDNEY_DISEA there is the common (in the Finnish population) variant at *PKHD1*, a series of variants mapped to *NKX6-1* (rs937539605, OR=119.88) and *TIGD7* (rs1447267563, OR=7.02). The *NKX6-1* variant is found in a regulatory region of the gene and is only seen in the Finnish population, at an allele frequency of $1.30x10^{-04}$. *NKX6-1* is a transcription factor with a crucial role in the regulation of pancreatic β cell development (Aigha and Abdelalim 2020) and has been investigated as potential novel biomarker in chromophobe renal cell carcinoma (Xie et al. 2022). There are no known links to CyKD in the literature. The *TIGD7* variant falls within the 5 prime UTR region of the gene and has an allele frequency of $1.22x10^{-03}$ in the Finnish population compared to an AF of $2.94x10^{-05}$. It is a DNA transposon with an unknown function and high expression in most tissues as per the human protein atlas. Given the lack of annotation for the non-*PKHD1* variants it is difficult to speculate further and given the lack of replication in other cohorts, these represent either Finnish specific variants, or noise. However, of note, in the "traditional" analysis looking at top SNVs in the association analysis based off the lowest P-value, the most significantly associated SNV in the GWAS map to *ZG16B* and is protective (OR 0.03) in this population. The gene itself is predominately expressed in salivary gland tissue with little clear relationship with CyKD, however, the enrichment of a protective allele mirrors that found in my 100KGP analysis.

In terms of subgroup analysis by driving variant type I did not detect any associations and this was almost certainly limited by sample size and consequent lack of power. It is likely that modifier variants will have small effect sizes and thus the power to detect them will be limited. Power calculations show that we would need roughly 1000 cases per cohort to have 80% statistical power and none of our subgroups meet that criterion.

3.4.6.2 Time to event analysis

The identification of specific factors that modify CyKD severity led me to conduct a time to ESRF analysis within CyKD both as a whole cohort and within molecular subgroups. Neither produced a result but again this was likely limited by low numbers

to detect modest effect sizes, for example a genome-wide association meta-analysis identified five modifier loci of lung disease severity in cystic fibrosis but required over 6000 cases with detailed phenotypic and functional information (Corvol et al. 2015).

One alternative method of analysis that has been used in CyKD recently is cryptic phenotype analysis which uses the principle that some Mendelian disorders represent the extreme end of a spectrum of pathologic variation such as is found in familial hypercholesterolaemia (Sturm et al. 2018) or long QT syndrome (Ingles and Semsarian 2020). These examples rely on the conditional mapping to a quantitative phenotype however, cryptic phenotype analysis replies on mapping Mendelian disorders to a map of high-dimensional arrays of disparate symptoms with the assumption that Mendelian disease patients represent the extreme end of the phenotypic spectrum (Blair, Hoffmann, and Shieh 2022). In this analysis performed in the UK Biobank and the University of California biobank the team discovered 30 associations between ADPKD and disease severity however, the have been no replication studies or further functional work. Of note, most of these loci have already been associated with kidney disease and blood pressure regulation. I am currently exploring applying this methodology to the 100KGP with the authors of the paper.

3.4.6.3 Heritability

A sequencing based approach to heritability estimation has never been performed in CyKD prior to this study and represents the most thorough analysis of the contribution of low frequency and common variants in this disease. I demonstrate that these variants, particularly those with a MAF between 1-5%, make up 8% contribution to phenotypic variance.

This seems like a surprisingly high number given the strong monogenic architecture of CyKD. Twin studies are seen as the gold standard of heritability estimation but very few studies have been performed in CyKD, and then only in ADPKD (Alexandre Persu et al. 2004) with a focus on influencers of severity. Other studies using large ADPKD populations looking at the intrafamilial variability in disease severity using variance componence analysis estimate that difference in genetic background account for 18-

59% of the phenotypic variance prior to ESRF (Paterson et al. 2005; Fain et al. 2005), a more recent example of such an analysis set a minimum variance of 11% (Lanktree et al. 2019), of note these would be a broad sense heritability analysis (H^2) as opposed to narrow sense that I have looked at using GCTA. Work by Blair et al. using a cryptic phenotype approach to ADPKD in the UKBB suggested a common variant (MAF>0.1) contribution to the heritability of disease severity as 9.7% (0.19% SD) using similar methodology (Blair, Hoffmann, and Shieh 2022).

In the paper mentioned above, Blair et al. used a cryptic phenotype approach meaning they included may more "cases," as they included a case to one with a set of signs and symptoms and would likely represent a diagnosis of ADPKD, many of whom did not have molecularly confirmed disease (n=308,095) allowing for a more confident assessment of heritability with a much smaller standard deviation. The two populations may differ, with the 100KGP enriched for clearly monogenic and atypical causes of CyKD which may have a smaller common variant contribution. Ideally, I would repeat this exact methodology in the UK Biobank, or with the >1000 new cases that have recently been sequenced via the NHS Genomics England platform who most likely have more typical disease. Nonetheless, my assessment of narrow sense heritability in CyKD is similar to that found in the UKBB.

In terms of identifying variants contributing to this signal, the absence of significant association in the metanalysis GWAS means that low-frequency variants are likely to have effect sizes below the threshold that be detected in a cohort of this size (OR <3). Larger, better powered studies are required to find these missing variants.

3.4.6.4 Polygenic association between monogenic and unsolved CyKD

By generating a PRS from the unsolved CyKD GWAS and applying it a European subset of the solved monogenic CyKD cohort I attempted to understand the role common variants play in the heritability of the monogenic CyKD phenotype. In both the 100KGP total CyKD cohort and Finngen CyKD GWAS the lead SNVs was protective, and it could be that common variants in fact have a protective effect on the cystic phenotype. It is known that up to 18% of patients with truncating *PKD1* variants display

mild disease, despite having the "worst" genotype (Lanktree, Guiard, et al. 2021). For such a relatively high rate of genotype-phenotype variability to occur the protective modifiers must be relatively common, be they genetic or environmental. It is feasible that the common variants that contribute to non-monogenic CyKD confer a protective modifier to monogenic CyKD, although unpicking the biology from this is difficult.

Whilst there are many examples of polygenic risk scores interacting additively with known disease risk (Fahed et al. 2020; Visscher et al. 2021), there are very few examples of common variants acting in a protective manner against monogenic disease. In Huntington's disease, GWAS of age of onset has revealed at least 21 common variant loci that confer protection against the onset of the disease by potentially modifying the genes involved in DNA maintenance (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019). Whilst the GWAS studies I performed did not reveal any significant associations at an individual SNV level bar one, the possibility that common variants play a protective role in CyKD remains. Irrespective it hints at the shared genetic architecture in cyst formation.

Given the recruitment criteria for CyKD in the 100KGP, some of the unsolved cohort will have monogenic disease that has either gone undetected in a known gene or is a novel monogenic disorder yet to be discovered, however, it is likely that there is shared biology between "simple" cyst formation and those associated with multiple cysts.

There is a paucity of common variant genetic association analysis in patients without clear monogenic causes of CyKD making it difficult to assess my findings in a broader context. Finding modifiers of CyKD has been challenged by the need for large cohorts to find loci of low effect size. By combining the total common variant contribution into a PRS I have performed a similar function to a collapsing rare variant association test and improved power at the cost of loss of granularity. These findings show common variant contribution but do not give us any further insights into molecular causation, a general issue with PRS. However, Peters et al in their review of the biological context affecting cyst development in ADPKD (Leonhard, Happe, and Peters 2016) provide a helpful diagram by which to contextualise these findings (Figure 3-33).

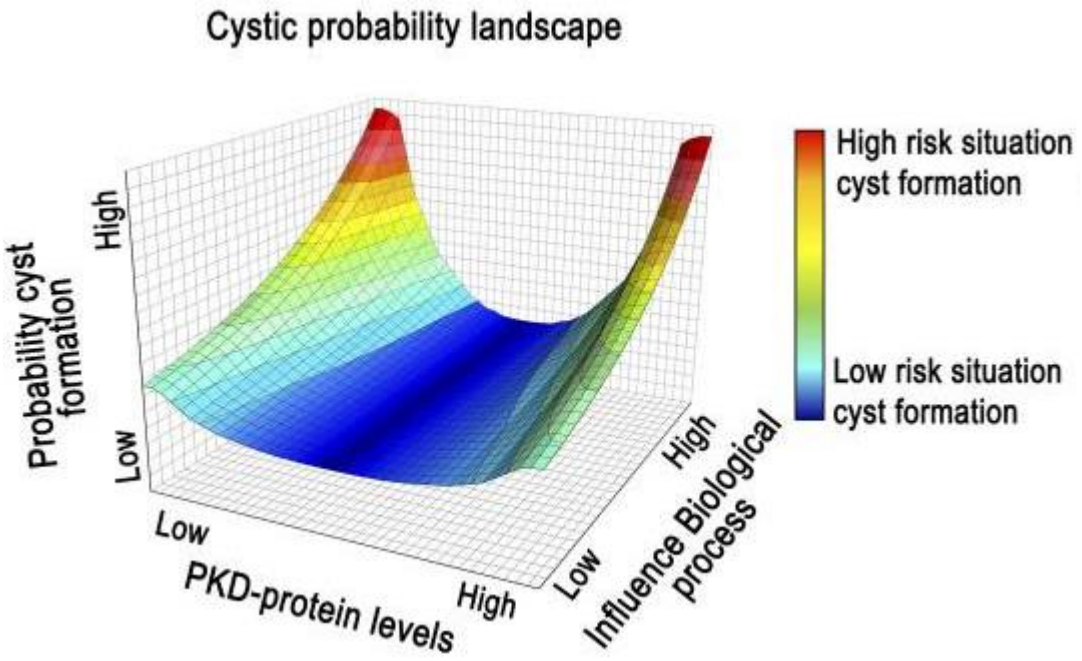

Figure 3-33 The cystic probability landscape

Taken from Leonhard et al 2016, this graph represents the cystic probability landscape of renal epithelial cells with the probability of cyst formation on the y-axis, the PKD-protein levels on the x-axis and the influence of biological processes on the z-axis. It could be that having a high polygenic risk score for CyKD in the no variant detected cohort could decrease the probability of cyst formation and either through a polycystin or non-polycystin mechanism by increasing the "Influence Biological process" level.

Whilst I did not find a difference between the PRS between each molecular subtype or stratified by the presence of ESRF within each subtype, I was limited by low numbers. It is conceivable that common variants play a role in the "biological context" of whether polycystin 1 is expressed. As larger patient numbers are obtained unpicking this will become a feasible target, however, in the first instance these results will need to be replicated in another biobank.

3.4.7 Strengths and Limitations

The main strength of this analysis lies in the use of WGS data which enables ancestry-independent variant detection and the association testing of variants. Furthermore, the uses of rigorous statistical approaches ensure adequate control for population structure in this mixed-ancestry cohort enabling an inclusive and better powered analysis. The lack of genomic inflation indicate that the detected associations are robust.

The main limitations of this study are the small sample size on a per variant basis limiting power to detect variants with small effects. Meta-analysis was attempted to overcome this issue in the CyKD cohort but was not possible in the subdivided cohorts as this information was not available from summary statistics. Finally, whilst I have shown that multi-ancestry GWAS does not lead to genomic inflation and population stratification, I have resorted to European only cohorts for heritability and PRS modelling. Methods are now being developed to apply PRS scores to mixed ancestry groups (Ruan et al. 2022) and I hope to use such tools in the future to expand our understanding of the common variant contribution to CyKD in all ancestries.

3.4.8 Conclusion

In this chapter I described the first mixed ancestry sequencing based GWAS for CyKD as well as the first meta-analysis of CyKD. The lack of association underlines that argument that the primary model for CyKD is monogenic although my findings regarding heritability and shared polygenicity between unsolved and solved cases highlights the importance more common variants potentially have in CyKD, expanding the spectrum of allelic contribution to the disease.

Chapter 4. Urinary stone disease

In this chapter I detail the work on urinary stone disease (USD) that I published in Kidney International (Sadeghi-Alavijeh et al. 2023). This work highlights the important contribution of low effect rare variants to disease heritability, a discovery made possible by access to large scale biobanks, WGS data and the latest in statistical genomics.

4.1 Introduction to USD

Urinary stone disease (USD) is a significant clinical and societal health burden affecting roughly 10% of the population at some point in their life (Scales et al. 2012). The prevalence is increasing and there are now over 80,000 hospital episodes per year in the UK(Geraghty et al. 2020). Consequently, the health economic burden is substantial, estimated around £250,000,000 in England per year for the initial stone treatment alone (Geraghty et al. 2020). In the USA, the annual cost for USD in 2000 was calculated as almost $3 billion and estimated to reach $4 billion by 2030(Antonelli et al. 2014). Moreover, there is a strong association between kidney stones and the development of chronic kidney disease (CKD), further adding to the burden from USD (AD Rule 2011; Rule 2009).

The aetiology of USD is multifactorial, with genetic and environmental factors implicated. There is a strong association between affluence of a society and the prevalence of USD, likely reflecting Western lifestyle habits that include a high salt and animal protein intake (Edvardsson et al. 2013). Yet, there is also a strong genetic contribution: a family history is seen in up to 65% of patients with USD with the heritability of stone disease estimated to be as high as 45% (DS Goldfarb 2005; DS Goldfarb 2019; Castro 1993; Monga et al. 2006). Indeed, a strong family history of kidney stone disease can confer a >50 times increased risk in an individual (Resnick, Pridgen, and Goodman 1968; Hemminki et al. 2017). At a polygenic level, multiple genome wide association studies (GWAS) have been conducted in multi-ancestry populations with greater than fifteen independent loci reported, accounting for roughly 5% of heritability (Howles et al. 2019). Moreover, there is increasing realization that the burden of monogenic causes of USD is considerable: in two recent studies up to 20% of

subjects with USD were considered to have a monogenic cause for their disease, although the rates are highly variable depending on whether a paediatric or adult population is used for analysis, and these studies may be subject to recruitment bias (Halbritter et al. 2015; Daga et al. 2018; Gale et al. 2020).

Identification of underlying genetic factors is important, as it facilitates targeted treatment and specific prognostic and genetic counselling1 (Bockenhauer et al. 2012). The gap between the contribution of the known polygenic risk factors and the observed heritability suggests that important genetic contributors to USD remain to be identified.

The 100,000 Genome Project (100KGP) is a pilot project to assess the utility of whole genome sequencing (WGS) in rare disease diagnosis in routine healthcare (100000 Genomes Project Pilot Investigators et al. 2021). This project's research arm provides an opportunity to correlate genomic information from participants with their clinical phenotype. I therefore aimed to investigate the contribution to USD of rare genetic variants (which are not ascertained by previous GWAS) by performing whole genome gene-based rare variant studies in participants with HPO codes for nephrocalcinosis and/or USD to identify and quantify genetic contributors to the missing heritability of stone disease.

4.2 Aims

1. To determine the prevalence of known monogenic disease in a cohort of patients with USD.
2. To discover novel candidate genes using an unbiased exome-wide rare variant association testing approach.
3. To combine this information with an assessment of the common variant polygenic risk to understand the missing heritability of USD

4.3 Methods

In this section I will detail the creation of the USD cohort and the novel methods used. I will not cover the rare variant collapsing gene-based analysis which uses the same methodology as the CyKD chapter.

4.3.1 Cohort creation

I searched for participants recruited with a primary diagnosis of USD or whose clinical information included HPO and/or Hospital Episode Statistics (HES) codes related to USD (a full list of searched codes is provided in the supplementary table). I only took cases and controls if their genomes had been aligned to build 38 of the human genome.

All cases recruited for USD had been assessed in the clinical interpretation arm of the 100KGP. This involved ascertainment of variants in an expert curated panel of 29 USD genes with multi-disciplinary review and application of American College of Molecular Genetics (ACMG) criteria to determine pathogenicity (Richards et al. 2015). The control cohort consisted of 27,660 unaffected relatives of non-renal rare disease participants, excluding those with HPO terms and/or hospital episode statistics (HES) data consistent with USD or secondary causes of USD, kidney disease or kidney failure. By utilizing a case-control cohort sequenced on the same platform, I aimed to minimize confounding by technical artefacts.

Whole genome sequencing was performed by Genomics England, with variant calling annotation and variant-level quality control described in more detail the methods chapter.

Given the small number of recruited cases, I chose to jointly analyse individuals from diverse ancestral backgrounds, thereby preserving sample size and boosting power. Ancestry matching, relatedness estimation, principal component analysis and the use of SAIGE's GRM were as detailed in the methods chapter. Figure 4-1 details the cohort creation workflow and figure 4-2 the process of ancestry matching cases to control.

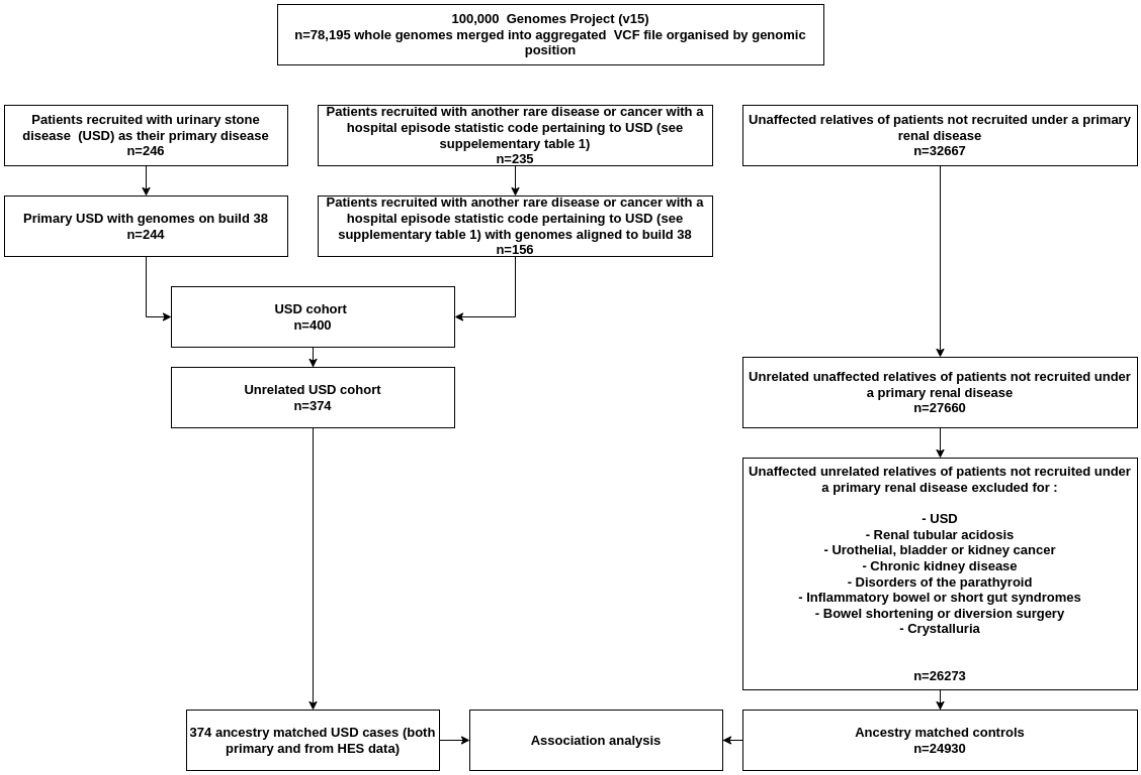

Figure 4-1 USD Study Workflow

The flowchart shows the number of samples included at each stage of filtering and the analytical strategies employed. USD – urinary stone disease

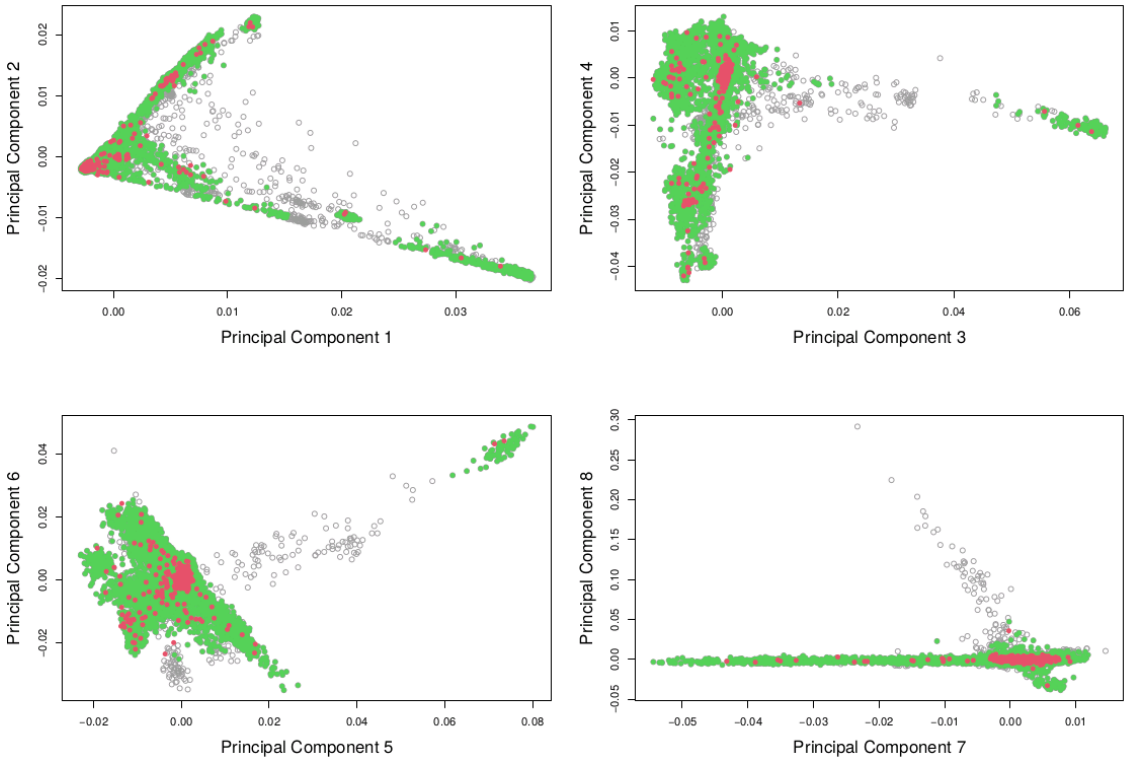

Figure 4-2 Ancestry Matching in USD

Principal component analysis showing the first eight principal components for matched cases (red) and controls (green) and unmatched controls (grey). This highlights that cases are taken from multiple different ancestries with the appropriate matched controls.

4.3.2 Validation of rare variant results in the UK Biobank

The AstraZeneca PheWAS portal (https://azphewas.com/) is a repository of gene-based phenotype associations derived from the UK Biobank, a prospective study over 500,000 individuals aged between 40-69 linking the health records with whole exome sequencing (Sudlow et al. 2015). Collapsing gene-based analyses similar to the SAIGE methodology was applied across thousands of phenotypes for each gene. Twelve different sets of qualifying variant filters (models: ten dominant models, one recessive model, and one synonymous "control" variant model) were applied to test the association between 18,762 genes and 18,780 phenotypes after extensive quality control filters (Q. Wang et al. 2021). I queried the PheWas portal for gene associations in USD across all models (tag: "Source of report of N20 (calculus of kidney and ureter)"). Results were given by Fisher's exact two-sided test P-values across each collapsed variant gene by model. The authors used a study-wide significance threshold of $P \leq 2x10^{-09}$.

4.3.3 Meta-analysis of rare variant collapsing tests

In order to boost power to detect signal I meta-analysed my collapsing rare variant study with the AstraZeneca PheWAS portal results. Ideally one would conduct a joint analysis with pooled individual level data, but this was not available to me as I did not have UK Biobank access nor did the AstraZeneca team respond to my requests for individual level data.

Meta-analysis requires only the sharing of summary statistics and has bene shown to be asymptotically equivalent to that of pooled analysis under reasonable conditions in common variant GWAS (Lin and Zeng 2010; D. J. Liu et al. 2014).

There are several methods to meta-analyse rare variant association studies (X. Li et al. 2023) however, they all require the storing of the covariances of individual variant test

statistics and/or the genetic relationship matrix, something I did not have access to from the UK Biobank summary statistics. It has been shown for fixed-effect models, that joint analyses of several variants in a gene or set can be performed with GWAS summary and a SNV-SNV correlation matrix obtained from a reference sample without the use of individual phenotype or genotype data (J. Yang et al. 2012). This principle has been implemented into the Fast Region-Based Association Tests on Summary Statistics (sumFREGAT) tool, an R package designed to take summary statistics and then use SNV-SNV correlation matrices calculated from the 1000 genomes project (Svishcheva et al. 2019).

Thus, I used the sumFREGAT package in R to combine summary statistics from the AstraZeneca/UK Biobank (UKBB) and 100KGP analyses on a per gene basis using the packages default settings.

4.3.4 Modelling the PRS and monogenic effect on heritability

The theory behind this approach has been discussed in section 3.4.3.4. In those cases, without a clear genetic diagnosis from the 100KGP clinical pipeline or a statistically significant gene association from the rare variant burden analysis, a polygenic risk score (PRS) was applied from a validated, multi-ancestry PRS of USD (Paranjpe et al. 2020) .

The PRS included 7,670,833 markers derived from the UK Biobank using LDpred (Privé, Arbel, and Vilhjálmsson 2021). This was lifted over using the UCSC Liftover tool (Hinrichs et al. 2006) and imported into the 100KGP where the scoring was performed using the “score” command in PLINK2 (C. C. Chang et al. 2015). PRS scores were standardized to controls using Z-score scaling.

To test the significance between the PRS of cases, controls, and those with *SLC34A3* qualifying variants I applied a Kruskal-Wallis test followed by a paired Wilcox test to differentiate the source of statistical significance using base R functions. All plotting was performed with ggplot2 in R (Wickham, 2016.).

Independence of PRS from the monoallelic *SLC34A3* signal was ascertained with a logistic model testing phenotype against PRS, the presence or absence of a qualifying *SLC34A3* variant, a multiplicative interaction between PRS and *SLC34A3*, sex and the first ten principal components as covariates (to control for differences in ancestry). A null model containing just our covariates against the phenotype (sex and the first ten principal components) was used to calculate the model contribution to the phenotype variance. A liability adjusted pseudoR^2 was calculated for the model with and without the presence or absence of *SLC34A3* monoallelic variants using an estimated prevalence of USD of 5% (S. H. Lee et al. 2011). Analysis of the generalized linear model outputs and associated performance statistics, including bootstrapping of confidence intervals for the AUC-ROC was performed with the pscl (Zeileis, Kleiber, and Jackman 2008), DescTools (Signorell 2023) and boot (Canty et al. 2022) and packages in R. Per ancestry PRS was also performed.

The PRS was applied to 336 cases and 24541 controls.

4.3.5 Burden heritability regression for rare variants

Estimating the contribution to heritability from rare variants is challenging (Seunggeun Lee et al. 2012). By their nature, rare variants are seen infrequently leading to low statistical power alongside issues with population stratification and cryptic relatedness that are discussed in the Methods section.

The burden heritability regression (BHR) method assesses the heritability explained by the gene-wise burden of rare and ultrarare coding alleles. Burden heritability is defined as the fraction of phenotypic variance explained by the minor allele burden in each gene under a random-effects model. The inputs to BHR are variant-level summary statistics. Burden test statistics are then regressed on burden scores with the regression slope giving the estimated heritability (Weiner et al. 2023).

BHR was applied to both the 100KGP and UK Biobank datasets using the recommended default settings altered to match the input settings for our SAIGE-GENE

analysis whereby within each gene, variants were stratified into two allele frequency bins (minor allele frequency (MAF) < 1×10^{-5} and 1×10^{-5} – 1×10^{-4}). The model was conditioned on the genome-wide burden model and fixed for effects of *SLC34A3* given the *SLC34A3* association. Heritability estimates were liability transformed as per the PRS methodology.

4.4 Results

4.4.1 Participants

Participants

After quality control, genome build and ancestry matching, I identified 374 unrelated probands with USD (244 recruited to 100KGP under “nephrocalcinosis/nephrolithiasis” as their primary diagnosis and an additional 130 participants with HES codes indicating USD) and 24930 controls recruited to the UK 100,000 Genomes Project (100KGP). Table 4-1 details their demographics and clinical characteristics in detail.

Table 4-1 Clinical and Demographic characteristics of the USD and *SLC34A3* cohort

		All USD (n [%])	***SLC34A3 cases (n [%])***	***P-Value***
Median age (range)		44 (6-92)	38 (10-75)	0.38
Males		211 (56.37%)	12 (57.14%)	0.82
Self-reported ethnicity	White	248 (70.25%)	20 (95.24%)	0.01
	South Asian	18 (5.10%)	0	0.61
	Other Asian	9 (2.55%)	0	1
	Mixed/Other	22(6.23%)	0	0.62
	Black	3 (0.85%)	1 (4.76%)	0.21
	Chinese	2 (0.57%)	0	1
	Unknown	51 (14.44%)	0	0.09
Family history		79(22.38%)	5(23.81%)	0.09
Reported consanguinity		15 (4.25%)	0	1
Stone type if stated	Calcium oxalate nephrolithiasis	35 (9.92%)	2 (9.52%)	1
	Calcium phosphate nephrolithiasis	33 (9.35%)	0	0.24
	Calcium nephrolithiasis	2 (0.57%)	1 (4.76%)	0.16
	Uric acid nephrolithiasis	2 (0.57%)	0	1
	Unknown	281 (79.60%)	18 (85.71%)	0.78
Relevant endocrine or electrolyte manifestations	Hypercalciuria	186(52.69%)	15(71.44%)	0.12

			Urinary stone disease	
	Hypercalcemia	176 (49.86%)	2 (9.52%)	$2.06x10^{-04}$
	Hyperoxaluria	174 (49.29%)	0	$1.22x10^{-06}$
	Hyperphosphaturia	172 (48.73%)	0	2.6410^{-06}
	Hypocitraturia	171 (48.44%)	0	$2.67x\ 10^{-06}$
	Hyperparathyroidism	172 (48.73%)	0	$2.64x10^{-06}$
	Hypomagnesuria	170 (48.16%)	0	$2.74x10^{-06}$
	Hypoparathyroidism	166 (47.03%)	4 (19.00%)	0.01
	Hypocalcaemia	169 (47.88%)	0	$2.84\ x10^{-06}$
Extra-renal manifestations	Diabetes Mellitus	65 (18.41%)	1 (4.76%)	0.14
	Hypertension	60 (17.00)	5 (23.81%)	0.38
	Gout	35 (9.92%)	1 (4.76%)	0.71
	Obesity	35 (9.92%)	1 (4.76%)	0.71
ESRF		36 (10.20%)	1 (4.76%)	0.70
Median age ESRF (range)		50 (7-78)	67	

ESRF – End Stage Renal Failure, USD – urinary stone disease

Of the 244 primary recruited cases 26 were previously solved by 100KGP, with the relevant genetic diagnoses being reported back to the participants, representing a diagnostic yield of 10.7% (see Table 4-2 for full breakdown). 21/244 (8.6%) had a primary diagnosis in keeping with stone forming disease whilst 5/244(2%) had other secondary diagnoses that were delivered to them via the clinical reporting pipeline that did not account for their stone disease. All disease-causing genes followed their established modes of inheritance.

Table 4-2 Solved USD cases

Primary Diagnosis	**Gene**	**No. of probands and zygosity**
Cystinuria	*SLC7A9* *SLC3A1*	2 biallelic, 4 monoallelic 2 biallelic, 3 monoallelic
Primary Hyperoxaluria type 1	*AGXT*	3 biallelic
Primary Hyperoxaluria type 2	*GRHPR*	2 biallelic
Infantile hypercalcaemia	*CYP24A1*	2 biallelic
Rain syndrome	*FAM20C*	1 biallelic
Hereditary Hypophosphataemic Rickets with Hypercalciuria	*SLC34A3*	1 biallelic
SLC34A1 associated USD	*SLC34A1*	1 biallelic
Secondary diagnoses	**Gene**	**No. of patients and zygosity**
CHARGE syndrome	*CHD2*	1 monoallelic
Alport syndrome	*COL4A4*	1 monoallelic
Mucopolysaccharidosis	*GALNS*	1 biallelic
Gitelman syndrome	*SLC12A3*	1 biallelic
Beta Thalassemia	*HBB*	1 monoallelic

USD – Urinary stone disease

4.4.2 Rare variant association testing

Two genes showed statistically significant enrichment of rare and predicted damaging variation in USD cases compared with controls: *SLC34A3* (P =2.61x10^{-07}, OR=3.75, 95% CI 2.27-5.91) and *OR9K2*, encoding an olfactory receptor (P =2.03x10^{-06}, OR = 8.47, 95 % CI 3.23-18.81) (Figure 4-3) under the "missense+" tag.

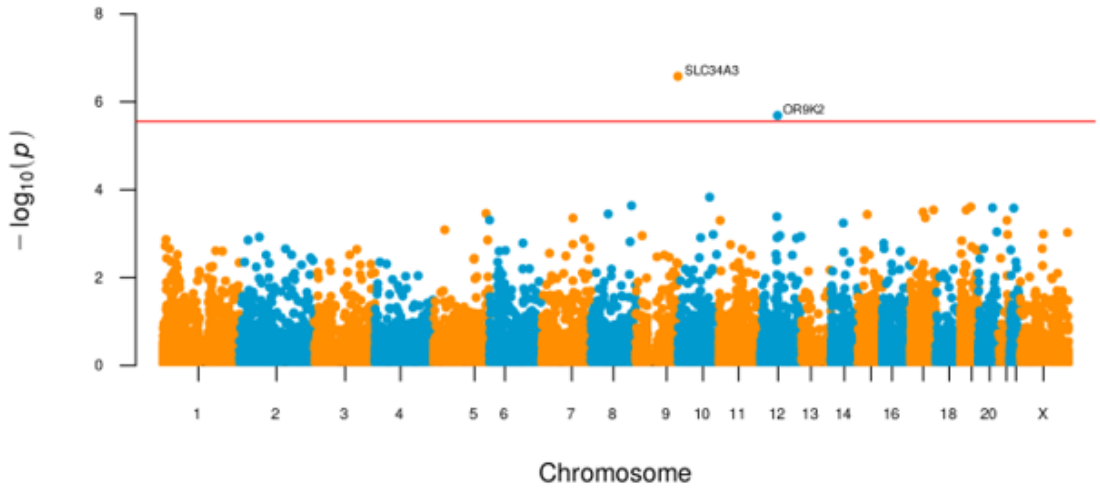

Figure 4-3 Gene based Manhattan for the association of likely damaging variants between USD cases and controls

Gene based Manhattan plot of the SAIGE-GENE analysis with the missense+ tag. Each point is a gene made up of variants that are predicted to be at least as damaging as missense, with a CADD score >20 and have a minor allele frequency (MAF) <0.01% in the gnomAD database. The horizontal line indicates the threshold for exome-wide significance. The only exome-wide significant associations were with *SLC34A3* (P =2.61x10^{-07}) and *OR9K2* (P = 2.03x10$^{-06)}$

No other genes were significantly enriched in the other tested collapsing tags: intronic, 5-UTR or 3-UTR, synonymous or splice site (see supplementary data for summary statistics).

4.4.3 Replication in UK Biobank

Association of USD with rare variation in *SLC34A3*, but not *OR9K2*, was replicated in publicly available analyses of whole exome sequencing data from 3,147 cases and 255,496 controls within the UKBB, an independent dataset (https://azphewas.com/). In

this analysis, multiple rare variants collapsing models were applied on a per gene basis and analysed with SKAT-O across all listed UKBB phenotypes. Under the "flexnonsynmtr" model, which equates to non-synonymous variants with a MAF<0.01% in both gnomAD and the UKBB with missense variants also having to fall within a region constrained for missense variation, USD was most strongly associated with *SLC34A3* ($p=3.67x10^{-10}$, OR = 2.01, see Figure 4-4). None of the cases in UKBB were homozygous for their qualifying *SLC34A3* variants (full list of variants in the supplementary data)

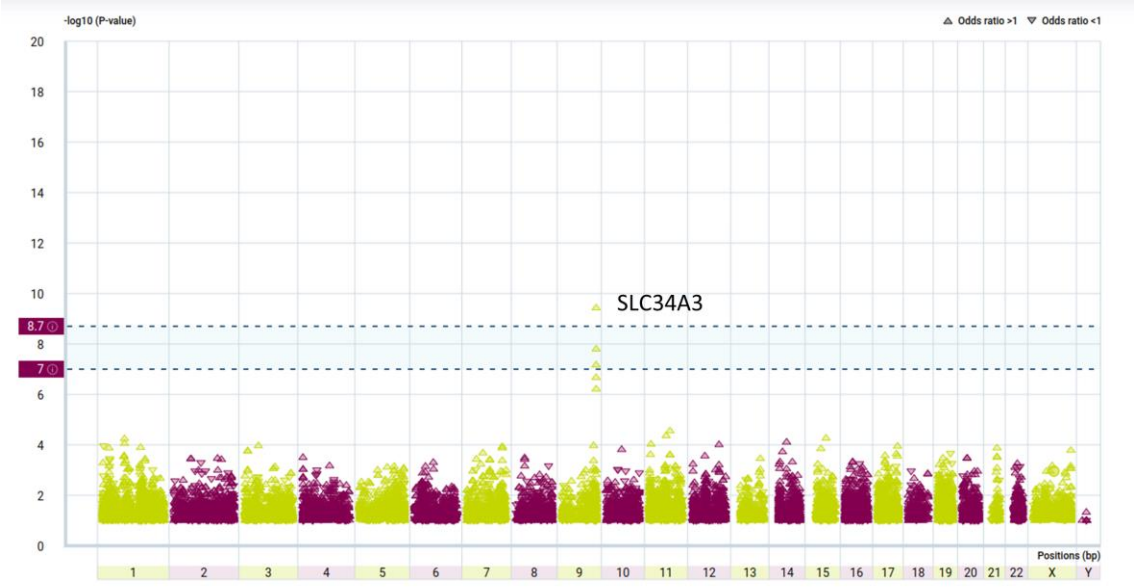

Figure 4-4 Gene based Manhattan of USD in the UKBB

Gene based Manhattan plot of the UK Biobank analysis obtained from the AstraZeneca PheWas portal. Each symbol represents variants in a gene that are predicted to be non-synonymous with a minor allele frequency (MAF) <0.1% in both gnomAD and the UK Biobank. The only exome wide significant association was *SLC34A3* ($p=3.67x10^{-10}$, OR = 2.01). The multiple symbols under *SLC34A3* (the "build-up") represent different analyses with respect to the predicted severity of the included variants. The lower dashed vertical line indicates exome-wide significance and the upper dashed line exome-wide significance corrected for the ~1500 different phenotypes analysed.

4.4.4 Metanalysis

Metanalysis of the two datasets confirmed a significant association in *SLC34A3* ($p=1.94x10^{-18}$). There were no other genome wide significant associations detected (full results in supplementary data).

4.4.5 Phenotype/Genotype analysis of SLC34A3

Qualifying variants were found in 6% of the ascertained stone population in the study (21/374) compared to 1.6% (389/24930) in the controls. Of the 21 cases with qualifying variants, 14 were recruited with stone disease, 1 with Congenital Anomalies of the Kidney or Urinary Tract, 4 with cystic kidney disease and two with intellectual disability. 19 cases were heterozygous, and 2 were compound heterozygous for qualifying *SLC34A3* variants with both cases' variants being confirmed in *trans* (Table 4-3). Qualifying variants in the control population were all heterozygous. Excluding the two compound heterozygous patients from the analysis and re-running the association led to a smaller but still significant association ($p=1.47x10^{-06}$).

Table 4-3 *SLC34A3* demographics and variant details

ID	Age	SEX	Recruited disease	FH	Solved?	Ancestry	CHR	POS	REF	ALT	Con	HGVSc	HGVSp	Clinvar	rsID
1	35-40	M	CyKD	N	N	WE	chr9	137233282	G	A	M	ENST00000361134.2:c.634G>A	ENSP00000355353.2:p.Ala212Thr	.	rs754447323
2	55-60	M	CyKD	N	Y	WE	chr9	137228902	C	T	M	ENST00000445101.3:c.287C>T	ENSP00000406665.2:p.Ala96Val	.	rs1290843983
3	45-50	M	USD	N	N	WE	chr9	137236048	G	A	M	ENST00000361134.2:c.1432G>A	ENSP00000355353.2:p.Gly478Arg	.	rs758689905
4	45-50	M	CyKD	1	Y	WE	chr9	137228893	C	T	M	ENST00000445101.3:c.278C>T	ENSP00000406665.2:p.Ala93Val	.	rs558338995
5	70-75	M	USD	N	N	WE	chr9	137234509	C	T	M	ENST00000361134.2:c.1187C>T	ENSP00000355353.2:p.Thr396Met	.	rs138798032
6	15-20	M	USD	N	N	WE	chr9	137233223	C	T	M	ENST00000361134.2:c.575C>T	ENSP00000355353.2:p.Ser192Leu	198610	rs199690076
7	25-30	F	USD	N	N	WE	chr9	137234509	C	T	M	ENST00000361134.2:c.1187C>T	ENSP00000355353.2:p.Thr396Met	.	rs138798032
8	05-10	M	ID	N	N	WE	chr9	137228657	A	C	Syn	ENST00000445101.3:c.42A>C	ENSP00000406665.2:p.Gly14%3D	.	.
9	25-30	M	USD	N	N	WE	chr9	137233051	G	A	M	ENST00000361134.2:c.496G>A	ENSP00000355353.2:p.Gly166Ser	.	rs200536604
10	65-70	M	CyKD	1	N	WE	chr9	137233223	C	T	M	ENST00000361134.2:c.575C>T	ENSP00000355353.2:p.Ser192Leu	198610	rs199690076
11	70-75	F	USD	1	N	WE	chr9	137232892	C	T	M	ENST00000361134.2:c.413C>T	ENSP00000355353.2:p.Ser138Phe	.	rs141734934
12	20-25	F	ID	N	N	WE	chr9	137236173	T	TC	FS	ENST00000361134.2:c.1561dup	ENSP00000355353.2:p.Leu521ProfsTer72	444095	rs765816079
13	15-20	M	ID	N	Y	WE	chr9	137228893	C	T	M	ENST00000445101.3:c.278C>T	ENSP00000406665.2:p.Ala93Val	.	rs558338995
14	40-45	F	CyKD	N	N	WE	chr9	137234630	C	T	M	ENST00000361134.2:c.1234C>T	ENSP00000355353.2:p.Arg412Trp	.	rs373242362
15	10-15	F	USD	N	N	AA	chr9	137234509	C	T	M	ENST00000361134.2:c.1187C>T	ENSP00000355353.2:p.Thr396Met	.	rs138798032
16	45-50	M	USD	N	N	WE	chr9	137228893	C	T	M	ENST00000445101.3:c.278C>T	ENSP00000406665.2:p.Ala93Val	.	rs558338995
17	40-45	M	USD	Y	N	WE	chr9	137232928	G	A	SD	ENST00000361134.2:c.448+1G>A	.	445687	rs150841256
18	40-45	F	USD	Y	N	WE	chr9	137236239	G	A	SG	ENST00000361134.2:c.1623G>A	ENSP00000355353.2:p.Trp541Ter	423400	rs762610288
19	25-30	F	CAKUT	N	N	WE	chr9	137233223	C	T	M	ENST00000361134.2:c.575C>T	ENSP00000355353.2:p.Ser192Leu	198610	rs199690076
Compound heterozygous patients															
20	20-25	F	USD	N	Y	WE	chr9	137233223	C	T	M	ENST00000361134.2:c.575C>T	ENSP00000355353.2:p.Ser192Leu	198610	rs199690076
20	20-25	F	USD	N	Y	WE	chr9	137236239	G	A	SG	ENST00000361134.2:c.1623G>A	ENSP00000355353.2:p.Trp541Ter	423400	rs762610288
21	10-15	F	USD	N	N	WE	chr9	137231705	G	A	SL	ENST00000361134.2:c.3G>A	ENSP00000355353.2:p.Met1?	.	rs369400414
21	10-15	F	USD	N	N	WE	chr9	137234505	T	C	M	ENST00000361134.2:c.1183T>C	ENSP00000355353.2:p.Phe395Leu	.	rs560440785

CyKD – cystic kidney disease, USD – urinary stone disease, ID – intellectual disability, CAKUT – congenital abnormalities of the kidney and urinary tract, WE – white European, AA – African-American, Con – consequence, M – missense, Syn – synonymous, FS – frameshift, SD - splice donor, SG – stop gain, SL – start loss

All 21 patients had gone through the Genomics England clinical pipeline with 4 cases receiving genetic diagnoses including: *PKD2*-associated cystic kidney disease (two patients), Kabuki syndrome (*KMT2D*) and biallelic *SLC34A3*-associated USD. In the solved biallelic *SLC34A3* case both variants were annotated as (likely) pathogenic by Clinvar (rs199690076 and rs762710288) whereas in the unsolved biallelic *SLC34A3* case the evidence for a clinical grade diagnosis was weaker (rs369400414 and rs560440785): whilst both variants met the inclusion criteria for the collapsing rare variant association, they did not meet the benchmark for clinically reportable results.

For both biallelic *SLC34A3* cases there was not enough phenotype data available to ascertain whether they met clinical diagnostic criteria for Hereditary Hypophosphataemic Rickets with Hypercalciuria, however, both had the hypercalciuria HPO code on record. The top ten HPO codes associated with the patients with *SLC34A3* associated USD are found in Figure 4-5.

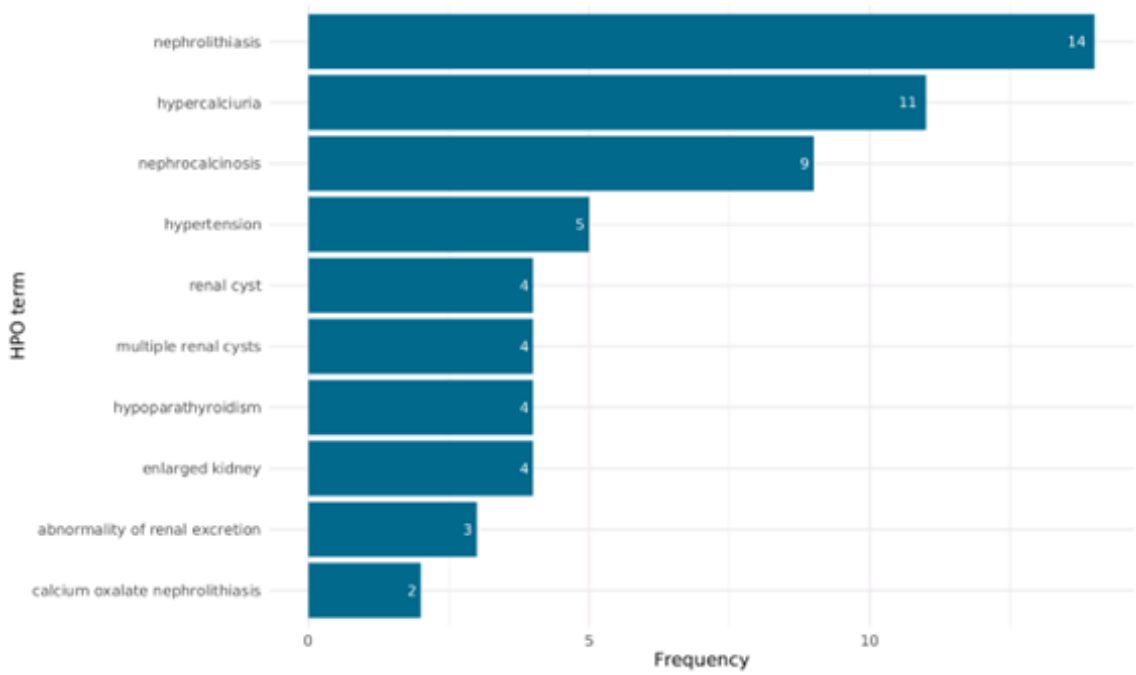

Figure 4-5 Top ten HPO codes associated with USD

The top ten HPO codes associated with cases with qualifying SLC34A3 variants. Nephrolithiasis makes up the most common associated clinical code followed by hypercalciuria and nephrocalcinosis.

4.4.6 Polygenic risk scoring

In a cohort of 336 unsolved cases and 24541 controls (both depleted for *SLC34A3* variants that would have qualified for inclusion in the SAIGE-GENE analysis), there was a significant elevation in USD PRS compared with controls (P =3.1 $x10^{-04}$). Initial analysis including a cohort of the cases with qualifying *SLC34A3* variants did identify statistically significant differences between the three cohorts (P =4.4 $x10^{-04}$), but this signal was driven by the difference between unsolved cases and controls (Figure 4-6).

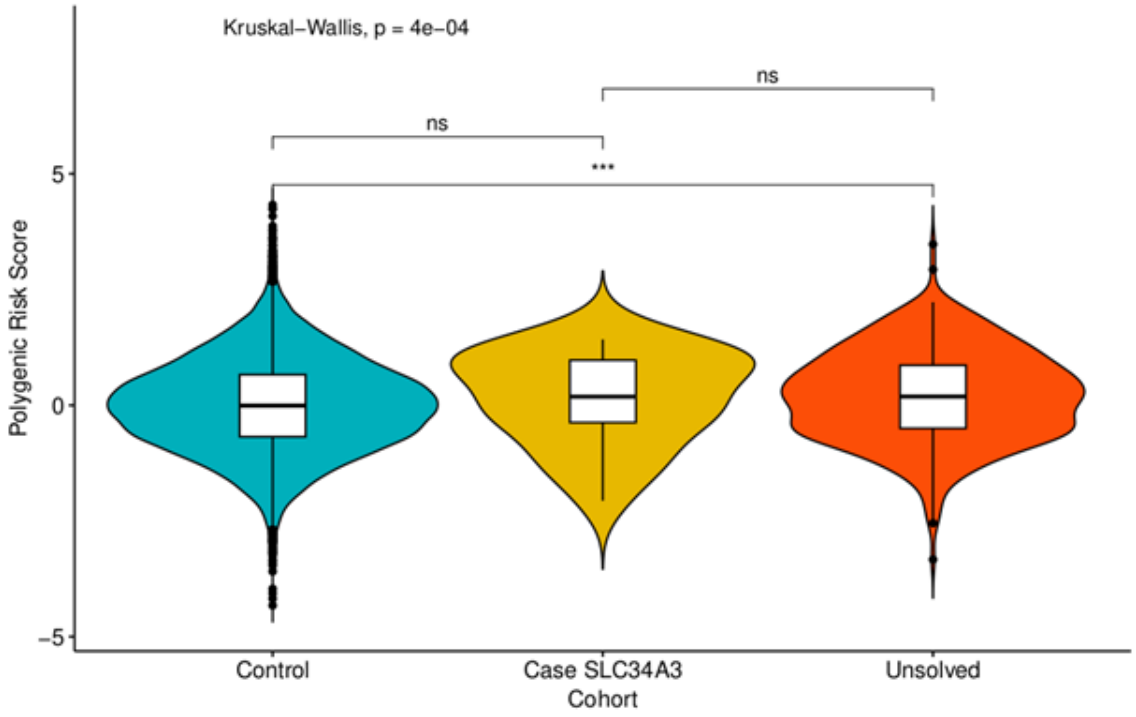

Figure 4-6 Violin and boxplot comparing polygenic risk score distribution across USD cohorts

Violin and boxplot showing the polygenic risk score (PRS) distributions between controls (those with qualifying *SLC34A3* variants removed), cases with qualifying *SLC34A3* variants and unsolved patients who have neither a reportable nor a qualifying variant in *SLC34A3*. The means of the three PRS were compared with a Kruskal-Wallis test (p=4.04 $x10^{-04}$) with the signal being driven by the difference between unsolved cases and controls (paired Willcox =3.6 $x10^{-04}$). **** = statistical significance, ns = no significant difference*

The difference between the control group and the *SLC34A3* cases did not reach statistical significance given the small number of *SLC34A3* cases. Adjusted odds of a USD diagnosis increased by a factor of 1.22 (95% CI: 1.10–1.36; P = 0.003) per standard deviation of PRS in an adjusted model including sex and the first 10 principal components. The area under the curve (AUC) was 0.62 (95% CI: 0.60–0.66).

4.4.7 PRS modelling with SLC34A3 risk

In the model, there was a significant association between phenotype and both PRS (P=3.8x10^{-04}) and the presence of a monoallelic *SLC34A3* variant (P =2.72x10^{-08}).

However, there was no log-additive (multiplicative) interaction between PRS and the *SLC34A3* binary with the phenotype (P =0.77), although this is likely to be underpowered to detect such an interaction. Of the other covariates, sex (P =1.34x10^{-05}) and the fourth principal component (P =4.72x10^{-08}) were also strongly associated with the phenotype. The presence of an *SLC34A3* variant increased the frequency of USD within 100KGP when plotted against polygenic risk score (Figure 4-7).

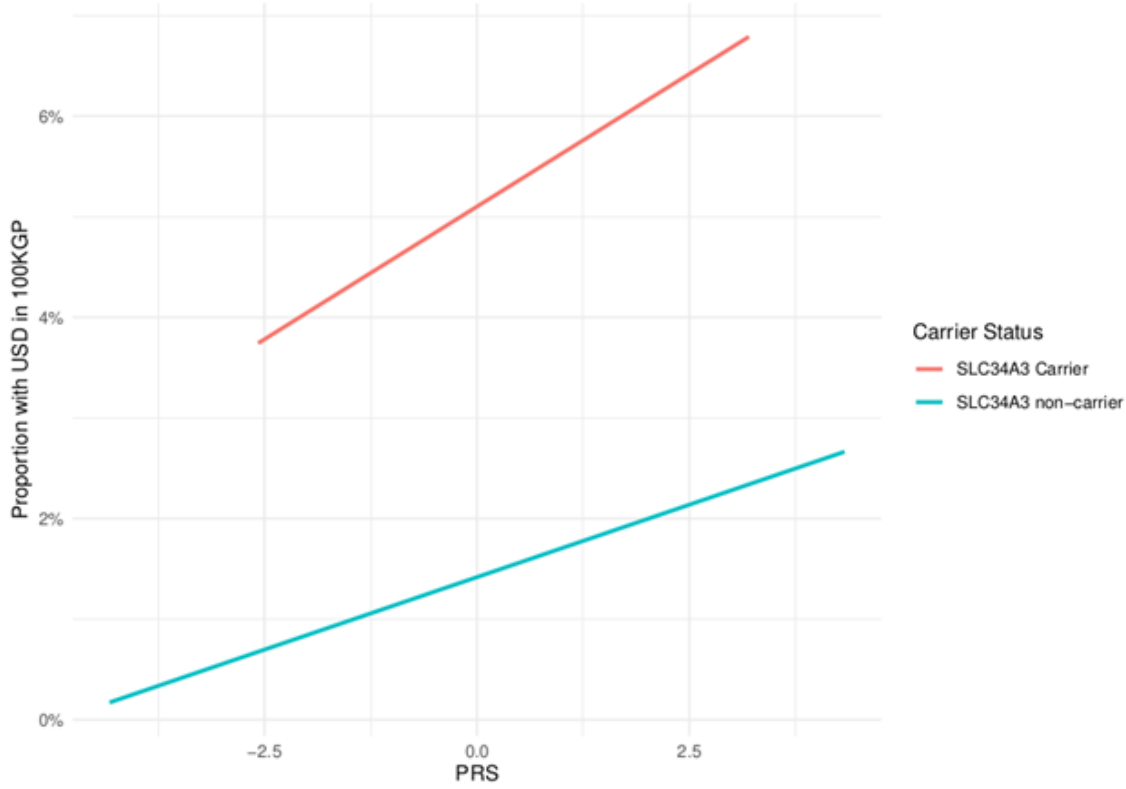

Figure 4-7 Frequency of urinary stone disease by centile of PRS

A line plot showing the frequency of USD against polygenic risk score stratified by the presence of absence of a qualifying SLC34A3 variant in the 100,000 genomes project cohort.

The addition of the *SLC34A3* variant binary to the linear model including PRS led to a significant rise in the estimated variance explained by the model (liability adjusted pseudo-R^2 rising from 5.1% to 14.2%) and a modest increase in the model's predictive

capability (AUC 0.64 95% CI 0.61-0.66). Implying a 9.1% contribution of *SLC34A3* to the heritability model.

4.4.8 Burden heritability regression

To confirm the heritability of the rare variants as well as the contribution of *SLC34A3* using orthogonal methodology I applied the burden heritability regression tool to both the 100KGP and UKBB datasets. The liability adjusted gene-wise burden heritability of rare and ultra-rare predicted loss-of-function (pLOF) and damaging missense variants explained 10.8% (95% CI = 7.7-13.9%) of phenotypic variance within the 100KGP dataset with variants in *SLC34A3* making up 7.6% (95% CI =5.64-9.6%) of this signal in total (lambda =1.076). In the UKBB analysis the liability adjusted gene-wise burden heritability of rare and ultrarare pLOF and damaging missense variants explained 5.4% (95% CI = 3.3%-8.4%) of the phenotypic variance with variants in *SLC34A3* making up 3.7% (95% CI= 1.3-6.1%) of this signal (lambda =1.018).

4.5 Summary

- Rare variants in *SLC34A3* are the most important genetic risk factor for USD
- These variants make up a significant proportion of the hitherto unexplained heritability of USD

4.6 Discussion

I identified rare variants in *SLC34A3* and *OR9K2* as significantly associated with USD among 100KGP participants.

4.6.1 SLC34A3

SLC34A3 encodes the sodium-dependent phosphate transport protein 2C expressed in the proximal tubule (NaPi-IIc). Our results highlight the importance of *SLC34A3* as a contributor to USD, with more than 5% of patients in this cohort from the 100KGP harbouring predicted damaging variants and independent replication of this association in the UKBB dataset (Q. Wang et al. 2021). Importantly, the odds ratio for stone disease with rare, predicted damaging variants in *SLC34A3* was comparable to that of the polygenic risk score derived from numerous common variants across the whole genome and a model combining PRS and *SLC34A3* monoallelic variants accounted for 14% of the genetic heritability of stone disease. This suggests that rare monoallelic variants in *SLC34A3* fall into an intermediate category of pathogenicity: they are insufficient to cause fully penetrant Mendelian disease but convey a higher disease risk than the aggregate effects of known common risk alleles. There is increasing recognition of such intermediate role for rare predicted damaging variants. For instance, approximately 1% of the general population carry such variants in *COL4A3* or *COL4A4* but they are not fully penetrant for the development of progressive chronic kidney disease (autosomal dominant Alport syndrome) and are therefore considered a risk factor (Gibson et al. 2021). Equally a recently described *UMOD* variant (p.Thr62Pro) seen in ~1/1000 individuals of European ancestry has been shown to confer an intermediate level of risk

of kidney failure, augmenting the known spectrum of *UMOD*-associated kidney disease (Olinger et al. 2022).

In the AstraZeneca UKBB rare variant collapsing analysis that used twelve different sets of qualifying variant filters (models: ten dominant models, one recessive model, and one synonymous "control" variant model) *SLC34A3* was the gene most significantly associated with USD and the association was strongest with models that include predicted damaging missense variation and weakened if the filter was constrained to only those variants predicted to cause protein truncation, suggesting that any predicted damaging variants in this gene can contribute to the risk of USD.

Our findings therefore highlight the importance of monoallelic variants in *SLC34A3* for USD. *SLC34A3* was reported in 2006 as a recessive disease gene for the rare disorder hereditary hypophosphataemic rickets with hypercalciuria (HHRH) (Bergwitz et al. 2006; Lorenz-Depiereux et al. 2006). While there was already recognition in the original publication that heterozygous carriers in the affected families were frequently affected by hypercalciuria, it remains listed as a recessive disease gene in OMIM (*609826). Yet, there is good evidence for the impact of monoallelic variants: an investigation in a cohort of affected families showed that the risk of USD was 46%, 16% and 6% in subjects with biallelic, monoallelic or no causative variants, respectively (Dasgupta et al. 2014). This is consistent with a paradigm in which identification of a rare, damaging monoallelic *SLC34A3* variants can be regarded as a risk factor for stone disease but is not a diagnostic finding.

The underlying mechanism is thought to be hypophosphataemia-mediated suppression of fibroblast growth factor-23 (FGF23) with consequent activation of the 1-a hydroxylase and increased 1,25 dihydroxy vitamin D levels, which in turn stimulates intestinal calcium absorption (Lorenz-Depiereux et al. 2006). The same mechanism is thought to apply in infantile hypercalcaemia due to biallelic loss-of-function variants in *SLC34A1* (Schlingmann et al. 2016). While the role of monoallelic *SLC34A1* variants in hypercalciuria has been controversial, large genome-wide studies have demonstrated a significant association between both coding and non-coding variants of *SLC34A1* and USD, consistent with the concept that a reduction in proximal tubular phosphate transport does increase the risk for kidney stones (Benjamin B. Sun et al. 2022; Oddsson et al. 2015).

Our study provides evidence of clinical relevance for coding variants in *SLC34A3*, with a significant enrichment of rare and predicted damaging variants in USD patients compared with controls, among participants of both the 100KGP and UKBB. While the 100KGP did not specifically encourage enrolment of patients with a family history of the respective disorders, it is possible that there may have been a recruitment bias that would have inflated the percentage of *SLC34A3*-related disease. Nevertheless, the additional identification of rare *SLC34A3* variation as the strongest rare variant association in UKBB participants provides independent replication and raises the question of whether identification of these risk variants in individual patients would provide utility in clinical practice.

While the modest risk effect precludes predictive use of such a test, the above pathophysiological mechanism suggests that phosphate supplementation may be a suitable treatment to stimulate FGF23 and thereby suppress 1-a hydroxylation of vitamin D in patients at risk of *SLC34A3*-related kidney stone disease. Indeed, successful use of this treatment has been reported (Schönauer et al. 2019; Dhir et al. 2017). However, clinical trial data would be needed to support such an intervention because there is a risk that large doses of phosphate supplementation increase the urinary phosphate concentrations with consequent increased risk of calcium phosphate precipitation. Indeed, nephrocalcinosis has been associated with phosphate supplementation in patients with *PHEX*-associated hypophosphataemic rickets, although these patients typically received enormous doses (Haffner et al. 2019). Thus, more data are needed before embarking on routine phosphate supplementation in *SLC34A3*-associated USD.

4.6.2 Olfactory associations

While the association with *SLC34A3* was independently replicated in the UKBB, a signal was not observed in this dataset at *OR9K2,* so the possibility exists that this finding is a type 1 error, which is well recognised with olfactory receptor genes owing to their enrichment for loss of function variation without clinical consequences (Karczewski et al. 2020; D. G. MacArthur et al. 2012). However, similar olfactory gene associations have not been observed in other studies using the 100KGP dataset analysed with similar methodology (Q. Wang et al. 2021), and there is increasing recognition that olfactory receptors regulate transport processes in many organ systems (Dalesio et al. 2018): *OR9K2* is expressed in the intestine (Uhlén et al. 2015), and it is conceivable that

it may be involved in the regulation of absorption of substrates with relevance to stone formation, such as oxalate or calcium. Therefore, further studies are needed to assess the relevance of *OR9K2* in USD.

4.7 Strengths and limitations

There were several notable limitations when undergoing this study. Firstly, I was underpowered in our discovery cohort to discover novel gene variants with either a weaker effect on risk or of greater rarity due to our small case number. Those that were recruited clearly may have had more severe USD, with potential ascertainment bias towards genes more likely to be involved with more severe disease. Following on, the addition of *SLC34A3* into the logistic model with PRS risks a "winner's curse" bias whereby its effect is overstated. This is supported by the BHR scores being higher in the 100KGP versus the UKBB cohort, although the confidence intervals do overlap for the analysis. However, the fact that the association between rare and predicted damaging variants in *SLC34A3* and USD is replicated in an independent cohort (as the "top gene") and that this signal is enriched on metanalysis is reassuring. The heritability and effect size of the PRS in the 100KGP is similar to the known common variant contribution to USD implying a similar underlying genetic architecture between cohorts. In terms of phenotyping, I was limited by the depth of information available for our cohort in 100KGP, which does not include the biochemical stone properties for most cases. As with all diseases that can be silently present, there is also the chance that our control population has been misclassified with a proportion of them having USD. Whilst all efforts were taken to remove any cause of potential USD as well as hospital codes directly pertaining to USD, I cannot completely exclude this possibility. While the

statistical evidence of enrichment of rare genetic variants in *SLC34A3* is strong and replicable, without functional analysis of each missense variant it is not possible to determine which of the observed variants play a causal role in disease.

4.8 Conclusion

This chapter highlights the substantial contribution of rare and predicted damaging variants in *SLC34A3* to the burden of USD, helping to close the missing heritability gap and supporting the idea of routine genetic testing in affected patients. Rare variants with intermediate effect sizes play an important role (comparable if not greater to the common variant contribution) in USD and likely in other diseases.

Chapter 5. Extreme-early onset hypertension

In this chapter I discuss work of Megha Manoj, an intercalating medical (iBSc) student, whom I supervised closely. This research project was awarded the best abstract at the British and Irish hypertension society ("Abstracts from the 2022 Annual Scientific Meeting of the British and Irish Hypertension Society (BIHS)" 2022), and Megha was invited to present the paper at the American Heart Association Hypertension meeting in Boston, MA in September 2023. I have included it in my thesis because although the work was performed jointly, it was under my close direction (I wrote almost all the code for her analysis), applying similar methodology to the previous results chapters on the topic of extreme early-onset hypertension (EEHTN). I have developed Megha's work by conducting further analyses to round this project out further namely the rare variant association testing in a primary hypertension cohort and the polygenic risk score modelling. I have delineated this by using "we" for work I supervised and "I" for work I undertook myself.

5.1 Introduction to extreme early onset hypertension

The global impact of hypertension is significant, affecting approximately a quarter of the world's population and representing the primary modifiable risk factor for cardiovascular disease and mortality (Ezzati et al. 2002). Traditionally, treatment efforts have prioritized individuals with the highest 10-year risk of cardiovascular events, often considering age as a significant determinant (Sundström et al. 2011). However, research indicates that hypertension at a young age substantially elevates the risk of cardiovascular events later in life (Yano et al. 2018). The presence of hypertension during early adulthood contributes to the premature onset of coronary heart disease, heart failure, stroke, and transient ischemic attacks. Notably, blood pressure tends to track strongly with age, meaning that elevated blood pressure in youth is likely to persist into later life (Barker et al. 1989). Studies have linked baseline blood pressure in young adults to cardiovascular mortality in follow-up assessments (Xiaoli Chen and Wang 2008) .

The evidence highlights the significance of early-life risk factors in determining long-term health outcomes and suggests that delaying the consideration of cardiovascular health until middle age may not be appropriate. With this there has been great interest in early diagnosis of hypertension particularly those secondary cases where a fixed and potentially treatable cause can be found (Rimoldi, Scherrer, and Messerli 2014).

In terms of hypertension as a whole there has been great interest in understanding the genetic architecture (Padmanabhan, Caulfield, and Dominiczak 2015). The high prevalence of hypertension and its associated morbidity present a clear justification for genetic studies. Multiple GWAS studies had implicated over 50 SNVs associated with primary hypertension, with the estimated heritability from these being roughly 2%, despite family studies estimating between 30-50%, with most of the loci yet to have their functional connotations understood. Monogenic syndromes and low frequency variants whilst rarer, have contributed to a larger functional characterisation of hypertension (>25 implicated genes), helping to elucidate the pathways behind salt retention and underlying the role the kidneys and adrenal glands have in blood pressure regulation. Most of the rare variant work has been conducted using linkage analysis with the few exome sequencing projects yielding mixed results probably due to low sample sizes (Albrechtsen et al. 2013; Paranjpe et al. 2019). However, to date there has not been a case/control whole genome sequencing analysis of secondary hypertension.

Investigating young adults for secondary causes of hypertension is strongly advocated by most guidelines with risk factors likely to yield a successful diagnosis including being less than 30 years of age or having a blood pressure >180/110mmHg (“Recommendations for Research | Hypertension in Adults: Diagnosis and Management | Guidance | NICE” 2018.d.; B. Williams et al. 2018; Whelton et al. 2018) . At present the process of diagnosis for these patients is both expensive and invasive necessitating a large battery of blood, imaging and functional tests at great cost (Chhabra et al. 2022). As a cohort they are enriched for monogenic causes and thus “extreme early-onset hypertension” (EEHTN) was included as a phenotypic cohort in the 100KGP. EEHTN is defined as a blood pressure in an adult >160/100mmHg in clinic and an average blood pressure of 150/95 on ambulatory blood pressure monitoring occurring below the

age of 30 (excluding patients with Primary hyperaldosteronism, phaeochromocytoma, Cushing's syndrome and hyper/hypothyroidism).

In this chapter I will discuss our investigation of the EEHTN cohort using techniques established thus far.

5.2 Aims

- To identify common variants that may contribute to EEHTN
- To determine the prevalence of known monogenic disease in a cohort of patients with EEHTN.
- To discover novel candidate genes using an unbiased exome-wide rare variant association testing approach.
- To understand the structural variant burden of known causative genes in an EEHTN cohort
- To understand the common variant contribution to EEHTN compared to primary hypertension and a control population

5.3 Methods

In this study we were limited by the time given to the iBSc project. The common and rare variant analyses used similar methodology to that described in previous chapters and will not be detailed again (Megha did this on the EEHTN cohorts and I on the primary hypertension cohort). The SV analysis was cut down to a candidate gene approach in the interests of time and will be detailed after the cohort creation. As part of my further analysis, I created a primary hypertension cohort and conducted rare variant analysis and polygenic risk scoring in this cohort also.

5.3.1 Cohort creation

For this analysis we created three separate ancestry matched, unrelated cohorts: The EEHTN cohort recruited as probands under the "EEHTN" tag (EETHN), a group including the EEHTN cohort and patients with a diagnosis of HTN prior to turning 30 as ascertained by querying their HES records (HES-EEHTN) and the same group as minus those patients with renal disease ("renal removed" or RR-EETHN). Controls were initially selected as the unaffected relatives of patients not recruited with EEHTN or renal disease and then further depleted them for primary and secondary hypertension using HES/HPO codes.

I created a primary hypertension cohort by searching the 100KGP for patients with "primary" or "essential" hypertension in the participant explorer (which searches HPO, HES and other affiliated terminology linked to the NHS data spine for all recruited participants). These were then further refined to be the unaffected relatives of non-renal, non-EEHTN probands recruited to the 100KGP. I then further removed those with any HPO or HES code pertaining to secondary causes of hypertension and to add a buffer to the criteria of EEHTN I only took those primary hypertension cases that were over the age of 40 at diagnosis of primary hypertension.

Each cohort then underwent relatedness filtering, quality control and ancestry matching as per the chapters above. This resulted in a final total of 179 cases and 20411 controls in the original EEHTN cohort, 901 cases and 20852 controls in the HES-EEHTN

generated cohort, 449 cases and 20852 controls in the RR-EETHN cohort and 7923 cases and 20695 controls in the primary hypertension cohort (P-HTN).

All recruited cases went through the clinical interpretation arm of the 100KGP whereby a panel of 27 “green” genes with known causative links to hypertension were applied virtually to the WGS data. These results were fed into a multidisciplinary team who applied the ACMG criteria and issued a diagnosis where appropriate.

5.3.2 SV analysis

SV were counted on a per gene basis as per the CyKD SV chapter (rare, exonic SV/CNVs) in cases and controls. However, only the green listed genes on the PanelApp entry for EEHTN were used (*CUL3, CYP11B1, CYP11B2, CYP17A1, HSD11B2, KCNJ5, KLHL3, MTX2, NR3C1, NR3C2, SCNN1B, SCNN1G, TTC21B, WNK1 and WNK4*). I applied a two tailed Fisher’s exact test to the cases and controls per gene and SV type.

5.3.3 Polygenic risk scoring

We applied a validated polygenic risk score for primary hypertension consisting of 186,726 variants derived from a European cohort within the UK Biobank (Sinnott-Armstrong et al. 2021). These were lifted over to build 38 using the UCSC liftover tool (Hinrichs et al. 2006) and then I scored these on our cohort using the PLINK2 score command (C. C. Chang et al. 2015). I then performed statistical tests for significance, modelling, goodness of fit testing and plotting using R.

5.4 Results

All summary statistics are available in the supplementary information.

5.4.1 Participants

200 proband were recruited to the 100KGP under the EEHTN tag, with 179 being used for analysis after ancestry matching and relatedness filtering. Table 5-1 details the demographics of the cohort:

Table 5-1 Demographics of the recruited EEHTN cohort

Demographics	Number	Percentage
Male	114	57%
Median Age	40(IQR 30-91)	NA
Recruitment		
Singleton	141	70.5%
Trio with Mother and Father	26	13%
Duo with Mother and Father	22	11%
Duo with other Biological Relative	5	2.5%
Families > 3 participants	4	2%
Trio with Mother or Father and other Biological Relationship	2	1%
Affected 1st degree relative	78	39%
Consanguinity in parents	8	4%
Self-reported ethnicity		
European	84	42%
African	17	8.5%
Other Asian	12	6%
South Asian	11	5.5%
East Asian	1	0.5%
Mixed	8	4%
Not stated/unknown	67	33.5%

IQR – Interquartile range

Of the 200 recruited cases two (1%) were solved by the 100KGP clinical interpretation arm. One case has a *PKD2* truncating variant (Clinvar 562248) whilst the other case had a heterozygous missense *COL4A5* variant (X_108686076) and was given a diagnosis of X-linked Alport syndrome (in a female).

5.4.2 Rare variant association testing

In the EEHTN cohort there were no significant genome-wide association in any gene compared to controls (179 vs 20411 cases) in either the LoF or missense categories of collapsing analysis (Figure 5-1).

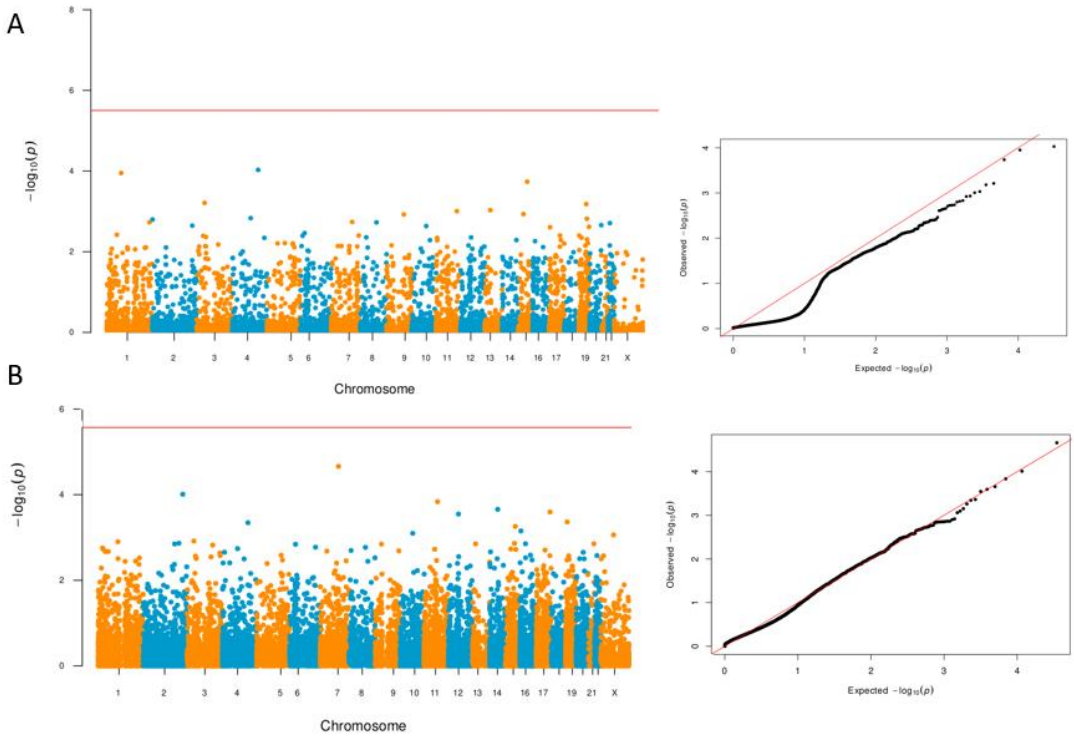

Figure 5-1 Gene based Manhattan plots for rare variant association testing in EEHTN

A – Likely damaging (missense+) B – Loss-of-function. There was no significant genome-wide association in any gene compared to controls (179 vs 20411 cases) in either the LoF or missense categories of collapsing analysis, The QQ plots did not reveal any genomic inflation.

In the HES-EEHTN cohort under the LoF tag there was significant enrichment of LoF variants in the *PKD1* gene ($P = 2.70 \times 10^{-13}$, OR=6.10 95%CI 3.82-9.44) compared to controls (901 cases versus 20852 controls) (Figure 5-2). There was no enrichment under the missense annotation (Figure 5-3).

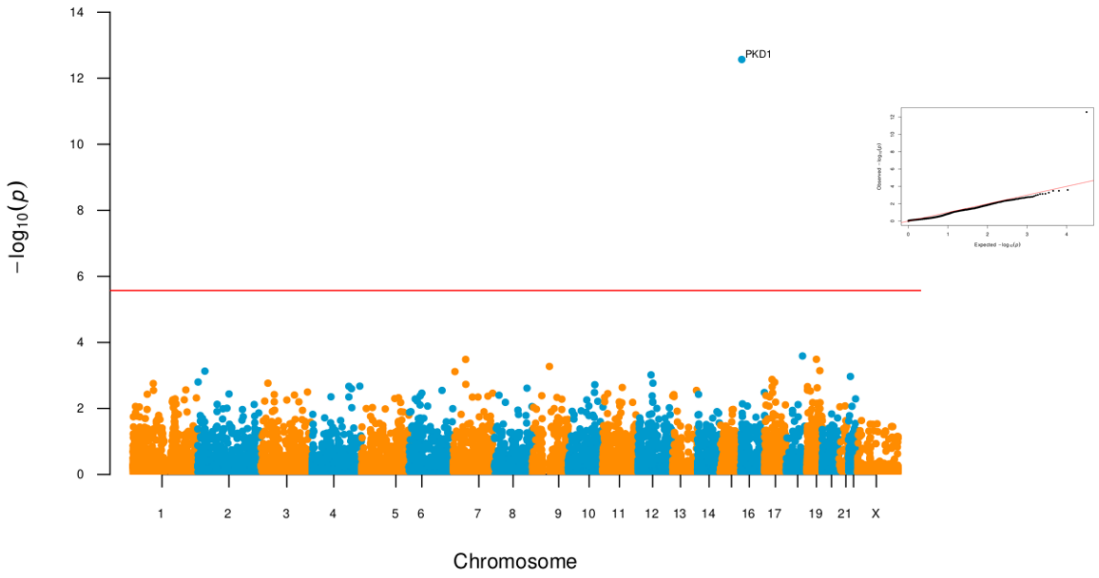

Figure 5-2 Gene based Manhattan for the association of loss-of-function variants between the HES-EEHTN cases and controls

Manhattan plot showing the results of the loss-of-function rare variant analysis, variants had a high confidence of being loss-of-function and were either rare (MAF < 0.001) or not seen in gnomAD in the HES generated EEHTN cohort, with the red line representing the -log10 of the P value (~2.6 x 10^{-6}). There is enrichment (P=2.70 x 10^{-13}) of LoF rare variants in *PKD1* affecting the EEHTN phenotype in this cohort of 901 cases v 20852 controls. The QQ-plot does not show any inflation.

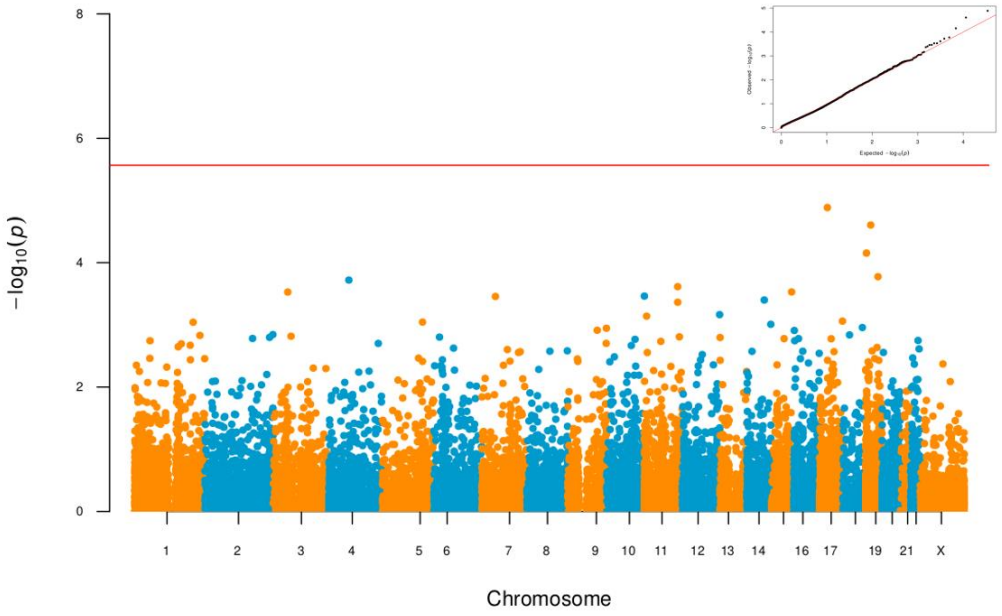

Figure 5-3 Gene based Manhattan for the association of likely damaging variants between the HES-EEHTN cases and controls

Manhattan plot showing the results of the likely damaging rare variant analysis (missense+). Variants were rare (MAF<0.001), at least a missense in consequence and had a CADD score of ≥ 20 in the HES generated EEHTN cohort, with the red line representing the -log10 of the P value (~2.6 x 10^{-6}). There is

no enrichment of any genes in this cohort of 901 cases v 20852 controls. The QQ-plot does not show any inflation.

Removing the cases with a renal diagnosis and running the analysis again results in abolishment of the *PKD1* signal under the LoF tag (Figure 5-4).

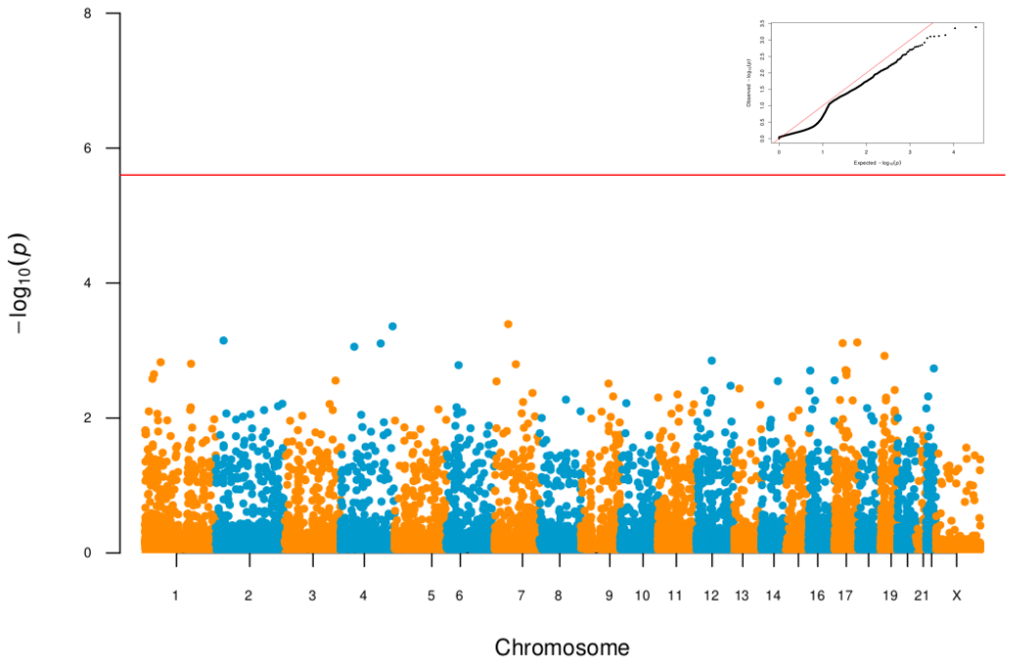

Figure 5-4 Gene based Manhattan for the association of loss-of-function variants between the RR-EEHTN cases and controls

Manhattan plot showing the results of the loss-of-function rare variant analysis, variants had a high confidence of being loss-of-function and were either rare (MAF < 0.001) or not seen in gnomAD in the HES generated EEHTN cohort with renal diagnoses removed, with the red line representing the -log10 of the P value (~2.6 x 10^{-6}). The previously seen enrichment in *PKD1* is lost in this cohort of 449 cases v 20852 controls. The QQ-plot does not show any inflation.

Rare variant association testing in the primary hypertension cohort did not show any enrichment in any tested mask (Figure 5-5)

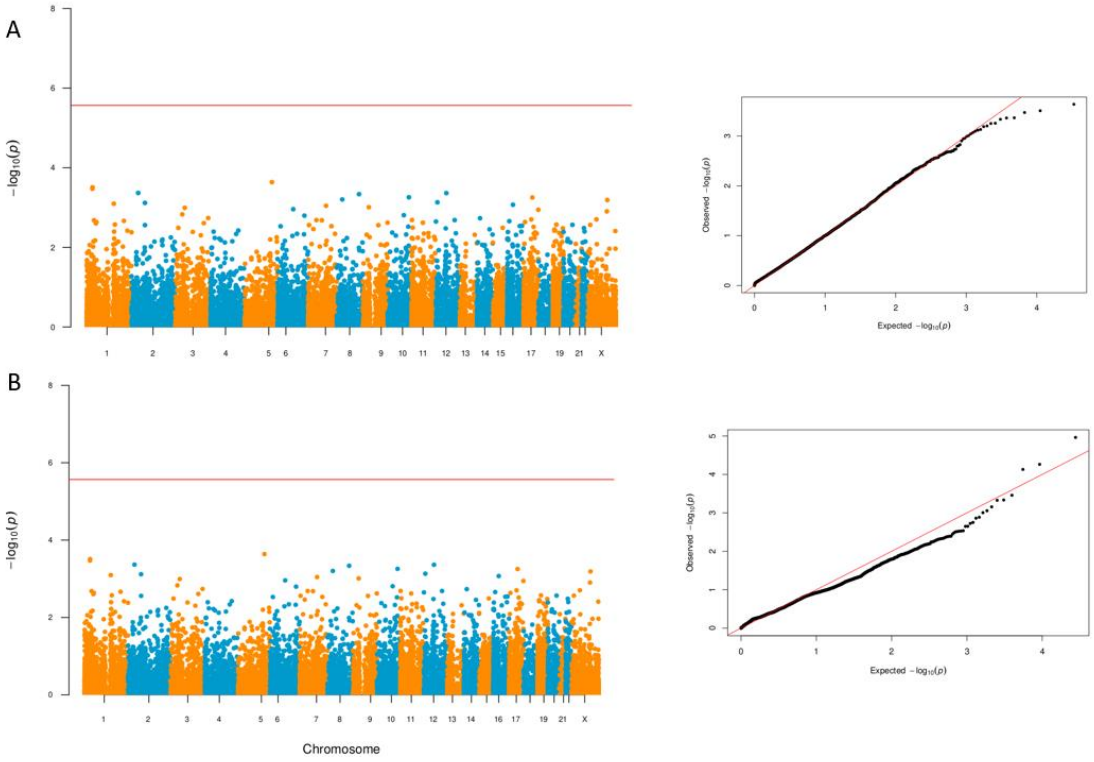

Figure 5-5 Gene based Manhattan plots for rare variant association testing in Primary hypertension

A – Likely damaging (missense+) B – Loss-of-function. There was no significant genome-wide association in any gene compared to controls (7923 vs 20695 cases) in either the LoF or missense categories of collapsing analysis. The QQ plots did not reveal any genomic inflation.

5.4.3 *PKD1* genotype/phenotype analysis

The *PKD1* signal was made up of 27 cases, all heterozygous for their respective variants. Of the 27 cases, 22 were recruited under the CyKD tag, 2 with epilepsy. 1 with congenital heart disease, 1 with CAKUT and another with unexplained kidney failure in young people. 13 out of the 27 patients had been solved by the 100KGP (12 of the CyKD and 1 of the CAKUT patients). A full breakdown of the 27 cases can be found in table 5-2.

Table 5-2 Demographic and variant details of individuals making up the *PKD1* signal

Patient #	*Age*	*Variant*	*Consequence*	*Solved*	*Consanguinity*	*FH*
1	5-10	16:2092140:G:A	Missense	No	No	No

2	30-35	16:2091153:CGGCGAACAGCA:G	Downstream	No	No	No
3	30-35	16:2114528:C:CG	Frameshift	Yes	No	No
4	10-15	16:2109342:C:CG	Frameshift	No	No	No
5	30-35	16:2117794:G:A	Stop gain	Yes	No	Yes
6	10-15	16:2108612:G:T	Stop gain	Yes	No	Yes
7	30-35	16:2105320:A:G	Splice donor	Yes	No	No
8	30-35	16:2135615:GC:G	Frameshift	No	No	No
9	20-25	16:2108680:G:A	Stop gain	No	No	Yes
10	5-10	16:2107947:ACGCCAGC:A	Frameshift	Yes	No	No
11	35-40	16:2111543:G:GACGC	Frameshift	Yes	No	No
12	50-55	16:2111514:TC:T	Frameshift	Yes	No	No
13	45-50	16:2108438:CTG:C	Frameshift	No	No	Yes
14	40-45	16:2110151:CCT:C	Frameshift	No	No	No
15	35-40	16:2114199:C:A	Stop gain	Yes	No	Yes
16	35-40	16:2091793:C:T	Downstream	Yes	No	No
17	35-40	16:2108680:G:A	Stop gain	No	No	Yes
18	30-34	16:2091861:G:T	Downstream	Yes	No	Yes
19	30-35	16:2090283:A:G	Downstream	No	No	Yes
20	40-45	16:2093886:CG:C	Downstream	No	Yes	Yes

21	40-45	16:2093824:C:T	Downstream	Yes	No	Yes
22	30-35	16:2097765:G:A	Stop gain	Yes	No	Yes
23	30-35	16:2090952:G:A	Downstream	Yes	No	No
24	50-55	16:2092078:TC:T	Downstream	No	No	Yes
25	35-40	16:2092140:G:A	Downstream	No	No	No
26	10-15	16:2135611:CG:C	Frameshift	No	No	Yes
27	5-10	16:2090983:G:GCGCA	Downstream	Partially	No	Yes

FH – Family history

5.4.4 Common variant association testing

Across all three cohorts (EEHTN, HES-EEHTN, RR-EEHTN) there were no significant genome-wide associations in variants with a MAF>0.01. There was no evidence of genomic inflation (Figure 5-6).

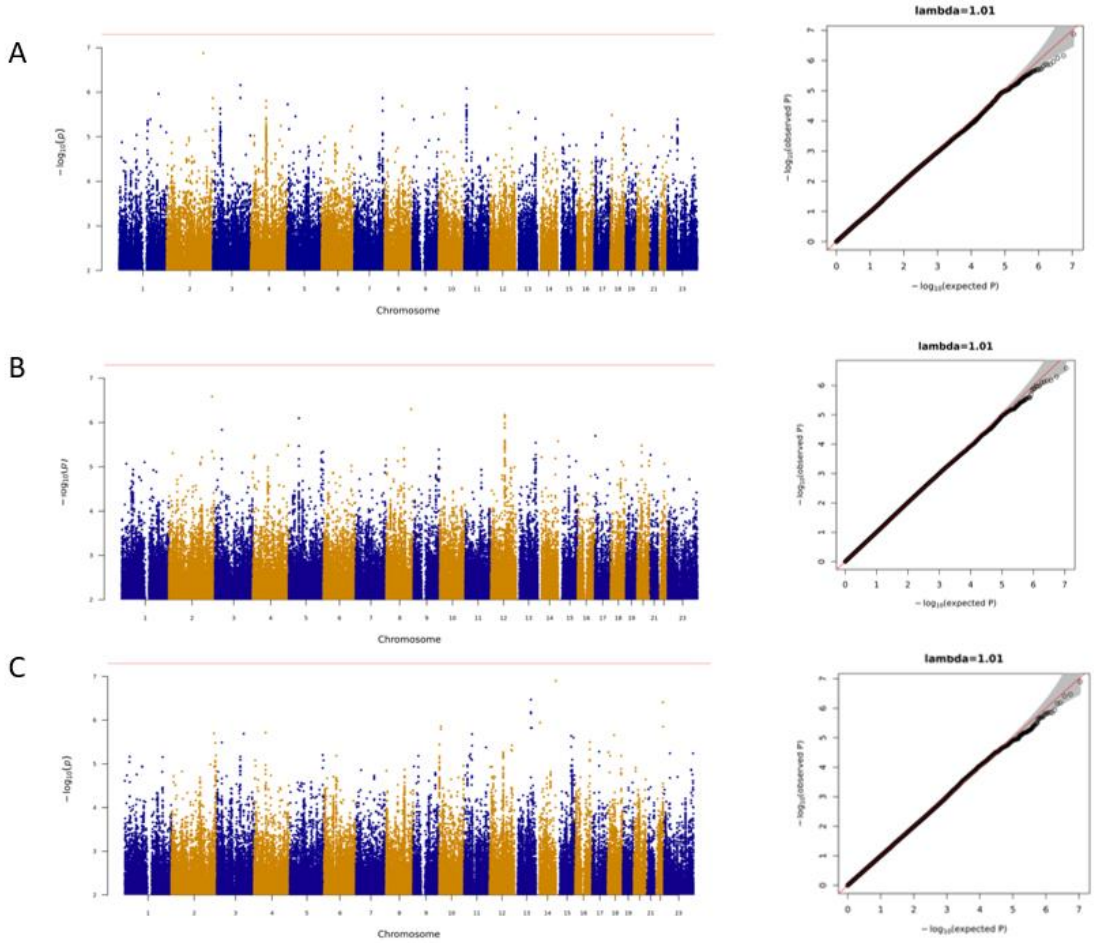

Figure 5-6 Manhattan plot of EEHTN GWAS

There is no significant association at a variant level in the seqGWAS of all three EEHTN cohorts. A – EEHTN (11019220 markers), B - HES-EEHTN (11010124 markers), C – RR-EEHTN (10999561 markers). There was no evidence of genomic inflation across the ancestry matched cohorts as evidence by the associated QQ plots.

5.4.5 Polygenic risk scoring and heritability

Polygenic risk scoring was performed in the RR-EEHTN, P-HTN and a control cohort. There was a significant difference in PRS between the control cohort and the two disease cohorts ($P<2.2x10^{-16}$) but no difference between the two hypertension cohorts (Figure 5-7). Liability adjusted R^2 (narrow sense heritability, h^2) explained 26.48% (SE 7.02%) and 22.10% (SE 1.35%) of the RR-EEHTN and P-HTN cohorts' phenotypic variance. The AUC of the RR-EETHN model was 0.91 (95% CI 0.90-0.93) and the P-HTN model was 0.61 (95% CI 0.59-0.63) respectively.

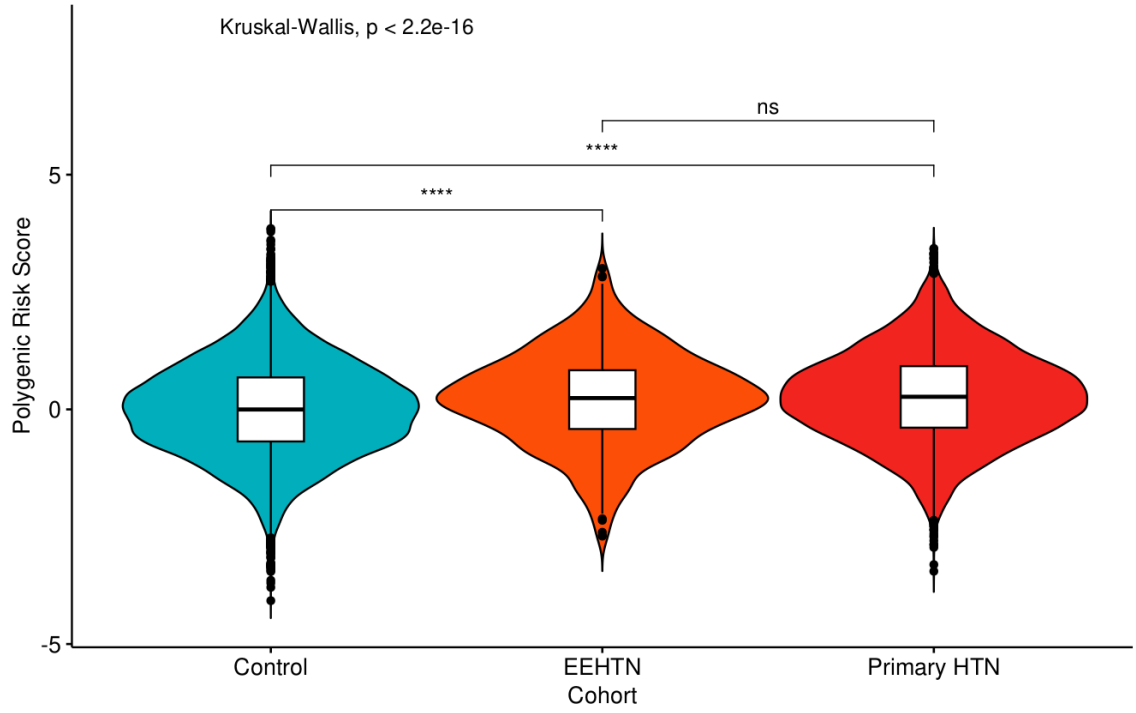

Figure 5-7 Violin and boxplot comparing polygenic risk score distribution across HTN cohorts

Violin and boxplot showing the polygenic risk score (PRS) distributions between controls (without primary or secondary hypertension), cases EEHTN and cases with primary HTN. The means of the three PRS were compared with a Kruskal-Wallis test (p=2.2 $x10^{164}$) with the signal being driven by the difference between unsolved cases and controls. *** = *statistical significance, ns = no significant difference*

5.4.6 Structural analysis

On a case-control basis looking at rare, exon crossing SVs there was no enrichment in any of the EEHTN associated genes. Of note there was a single *WNK1* exon spanning CNV loss in a case (Chr12:765990-880472), with none seen in controls. This variant was not seen in the ClinVar database, however, there was a comparable CNV loss seen that was linked to Pseudohypoaldosteronism type II (ClinVar: #5161, Figure 5-8).

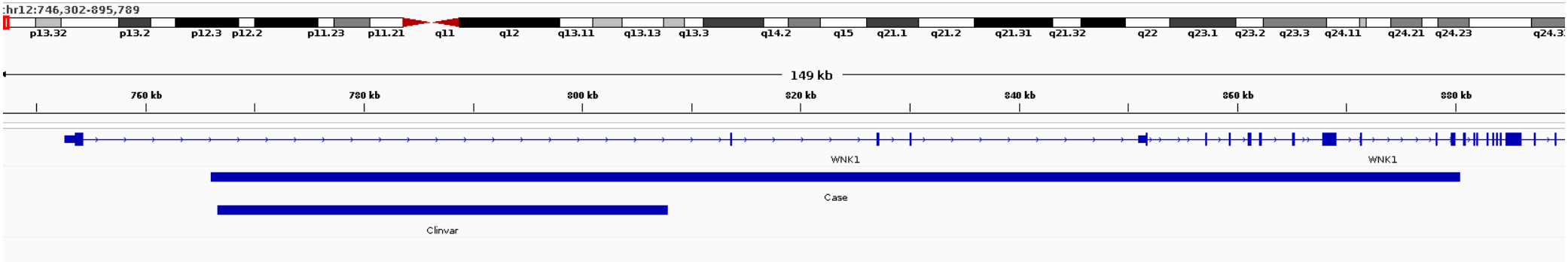

Figure 5-8 Plot of novel and Clinvar *WNK1* deletions

A plot from the IGV browser demonstrating the similarities between the *WNK1* deletion detected in a single EEHTN case (top bar) and a Clinvar deletion (bottom black bar) that has been determined as pathogenic with *WNK1* along the top. It is likely the 100KGP variant is pathogenic also.

5.5 Summary

- Rare predicted damaging variants in *PKD1* are associated genome-wide with extreme early onset hypertension.
- I am underpowered to detect common variant contributions to the phenotype and our subsequent analysis confirmed this.
- Polygenic risk scoring and modelling imply EEHTN may represent a phenotypic extreme of primary hypertension.
- There is no evidence of structural variants playing a genome wide role in any of the candidate genes for EEHTN.

5.6 Discussion

5.6.1 EEHTN as a complex polygenic disorder

These findings replicate those found in the UKBB whereby *PKD1* is the top genome wide significant gene for causes of secondary hypertension under a rare variant collapsing gene model (*P=8.16x10-13*, OR=116.65 95%CI = 53.38-254.90) (Q. Wang et al. 2021). This serves to confirm the workings of this methodology as well as underline the complexities of studying a poorly defined phenotype with an uncertain underlying genetic architecture.

Whilst CyKD has a strong monogenic architecture, hypertension does not. 95% of hypertensive patients have primary hypertension which is a highly heterogenous condition with a large amount of environmental influence. Many of the EEHTN cohort likely have primary hypertension, given the fact that to be eligible for the 100KGP recruitment patients had to have been tested for many of the known secondary causes of hypertension (Primary hyperaldosteronism, phaeochromocytoma, Cushing's syndrome and hyper/hypothyroidism) implying a relatively thorough work-up. Our findings that there is no difference in the hypertension polygenic risk score between the EEHTN and primary hypertension cohorts serves to underline this point. It took a huge number of cases before any genetic signal became apparent from common variant genome wide association studies in hypertension and this is likely due to the heterogenous nature of the disease and it's underlying polygenic architecture as well as environmental factors. Many different phenotypes can result in hypertension and thus thousands of cases are needed to power studies at a population level, or the cohorts need to be very well characterised to allow for cleaner case definitions. In fact, in the same AstraZeneca led UK Biobank analysis, *PKD1* is the lead gene in the primary hypertension cohort also ($P=5.23x10^{-13}$) followed by the genes *PKD2, COL4A4, UMOD, CACNA1D* and *NR3C2*, the latter two being known monogenic causes of hypertension (primary aldosteronism and Geller syndrome respectively). This highlights that a) renal disease is a key cause of hypertension, the first four genes being linked to CyKD, Alport and ADTKD respectively, b) many patients with primary hypertension are misclassified and that c) age may not be the best metric to identify causes of secondary hypertension as despite

the median age of UKBB participants being 58 there are still secondary causes being discovered, finally the UKBB had 149,306 primary hypertension cases against 194,956 controls highlighting the number of cases required to get tractable genetic signals at a population level in a hypertension cohort. Interestingly, the AUC for the EEHTN cohort was much higher than the primary hypertension cohort despite a smaller sample size, whilst this doesn't give any information on the effect size it does hint that genetic factors have a larger role in younger versus older patients with hypertension.

Heritability estimates for hypertension also serve to highlight that complex phenotyping is present. Heritability from the UK Biobank generated and made publicly available by the Neale Lab (https://nealelab.github.io/UKBB_ldsc/h2_browser.html) estimate a liability adjusted heritability for primary hypertension at around 24%, a finding comparable to mine. Even with a small amount of refinement this can be drastically altered, "hypertensive renal disease" i.e. one where the hypertension is secondary to a renal disease in the UK Biobank has an adjusted heritability estimate of 68% (Ojavee, Kutalik, and Robinson 2022) highlighting the complexities with estimating the heritability of what is effectively multiple different diseases.

5.6.2 *PKD1* as an early marker of severity in CyKD

Hypertension is commonly associated with CyKD, present in 60% of ADPKD patients before renal impairment occurs (Ecder and Schrier 2001). In young adults hypertension is present in roughly 17%, being associated with increased severity of ADKPD including the presence of a *PKD1* variant, a family history of early renal replacement therapy, lower renal function, larger kidney size and a history of haematuria (Schrier et al. 2014; Cornec-Le Gall et al. 2016) . While hypertension in ADPKD is often due to kidney enlargement and cyst compression leading to an induction of a high renin state in ADPKD (Loghman-Adham et al. 2004), *PKD1* is expressed in the major vessels, cilia of endothelial cells and vascular smooth muscle cells, meaning that absence or an insufficiency of *PKD1* expression is linked to vascular structural and functional abnormalities via polycystin mediated modulation of vasodilatory pathways including nitric oxide (Sharif-Naeini et al. 2009). Knockout mice for both *PKD1* and *PKD2* reveal

disturbance in flow mediated vasodilation and an increased blood pressure confirming *PKDs* mechanistic role in hypertension separate to upregulation of the renin-angiotensin-aldosterone system(RAAS) (MacKay et al. 2022, 2020; Hamzaoui et al. 2022).

In our cohort, three of the patients with qualifying variants in *PKD1* had developed hypertension prior to their diagnosis underlying the fact that hypertension can develop prior to clinically appreciable disease. A deleterious, likely truncating *PKD1* variant will lead to larger kidneys and cysts with associated hypertension from the ensuing compression and RAAS induction, however, it is also likely that these patients have a higher degree of associated endothelial dysfunction.

5.6.3 Role of WGS in EEHTN

There has been hope that WGS can replace some elements of the diagnostic odyssey that many patients with rare or complex disorders face (Vrijenhoek et al. 2015). In secondary hypertension the workup requires multiple blood, urine and imaging tests and can occasionally involve invasive adrenal vein sampling (Omura et al. 2004). The clinical arm of the 100KGP has yielded a diagnosis in only 1% of cases (2 participants), a very low return especially when compared with published panel and WES cohorts of between 6.79%-11.2% (Paranjpe et al. 2019; Bao et al. 2020). This is despite the good evidence that WGS is superior to WES, even when detecting protein coding variants (Belkadi et al. 2015). In the above studies the cohorts were either enriched for likely monogenic hypertension or unselected as opposed to the 100KGP cohort which was depleted for likely secondary causes of hypertension. Whilst it would be difficult to recommend WGS as a key diagnostic step given the current evidence a better study to assess this would be to perform WGS and analysis on patients being referred to a secondary hypertension clinic and then comparing the diagnostic WGS rates with that of a traditional work-up. The results discussed from the UKBB analysis above in 5.6.1 hint that an unselected population would likely yield better diagnostic rates using WGS. This would help more robustly answer the question of the role WGS has in diagnosing secondary hypertension.

5.7 Strengths and limitations

This study is the first WGS analysis of an EEHTN cohort and helps define further experiments in this area. This analysis helps underline the severity of *PKD1* variants in causing early onset hypertension, highlights the numbers required to get tractable genetic signals in a EEHTN cohort and gives an assessment of WGS for diagnosis. The primary limitation as alluded to above is the lack of power in our cohort to detect meaningful signal. Whilst it is easy to call for bigger numbers, the proposal to conduct a trial of WGS against traditional diagnosis for secondary hypertension would represent a way of boosting power by increasing the likely diagnostic yield. Bar the *PKD1* signal that is found in a predominantly renal cohort we did not find any enrichment at a variant, gene, or SV level. Of note the SV analysis focussed on only known genes and the next step would be to conduct a genome-wide SV analysis similar to the work in the CyKD chapter to assess the genome-wide SV burden in CyKD.

5.8 Conclusion

Rare variants in *PKD1* were significantly associated with the EEHTN phenotype, predominately made up of patients with CyKD. The lack of signal in other analytical methods was hampered by small sample size as evidenced by the presence of signal in larger biobanks using similar methodology. WGS when used in a subset of EEHTN patients who have been assessed in a specialised clinic does not seem to yield adequate diagnostic rates, the patients who remain are likely to have primary hypertension as ascertained by polygenic risk scoring.

Chapter 6. Discussion

The primary objective of this thesis was to better characterize several renal disorders using the largest WGS biobank of rare renal diseases to date with the aim of providing insights into their pathogenesis and underlying genetic architecture. It is the first time the non-coding and coding variations across the entire allele frequency spectrum has been studied in an unbiased manner across all three diseases. The results of this work have been discussed in detail in previous chapters and the
key findings are summarised below:

- In CyKD monogenic architecture prevails at a coding, non-coding, and structural variant level with rare monogenic causes from all three cohorts explaining disease in the bulk of the cohort.
- The largest meta-analysis of common variants to date does not reveal any significant associations in CyKD, however, enrichment of rare alleles in the Finnish population help confirm our findings in the 100KGP that monoallelic *PKHD1* variants make an important contribution to CyKD.
- Variants with a MAF >0.1% make a roughly 9% contribution to the heritability of CyKD and common variants from those patients with unsolved CyKD make a roughly 9% contribution to the heritability of those patients with solved CyKD. There is some evidence that these variants are protective. This highlights the role more common variants have even in very monogenic disease.
- Rare *SLC34A3* variants of moderate effect size account for a large proportion of the missing heritability of USD.
- In EEHTN, larger cohorts are required to obtain tractable genetic signals within the rare variant space, however, rare, damaging *PKD1* variants are an important cause of secondary hypertension.
- The primary hypertension and EEHTN cohorts have similar genetic architecture as evidenced by overlap in their PRS.

Going into this thesis, I saw it as a series of siloed experiments on the “rare” and “common” variant space. However, as my work evolved over time and my

understanding of the resources and techniques improved, I realised that such a viewpoint has mainly been fostered by our inability to interrogate the full allelic spectrum at once. I have been able to probe the area of low frequency variants with small to intermediate effect size as well as assess the polygenic contribution within three diseases, using modelling to understand their interplay to genetic architecture.

As discussed in Section 1.2, genetic architecture has historically been thought of as either monogenic, oligogenic or polygenic with the more recent addition of the omnigenic model. CyKD and secondary hypertension are seen as monogenic whilst USD is seen as polygenic. However, what is common between all of them in this thesis is that as cohort sizes have increased and sequencing coverage has improved genetic contributions across the allele spectrum and variant types, have been shown to contribute to their respective phenotypes. I will discuss insights into genetic architecture for all three diseases below:

6.1 The genetic architecture of Cystic kidney disease

CyKD is a very monogenic disease. 994 of the 1209 (82%) tested cystic kidney disease cases had a likely explaining monogenic or single structural variants cause for their disease identified in an unbiased way. The known causative genes all have tractable pathophysiology in or along the polycystin pathway and there is a clear precedent to view the disease as such. However, intrafamilial variability and general phenotypic variability amongst patients with the same class of variants hint at other factors involved such as epistasis or gene/environment interactions.

Historically, this has focused on an oligogenic model with studies looking for a small number of rare of modest frequency variants of moderate to large effect in modifier genes (Fain et al. 2005; A. Persu et al. 2002; D. Walker et al. 2003; Y.-H. Hwang et al. 2016). Such studies have largely failed to find convincing oligogenic markers, although patients with a second hypomorphic *PKD1* variant alongside their driving variant do have a phenotype more akin to ARPKD (Durkie et al. 2021) and the *TSC2* gene disruptions encompassing *PKD1* (and *IFT140*, which lies between the two) lead to a more severe phenotype (Sampson et al. 1997).

A polygenic model of CyKD has not been seriously examined prior to the publication of two papers referenced in section 3.4.6 (Blair, Hoffmann, and Shieh 2022; Khan et al. 2023). My work has shown a ~9% contribution of common and low-frequency variants to the phenotypic variance of CyKD as well as variants in patients without monogenic CyKD contributing roughly 9% heritability to those with a monogenic cause of CyKD. Clearly this work needs to be validated in another cohort but taken together, gives strong evidence that polygenic architecture has an important role in a predominantly monogenic disease. A model where common and low-frequency variants alter the penetrance or rarer pathogenic alleles could explain the variability seen in the CyKD phenotypes. This is increasingly being recognised as a model in disorders such as hypercholesterolaemia, neurodevelopmental disorders and inherited cancer syndromes (Paquette et al. 2017; Niemi et al. 2018; Oetjens et al. 2019; Fahed et al. 2020; Mars et al. 2020; Weiner et al. 2023). The gold standard experiment would be to take families with a known monogenic cause of CyKD but with phenotypic variability to analyse further.

The omnigenic model stipulates that a small number of core genes with biologically interpretable effects interact with a larger set of peripheral genes with the effects between the two systems being mediated by connected intracellular networks (Boyle, Li, and Pritchard 2017). Here the boundary between polygenic and omnigenic becomes blurred. The finding that common and low frequency variations contributes to the phenotypic variance of CyKD supports this theory. Work by Gazal et al found coding variants explain a much larger fraction of heritability for low-frequency variants (~26%) compared to common variants (~8%), due to the effects of negative selection preventing deleterious variants from becoming common (Gazal et al. 2018). In CyKD this is supported by the idea that well-described “core gene” such as *PKD1, PKD2* etc explain a large proportion of the disease risk but many other more common variants with non-zero effects modify these networks. The word modify is key here, as my PRS work shows a protective effect of common variants. However, whether identifying the large number of variants with small effects sizes will lead to translatable insights into underlying biology remains to be seen.

Finally, the role that more newly discovered genes have in elucidating genetic architecture needs to be understood. Their odds ratios for causing CyKD are nowhere near those of *PKD1/PKD2*, yet their presentation is of a monogenic cause of CyKD potentially due to somatic second hits. Table 6-1 represents the odds ratio of developing CyKD in different genes across both the 100KGP and the UK Biobank (odds ratios taken from the AstraZeneca analysis using the model closest to my analysis with the 100KGP, the odds ratios are adjusted for those cases that were between 40-70 years old at recruitment to match the UKBB criteria), highlighting that some of the newer genes reported in CyKD are likely to have such a low penetrance that they may seldom exhibit Mendelian patterns of inheritance in families and may be more usefully be regarded clinically as intermediate risk factors for developing CyKD. Communicating this information clearly to patients and their relatives is likely to be important when counselling them about the pros and cons of predictive testing for these disorders.

Table 6-1 Age adjusted odds ratio of developing CyKD in the 100KGP (n=741) and UKBB (n=825)

Gene	100KGP OR (95%CI)	UKBB OR (95%CI)
PKD1 truncating	264 (218-329)	658 (451-959)
PKD1 non-truncating	7.90 (6.84 -9.10)	8.97(7.44-10.82)
PKD2 truncating	931 (525-1600)	1310 (697-2460)
PKD2 non-truncating	13.36 (9.91-17.91)	12.92 (9.61-17.36)
GANAB non-truncating	1.63 (0.79-3.02)**	Not seen
GANAB truncating	5.40 (0.11-54.56)**	Not seen
DNAJB11 truncating	1.07 (0.94-1.24)	30.05 (7.12-126.57)**
IFT140 truncating	12.21 (5.85-24.50)	14.99(9.92-22.66)
ALG5 truncating	1.00 (0.12-3.77)**	Not seen
ALG9 truncating	7.20 (0.71-40.35)**	22.03 (9.66-50.23)**
COL4A3 non-truncating	2.72 (1.71-4.15)	Not seen
Monoallelic *PKHD1* truncating	3.23 (1.36-6.56)	2.13 (1.01-4.50)**

** - Gene not significantly associated with the CyKD phenotype in the association analysis. The 100KGP results are presented for those individuals who were between 40-70 years old at the time of recruitment to match the UKBB recruitment analysis.

Whilst this may hold true, the potential for them to be a modifier alongside other unascertained genetic or environmental factors remains to be investigated.

6.2 The genetic architecture of urinary stone disease

USD has equally been divided into those with monogenic disorders (Halbritter et al. 2015) and then at a population level, the common variants risk loci derived from GWAS (Howles et al. 2019). Our study, alongside a few more published around the same time (Benjamin B. Sun et al. 2022) have interrogated the space between these two groups and convincingly shown that low-frequency variants of intermediate effect size such as in *SLC34A3* and *SLC34A1* increase the risk of USD combined with a combination of polygenic risk and known environmental factors such as obesity and smoking. With a much smaller proportion of case USD cases explained by monogenic causes, as opposed to CyKD, the likelihood of an omnigenic model seems much more plausible.

6.3 The genetic architecture of Early onset hypertension

Hypertension for clinical reasons has been divided into primary and secondary causes. My work has shown an interplay between those labelled as likely to have secondary hypertension and those with primary hypertension, with the caveat being that our cohort was depleted for solved secondary causes prior to recruitment. GWAS of primary hypertension have uncovered >1000 loci throughout the genome associated with the phenotype (Padmanabhan and Dominiczak 2021). This fits the omnigenic model of disease, whereby whilst the variants with the largest effect are fairly enriched in related genes and pathways that are related to the pathogenesis of the disease, variants contributing the most to the heritability of a trait are found across the genome (Boyle, Li, and Pritchard 2017). Interestingly, in the AstraZeneca rare variant gene based PheWAS for essential hypertension, rare variation was linked to primary hypertension in seven previously linked genes from GWAS (*CACNA1D, NR3C2, NOS3, DNMT3A, ENPEP, GUCY1A1, and UMOD*) (Zöller et al. 2023) highlighting the link between common and rare variants. My findings that those patients with early onset hypertension, without a renal cause, have a shared polygenic risk with a primary

hypertension cohort highlights that multiple factors must be at play. Whilst lifestyle and environmental factors pay a huge part in the pathogenesis of hypertension, it is clear that multiple genetic associations also play a role, cementing the likely "omnigenic" architecture of EEHTN.

6.4 Impact and implications

From an immediate clinical perspective, many of the variants in CyKD, USD and one in EEHTN are in patients who are unsolved by the clinical arm of the 100KGP. These are in the process of being fed back to the relevant patient's clinicians with a view to potentially offering them a molecular diagnosis. More generally these results will be of interest to researchers in nephrogenetics as well as clinicians involved in rare renal disease. I hope that these results will be hypothesis forming for both *in silico* and functional analyses. For the wider genomics community, I have used a mixed ancestry in nearly all of my analyses without major genomic confounding. This demonstrates the scientific advantages to including a wider cohort for genomic analysis and normalises the representation of individuals from diverse ancestral background. At the start of my thesis, I was only using Europeans for my analyses, which as a researcher not of European ancestry, was odd to me. I am pleased that our group has developed methods to improve representation in rare disease analyses.

Finally, attempting to marry the common and rare variants domains via the analysis of low frequency variants in rare disease has great implications for the future of rare disease genomics. As rare disease cohorts become larger and sequencing improves, we really are at an exciting time to tease out the "missing heritability" of diseases. This will help guide understanding of biology and more importantly offer new avenues for therapeutics for a series of diseases that really lack personalised approaches.

6.5 Future directions

This thesis has generated reams of data that I have attempted to unpack, however, future work and collaboration is required to unpick many of the hypotheses generated. Some of the key questions and their potential approaches are discussed below:

- How does the polygenic risk of CyKD influence the final phenotype? What is the mechanism by which this occurs?
- What are the mechanisms by which monoallelic *COL4A3, PKHD1* and *IFT140* variants cause CyKD?
- Why is there such a disparity in the size of likely disease-causing structural variants between *PKD1* and *PKD2*?
- Are there other undiscovered genes that cause CyKD?
- In USD, what is the mechanism for *SLC34A3*'s contribution to the disease.
- What are the other genes that increase the risk of USD?
- In hypertension, could WGS be used as an adjunct to traditional screening in secondary hypertension?

To answer these questions a combination of *in silico, in vitro* and *in vivo* approaches will be needed. It is beyond the scope of this thesis to give a detailed description of all of the functional assays across the three diseases that could be of use, and I will focus primarily on bioinformatic approaches to further this work.

With the inclusion of CyKD to the NHS Genomics Medicine directory, I will soon have access to an even larger dataset of CyKD patient with WGS data (at least another 1300 cases have been added since I completed my analysis). This will tease out the question of whether rare and structural variants in as yet undetected genes are involved in CyKD and allow for more powerful common variant analysis. I am also confident that with larger numbers, the time-to-event GWAS analysis, especially when stratified by the driving variant, will yield useful findings to understand inter-patient variability in phenotype. Whilst USD and EEHTN are not part of the WGS NHS programme, recent advances in methodology that allow for meta-analysis of rare variant studies look very promising to increase power in these cohorts to further uncover genetic loci within the rare to low frequency allelic frequency (X. Li et al. 2023).

The availability of large-scale proteomic datasets, particularly via the UK Biobanks will also enable me to look for biomarkers or protein signatures that could account for

phenotype severity across all three diseases using protein quantitative trait loci analysis (B. B. Sun et al. 2022; Gadd et al. 2023).

6.6 Lessons Learnt

Working with the 100KGP dataset has at times been challenging. Having worked with the project since its inception the resource has been constantly evolving. At the beginning the WGS data was unfiltered and analysis at scale was challenging, particularly as I my bioinformatic skills were still nascent. SAIGE and SAIGE-GENE were published and established as tools during my PhD and my initial attempts to perform analyses was based on PLINK for GWAS and RVTESTS for rare variant analyses, both of which suffered greatly from high type I error rate with large case-control imbalances (as found in biobanks). Over time, with my growing abilities, help from online bioinformatics communities (I am now a moderator on stackoverflow!) and the 100KGP bioinformaticians the data became cleaner and easier to use and the latest cutting-edge tools were made available.

If I were to begin this project again I would:

1. Use the latest set based rare variant association testing tools such as SAIGE-GENE+ (W. Zhou et al. 2022) which greatly improves type I error rates.
2. Perform joint SNV-SV analysis to examine the burden of these variants across the exome, particularly in CyKD.
3. Use a tool such as parliment2 (Zarate et al. 2020) to call SVs with multiple callers to ensure better call accuracy for variants.
4. Version control my code using tools such as GitHub or Jupiter notebooks.
5. Perform combined rare and common variants meta-analysis with other largescale biobanks to improve power.

6.7 Conclusion

In this thesis, I have used WGS to investigate the genetic architecture of three renal disorders at an unprecedented level using well matched controls. I have demonstrated that variants across the allelic spectrum play a role in all three disorders including novel associations in CyKD and USD. I have shown that WGS is a useful tool beyond monogenic gene discovery and that inclusion of individuals from diverse ancestral backgrounds is possible and adds power to disease loci.

Taken together these should prompt us to rethink what "Mendelian" disorders are. Whilst some diseases tend towards the monogenic and other the polygenic, the more our ability to analyse the genetics of a condition improve, the more we find the definition of "monogenic" blurs. As technology and biobanks improve, I am confident these paradigms will be challenged even further.

Reference List

100000 Genomes Project Pilot Investigators, Damian Smedley, Katherine R. Smith, Antonio Martin, Ellen A. Thomas, Ellen M. McDonagh, Valentina Cipriani, et al. 2021. "100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report." *The New England Journal of Medicine* 385 (20): 1868–80.

Abdellaoui, Abdel, Loic Yengo, Karin J. H. Verweij, and Peter M. Visscher. 2023. "15 Years of GWAS Discovery: Realizing the Promise." *American Journal of Human Genetics* 110 (2): 179–94.

Abel, Haley J., David E. Larson, Allison A. Regier, Colby Chiang, Indraniel Das, Krishna L. Kanchi, Ryan M. Layer, et al. 2020. "Mapping and Characterization of Structural Variation in 17,795 Human Genomes." *Nature* 583 (7814): 83–89.

Abifadel, Marianne, Mathilde Varret, Jean-Pierre Rabès, Delphine Allard, Khadija Ouguerram, Martine Devillers, Corinne Cruaud, et al. 2003. "Mutations in PCSK9 Cause Autosomal Dominant Hypercholesterolemia." *Nature Genetics* 34 (2): 154–56.

"Abstracts from the 2022 Annual Scientific Meeting of the British and Irish Hypertension Society (BIHS)." 2022. *Journal of Human Hypertension* 36 (1): 1–22.

AD Rule, A. E. Krambeck J. C. Lieske. 2011. "Chronic Kidney Disease in Kidney Stone Formers." *Clinical Journal of the American Society of Nephrology: CJASN* 6: 2069–75.

Aigha, Idil I., and Essam M. Abdelalim. 2020. "NKX6.1 Transcription Factor: A Crucial Regulator of Pancreatic β Cell Development, Identity, and Proliferation." *Stem Cell Research & Therapy* 11 (1): 459.

Akbari, Parsa, Ankit Gilani, Olukayode Sosina, Jack A. Kosmicki, Lori Khrimian, Yi-Ya Fang, Trikaldarshi Persaud, et al. 2021. "Sequencing of 640,000 Exomes Identifies GPR75 Variants Associated with Protection from Obesity." *Science (New York, N.Y.)* 373 (6550): eabf8683.

Albrechtsen, A., N. Grarup, Y. Li, T. Sparsø, G. Tian, H. Cao, T. Jiang, et al. 2013. "Exome Sequencing-Driven Discovery of Coding Polymorphisms Associated with Common Metabolic Phenotypes." *Diabetologia* 56 (2): 298–310.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101.

Ali, Hamad, Fahd Al-Mulla, Naser Hussain, Medhat Naim, Akram M. Asbeutah, Ali AlSahow, Mohamed Abu-Farha, et al. 2019. "PKD1 Duplicated Regions Limit Clinical Utility of Whole Exome Sequencing for Genetic Diagnosis of Autosomal Dominant Polycystic Kidney Disease." *Scientific Reports* 9 (1): 4141.

Antonelli, Jodi A., Naim M. Maalouf, Margaret S. Pearle, and Yair Lotan. 2014. "Use of the National Health and Nutrition Examination Survey to Calculate the Impact of Obesity and Diabetes on Cost and Prevalence of Urolithiasis in 2030." *European Urology* 66 (4): 724–29.

Apple, Benjamin, Gino Sartori, Bryn Moore, Kiran Chintam, Gurmukteshwar Singh, Prince Mohan Anand, Natasha T. Strande, Tooraj Mirshahi, William Triffo, and Alexander R. Chang. 2023. "Individuals Heterozygous for ALG8 Protein-

Truncating Variants Are at Increased Risk of a Mild Cystic Kidney Disease." *Kidney International* 103 (3): 607–15.

Attanasio, Massimo, N. Henriette Uhlenhaut, Vitor H. Sousa, John F. O'Toole, Edgar Otto, Katrin Anlag, Claudia Klugmann, et al. 2007. "Loss of GLIS2 Causes Nephronophthisis in Humans and Mice by Increased Apoptosis and Fibrosis." *Nature Genetics* 39 (8): 1018–24.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." *Cell* 176 (3): 663-675.e19.

Baboolal, K., D. Ravine, J. Daniels, N. Williams, P. Holmans, G. A. Coles, and J. D. Williams. 1997. "Association of the Angiotensin I Converting Enzyme Gene Deletion Polymorphism with Early Onset of ESRF in PKD1 Adult Polycystic Kidney Disease." *Kidney International* 52 (3): 607–13.

Badano, Jose L., and Nicholas Katsanis. 2002. "Beyond Mendel: An Evolving View of Human Genetic Disease Transmission." *Nature Reviews. Genetics* 3 (10): 779–89.

Bailey, Jeffrey A., Zhiping Gu, Royden A. Clark, Knut Reinert, Rhea V. Samonte, Stuart Schwartz, Mark D. Adams, Eugene W. Myers, Peter W. Li, and Evan E. Eichler. 2002. "Recent Segmental Duplications in the Human Genome." *Science* 297 (5583): 1003–7.

Bao, Minghui, Ping Li, Qifu Li, Hui Chen, Ying Zhong, Shuangyue Li, Ling Jin, et al. 2020. "Genetic Screening for Monogenic Hypertension in Hypertensive Individuals in a Clinical Setting." *Journal of Medical Genetics* 57 (8): 571–80.

Barker, D. J., C. Osmond, J. Golding, D. Kuh, and M. E. Wadsworth. 1989. "Growth in Utero, Blood Pressure in Childhood and Adult Life, and Mortality from Cardiovascular Disease." *BMJ* 298 (6673): 564–67.

Belkadi, Aziz, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B. Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. 2015. "Whole-Genome Sequencing Is More Powerful than Whole-Exome Sequencing for Detecting Exome Variants." *Proceedings of the National Academy of Sciences of the United States of America* 112 (17): 5473–78.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bergmann, Carsten, Lisa M. Guay-Woodford, Peter C. Harris, Shigeo Horie, Dorien J. M. Peters, and Vicente E. Torres. 2018. "Polycystic Kidney Disease." *Nature Reviews. Disease Primers* 4 (1): 50.

Bergwitz, Clemens, Nicole M. Roslin, Martin Tieder, J. C. Loredo-Osti, Murat Bastepe, Hilal Abu-Zahra, Danielle Frappier, et al. 2006. "SLC34A3 Mutations in Patients with Hereditary Hypophosphatemic Rickets with Hypercalciuria Predict a Key Role for the Sodium-Phosphate Cotransporter NaPi-IIc in Maintaining Phosphate Homeostasis." *American Journal of Human Genetics* 78 (2): 179–92.

Besse, Whitney, Alex R. Chang, Jonathan Z. Luo, William J. Triffo, Bryn S. Moore, Ashima Gulati, Dustin N. Hartzel, et al. 2019. "ALG9 Mutation Carriers Develop Kidney and Liver Cysts." *Journal of the American Society of Nephrology: JASN* 30 (11): 2091–2102.

Besse, Whitney, Ke Dong, Jungmin Choi, Sohan Punia, Sorin V. Fedeles, Murim Choi, Anna-Rachel Gallagher, et al. 2017. “Isolated Polycystic Liver Disease Genes Define Effectors of Polycystin-1 Function.” *The Journal of Clinical Investigation* 127 (5): 1772–85.

Blair, David R., Thomas J. Hoffmann, and Joseph T. Shieh. 2022. “Common Genetic Variation Associated with Mendelian Disease Severity Revealed through Cryptic Phenotype Analysis.” *Nature Communications* 13 (1): 3675.

Blake, Judith A., Joel E. Richardson, Carol J. Bult, Jim A. Kadin, Janan T. Eppig, and Mouse Genome Database Group. 2003. “MGD: The Mouse Genome Database.” *Nucleic Acids Research* 31 (1): 193–95.

Boca, Manila, Gianfranco Distefano, Feng Qian, Anil K. Bhunia, Gregory G. Germino, and Alessandra Boletta. 2006. “Polycystin-1 Induces Resistance to Apoptosis through the Phosphatidylinositol 3-Kinase/Akt Signaling Pathway.” *Journal of the American Society of Nephrology: JASN* 17 (3): 637–47.

Bocher, Ozvan, and Emmanuelle Génin. 2020. “Rare Variant Association Testing in the Non-Coding Genome.” *Human Genetics* 139 (11): 1345–62.

Bockenhauer, Detlef, Alan J. Medlar, Emma Ashton, Robert Kleta, and Nick Lench. 2012. “Genetic Testing in Renal Disease.” *Pediatric Nephrology* 27 (6): 873–83.

“Bootstrap Functions (Originally by Angelo Canty for S) [R Package Boot Version 1.3-28.1].” 2022, November. https://cran.r-project.org/web/packages/boot/index.html.

Borràs, Daniel M., Rolf H. A. M. Vossen, Michael Liem, Henk P. J. Buermans, Hans Dauwerse, Dave van Heusden, Ron T. Gansevoort, et al. 2017. “Detecting PKD1 Variants in Polycystic Kidney Disease Patients by Single-Molecule Long-Read Sequencing.” *Human Mutation* 38 (7): 870–79.

Botstein, David, and Neil Risch. 2003. “Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease.” *Nature Genetics* 33 Suppl (March): 228–37.

Bouaziz, Matthieu, Jimmy Mullaert, Benedetta Bigio, Yoann Seeleuthner, Jean-Laurent Casanova, Alexandre Alcais, Laurent Abel, and Aurélie Cobat. 2021. “Controlling for Human Population Stratification in Rare Variant Association Studies.” *Scientific Reports* 11 (1): 19015.

Boycott, Kym M., Megan R. Vanstone, Dennis E. Bulman, and Alex E. MacKenzie. 2013. “Rare-Disease Genetics in the Era of next-Generation Sequencing: Discovery to Translation.” *Nature Reviews. Genetics* 14 (10): 681–91.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* 169 (7): 1177–86.

Brasier, J. L., and E. P. Henske. 1997. “Loss of the Polycystic Kidney Disease (PKD1) Region of Chromosome 16p13 in Renal Cyst Cells Supports a Loss-of-Function Model for Cyst Pathogenesis.” *The Journal of Clinical Investigation* 99 (2): 194–99.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3): 291–95.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9.

Cabezas, Oscar Rubio, Sarah E. Flanagan, Horia Stanescu, Elena García-Martínez, Richard Caswell, Hana Lango-Allen, Montserrat Antón-Gamero, et al. 2017. “Polycystic Kidney Disease with Hyperinsulinemic Hypoglycemia Caused by a Promoter Mutation in Phosphomannomutase 2.” *Journal of the American Society of Nephrology: JASN* 28 (8): 2529–39.

Cai, Yiqiang, Sorin V. Fedeles, Ke Dong, Georgia Anyatonwu, Tamehito Onoe, Michihiro Mitobe, Jian-Dong Gao, et al. 2014. “Altered Trafficking and Stability of Polycystins Underlie Polycystic Kidney Disease.” *The Journal of Clinical Investigation* 124 (12): 5129–44.

Cameron, Daniel L., Leon Di Stefano, and Anthony T. Papenfuss. 2019. “Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software.” *Nature Communications 2019 10:1* 10 (1): 1–11.

Carvalho, Claudia M. B., and James R. Lupski. 2016. “Mechanisms Underlying Structural Variant Formation in Genomic Disorders.” *Nature Reviews. Genetics* 17 (4): 224–38.

Carvalho, Claudia M. B., Davut Pehlivan, Melissa B. Ramocki, Ping Fang, Benjamin Alleva, Luis M. Franco, John W. Belmont, P. J. Hastings, and James R. Lupski. 2013. “Replicative Mechanisms for CNV Formation Are Error Prone.” *Nature Genetics* 45 (11): 1319–26.

Castro, J. M. 1993. “A Twin Study of Genetic and Environmental Influences on the Intake of Fluids and Beverages.” *Physiology & Behavior* 54: 677–87.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. “Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes.” *Nature Communications* 10 (1): 1784.

Chan, Melanie M. Y., Omid Sadeghi-Alavijeh, Filipa M. Lopes, Alina C. Hilger, Horia C. Stanescu, Catalin D. Voinescu, Glenda M. Beaman, et al. 2022. “Diverse Ancestry Whole-Genome Sequencing Association Study Identifies TBX5 and PTK7 as Susceptibility Genes for Posterior Urethral Valves.” *ELife* 11 (September). https://doi.org/10.7554/eLife.74777.

Chang, Alexander R., Bryn S. Moore, Jonathan Z. Luo, Gino Sartori, Brian Fang, Steven Jacobs, Yoosif Abdalla, et al. 2022. “Exome Sequencing of a Clinical Population for Autosomal Dominant Polycystic Kidney Disease.” *JAMA: The Journal of the American Medical Association* 328 (24): 2412–21.

Chang, Christopher C., Carson C. Chow, Laurent C. A. M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1): 7.

Chen, Xiaoli, and Youfa Wang. 2008. “Tracking of Blood Pressure from Childhood to Adulthood: A Systematic Review and Meta-Regression Analysis.” *Circulation* 117 (25): 3171–80.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. “Manta: Rapid Detection of Structural Variants and Indels for

Germline and Cancer Sequencing Applications." *Bioinformatics* 32 (8): 1220–22.
Chhabra, Roohi, Reecha Sofat, Aroon Hingorani, and Jennifer Cross. 2022. "Approach to Hypertension: Diagnosis and Investigation." In *Primer on Nephrology*, edited by Mark Harber, 317–33. Cham: Springer International Publishing.
Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O'Reilly. 2020. "Tutorial: A Guide to Performing Polygenic Risk Score Analyses." *Nature Protocols* 15 (9): 2759–72.
Choi, Shing Wan, and Paul F. O'Reilly. 2019. "PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data." *GigaScience* 8 (7). https://doi.org/10.1093/GIGASCIENCE/GIZ082.
Choi, Yun-Hee, Akira Suzuki, Sachin Hajarnis, Zhendong Ma, Hannah C. Chapin, Michael J. Caplan, Marco Pontoglio, Stefan Somlo, and Peter Igarashi. 2011. "Polycystin-2 and Phosphodiesterase 4C Are Components of a Ciliary A-Kinase Anchoring Protein Complex That Is Disrupted in Cystic Kidney Diseases." *Proceedings of the National Academy of Sciences of the United States of America* 108 (26): 10679–84.
Claes, Kathleen B. M., and Kim De Leeneer. 2014. "Dealing with Pseudogenes in Molecular Diagnostics in the Next-Generation Sequencing Era." *Methods in Molecular Biology* 1167: 303–15.
Claus, Laura R., Rozemarijn Snoek, Nine V. A. M. Knoers, and Albertien M. van Eerde. 2022. "Review of Genetic Testing in Kidney Disease Patients: Diagnostic Yield of Single Nucleotide Variants and Copy Number Variations Evaluated across and within Kidney Phenotype Groups." *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics* 190 (3): 358–76.
Claverie-Martin, Felix, Francisco J. Gonzalez-Paredes, and Elena Ramos-Trujillo. 2015. "Splicing Defects Caused by Exonic Mutations in PKD1 as a New Mechanism of Pathogenesis in Autosomal Dominant Polycystic Kidney Disease." *RNA Biology* 12 (4): 369–74.
Cochran, J. Nicholas, Ethan G. Geier, Luke W. Bonham, J. Scott Newberry, Michelle D. Amaral, Michelle L. Thompson, Brittany N. Lasseigne, et al. 2020. "Non-Coding and Loss-of-Function Coding Variants in TET2 Are Associated with Multiple Neurodegenerative Diseases." *American Journal of Human Genetics* 106 (5): 632–45.
Collins, F. S. 1990. "Identifying Human Disease Genes by Positional Cloning." *Harvey Lectures* 86: 149–64.
Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. "A Structural Variation Reference for Medical and Population Genetics." *Nature* 581 (7809): 444–51.
Cornec-Le Gall, Emilie, Ahsan Alam, and Ronald D. Perrone. 2019. "Autosomal Dominant Polycystic Kidney Disease." *The Lancet* 393 (10174): 919–35.
Cornec-Le Gall, Emilie, Marie-Pierre Audrézet, Annick Rousseau, Maryvonne Hourmant, Eric Renaudineau, Christophe Charasse, Marie-Pascale Morin, et al. 2016. "The PROPKD Score: A New Algorithm to Predict Renal Survival in Autosomal Dominant Polycystic Kidney Disease." *Am Soc Nephrol* 27: 942–51.
Cornec-Le Gall, Emilie, Rory J. Olson, Whitney Besse, Christina M. Heyer, Vladimir G. Gainullin, Jessica M. Smith, Marie Pierre Audrézet, et al. 2018. "Monoallelic

Mutations to DNAJB11 Cause Atypical Autosomal-Dominant Polycystic Kidney Disease." *American Journal of Human Genetics* 102 (5): 832.
Corvol, Harriet, Scott M. Blackman, Pierre-Yves Boëlle, Paul J. Gallins, Rhonda G. Pace, Jaclyn R. Stonebraker, Frank J. Accurso, et al. 2015. "Genome-Wide Association Meta-Analysis Identifies Five Modifier Loci of Lung Disease Severity in Cystic Fibrosis." *Nature Communications* 6 (September): 8382.
Costain, Gregory, Rebekah Jobling, Susan Walker, Miriam S. Reuter, Meaghan Snell, Sarah Bowdin, Ronald D. Cohn, et al. 2018. "Periodic Reanalysis of Whole-Genome Sequencing Data Enhances the Diagnostic Advantage over Standard Clinical Genetic Testing." *European Journal of Human Genetics: EJHG* 26 (5): 740–44.
Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert. 1978. "Molecular Basis of Base Substitution Hotspots in Escherichia Coli." *Nature* 274 (5673): 775–80.
Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society* 34 (2): 187–202.
D. Turner, Stephen. 2018. "Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots." *Journal of Open Source Software* 3 (25): 731.
Daga, Ankana, Amar J. Majmundar, Daniela A. Braun, Heon Yung Gee, Jennifer A. Lawson, Shirlee Shril, Tilman Jobst-Schwan, et al. 2018. "Whole Exome Sequencing Frequently Detects a Monogenic Cause in Early Onset Nephrolithiasis and Nephrocalcinosis." *Kidney International* 93 (1): 204–13.
Dalesio, Nicholas M., Sebastian F. Barreto Ortiz, Jennifer L. Pluznick, and Dan E. Berkowitz. 2018. "Olfactory, Taste, and Photo Sensory Receptors in Non-Sensory Organs: It Just Makes Sense." *Frontiers in Physiology* 9 (November): 1673.
Dandine-Roulland, Claire, and Hervé Perdry. 2015. "The Use of the Linear Mixed Model in Human Genetics." *Human Heredity* 80 (4): 196–206.
Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). https://doi.org/10.1093/GIGASCIENCE/GIAB008.
Dasgupta, Debayan, Mark J. Wee, Monica Reyes, Yuwen Li, Peter J. Simm, Amita Sharma, Karl Peter Schlingmann, et al. 2014. "Mutations in SLC34A3/NPT2c Are Associated with Kidney Stones and Nephrocalcinosis." *Journal of the American Society of Nephrology: JASN* 25 (10): 2366–75.
Deaton, Aimee M., Margaret M. Parker, Lucas D. Ward, Alexander O. Flynn-Carroll, Lucas BonDurant, Gregory Hinkle, Parsa Akbari, et al. 2021. "Gene-Level Analysis of Rare Variants in 379,066 Whole Exome Sequences Identifies an Association of GIGYF1 Loss of Function with Type 2 Diabetes." *Scientific Reports* 11 (1): 21565.
Deininger, Prescott. 2011. "Alu Elements: Know the SINEs." *Genome Biology* 12 (12): 236.
Delling, M., A. A. Indzhykulian, X. Liu, Y. Li, T. Xie, D. P. Corey, and D. E. Clapham. 2016. "Primary Cilia Are Not Calcium-Responsive Mechanosensors." *Nature* 531 (7596): 656–60.
Dennis, Megan Y., and Evan E. Eichler. 2016. "Human Adaptation and Evolution by Segmental Duplication." *Current Opinion in Genetics & Development* 41 (December): 44–52.

Derkach, Andriy, Haoyu Zhang, and Nilanjan Chatterjee. 2018. “Power Analysis for Genetic Association Test (PAGEANT) Provides Insights to Challenges for Rare Variant Association Studies.” *Bioinformatics* 34 (9): 1506–13.

Devlin, B., and K. Roeder. 1999. “Genomic Control for Association Studies.” *Biometrics* 55 (4): 997–1004.

Dewey, Frederick E., Megan E. Grove, Cuiping Pan, Benjamin A. Goldstein, Jonathan A. Bernstein, Hassan Chaib, Jason D. Merker, et al. 2014. “Clinical Interpretation and Implications of Whole-Genome Sequencing.” *JAMA: The Journal of the American Medical Association* 311 (10): 1035–45.

Dey, Rounak, Ellen M. Schmidt, Goncalo R. Abecasis, and Seunggeun Lee. 2017. “A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS.” *American Journal of Human Genetics* 101 (1): 37–49.

Dey, Rounak, Wei Zhou, Tuomo Kiiskinen, Aki Havulinna, Amanda Elliott, Juha Karjalainen, Mitja Kurki, et al. 2022. “Efficient and Accurate Frailty Model Approach for Genome-Wide Survival Association Analysis in Large-Scale Biobanks.” *Nature Communications* 13 (1): 5437.

Dhir, Gauri, Dong Li, Hakon Hakonarson, and Michael A. Levine. 2017. “Late-Onset Hereditary Hypophosphatemic Rickets with Hypercalciuria (HHRH) Due to Mutation of SLC34A3/NPT2c.” *Bone* 97 (April): 15.

DiCorpo, Daniel, Sheila M. Gaynor, Emily M. Russell, Kenneth E. Westerman, Laura M. Raffield, Timothy D. Majarian, Peitao Wu, et al. 2022. “Whole Genome Sequence Association Analysis of Fasting Glucose and Fasting Insulin Levels in Diverse Cohorts from the NHLBI TOPMed Program.” *Communications Biology* 5 (1): 756.

DS Goldfarb, A. R. Avery L. Beara-Lasic G. E. Duncan J. Goldberg. 2019. “A Twin Study of Genetic Influences on Nephrolithiasis in Women and Men.” *Kidney Int. Rep.* 4: 535–40.

DS Goldfarb, M. E. Fischer Y. Keich J. Goldberg. 2005. “A Twin Study of Genetic and Dietary Influences on Nephrolithiasis: A Report from the Vietnam Era Twin (VET) Registry.” *Kidney International* 67: 1053–61.

Dubois-Laforgue, Danièle, Erika Cornu, Cécile Saint-Martin, Joël Coste, Christine Bellanné-Chantelot, José Timsit, and Monogenic Diabetes Study Group of the Société Francophone du Diabète. 2017. “Diabetes, Associated Clinical Spectrum, Long-Term Prognosis, and Genotype/Phenotype Correlations in 201 Adult Patients With Hepatocyte Nuclear Factor 1B (HNF1B) Molecular Defects.” *Diabetes Care* 40 (11): 1436–43.

Dudbridge, Frank. 2013. “Power and Predictive Accuracy of Polygenic Risk Scores.” *PLoS Genetics* 9 (3): e1003348.

Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. 2019. “Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations.” *Nature Communications* 10 (1): 3328.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Durkie, Miranda, Jiehan Chong, Manoj K. Valluru, Peter C. Harris, and Albert C. M. Ong. 2021. “Biallelic Inheritance of Hypomorphic PKD1 Variants Is Highly Prevalent in Very Early Onset Polycystic Kidney Disease.” *Genetics in*

Medicine: Official Journal of the American College of Medical Genetics 23 (4): 689–97.

Dutta, Diptavo, Peter VandeHaar, Lars G. Fritsche, Sebastian Zöllner, Michael Boehnke, Laura J. Scott, and Seunggeun Lee. 2021. "A Powerful Subset-Based Method Identifies Gene Set Associations and Improves Interpretation in UK Biobank." *The American Journal of Human Genetics* 108 (4): 669–81.

Ecder, Tevfik, and Robert W. Schrier. 2001. "Hypertension in Autosomal-Dominant Polycystic Kidney Disease: Early Occurrence and Unique Aspects." *Journal of the American Society of Nephrology: JASN* 12 (1): 194–200.

Edghill, Emma L., Richard A. Oram, Martina Owens, Karen L. Stals, Lorna W. Harries, Andrew T. Hattersley, Sian Ellard, and Coralie Bingham. 2007. "Hepatocyte Nuclear Factor-1β Gene Deletions—a Common Cause of Renal Disease." *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 23 (2): 627–35.

Edvardsson, Vidar O., Olafur S. Indridason, Gudjon Haraldsson, Olafur Kjartansson, and Runolfur Palsson. 2013. "Temporal Trends in the Incidence of Kidney Stone Disease." *Kidney International* 83 (1): 146–52.

Egbuna, Ogo, Brandon Zimmerman, George Manos, Anne Fortier, Madalina C. Chirieac, Leslie A. Dakin, David J. Friedman, et al. 2023. "Inaxaplin for Proteinuric Kidney Disease in Persons with Two APOL1 Variants." *The New England Journal of Medicine* 388 (11): 969–79.

Evangelou, Evangelos, and John P. A. Ioannidis. 2013. "Meta-Analysis Methods for Genome-Wide Association Studies and Beyond." *Nature Reviews. Genetics* 14 (6): 379–89.

Evans, Katharine, Rhodri Pyart, Retha Steenkamp, Tim Whitlock, Catherine Stannard, Rachel Gair, James Mccann, Julie Slevin, James Medcalf, and Fergus Caskey. n.d. "UK Renal Registry 20th Annual Report: Introduction." *Karger.Com*. https://doi.org/10.1159/000490958.

Evans, Luke M., Rasool Tahmasbi, Scott I. Vrieze, Gonçalo R. Abecasis, Sayantan Das, Steven Gazal, Douglas W. Bjelland, et al. 2018. "Comparison of Methods That Use Whole Genome Data to Estimate the Heritability and Genetic Architecture of Complex Traits." *Nature Genetics* 50 (5): 737–45.

Ezzati, Majid, Alan D. Lopez, Anthony Rodgers, Stephen Vander Hoorn, Christopher J. L. Murray, and Comparative Risk Assessment Collaborating Group. 2002. "Selected Major Risk Factors and Global and Regional Burden of Disease." *The Lancet* 360 (9343): 1347–60.

Fahed, Akl C., Minxian Wang, Julian R. Homburger, Aniruddh P. Patel, Alexander G. Bick, Cynthia L. Neben, Carmen Lai, et al. 2020. "Polygenic Background Modifies Penetrance of Monogenic Variants for Tier 1 Genomic Conditions." *Nature Communications* 11 (1): 3635.

Fain, Pamela R., Kimberly K. McFann, Matthew R. G. Taylor, Maryellyn Tison, Ann M. Johnson, Berenice Reed, and Robert W. Schrier. 2005. "Modifier Genes Play a Significant Role in the Phenotypic Expression of PKD1." *Kidney International* 67 (4): 1256–67.

Fallerini, C., L. Dosa, R. Tita, D. Del Prete, S. Feriozzi, G. Gai, M. Clementi, et al. 2014. "Unbiased next Generation Sequencing Analysis Confirms the Existence

of Autosomal Dominant Alport Syndrome in a Relevant Fraction of Cases." *Clinical Genetics* 86 (3): 252–57.
Fatumo, Segun, Tinashe Chikowore, Ananyo Choudhury, Muhammad Ayub, Alicia R. Martin, and Karoline Kuchenbaecker. 2022. "A Roadmap to Increase Diversity in Genomic Studies." *Nature Medicine* 28 (2): 243–50.
Fedeles, Sorin V., Xin Tian, Anna-Rachel Gallagher, Michihiro Mitobe, Saori Nishio, Seung Hun Lee, Yiqiang Cai, Lin Geng, Craig M. Crews, and Stefan Somlo. 2011. "A Genetic Interaction Network of Five Genes for Human Polycystic Kidney and Liver Diseases Defines Polycystin-1 as the Central Determinant of Cyst Formation." *Nature Genetics* 43 (7): 639–47.
Fernando, Rohan L., and Dorian Garrick. 2013. "Bayesian Methods Applied to GWAS." *Methods in Molecular Biology* 1019: 237–74.
Ferrè, Silvia, and Peter Igarashi. 2019. "New Insights into the Role of HNF-1β in Kidney (Patho)Physiology." *Pediatric Nephrology* 34 (8): 1325–35.
Findlay, Gregory M., Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 2018. "Accurate Classification of BRCA1 Variants with Saturation Genome Editing." *Nature* 562 (7726): 217–22.
Fischer, Dagmar-Christiane, Ulrike Jacoby, Lars Pape, Christopher J. Ward, Eberhard Kuwertz-Broeking, Catharina Renken, Horst Nizze, et al. 2009. "Activation of the AKT/MTOR Pathway in Autosomal Recessive Polycystic Kidney Disease (ARPKD)." *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 24 (6): 1819–27.
Fisher, R. A. 1919. "XV.—the Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52 (2): 399–433.
Fokkema, Ivo F. A. C., Peter E. M. Taschner, Gerard C. P. Schaafsma, J. Celli, Jeroen F. J. Laros, and Johan T. den Dunnen. 2011. "LOVD v.2.0: The next Generation in Gene Variant Databases." *Human Mutation* 32 (5): 557–63.
Follit, John A., Lixia Li, Yvonne Vucica, and Gregory J. Pazour. 2010. "The Cytoplasmic Tail of Fibrocystin Contains a Ciliary Targeting Sequence." *The Journal of Cell Biology* 188 (1): 21–28.
French, J. D., and S. L. Edwards. 2020. "The Role of Noncoding Variants in Heritable Disease." *Trends in Genetics: TIG* 36 (11): 880–91.
Fujimaru, T., T. Mori, A. Sekine, S. Mandai, M. Chiga, H. Kikuchi, F. Ando, et al. 2018. "Kidney Enlargement and Multiple Liver Cyst Formation Implicate Mutations in PKD1/2 in Adult Sporadic Polycystic Kidney Disease." *Clinical Genetics* 94 (1): 125–31.
Gadd, Danni A., Robert F. Hillary, Zhana Kuncheva, Tasos Mangelis, Yipeng Cheng, Manju Dissanayake, Romi Admanit, et al. 2023. "Blood Protein Levels Predict Leading Incident Diseases and Mortality in UK Biobank." *BioRxiv*. https://doi.org/10.1101/2023.05.01.23288879.
Gale, Daniel P., Andrew Mallett, Chirag Patel, Tam P. Sneddon, Heidi L. Rehm, Matthew G. Sampson, and Detlef Bockenhauer. 2020. "Diagnoses of Uncertain Significance: Kidney Genetics in the 21st Century." *Nature Reviews. Nephrology* 16 (11): 616–18.

Gallagher, Anna Rachel, Gregory G. Germino, and Stefan Somlo. 2010. “Molecular Advances in Autosomal Dominant Polycystic Kidney Disease.” *Advances in Chronic Kidney Disease* 17 (2): 118–30.

Garrelfs, Sander F., Yaacov Frishberg, Sally A. Hulton, Michael J. Koren, William D. O’Riordan, Pierre Cochat, Georges Deschênes, et al. 2021. “Lumasiran, an RNAi Therapeutic for Primary Hyperoxaluria Type 1.” *The New England Journal of Medicine* 384 (13): 1216–26.

Gazal, Steven, Po-Ru Loh, Hilary K. Finucane, Andrea Ganna, Armin Schoech, Shamil Sunyaev, and Alkes L. Price. 2018. “Functional Architecture of Low-Frequency Variants Highlights Strength of Negative Selection across Coding and Non-Coding Annotations.” *Nature Genetics* 50 (11): 1600–1607.

Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium. Electronic address: gusella@helix.mgh.harvard.edu, and Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium. 2019. “CAG Repeat Not Polyglutamine Length Determines Timing of Huntington’s Disease Onset.” *Cell* 178 (4): 887-900.e14.

Geng, L., Y. Segal, A. Pavlova, E. J. Barros, C. Löhning, W. Lu, S. K. Nigam, A. M. Frischauf, S. T. Reeders, and J. Zhou. 1997. “Distribution and Developmentally Regulated Expression of Murine Polycystin.” *The American Journal of Physiology* 272 (4 Pt 2): F451-9.

GenomeAsia100K Consortium. 2019. “The GenomeAsia 100K Project Enables Genetic Discoveries across Asia.” *Nature* 576 (7785): 106–11.

Geraghty, Robert M., Paul Cook, Valerie Walker, and Bhaskar K. Somani. 2020. “Evaluation of the Economic Burden of Kidney Stone Disease in the UK: A Retrospective Cohort Study with a Mean Follow-up of 19 Years.” *BJU International* 125 (4): 586–94.

Gibson, J., R. Fieldhouse, M. M. Y. Chan, O. Sadeghi-Alavijeh, L. Burnett, V. Izzi, A. V. Persikov, D. P. Gale, H. Storey, and J. Savige. 2021. “Prevalence Estimates of Predicted Pathogenic Col4a3-Col4a5 Variants in a Population Sequencing Database and Their Implications for Alport Syndrome.” *Journal of the American Society of Nephrology: JASN* 32 (9). https://doi.org/10.1681/ASN.2020071065.

Gilissen, Christian, Jayne Y. Hehir-Kwa, Djie Tjwan Thung, Maartje van de Vorst, Bregje W. M. van Bon, Marjolein H. Willemsen, Michael Kwint, et al. 2014. “Genome Sequencing Identifies Major Causes of Severe Intellectual Disability.” *Nature* 511 (7509): 344–47.

Gong, Yimei, Zhendong Ma, Vishal Patel, Evelyne Fischer, Thomas Hiesberger, Marco Pontoglio, and Peter Igarashi. 2009. “HNF-1beta Regulates Transcription of the PKD Modifier Gene Kif12.” *Journal of the American Society of Nephrology: JASN* 20 (1): 41–47.

Grantham, J. J. 1996. “The Etiology, Pathogenesis, and Treatment of Autosomal Dominant Polycystic Kidney Disease: Recent Advances.” *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 28 (6): 788–803.

Gresh, Lionel, Evelyne Fischer, Andreas Reimann, Myriam Tanguy, Serge Garbay, Xinli Shao, Thomas Hiesberger, et al. 2004. “A Transcriptional Network in Polycystic Kidney Disease.” *The EMBO Journal* 23 (7): 1657–68.

Groopman, Emily E., Maddalena Marasa, Sophia Cameron-Christie, Slavé Petrovski, Vimla S. Aggarwal, Hila Milo-Rasouly, Yifu Li, et al. 2019. “Diagnostic Utility

of Exome Sequencing for Kidney Disease." *The New England Journal of Medicine* 380 (2): 142–51.
Groza, Tudor, Sebastian Köhler, Dawid Moldenhauer, Nicole Vasilevsky, Gareth Baynam, Tomasz Zemojtel, Lynn Marie Schriml, et al. 2015. "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease." *The American Journal of Human Genetics* 97 (1): 111–24.
Grünfeld, Jean-Pierre, and Bernard C. Rossier. 2009. "Lithium Nephrotoxicity Revisited." *Nature Reviews. Nephrology* 5 (5): 270–76.
Gudbjartsson, Daniel F., Hannes Helgason, Sigurjon A. Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, et al. 2015. "Large-Scale Whole-Genome Sequencing of the Icelandic Population." *Nature Genetics* 47 (5): 435–44.
Gulati, Ashima, Angel M. Sevillano, Manuel Praga, Eduardo Gutierrez, Ignacio Alba, Neera K. Dahl, Whitney Besse, Jungmin Choi, and Stefan Somlo. 2020. "Collagen IV Gene Mutations in Adults With Bilateral Renal Cysts and CKD." *Kidney International Reports* 5 (1): 103–8.
Gunay-Aygun, Meral, Baris I. Turkbey, Joy Bryant, Kailash T. Daryanani, Maya Tuchman Gerstein, Katie Piwnica-Worms, Peter Choyke, Theo Heller, and William A. Gahl. 2011. "Hepatorenal Findings in Obligate Heterozygotes for Autosomal Recessive Polycystic Kidney Disease." *Molecular Genetics and Metabolism* 104 (4): 677–81.
H3Africa Consortium, Charles Rotimi, Akin Abayomi, Alash'le Abimiku, Victoria May Adabayeri, Clement Adebamowo, Ezekiel Adebiyi, et al. 2014. "Research Capacity. Enabling the Genomic Revolution in Africa." *Science (New York, N.Y.)* 344 (6190): 1346–48.
Ha, Kotdaji, Mai Nobuhara, Qinzhe Wang, Rebecca V. Walker, Feng Qian, Christoph Schartner, Erhu Cao, and Markus Delling. 2020. "The Heteromeric PC-1/PC-2 Polycystin Complex Is Activated by the PC-1 N-Terminus." *ELife* 9 (November). https://doi.org/10.7554/eLife.60684.
Haffner, Dieter, Francesco Emma, Deborah M. Eastwood, Martin Biosse Duplan, Justine Bacchetta, Dirk Schnabel, Philippe Wicart, et al. 2019. "Clinical Practice Recommendations for the Diagnosis and Management of X-Linked Hypophosphataemia." *Nature Reviews. Nephrology* 15 (7): 435–55.
Halbritter, Jan, Michelle Baum, Ann Marie Hynes, Sarah J. Rice, David T. Thwaites, Zoran S. Gucev, Brittany Fisher, et al. 2015. "Fourteen Monogenic Genes Account for 15% of Nephrolithiasis/Nephrocalcinosis." *Am Soc Nephrol* 26: 543–51.
Halldorsson, Bjarni V., Hannes P. Eggertsson, Kristjan H. S. Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O. Ulfarsson, Gunnar Palsson, et al. 2022. "The Sequences of 150,119 Genomes in the UK Biobank." *Nature* 607 (7920): 732–40.
Hamzaoui, Mouad, Deborah Groussard, Dorian Nezam, Zoubir Djerada, Gaspard Lamy, Virginie Tardif, Anais Dumesnil, et al. 2022. "Endothelium-Specific Deficiency of Polycystin-1 Promotes Hypertension and Cardiovascular Disorders." *Hypertension* 79 (11): 2542–51.
Harris, Peter C. 2010. "What Is the Role of Somatic Mutation in Autosomal Dominant Polycystic Kidney Disease?" *Journal of the American Society of Nephrology: JASN* 21 (7): 1073–76.

Harris, Peter C., and Sandro Rossetti. 2010. “Determinants of Renal Disease Variability in ADPKD.” *Advances in Chronic Kidney Disease* 17 (2): 131–39.
Harris, Peter C., and Vicente E. Torres. 2014. “Genetic Mechanisms and Signaling Pathways in Autosomal Dominant Polycystic Kidney Disease.” *The Journal of Clinical Investigation* 124 (6): 2315–24.
Heather, James M., and Benjamin Chain. 2016. “The Sequence of Sequencers: The History of Sequencing DNA.” *Genomics* 107 (1): 1–8.
Hehir-Kwa, Jayne Y., Rolph Pfundt, and Joris A. Veltman. 2015. “Exome Sequencing and Whole Genome Sequencing for the Detection of Copy Number Variation.” *Expert Review of Molecular Diagnostics* 15 (8): 1023–32.
Heinzl, Harald, Thomas Waldhör, and Martina Mittlböck. 2005. “Careful Use of Pseudo R-Squared Measures in Epidemiological Studies.” *Statistics in Medicine* 24 (18): 2867–72.
Hemminki, Kari, Otto Hemminki, Kristina Sundquist, Jan Sundquist, and Xinjun Li. 2017. “Familial Risks in Urolithiasis in the Population of Sweden.” *Wiley Online Library* 121 (3): 479–85.
Hiesberger, Thomas, Yun Bai, Xinli Shao, Brian T. McNally, Angus M. Sinclair, Xin Tian, Stefan Somlo, and Peter Igarashi. 3 2004. “Mutation of Hepatocyte Nuclear Factor–1β Inhibits Pkhd1 Gene Expression and Produces Renal Cysts in Mice.” *The Journal of Clinical Investigation* 113 (6): 814–25.
Hiesberger, Thomas, Eric Gourley, Andrea Erickson, Peter Koulen, Christopher J. Ward, Tatyana V. Masyuk, Nicholas F. Larusso, Peter C. Harris, and Peter Igarashi. 2006. “Proteolytic Cleavage and Nuclear Translocation of Fibrocystin Is Regulated by Intracellular Ca2+ and Activation of Protein Kinase C.” *The Journal of Biological Chemistry* 281 (45): 34357–64.
Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research* 34 (Database issue). https://doi.org/10.1093/NAR/GKJ144.
Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. 2020. “Structural Variation in the Sequencing Era.” *Nature Reviews. Genetics* 21 (3): 171–89.
Hofmeister, Robin J., Diogo M. Ribeiro, Simone Rubinacci, and Olivier Delaneau. 2023. “Accurate Rare Variant Phasing of Whole-Genome and Whole-Exome Sequencing Data in the UK Biobank.” *Nature Genetics* 55 (7): 1243–49.
Hopp, Katharina, Christopher J. Ward, Cynthia J. Hommerding, Samih H. Nasr, Han-Fang Tuan, Vladimir G. Gainullin, Sandro Rossetti, Vicente E. Torres, and Peter C. Harris. 2012. “Functional Polycystin-1 Dosage Governs Autosomal Dominant Polycystic Kidney Disease Severity.” *The Journal of Clinical Investigation* 122 (11): 4257–73.
Horikawa, Yukio, Naoko Iwasaki, Manami Hara, Hiroto Furuta, Yoshinori Hinokio, Brian N. Cockburn, Tom Lindner, et al. 1997. “Mutation in Hepatocyte Nuclear Factor–1β Gene (TCF2) Associated with MODY.” *Nature Genetics* 17 (4): 384–85.
Hort, Yvonne, Patricia Sullivan, Laura Wedd, Lindsay Fowles, Igor Stevanovski, Ira Deveson, Cas Simons, et al. 2023. “Atypical Splicing Variants in PKD1 Explain Most Undiagnosed Typical Familial ADPKD.” *NPJ Genomic Medicine* 8 (1): 16.

Hougaard, P. 1995. “Frailty Models for Survival Data.” *Lifetime Data Analysis* 1 (3): 255–73.
Howles, Sarah A., Akira Wiberg, Michelle Goldsworthy, Asha L. Bayliss, Anna K. Gluck, Michael Ng, Emily Grout, et al. 2019. “Genetic Variants of Calcium and Vitamin D Metabolism in Kidney Stone Disease.” *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13145-x.
Hu, Yao, Adrienne M. Stilp, Caitlin P. McHugh, Shuquan Rao, Deepti Jain, Xiuwen Zheng, John Lane, et al. 2021. “Whole-Genome Sequencing Association Analysis of Quantitative Red Blood Cell Phenotypes: The NHLBI TOPMed Program.” *American Journal of Human Genetics* 108 (6): 1165.
Hughes, Jim, Christopher J. Ward, Belén Peral, Richard Aspinwall, Kevin Clark, José L. San Millán, Vicki Gamble, and Peter C. Harris. 1995. “The Polycystic Kidney Disease 1 (PKD1) Gene Encodes a Novel Protein with Multiple Cell Recognition Domains.” *Nature Genetics* 10 (2): 151–60.
Hwang, Daw-Yang, Gabriel C. Dworschak, Stefan Kohl, Pawaree Saisawat, Asaf Vivante, Alina C. Hilger, Heiko M. Reutter, et al. 2014. “Mutations in 12 Known Dominant Disease-Causing Genes Clarify Many Congenital Anomalies of the Kidney and Urinary Tract.” *Kidney International* 85 (6): 1429–33.
Hwang, Young-Hwan, John Conklin, Winnie Chan, Nicole M. Roslin, Jannel Liu, Ning He, Kairong Wang, et al. 2016. “Refining Genotype-Phenotype Correlation in Autosomal Dominant Polycystic Kidney Disease.” *Am Soc Nephrol* 27: 1861–68.
Ingles, Jodie, and Christopher Semsarian. 2020. “Time to Rethink the Genetic Architecture of Long QT Syndrome.” *Circulation*.
International HapMap 3 Consortium, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, et al. 2010. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467 (7311): 52–58.
International Schizophrenia Consortium, Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O’Donovan, Patrick F. Sullivan, and Pamela Sklar. 2009. “Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder.” *Nature* 460 (7256): 748–52.
Ishigaki, Kazuyoshi, Saori Sakaue, Chikashi Terao, Yang Luo, Kyuto Sonehara, Kensuke Yamaguchi, Tiffany Amariuta, et al. 2022. “Multi-Ancestry Genome-Wide Association Analyses Identify Novel Genetic Mechanisms in Rheumatoid Arthritis.” *Nature Genetics* 54 (11): 1640–51.
Izzi, Claudia, Chiara Dordoni, Laura Econimo, Elisa Delbarba, Francesca Romana Grati, Eva Martin, Cinzia Mazza, et al. 2020. “Variable Expressivity of HNF1B Nephropathy, From Renal Cysts and Diabetes to Medullary Sponge Kidney Through Tubulo-Interstitial Kidney Disease.” *Kidney International Reports* 5 (12): 2341–50.
Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, et al. 2019. “Predicting Splicing from Primary Sequence with Deep Learning.” *Cell* 176 (3): 535-548.e24.
Jeffares, Daniel C., Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J. Sedlazeck. 2017. “Transient Structural Variations Have Strong Effects

on Quantitative Traits and Reproductive Isolation in Fission Yeast." *Nature Communications* 8 (January): 14061.

Jiang, Meiqin, Vivek Reddy Palicharla, Darcie Miller, Sun-Hee Hwang, Hanwen Zhu, Patricia Hixson, Saikat Mukhopadhyay, and Ji Sun. 2023. "Human IFT-A Complex Structures Provide Molecular Insights into Ciliary Transport." *Cell Research* 33 (4): 288–98.

Jiang, Si-Tse, Yuan-Yow Chiou, Ellian Wang, Hsiu-Kuan Lin, Yuan-Ta Lin, Ying-Chih Chi, Chi-Kuang Leo Wang, Ming-Jer Tang, and Hung Li. 2006. "Defining a Link with Autosomal-Dominant Polycystic Kidney Disease in Mice with Congenitally Low Expression of Pkd1." *The American Journal of Pathology* 168 (1): 205–20.

Johnstone, Iain M., and Arthur Yu Lu. 2009. "On Consistency and Sparsity for Principal Components Analysis in High Dimensions." *Journal of the American Statistical Association* 104 (486): 682–93.

Kaimori, Jun-Ya, Yasuyuki Nagasawa, Luis F. Menezes, Miguel A. Garcia-Gonzalez, Jie Deng, Enyu Imai, Luiz F. Onuchic, Lisa M. Guay-Woodford, and Gregory G. Germino. 2007. "Polyductin Undergoes Notch-like Processing and Regulated Release from Primary Cilia." *Human Molecular Genetics* 16 (8): 942–56.

Kamiyoshi, Naohiro, Kandai Nozu, Xue Jun Fu, Naoya Morisada, Yoshimi Nozu, Ming Juan Ye, Aya Imafuku, et al. 2016. "Genetic, Clinical, and Pathologic Backgrounds of Patients with Autosomal Dominant Alport Syndrome." *Clinical Journal of the American Society of Nephrology: CJASN* 11 (8): 1441–49.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43.

Kebschull, Justus M., and Anthony M. Zador. 2015. "Sources of PCR-Induced Distortions in High-Throughput Sequencing Data Sets." *Nucleic Acids Research* 43 (21): e143.

Khan, Atlas, Ning Shang, Jordan G. Nestor, Chunhua Weng, George Hripcsak, Peter C. Harris, Ali G. Gharavi, and Krzysztof Kiryluk. 2023. "Polygenic Risk Affects the Penetrance of Monogenic Kidney Disease." *MedRxiv : The Preprint Server for Health Sciences*, May. https://doi.org/10.1101/2023.05.07.23289614.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations." *Nature Genetics* 50 (9): 1219–24.

Kim, Hyunho, Hangxue Xu, Qin Yao, Weizhe Li, Qiong Huang, Patricia Outeda, Valeriu Cebotaru, et al. 2014. "Ciliary Membrane Proteins Traffic through the Golgi via a Rabep1/GGA1/Arl3-Dependent Mechanism." *Nature Communications* 5 (November): 5482.

Kim, Ingyu, Yulong Fu, Kwokyin Hui, Gilbert Moeckel, Weiyi Mai, Cunxi Li, Dan Liang, et al. 2008. "Fibrocystin/Polyductin Modulates Renal Tubular Formation by Regulating Polycystin-2 Expression and Function." *Journal of the American Society of Nephrology: JASN* 19 (3): 455–68.

King, Kathy, Frances A. Flinter, Vandana Nihalani, and Peter M. Green. 2002. "Unusual Deep Intronic Mutations in the COL4A5 Gene Cause X Linked Alport Syndrome." *Human Genetics* 111 (6): 548–54.

Kinoshita, Moritoshi, Eiji Higashihara, Haruna Kawano, Ryo Higashiyama, Daisuke Koga, Takafumi Fukui, Nobuhisa Gondo, et al. 2016. “Technical Evaluation: Identification of Pathogenic Mutations in PKD1 and PKD2 in Patients with Autosomal Dominant Polycystic Kidney Disease by Next-Generation Sequencing and Use of a Comprehensive New Classification System.” *PloS One* 11 (11): e0166288.

Kirsch, Stefan, Juanjo Pasantes, Andreas Wolf, Nadia Bogdanova, Claudia Münch, Arseni Markoff, Petra Pennekamp, Michael Krawczak, Bernd Dworniczak, and Werner Schempp. 2008. “Chromosomal Evolution of the PKD1 Gene Family in Primates.” *BMC Evolutionary Biology* 8 (September): 263.

Kiryluk, Krzysztof, Elena Sanchez-Rodriguez, Xu-Jie Zhou, Francesca Zanoni, Lili Liu, Nikol Mladkova, Atlas Khan, et al. 2023. “Genome-Wide Association Analyses Define Pathogenic Signaling Pathways and Prioritize Drug Targets for IgA Nephropathy.” *Nature Genetics* 55 (7): 1091–1105.

Kolatsi-Joannou, Maria, Coralie Bingham, Sian Ellard, Michael P. Bulman, Lisa I. S. Allen, Andrew T. Hattersley, and Adrian S. Woolf. 2001. “Hepatocyte Nuclear Factor-1β: A New Kindred with Renal Cysts and Diabetes and Gene Expression in Normal Human Development.” *Journal of the American Society of Nephrology: JASN* 12 (10): 2175.

Kremer, B., P. Goldberg, S. E. Andrew, J. Theilmann, H. Telenius, J. Zeisler, F. Squitieri, B. Lin, A. Bassett, and E. Almqvist. 1994. “A Worldwide Study of the Huntington’s Disease Mutation. The Sensitivity and Specificity of Measuring CAG Repeats.” *The New England Journal of Medicine* 330 (20): 1401–6.

Kurbegovic, Almira, Hyunho Kim, Hangxue Xu, Shengqiang Yu, Julie Cruanès, Robin L. Maser, Alessandra Boletta, Marie Trudel, and Feng Qian. 2014. “Novel Functional Complexity of Polycystin-1 by GPS Cleavage in Vivo: Role in Polycystic Kidney Disease.” *Molecular and Cellular Biology* 34 (17): 3341–53.

Kurki, Mitja I., Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati Donner, Mary P. Reeve, et al. 2022. “FinnGen: Unique Genetic Insights from Combining Isolated Population and National Health Register Data.” *BioRxiv*. https://doi.org/10.1101/2022.03.03.22271360.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Lanktree, Matthew B., Elsa Guiard, Pedram Akbari, Marina Pourafkari, Ioan-Andrei Iliuta, Syed Ahmed, Amirreza Haghighi, et al. 2021. “Patients with Protein-Truncating PKD1 Mutations and Mild ADPKD.” *Clinical Journal of the American Society of Nephrology: CJASN* 16 (3): 374–83.

Lanktree, Matthew B., Elsa Guiard, Weili Li, Pedram Akbari, Amirreza Haghighi, Ioan-Andrei Iliuta, Belili Shi, et al. 2019. “Intrafamilial Variability of ADPKD.” *Kidney International Reports* 4 (7): 995–1003.

Lanktree, Matthew B., Amirreza Haghighi, Ighli di Bari, Xuewen Song, and York Pei. 2021. “Insights into Autosomal Dominant Polycystic Kidney Disease from Genetic Studies.” *Clinical Journal of the American Society of Nephrology: CJASN* 16 (5): 790–99.

Lantinga-van Leeuwen, Irma S., Johannes G. Dauwerse, Hans J. Baelde, Wouter N. Leonhard, Annemieke van de Wal, Christopher J. Ward, Sjef Verbeek, et al.

2004. “Lowering of Pkd1 Expression Is Sufficient to Cause Polycystic Kidney Disease.” *Human Molecular Genetics* 13 (24): 3069–77.
Lappalainen, Tuuli, Alexandra J. Scott, Margot Brandt, and Ira M. Hall. 2019. “Genomic Analysis in the Age of Human Genome Sequencing.” *Cell* 177 (1): 70–84.
Lata, Sneh, Maddalena Marasa, Yifu Li, David A. Fasel, Emily Groopman, Vaidehi Jobanputra, Hila Rasouly, et al. 2018. “Whole-Exome Sequencing in Adults with Chronic Kidney Disease: A Pilot Study.” *Annals of Internal Medicine* 168 (2): 100–109.
Lee, Sang Hong, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. 2011. “Estimating Missing Heritability for Disease from Genome-Wide Association Studies.” *American Journal of Human Genetics* 88 (3): 294–305.
Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. 2012. “Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies.” *American Journal of Human Genetics* 91 (2): 224–37.
Lee, Seunggeun, Christian Fuchsberger, Sehee Kim, and Laura Scott. 2016. “An Efficient Resampling Method for Calibrating Single and Gene-Based Rare Variant Association Analysis in Case-Control Studies.” *Biostatistics* 17 (1): 1–15.
Lee, Seunggeung, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. “Rare-Variant Association Analysis: Study Designs and Statistical Tests.” *American Journal of Human Genetics* 95 (1): 5.
Legué, Emilie, and Karel F. Liem Jr. 2019. “Tulp3 Is a Ciliary Trafficking Gene That Regulates Polycystic Kidney Disease.” *Current Biology: CB* 29 (5): 803-812.e5.
Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91.
Lelieveld, Stefan H., Malte Spielmann, Stefan Mundlos, Joris A. Veltman, and Christian Gilissen. 2015. “Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions.” *Human Mutation* 36 (8): 815–22.
Lemoine, Hugo, Loann Raud, François Foulquier, John A. Sayer, Baptiste Lambert, Eric Olinger, Siriane Lefèvre, et al. 2022. “Monoallelic Pathogenic ALG5 Variants Cause Atypical Polycystic Kidney Disease and Interstitial Fibrosis.” *The American Journal of Human Genetics* 109 (8): 1484–99.
Leonhard, Wouter N., Hester Happe, and Dorien J. M. Peters. 2016. “Variable Cyst Development in Autosomal Dominant Polycystic Kidney Disease: The Biologic Context.” *Journal of the American Society of Nephrology: JASN* 27 (12): 3530–38.
Li, Bingshan, and Suzanne M. Leal. 2008. “Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data.” *American Journal of Human Genetics* 83 (3): 311–21.
Li, Gengxin, and Hongjiang Zhu. 2013. “Genetic Studies: The Linear Mixed Models in Genome-Wide Association Studies.” *The Open Bioinformatics Journal* 7 (Suppl-1, M2): 27–33.

Li, Xihao, Corbin Quick, Hufeng Zhou, Sheila M. Gaynor, Yaowu Liu, Han Chen, Margaret Sunitha Selvaraj, et al. 2023. “Powerful, Scalable and Resource-Efficient Meta-Analysis of Rare Variant Associations in Large Whole Genome Sequencing Studies.” *Nature Genetics* 55 (1): 154–64.

Li, Zilin, Xihao Li, Hufeng Zhou, Sheila M. Gaynor, Margaret Sunitha Selvaraj, Theodore Arapoglou, Corbin Quick, et al. 2022. “A Framework for Detecting Noncoding Rare-Variant Associations of Large-Scale Whole-Genome Sequencing Studies.” *Nature Methods* 19 (12): 1599–1611.

Lin, D. Y., and D. Zeng. 2010. “Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data.” *Genetic Epidemiology* 34 (1): 60–66.

Lionel, Anath C., Gregory Costain, Nasim Monfared, Susan Walker, Miriam S. Reuter, S. Mohsen Hosseini, Bhooma Thiruvahindrapuram, et al. 2018. “Improved Diagnostic Yield Compared with Targeted Gene Sequencing Panels Suggests a Role for Whole-Genome Sequencing as a First-Tier Genetic Test.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (4): 435–43.

Lipner, Ettie M., and David A. Greenberg. 2018. “The Rise and Fall and Rise of Linkage Analysis as a Technique for Finding and Characterizing Inherited Influences on Disease Expression.” *Methods in Molecular Biology* 1706: 381–97.

Liu, Dajiang J., and Suzanne M. Leal. 2010. “A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions.” *PLoS Genetics* 6 (10): e1001156.

Liu, Dajiang J., Gina M. Peloso, Xiaowei Zhan, Oddgeir L. Holmen, Matthew Zawistowski, Shuang Feng, Majid Nikpay, et al. 2014. “Meta-Analysis of Gene-Level Tests for Rare Variant Association.” *Nature Genetics* 46 (2): 200–204.

Liu, Qianying, Dan L. Nicolae, and Lin S. Chen. 2013. “Marbled Inflation from Population Structure in Gene-Based Association Studies with Rare Variants.” *Genetic Epidemiology* 37 (3): 286–92.

Lo, Yi-Fen, Kandai Nozu, Kazumoto Iijima, Takahiro Morishita, Che-Chung Huang, Sung-Sen Yang, Huey-Kang Sytwu, Yu-Wei Fang, Min-Hua Tseng, and Shih-Hua Lin. 2011. “Recurrent Deep Intronic Mutations in the SLC12A3 Gene Responsible for Gitelman’s Syndrome.” *Clinical Journal of the American Society of Nephrology: CJASN* 6 (3): 630–39.

Loghman-Adham, Mahmoud, Carlos E. Soto, Tadashi Inagami, and Lisa Cassis. 2004. “The Intrarenal Renin-Angiotensin System in Autosomal Dominant Polycystic Kidney Disease.” *American Journal of Physiology. Renal Physiology* 287 (4): F775-88.

Lorenz-Depiereux, Bettina, Anna Benet-Pages, Gertrud Eckstein, Yardena Tenenbaum-Rakover, Janine Wagenstaller, Dov Tiosano, Ruth Gershoni-Baruch, et al. 2006. “Hereditary Hypophosphatemic Rickets with Hypercalciuria Is Caused by Mutations in the Sodium-Phosphate Cotransporter Gene SLC34A3.” *American Journal of Human Genetics* 78 (2): 193–201.

Lu, Hao, Maria C. Rondón Galeano, Elisabeth Ott, Geraldine Kaeslin, P. Jaya Kausalya, Carina Kramer, Nadina Ortiz-Brüchle, et al. 2017. “Mutations in DZIP1L,

Which Encodes a Ciliary-Transition-Zone Protein, Cause Autosomal Recessive Polycystic Kidney Disease." *Nature Genetics* 49 (7): 1025–34.
Lu, Zeyun, Shyamalika Gopalan, Dong Yuan, David V. Conti, Bogdan Pasaniuc, Alexander Gusev, and Nicholas Mancuso. 2022. "Multi-Ancestry Fine-Mapping Improves Precision to Identify Causal Genes in Transcriptome-Wide Association Studies." *American Journal of Human Genetics* 109 (8): 1388–1404.
MacArthur, Daniel G., Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science* 335 (6070): 823–28.
MacArthur, Jacqueline A. L., Timothy D. Spector, Sarah J. Lindsay, Massimo Mangino, Raj Gill, Kerrin S. Small, and Matthew E. Hurles. 2014. "The Rate of Nonallelic Homologous Recombination in Males Is Highly Variable, Correlated between Monozygotic Twins and Independent of Age." *PLoS Genetics* 10 (3): e1004195.
MacKay, Charles E., Miranda Floen, M. Dennis Leo, Raquibul Hasan, Tessa A. C. Garrud, Carlos Fernández-Peña, Purnima Singh, Kafait U. Malik, and Jonathan H. Jaggar. 2022. "A Plasma Membrane-Localized Polycystin-1/Polycystin-2 Complex in Endothelial Cells Elicits Vasodilation." *ELife* 11 (March). https://doi.org/10.7554/eLife.74765.
MacKay, Charles E., M. Dennis Leo, Carlos Fernández-Peña, Raquibul Hasan, Wen Yin, Alejandro Mata-Daboin, Simon Bulley, Jesse Gammons, Salvatore Mancarella, and Jonathan H. Jaggar. 2020. "Intravascular Flow Stimulates PKD2 (Polycystin-2) Channels in Endothelial Cells to Reduce Blood Pressure." *ELife* 9 (May). https://doi.org/10.7554/eLife.56655.
Mackay, T. F. 2001. "The Genetic Architecture of Quantitative Traits." *Annual Review of Genetics* 35: 303–39.
Madsen, Bo Eskerod, and Sharon R. Browning. 2009. "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic." *PLoS Genetics* 5 (2): e1000384.
Mahajan, Anubha, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Grace Z. Yu, et al. 2022. "Multi-Ancestry Genetic Study of Type 2 Diabetes Highlights the Power of Diverse Populations for Discovery and Translation." *Nature Genetics* 54 (5): 560–72.
Mak, Timothy Shin Heng, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. 2017. "Polygenic Scores via Penalized Regression on Summary Statistics." *Genetic Epidemiology* 41 (6): 469–80.
Mallawaarachchi, Amali C., Ben Lundie, Yvonne Hort, Nicole Schonrock, Sarah R. Senum, Velimir Gayevskiy, Andre E. Minoche, et al. 2021. "Genomic Diagnostics in Polycystic Kidney Disease: An Assessment of Real-World Use of Whole-Genome Sequencing." *European Journal of Human Genetics 2021 29:5* 29 (5): 760–70.
Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics (Oxford, England)* 26 (22): 2867–73.
Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53.

Mars, Nina, Jukka T. Koskela, Pietari Ripatti, Tuomo T. J. Kiiskinen, Aki S. Havulinna, Joni V. Lindbohm, Ari Ahola-Olli, et al. 2020. “Polygenic and Clinical Risk Scores and Their Impact on Age at Onset and Prediction of Cardiometabolic Diseases and Common Cancers.” *Nature Medicine* 26 (4): 549–57.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2020. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *American Journal of Human Genetics* 107 (4): 788–89.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. “Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities.” *Nature Genetics* 51 (4): 584–91.

Martin, Antonio Rueda, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, et al. 2019. “PanelApp Crowdsources Expert Knowledge to Establish Consensus Diagnostic Gene Panels.” *Nature Genetics* 51 (11): 1560–65.

Mayle, Ryan, Ian M. Campbell, Christine R. Beck, Yang Yu, Marenda Wilson, Chad A. Shaw, Lotte Bjergbaek, James R. Lupski, and Grzegorz Ira. 2015. “DNA REPAIR. Mus81 and Converging Forks Limit the Mutagenicity of Replication Fork Breakage.” *Science* 349 (6249): 742–47.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 1–14.

Mele, Caterina, Mathieu Lemaire, Paraskevas Iatropoulos, Rossella Piras, Elena Bresin, Serena Bettoni, David Bick, et al. 2015. “Characterization of a New DGKE Intronic Mutation in Genetically Unsolved Cases of Familial Atypical Hemolytic Uremic Syndrome.” *Clinical Journal of the American Society of Nephrology: CJASN* 10 (6): 1011–19.

Menezes, Luís F. C., Yiqiang Cai, Yasuyuki Nagasawa, Ana M. G. Silva, Mary L. Watkins, Aline M. Da Silva, Stefan Somlo, Lisa M. Guay-Woodford, Gregory G. Germino, and Luiz F. Onuchic. 2004. “Polyductin, the PKHD1 Gene Product, Comprises Isoforms Expressed in Plasma Membrane, Primary Cilium, and Cytoplasm.” *Kidney International* 66 (4): 1345–55.

Minoche, Andre E., Ben Lundie, Greg B. Peters, Thomas Ohnesorg, Mark Pinese, David M. Thomas, Andreas Zankl, et al. 2021. “ClinSV: Clinical Grade Structural and Copy Number Variant Detection from Whole Genome Sequencing Data.” *Genome Medicine* 13 (1): 32.

Mitchel, M. W., D. Moreno-De-Luca, S. M. Myers, M. P. Adam, H. H. Ardinger, and R. A. Pagon. 2016. “GeneReviews.” University of Washington, Seattle,.

Mochizuki, T., G. Wu, T. Hayashi, … S. L. Xenophontos -., and Undefined 1996. n.d. “PKD2, a Gene for Polycystic Kidney Disease That Encodes an Integral Membrane Protein.” *Science.Sciencemag.Org*. https://science.sciencemag.org/content/272/5266/1339.abstract.

Mochizuki, Toshio, Atsuko Teraoka, Hiroyuki Akagawa, Shiho Makabe, Taro Akihisa, Masayo Sato, Hiroshi Kataoka, et al. 2019. “Mutation Analyses by Next-Generation Sequencing and Multiplex Ligation-Dependent Probe Amplification in Japanese Autosomal Dominant Polycystic Kidney Disease Patients.” *Clinical and Experimental Nephrology* 23 (8): 1022–30.

Monga, Manoj, Brandon Macias, Eli Groppo, and Alan Hargens. 2006. “Genetic Heritability of Urinary Stone Risk in Identical Twins.” *The Journal of Urology* 175 (6): 2125–28.

Moore, Camille M., Sean A. Jacobson, and Tasha E. Fingerlin. 2019. “Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification.” *Human Heredity* 84 (6): 256–71.

Morgenthaler, Stephan, and William G. Thilly. 2007. “A Strategy to Discover Genes That Carry Multi-Allelic or Mono-Allelic Risk for Common Diseases: A Cohort Allelic Sums Test (CAST).” *Mutation Research* 615 (1–2): 28–56.

Morris, Andrew P., and Eleftheria Zeggini. 2010. “An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies.” *Genetic Epidemiology* 34 (2): 188–93.

Moser, Markus, Sonja Matthiesen, Jutta Kirfel, Hubert Schorle, Carsten Bergmann, Jan Senderek, Sabine Rudnik-Schöneborn, Klaus Zerres, and Reinhard Buettner. 2005. “A Mouse Model for Cystic Biliary Dysgenesis in Autosomal Recessive Polycystic Kidney Disease (ARPKD).” *Hepatology* 41 (5): 1113–21.

Müller, Roman-Ulrich, A. Lianne Messchendorp, Henrik Birn, Giovambattista Capasso, Emilie Cornec-Le Gall, Olivier Devuyst, Albertien van Eerde, et al. 2022. “An Update on the Use of Tolvaptan for Autosomal Dominant Polycystic Kidney Disease: Consensus Statement on Behalf of the ERA Working Group on Inherited Kidney Disorders, the European Rare Kidney Disease Reference Network and Polycystic Kidney Disease International.” *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 37 (5): 825–39.

Nagai, Akiko, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, et al. 2017. “Overview of the BioBank Japan Project: Study Design and Profile.” *Journal of Epidemiology* 27 (3S): S2–8.

Nagano, China, Naoya Morisada, Kandai Nozu, Koichi Kamei, Ryojiro Tanaka, Shoichiro Kanda, Shinichi Shiona, et al. 2019. “Clinical Characteristics of HNF1B-Related Disorders in a Japanese Population.” *Clinical and Experimental Nephrology* 23 (9): 1119–29.

Nelson, Elizabeth K., Britt Piehler, Josh Eckels, Adam Rauch, Matthew Bellew, Peter Hussey, Sarah Ramsay, et al. 2011. “LabKey Server: An Open Source Platform for Scientific Data Integration, Analysis and Collaboration.” *BMC Bioinformatics* 12 (1): 1–23.

Ng, Sarah B., Kati J. Buckingham, Choli Lee, Abigail W. Bigham, Holly K. Tabor, Karin M. Dent, Chad D. Huff, et al. 2010. “Exome Sequencing Identifies the Cause of a Mendelian Disorder.” *Nature Genetics* 42 (1): 30–35.

Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, et al. 2009. “Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes.” *Nature* 461 (7261): 272–76.

Niemi, Mari E. K., Hilary C. Martin, Daniel L. Rice, Giuseppe Gallone, Scott Gordon, Martin Kelemen, Kerrie McAloney, et al. 2018. “Common Genetic Variants Contribute to Risk of Rare Severe Neurodevelopmental Disorders.” *Nature* 562 (7726): 268–71.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. “The Complete Sequence of a Human Genome.” *Science (New York, N.Y.)* 376 (6588): 44–53.

Oddsson, Asmundur, Patrick Sulem, Hannes Helgason, Vidar O. Edvardsson, Gudmar Thorleifsson, Gardar Sveinbjörnsson, Eik Haraldsdottir, et al. 2015. “Common and Rare Variants Associated with Kidney Stones and Biochemical Traits.” *Nature Communications 2015 6:1* 6 (1): 1–9.

Oetjens, M. T., M. A. Kelly, A. C. Sturm, C. L. Martin, and D. H. Ledbetter. 2019. “Quantifying the Polygenic Contribution to Variable Expressivity in Eleven Rare Genetic Disorders.” *Nature Communications* 10 (1): 1–10.

Ojavee, Sven E., Zoltan Kutalik, and Matthew R. Robinson. 2022. “Liability-Scale Heritability Estimation for Biobank Studies of Low-Prevalence Disease.” *American Journal of Human Genetics* 109 (11): 2009–17.

Olinger, Eric, Céline Schaeffer, Kendrah Kidd, Elhussein A. E. Elhassan, Yurong Cheng, Inès Dufour, Guglielmo Schiano, et al. 2022. “An Intermediate-Effect Size Variant in *UMOD* Confers Risk for Chronic Kidney Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (33): e2114734119.

Olson, Rory J., Katharina Hopp, Harrison Wells, Jessica M. Smith, Jessica Furtado, Megan M. Constans, Diana L. Escobar, Aron M. Geurts, Vicente E. Torres, and Peter C. Harris. 2019. “Synergistic Genetic Interactions between Pkhd1 and Pkd1 Result in an ARPKD-Like Phenotype in Murine Models.” *Journal of the American Society of Nephrology: JASN* 30 (11): 2113–27.

Omura, Masao, Jun Saito, Kunio Yamaguchi, Yukio Kakuta, and Tetsuo Nishikawa. 2004. “Prospective Study on the Prevalence of Secondary Hypertension among Hypertensive Patients Visiting a General Outpatient Clinic in Japan.” *Hypertension Research: Official Journal of the Japanese Society of Hypertension* 27 (3): 193–202.

Ong, A. C. 2000. “Polycystin Expression in the Kidney and Other Tissues: Complexity, Consensus and Controversy.” *Experimental Nephrology* 8 (4–5): 208–14.

Ong, Albert C. M., and Peter C. Harris. 2005. “Molecular Pathogenesis of ADPKD: The Polycystin Complex Gets Complex.” *Kidney International* 67 (4): 1234–47.

Onuchic, Laura, Valeria Padovano, Giorgia Schena, Vanathy Rajendran, Ke Dong, Xiaojian Shi, Raj Pandya, et al. 2023. “The C-Terminal Tail of Polycystin-1 Suppresses Cystic Disease in a Mitochondrial Enzyme-Dependent Fashion.” *Nature Communications* 14 (1): 1790.

Onuchic, Luiz F., Laszlo Furu, Yasuyuki Nagasawa, Xiaoying Hou, Thomas Eggermann, Zhiyong Ren, Carsten Bergmann, et al. 2002. “PKHD1, the Polycystic Kidney and Hepatic Disease 1 Gene, Encodes a Novel Large Protein Containing Multiple Immunoglobulin-like Plexin-Transcription-Factor Domains and Parallel Beta-Helix 1 Repeats.” *American Journal of Human Genetics* 70 (5): 1305–17.

Orth, S. R., A. Stöckmann, C. Conradt, E. Ritz, M. Ferro, W. Kreusser, G. Piccoli, et al. 1998. “Smoking as a Risk Factor for End-Stage Renal Failure in Men with Primary Renal Disease.” *Kidney International* 54 (3): 926–31.

Padmanabhan, Sandosh, Mark Caulfield, and Anna F. Dominiczak. 2015. “Genetic and Molecular Aspects of Hypertension.” *Circulation Research* 116 (6): 937–59.

Padmanabhan, Sandosh, and Anna F. Dominiczak. 2021. “Genomics of Hypertension: The Road to Precision Medicine.” *Nature Reviews. Cardiology* 18 (4): 235–50.
Palla, Luigi, and Frank Dudbridge. 2015. “A Fast Method That Uses Polygenic Scores to Estimate the Variance Explained by Genome-Wide Marker Panels and the Proportion of Variants Affecting a Trait.” *American Journal of Human Genetics* 97 (2): 250–59.
Paquette, Martine, Michael Chong, Sébastien Thériault, Robert Dufour, Guillaume Paré, and Alexis Baass. 2017. “Polygenic Risk Score Predicts Prevalence of Cardiovascular Disease in Patients with Familial Hypercholesterolemia.” *Journal of Clinical Lipidology* 11 (3): 725-732.e5.
Paranjpe, Ishan, Kipp Johnson, Steven G. Coca, Lili Chan, Cijiang He, Barbara Murphy, Jagat Narula, Ron Do, and Girish N. Nadkarni. 2019. “Abstract P2068: Whole Exome Sequencing And Monogenic Hypertension In A Multiethnic Cohort Of 28,000 Individuals.” *Hypertension* 74 (Suppl_1): AP2068–AP2068.
Paranjpe, Ishan, Noah Tsao, Renae Judy, Manish Paranjpe, Kumardeep Chaudhary, Derek Klarin, Iain Forrest, et al. 2020. “Derivation and Validation of Genome-Wide Polygenic Score for Urinary Tract Stone Diagnosis.” *Kidney International* 98 (5): 1323–30.
Park, Jason Y., Peter Clark, Eric Londin, Marialuisa Sponziello, Larry J. Kricka, and Paolo Fortina. 2015. “Clinical Exome Performance for Reporting Secondary Genetic Findings.” *Clinical Chemistry* 61 (1): 213–20.
Paterson, Andrew D., Riccardo Magistroni, Ning He, Kairong Wang, Ann Johnson, Pamela R. Fain, Elizabeth Dicks, Patrick Parfrey, Peter St George-Hyslop, and York Pei. 2005. “Progressive Loss of Renal Function Is an Age-Dependent Heritable Trait in Type 1 Autosomal Dominant Polycystic Kidney Disease.” *Journal of the American Society of Nephrology: JASN* 16 (3): 755–62.
Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” *PLoS Genetics* 2 (12): e190.
Paul, Binu M., Mark B. Consugar, Moonnoh Ryan Lee, Jamie L. Sundsbak, Christina M. Heyer, Sandro Rossetti, Vickie J. Kubly, et al. 2014. “Evidence of a Third ADPKD Locus Is Not Supported by Re-Analysis of Designated PKD3 Families.” *Kidney International* 85 (2): 383–92.
Pereira, Tiago Veiga, Ane Cláudia Fernandes Nunes, Martina Rudnicki, Ricardo Magistroni, Alberto Albertazzi, Alexandre Costa Pereira, and José Eduardo Krieger. 2006. “Influence of ACE I/D Gene Polymorphism in the Progression of Renal Failure in Autosomal Dominant Polycystic Kidney Disease: A Meta-Analysis.” *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 21 (11): 3155–63.
Pers, Tune H. 2016. “Gene Set Analysis for Interpreting Genetic Studies.” *Human Molecular Genetics* 25 (R2): R133–40.
Persu, A., M. S. Stoenoiu, T. Messiaen, S. Davila, C. Robino, O. El-Khattabi, M. Mourad, et al. 2002. “Modifier Effect of ENOS in Autosomal Dominant Polycystic Kidney Disease.” *Human Molecular Genetics* 11 (3): 229–41.
Persu, Alexandre, Michel Duyme, Yves Pirson, Xosé M. Lens, Thierry Messiaen, Martijn H. Breuning, Dominique Chauveau, Micheline Levy, Jean-Pierre Grünfeld, and Olivier Devuyst. 2004. “Comparison between Siblings and Twins

Supports a Role for Modifier Genes in ADPKD." *Kidney International* 66 (6): 2132–36.
Peterson, Roseann E., Karoline Kuchenbaecker, Raymond K. Walters, Chia-Yen Chen, Alice B. Popejoy, Sathish Periyasamy, Max Lam, et al. 2019. "Genome-Wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations." *Cell* 179 (3): 589–603.
Piontek, Klaus, Luis F. Menezes, Miguel A. Garcia-Gonzalez, David L. Huso, and Gregory G. Germino. 2007. "A Critical Developmental Switch Defines the Kinetics of Kidney Cyst Formation after Loss of Pkd1." *Nature Medicine* 13 (12): 1490–95.
Plaisier, Emmanuelle, Olivier Gribouval, Sonia Alamowitch, Béatrice Mougenot, Catherine Prost, Marie Christine Verpont, Béatrice Marro, et al. 2007. "COL4A1 Mutations and Hereditary Angiopathy, Nephropathy, Aneurysms, and Muscle Cramps." *The New England Journal of Medicine* 357 (26): 2687–95.
Porath, Binu, Vladimir G. Gainullin, Emilie Cornec-Le Gall, Elizabeth K. Dillinger, Christina M. Heyer, Katharina Hopp, Marie E. Edwards, et al. 2016. "Mutations in GANAB, Encoding the Glucosidase IIα Subunit, Cause Autosomal-Dominant Polycystic Kidney and Liver Disease." *American Journal of Human Genetics* 98 (6): 1193–1207.
Pritchard, J. K. 2001. "Are Rare Variants Responsible for Susceptibility to Complex Diseases?" *American Journal of Human Genetics* 69 (1): 124–37.
Privé, Florian, Julyan Arbel, and Bjarni J. Vilhjálmsson. 2021. "LDpred2: Better, Faster, Stronger." *Bioinformatics* 36 (22–23): 5424–31.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.
Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
"Recommendations for Research | Hypertension in Adults: Diagnosis and Management | Guidance | NICE." n.d. Accessed July 24, 2023. https://www.nice.org.uk/guidance/ng136/chapter/Recommendations-for-research.
Reich, D. E., and E. S. Lander. 2001. "On the Allelic Spectrum of Human Disease." *Trends in Genetics: TIG* 17 (9): 502–10.
Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94.
Resnick, Martin, Durward B. Pridgen, and Harold O. Goodman. 1968. "Genetic Predisposition to Formation of Calcium Oxalate Renal Calculi." *The New England Journal of Medicine* 278 (24): 1313–18.
Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.

Rimoldi, Stefano F., Urs Scherrer, and Franz H. Messerli. 2014. “Secondary Arterial Hypertension: When, Who, and How to Screen?” *European Heart Journal* 35 (19): 1245–54.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539): 317–30.

Roller, E., S. Ivakhno, S. Lee, T. Royce, S. Tanner -. Bioinformatics, and Undefined 2016. n.d. “Canvas: Versatile and Scalable Detection of Copy Number Variants.” *Academic.Oup.Com*. https://academic.oup.com/bioinformatics/article-abstract/32/15/2375/1743834.

Rosado, Consolación, Elena Bueno, Carmen Felipe, Sebastián Valverde, and Rogelio González-Sarmiento. 2015. “Study of the True Clinical Progression of Autosomal Dominant Alport Syndrome in a European Population.” *Kidney & Blood Pressure Research* 40 (4): 435–42.

Rossetti, Sandro, Vickie J. Kubly, Mark B. Consugar, Katharina Hopp, Sushmita Roy, Sharon W. Horsley, Dominique Chauveau, et al. 2009. “Incompletely Penetrant PKD1 Alleles Suggest a Role for Gene Dosage in Cyst Initiation in Polycystic Kidney Disease.” *Kidney International* 75 (8): 848–55.

Ruan, Yunfeng, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, et al. 2022. “Improving Polygenic Prediction in Ancestrally Diverse Populations.” *Nature Genetics* 54 (5): 573–80.

Rule, A. D. 2009. “Kidney Stones and the Risk for Chronic Kidney Disease.” *Clinical Journal of the American Society of Nephrology: CJASN* 4: 804–11.

Sadeghi-Alavijeh, Omid, Melanie My Chan, Shabbir H. Moochhala, Genomics England Research Consortium, Sarah Howles, Daniel P. Gale, and Detlef Böckenhauer. 2023. “Rare Variants in in the Sodium-Dependent Phosphate Transporter Gene SLC34A3 Explain Missing Heritability of Urinary Stone Disease.” *Kidney International*, July. https://doi.org/10.1016/j.kint.2023.06.019.

Sakaue, Saori, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, et al. 2021. “A Cross-Population Atlas of Genetic Associations for 220 Human Phenotypes.” *Nature Genetics* 53 (10): 1415–24.

Salmaso, Luigi, Rosa Arboretti, Livio Corain, and Dario Mazzaro. 2011. “Association Studies in Genetics.” In *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, 5–17. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sampson, J. R., M. M. Maheshwar, R. Aspinwall, P. Thompson, J. P. Cheadle, D. Ravine, S. Roy, E. Haan, J. Bernstein, and P. C. Harris. 1997. “Renal Cystic Disease in Tuberous Sclerosis: Role of the Polycystic Kidney Disease 1 Gene.” *American Journal of Human Genetics* 61 (4): 843–51.

Savige, Judy, and Philip Harraka. 2021. “Pathogenic Variants in the Genes Affected in Alport Syndrome (COL4A3–COL4A5) and Their Association With Other Kidney Conditions: A Review.” *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, July. https://doi.org/10.1053/J.AJKD.2021.04.017.

Scales, Charles D., Alexandria C. Smith, Janet M. Hanley, and Christopher S. Saigal. 2012. “Prevalence of Kidney Stones in the United States.” *European Urology* 62 (1): 160–65.

Schlingmann, Karl P., Justyna Ruminska, Martin Kaufmann, Ismail Dursun, Monica Patti, Birgitta Kranz, Ewa Pronicka, et al. 2016. “Autosomal-Recessive Mutations in SLC34A1 EncoDing Sodium-Phosphate Cotransporter 2a Cause Idiopathic Infantile Hypercalcemia.” *Journal of the American Society of Nephrology: JASN* 27 (2): 604–14.

Schmutz, Jeremy, Jeremy Wheeler, Jane Grimwood, Mark Dickson, Joan Yang, Chenier Caoile, Eva Bajorek, et al. 2004. “Quality Assessment of the Human Genome Sequence.” *Nature* 429 (6990): 365–68.

Schönauer, Ria, Sebastian Baatz, Melanie Nemitz-Kliemchen, Valeska Frank, Friederike Petzold, Sebastian Sewerin, Bernt Popp, et al. 2020. “Matching Clinical and Genetic Diagnoses in Autosomal Dominant Polycystic Kidney Disease Reveals Novel Phenocopies and Potential Candidate Genes.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 22 (8): 1374–83.

Schönauer, Ria, Friederike Petzold, Wilhelmina Lucinescu, Anna Seidel, Luise Müller, Steffen Neuber, Carsten Bergmann, John A. Sayer, Andreas Werner, and Jan Halbritter. 2019. “Evaluating Pathogenicity of SLC34A3-Ser192Leu, a Frequent European Missense Variant in Disorders of Renal Phosphate Wasting.” *Urolithiasis* 47 (6): 511–19.

Schrier, Robert W., Godela Brosnahan, Melissa A. Cadnapaphornchai, Michel Chonchol, Keith Friend, Berenice Gitomer, and Sandro Rossetti. 2014. “Predictors of Autosomal Dominant Polycystic Kidney Disease Progression.” *Journal of the American Society of Nephrology: JASN* 25 (11): 2399–2418.

Sekine, Akinari, Sumi Hidaka, Tomofumi Moriyama, Yasuto Shikida, Keiji Shimazu, Eiji Ishikawa, Kiyotaka Uchiyama, et al. 2022. “Cystic Kidney Diseases That Require a Differential Diagnosis from Autosomal Dominant Polycystic Kidney Disease (ADPKD).” *Journal of Clinical Medicine Research* 11 (21). https://doi.org/10.3390/jcm11216528.

Selvaraj, Margaret Sunitha, Xihao Li, Zilin Li, Akhil Pampana, David Y. Zhang, Joseph Park, Stella Aslibekyan, et al. 2022. “Whole Genome Sequence Analysis of Blood Lipid Levels in >66,000 Individuals.” *Nature Communications* 13 (1): 5995.

Senum, Sarah R., Ying (sabrina) M. Li, Katherine A. Benson, Giancarlo Joli, Eric Olinger, Sravanthi Lavu, Charles D. Madsen, et al. 2022. “Monoallelic IFT140 Pathogenic Variants Are an Important Cause of the Autosomal Dominant Polycystic Kidney-Spectrum Phenotype.” *American Journal of Human Genetics* 109 (1): 136–56.

Sevillano, Angel M., Eduardo Gutierrez, Enrique Morales, Eduardo Hernandez, Maria Molina, Ester Gonzalez, and Manuel Praga. 2014. “Multiple Kidney Cysts in Thin Basement Membrane Disease with Proteinuria and Kidney Function Impairment.” *Clinical Kidney Journal* 7 (3): 251–56.

Shan, Dan, Gabriel Rezonzew, Sean Mullen, Ronald Roye, Juling Zhou, Phillip Chumley, Dustin Z. Revell, et al. 2019. “Heterozygous Pkhd1C642* Mice Develop Cystic Liver Disease and Proximal Tubule Ectasia That Mimics Radiographic Signs of Medullary Sponge Kidney.” *American Journal of Physiology. Renal Physiology* 316 (3): F463–72.

Sharif-Naeini, Reza, Joost H. A. Folgering, Delphine Bichet, Fabrice Duprat, Inger Lauritzen, Malika Arhatte, Martine Jodar, et al. 2009. “Polycystin-1 and -2 Dosage Regulates Pressure Sensing.” *Cell* 139 (3): 587–96.

Shen, Peter S., Xiaoyong Yang, Paul G. DeCaen, Xiaowen Liu, David Bulkley, David E. Clapham, and Erhu Cao. 2016. “The Structure of the Polycystic Kidney Disease Channel PKD2 in Lipid Nanodiscs.” *Cell* 167 (3): 763-773.e11.

Signorell, Andri. 2023. “Tools for Descriptive Statistics [R Package DescTools Version 0.99.48],” February. https://CRAN.R-project.org/package=DescTools.

Sinnott-Armstrong, Nasa, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, et al. 2021. “Genetics of 35 Blood and Urine Biomarkers in the UK Biobank.” *Nature Genetics* 53 (2): 185–94.

Song, Xuewen, Valeria Di Giovanni, Ning He, Kairong Wang, Alistair Ingram, Norman D. Rosenblum, and York Pei. 2009. “Systems Biology of Autosomal Dominant Polycystic Kidney Disease (ADPKD): Computational Identification of Gene Expression Pathways and Integrated Regulatory Networks.” *Human Molecular Genetics* 18 (13): 2328–43.

Spielmann, Malte, and Stefan Mundlos. 2016. “Looking beyond the Genes: The Role of Non-Coding Variants in Human Disease.” *Human Molecular Genetics* 25 (R2): R157–65.

Staley, James R., Edmund Jones, Stephen Kaptoge, Adam S. Butterworth, Michael J. Sweeting, Angela M. Wood, and Joanna M. M. Howson. 2017. “A Comparison of Cox and Logistic Regression for Use in Genome-Wide Association Studies of Cohort and Case-Cohort Design.” *European Journal of Human Genetics: EJHG* 25 (7): 854–62.

Stankiewicz, Paweł, and James R. Lupski. 2002. “Genome Architecture, Rearrangements and Genomic Disorders.” *Trends in Genetics: TIG* 18 (2): 74–82.

Stavropoulos, Dimitri J., Daniele Merico, Rebekah Jobling, Sarah Bowdin, Nasim Monfared, Bhooma Thiruvahindrapuram, Thomas Nalpathamkalam, et al. 2016. “Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine.” *NPJ Genomic Medicine* 1 (January): 15012-.

Stoltzfus, Jill C. 2011. “Logistic Regression: A Brief Primer.” *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine* 18 (10): 1099–1104.

Sturm, Amy C., Joshua W. Knowles, Samuel S. Gidding, Zahid S. Ahmad, Catherine D. Ahmed, Christie M. Ballantyne, Seth J. Baum, et al. 2018. “Clinical Genetic Testing for Familial Hypercholesterolemia: JACC Scientific Expert Panel.” *Journal of the American College of Cardiology* 72 (6): 662–80.

Su, Qiang, Feizhuo Hu, Xiaofei Ge, Jianlin Lei, Shengqiang Yu, Tingliang Wang, Qiang Zhou, Changlin Mei, and Yigong Shi. 2018. “Structure of the Human PKD1-PKD2 Complex.” *Science* 361 (6406). https://doi.org/10.1126/science.aat9819.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-

Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.
Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLoS Medicine* 12 (3): e1001779.
Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.
Sun, B. B., J. Chiou, M. Traylor, C. Benner, and Y. H. Hsu. 2022. "Genetic Regulation of the Human Plasma Proteome in 54,306 UK Biobank Participants." *BioRxiv*. https://www.biorxiv.org/content/10.1101/2022.06.17.496443.abstract.
Sun, Benjamin B., Mitja I. Kurki, Christopher N. Foley, Asma Mechakra, Chia-Yen Chen, Eric Marshall, Jemma B. Wilk, et al. 2022. "Genetic Associations of Protein-Coding Variants in Human Disease." *Nature* 603 (7899): 95–102.
Sundström, Johan, Martin Neovius, Per Tynelius, and Finn Rasmussen. 2011. "Association of Blood Pressure in Late Adolescence with Subsequent Mortality: Cohort Study of Swedish Male Conscripts." *BMJ (Clinical Research Ed.)* 342 (feb22 2): d643.
Svishcheva, Gulnara R., Nadezhda M. Belonogova, Irina V. Zorkoltseva, Anatoly V. Kirichenko, and Tatiana I. Axenovich. 2019. "Gene-Based Association Tests Using GWAS Summary Statistics." *Bioinformatics* 35 (19): 3701–8.
Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2019. "Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program." *BioRxiv*. bioRxiv. https://doi.org/10.1101/563866.
———. 2021. "Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program." *Nature* 590 (7845): 290–99.
Tan, Adrian Y., Tuo Zhang, Alber Michaeel, Jon Blumenfeld, Genyan Liu, Wanying Zhang, Zhengmao Zhang, et al. 2018. "Somatic Mutations in Renal Cyst Epithelium in Autosomal Dominant Polycystic Kidney Disease." *Journal of the American Society of Nephrology: JASN* 29 (8): 2139–56.
Tan, Renjie, Yadong Wang, Sarah E. Kleinstein, Yongzhuang Liu, Xiaolin Zhu, Hongzhe Guo, Qinghua Jiang, Andrew S. Allen, and Mingfu Zhu. 2014. "An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data." *Human Mutation* 35 (7): 899–907.
Tanner, G. A., and J. A. Tanner. 2001. "Chronic Caffeine Consumption Exacerbates Hypertension in Rats with Polycystic Kidney Disease." *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 38 (5): 1089–95.
Taylor, Jenny C., Hilary C. Martin, Stefano Lise, John Broxholme, Jean-Baptiste Cazier, Andy Rimmer, Alexander Kanapin, et al. 2015. "Factors Influencing Success of Clinical Genome Sequencing across a Broad Spectrum of Disorders." *Nature Genetics* 47 (7): 717–26.
The GTEx Consortium, François Aguet, Shankara Anand, Kristin G. Ardlie, Stacey Gabriel, Gad A. Getz, Aaron Graubert, et al. 2020. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues." *Science (New York, N.Y.)* 369 (6509): 1318–30.

"The Polycystic Kidney Disease 1 Gene Encodes a 14 Kb Transcript and Lies within a Duplicated Region on Chromosome 16. The European Polycystic Kidney Disease Consortium." 1994. *Cell* 77 (6): 881–94.
Therneau, Terry M. 2023. "Survival Analysis [R Package Survival Version 3.5-5]," March. https://CRAN.R-project.org/package=survival.
Therneau, Terry M., Patricia M. Grambsch, and V. Shane Pankratz. 2003. "Penalized Survival Models and Frailty." *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 12 (1): 156–75.
Torres, Vicente E., Arlene B. Chapman, Olivier Devuyst, Ron T. Gansevoort, Jared J. Grantham, Eiji Higashihara, Ronald D. Perrone, et al. 2012. "Tolvaptan in Patients with Autosomal Dominant Polycystic Kidney Disease." *The New England Journal of Medicine* 367 (25): 2407–18.
Tsiokas, L., E. Kim, T. Arnould, V. P. Sukhatme, and G. Walz. 1997. "Homo- and Heterodimeric Interactions between the Gene Products of PKD1 and PKD2." *Proceedings of the National Academy of Sciences of the United States of America* 94 (13): 6965–70.
Turro, Ernest, William J. Astle, Karyn Megy, Stefan Gräf, Daniel Greene, Olga Shamardina, Hana Lango Allen, et al. 2020. "Whole-Genome Sequencing of Patients with Rare Diseases in a National Health System." *Nature* 583 (7814): 96–102.
Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419.
Ulinski, Tim, Sandra Lescure, Sandrine Beaufils, Vincent Guigonis, Stéphane Decramer, Denis Morin, Séverine Clauin, et al. 2006. "Renal Phenotypes Related to Hepatocyte Nuclear Factor-1β (TCF2) Mutations in a Pediatric Cohort." *Journal of the American Society of Nephrology: JASN* 17 (2): 497.
Van Buren, Jacob D., Jeremy T. Neuman, and Richard Sidlow. 2023. "Predominant Liver Cystic Disease in a New Heterozygotic PKHD1 Variant: A Case Report." *The American Journal of Case Reports* 24 (January): e938507.
Vilhjálmsson, Bjarni J., Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *American Journal of Human Genetics* 97 (4): 576.
Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era--Concepts and Misconceptions." *Nature Reviews. Genetics* 9 (4): 255–66.
Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics* 101 (1): 5–22.
Visscher, Peter M., Loic Yengo, Nancy J. Cox, and Naomi R. Wray. 2021. "Discovery and Implications of Polygenicity of Common Diseases." *Science* 373 (6562): 1468–73.
Vrijenhoek, Terry, Ken Kraaijeveld, Martin Elferink, Joep de Ligt, Elcke Kranendonk, Gijs Santen, Isaac J. Nijman, et al. 2015. "Next-Generation Sequencing-Based Genome Diagnostics across Clinical Genetics Centers: Implementation Choices

and Their Effects." *European Journal of Human Genetics: EJHG* 23 (9): 1142–50.

Walker, Denise, Mark Consugar, Jeff Slezak, Sandro Rossetti, Vicente E. Torres, Christopher G. Winearls, and Peter C. Harris. 2003. "The ENOS Polymorphism Is Not Associated with Severity of Renal Disease in Polycystic Kidney Disease 1." *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 41 (1): 90–94.

Walker, Rebecca, Qin Yao, Hangxue Xu, Anthony Maranto, Kristen Swaney, Sreekumar Ramachandran, Rong Li, et al. 2022. "Fibrocystin/Polyductin Releases a C-Terminal Fragment That Translocates into Mitochondria and Prevents Cystogenesis," September. https://doi.org/10.21203/rs.3.rs-2016158/v1.

Wang, Dong, Jie Li, Yadong Wang, and Edwin Wang. 2022. "A Comparison on Predicting Functional Impact of Genomic Variants." *NAR Genomics and Bioinformatics* 4 (1): lqab122.

Wang, Jiaru, Huayu Yang, Ruohan Guo, Xinting Sang, and Yilei Mao. 2021. "Association of a Novel PKHD1 Mutation in a Family with Autosomal Dominant Polycystic Liver Disease." *Annals of Translational Medicine* 9 (2): 120.

Wang, Kiarong, Xiao Zhao, Shelly Chan, Onur Cil, Ning He, Xuewen Song, Andrew D. Paterson, and York Pei. 2009. "Evidence for Pathogenicity of Atypical Splice Mutations in Autosomal Dominant Polycystic Kidney Disease." *Clinical Journal of the American Society of Nephrology: CJASN* 4 (2): 442–49.

Wang, Quanli, Ryan S. Dhindsa, Keren Carss, Andrew R. Harper, Abhishek Nag, Ioanna Tachmazidou, Dimitrios Vitsios, et al. 2021. "Rare Variant Contribution to Human Disease in 281,104 UK Biobank Exomes." *Nature* 597 (7877): 527–32.

Wang, Xiaofang, Vincent Gattone 2nd, Peter C. Harris, and Vicente E. Torres. 2005. "Effectiveness of Vasopressin V2 Receptor Antagonists OPC-31260 and OPC-41061 on Polycystic Kidney Disease Development in the PCK Rat." *Journal of the American Society of Nephrology: JASN* 16 (4): 846–51.

Ward, Christopher J., Marie C. Hogan, Sandro Rossetti, Denise Walker, Tam Sneddon, Xiaofang Wang, Vicky Kubly, et al. 2002. "The Gene Mutated in Autosomal Recessive Polycystic Kidney Disease Encodes a Large, Receptor-like Protein." *Nature Genetics* 30 (3): 259–69.

Ward, Christopher J., David Yuan, Tatyana V. Masyuk, Xiaofang Wang, Rachaneekorn Punyashthiti, Shelly Whelan, Robert Bacallao, et al. 2003. "Cellular and Subcellular Localization of the ARPKD Protein; Fibrocystin Is Expressed on Primary Cilia." *Human Molecular Genetics* 12 (20): 2703–10.

Watnick, Terry J., Vicente E. Torres, Michael A. Gandolph, Feng Qian, Luiz F. Onuchic, Katherine W. Klinger, Gregory Landes, and Gregory G. Germino. 1998. "Somatic Mutation in Individual Liver Cysts Supports a Two-Hit Model of Cystogenesis in Autosomal Dominant Polycystic Kidney Disease." *Molecular Cell* 2 (2): 247–51.

Weiner, Daniel J., Ajay Nadig, Karthik A. Jagadeesh, Kushal K. Dey, Benjamin M. Neale, Elise B. Robinson, Konrad J. Karczewski, and Luke J. O'Connor. 2023. "Polygenic Architecture of Rare Coding Variation across 394,783 Exomes." *Nature* 614 (7948): 492–99.

Whelton, Paul K., Robert M. Carey, Wilbert S. Aronow, Jr DE Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, et al. 2018. “2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NM CNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines.” *Hypertension* 71 (6): E13–115.

Wickham, Hadley. n.d. *Ggplot2*. Springer New York. Accessed June 16, 2023.

Willer, Cristen J., Yun Li, and Gonçalo R. Abecasis. 2010. “METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans.” *Bioinformatics* 26 (17): 2190–91.

Williams, Bryan, Giuseppe Mancia, Wilko Spiering, Agabiti Rosei, Michel Azizi, Michel Burnier, Denis L. Clement, et al. 2018. “2018 ESC/ESH Guidelines for the Management of Arterial Hypertension: The Task Force for the Management of Arterial Hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH).” *European Heart Journal* 39 (33): 3021–3104.

Williams, Scott S., Patricia Cobo-Stark, Leighton R. James, Stefan Somlo, and Peter Igarashi. 2008. “Kidney Cysts, Pancreatic Cysts, and Biliary Disease in a Mouse Model of Autosomal Recessive Polycystic Kidney Disease.” *Pediatric Nephrology* 23 (5): 733–41.

Wilson, Patricia D. 2008. “Chapter 6 Mouse Models of Polycystic Kidney Disease.” In *Current Topics in Developmental Biology*, 84:311–50. Academic Press.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *American Journal of Human Genetics* 89 (1): 82–93.

Xie, Bin, Kuo Tong, Jiao Yang, Taoli Wang, Lingchao Cheng, Suimin Zeng, and Zhongliang Hu. 2022. “NKX6-1 Is a Less Sensitive But Specific Biomarker of Chromophobe Renal Cell Carcinoma.” *The American Journal of Surgical Pathology* 46 (6): 809–15.

Xu, Dechao, Yiyi Ma, Xiangchen Gu, Rongrong Bian, Yunhui Lu, Xiaohong Xing, and Changlin Mei. 2018. “Novel Mutations in the PKD1 and PKD2 Genes of Chinese Patients with Autosomal Dominant Polycystic Kidney Disease.” *Kidney & Blood Pressure Research* 43 (2): 297–309.

Xu, Yaoxian, Andrew J. Streets, Andrea M. Hounslow, Uyen Tran, Frederic Jean-Alphonse, Andrew J. Needham, Jean-Pierre Vilardaga, Oliver Wessely, Michael P. Williamson, and Albert C. M. Ong. 2016. “The Polycystin-1, Lipoxygenase, and α-Toxin Domain Regulates Polycystin-1 Trafficking.” *Journal of the American Society of Nephrology: JASN* 27 (4): 1159–73.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. “Common SNPs Explain a Large Proportion of the Heritability for Human Height.” *Nature Genetics* 42 (7): 565–69.

Yang, Jian, Teresa Ferreira, Andrew P. Morris, Sarah E. Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela A. F. Madden, et al. 2012. “Conditional and Joint Multiple-SNP Analysis of GWAS Summary

Statistics Identifies Additional Variants Influencing Complex Traits." *Nature Genetics* 44 (4): 369–75, S1-3.
Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A Tool for Genome-Wide Complex Trait Analysis." *The American Journal of Human Genetics* 88 (1): 76–82.
Yang, Jian, S. Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. 2016. "GCTA-GREML Accounts for Linkage Disequilibrium When Estimating Genetic Variance from Genome-Wide SNPs." *Proceedings of the National Academy of Sciences of the United States of America*.
Yang, Yaping, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, et al. 2013. "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders." *The New England Journal of Medicine* 369 (16): 1502–11.
Yano, Yuichiro, Jared P. Reis, Laura A. Colangelo, Daichi Shimbo, Anthony J. Viera, Norrina B. Allen, Samuel S. Gidding, et al. 2018. "Association of Blood Pressure Classification in Young Adults Using the 2017 American College of Cardiology/American Heart Association Blood Pressure Guideline With Cardiovascular Events Later in Life." *JAMA: The Journal of the American Medical Association* 320 (17): 1774–82.
Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, et al. 2022. "A Saturated Map of Common Genetic Variants Associated with Human Height." *Nature* 610 (7933): 704–12.
Zarate, Samantha, Andrew Carroll, Medhat Mahmoud, Olga Krasheninina, Goo Jun, William J. Salerno, Michael C. Schatz, Eric Boerwinkle, Richard A. Gibbs, and Fritz J. Sedlazeck. 2020. "Parliament2: Accurate Structural Variant Calling at Scale." *GigaScience* 9 (12): 1–9.
Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. "Regression Models for Count Data in R." *Journal of Statistical Software* 27 (July): 1–25.
Zhang, Mingchao, Shuaimei Liu, Xinyi Xia, Yingxia Cui, and Xiaojun Li. 2019. "Identification of Novel Mutations and Risk Assessment of Han Chinese Patients with Autosomal Dominant Polycystic Kidney Disease." *Nephrology* 24 (5): 504–10.
Zhang, Zhiwu, Edward S. Buckler, Terry M. Casstevens, and Peter J. Bradbury. 2009. "Software Engineering the Mixed Model for Genome-Wide Association Studies on Large Samples." *Briefings in Bioinformatics* 10 (6): 664–75.
Zhou, Julie Xia, and Vicente E. Torres. 2023. "Autosomal Dominant Polycystic Kidney Disease Therapies on the Horizon." *Advances in Kidney Disease and Health* 30 (3): 245–60.
Zhou, Wei, Wenjian Bi, Zhangchen Zhao, Kushal K. Dey, Karthik A. Jagadeesh, Konrad J. Karczewski, Mark J. Daly, Benjamin M. Neale, and Seunggeun Lee. 2022. "SAIGE-GENE+ Improves the Efficiency and Accuracy of Set-Based Rare Variant Association Tests." *Nature Genetics* 54 (10): 1466–69.
Zhou, Wei, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, et al. 2018. "Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies." *Nature Genetics* 50 (9): 1335–41.
Zhou, Wei, Zhangchen Zhao, Jonas B. Nielsen, Lars G. Fritsche, Jonathon LeFaive, Sarah A. Gagliano Taliun, Wenjian Bi, et al. 2020. "Scalable Generalized Linear

Mixed Model for Region-Based Association Tests in Large Biobanks and Cohorts." *Nature Genetics* 52 (6): 634.
Zhu, Huanhuan, and Xiang Zhou. 2020. "Statistical Methods for SNP Heritability Estimation and Partition: A Review." *Computational and Structural Biotechnology Journal* 18 (June): 1557–68.
Zöller, Bengt, Eric Manderstedt, Christina Lind-Halldén, and Christer Halldén. 2023. "Rare-Variant Collapsing Analyses of Arterial Hypertension in the UK Biobank." *Journal of Human Hypertension*, April. https://doi.org/10.1038/s41371-023-00829-7.