JCI The Journal of Clinical Investigation

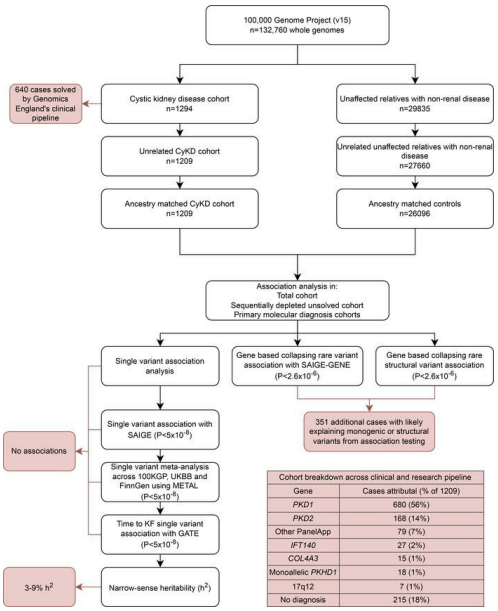
Quantifying variant contributions in cystic kidney disease using national-scale whole genome sequencing

Omid Sadeghi-Alavijeh, … , Adam P. Levine, Daniel P. Gale

J Clin Invest. 2024. https://doi.org/10.1172/JCI181467.

Clinical Medicine In-Press Preview **Genetics** **Nephrology**

Graphical abstract



Cohort breakdown across clinical and research pipeline	
Gene	Cases attributal (% of 1209)
PKD1	680 (56%)
PKD2	168 (14%)
Other PanelApp	79 (7%)
IFT140	27 (2%)
COL4A3	15 (1%)
Monoallelic *PKHD1*	18 (1%)
17q12	7 (1%)
No diagnosis	215 (18%)

Title: Quantifying variant contributions in cystic kidney disease using national-scale whole genome sequencing.

Authors: Omid Sadeghi-Alavijeh[1], Melanie MY. Chan[1], Gabriel Doctor[1], Catalin D. Voinescu[1], Alexander Stuckey[2], Athanasios Kousathanas[2], Alexander Ho[2], Horia C. Stanescu[1], Detlef Bockenhauer[1,5], Richard N. Sandford[4], Adam P. Levine[1,3], Daniel P. Gale[1]

Affiliations: 1- Centre for Kidney and Bladder Health, University College London, London, NW3 2PF

2- Genomics England, Queen Mary University of London, London, United Kingdom

3- Research Department of Pathology, University College London, London, United Kingdom

4- Academic Department of Medical Genetics, Cambridge University, Cambridge

5- University Hospital and Katholic University Leuven, Leuven, Belgium

Corresponding author: Daniel P Gale d.gale@ucl.ac.uk

Abstract

Background

Cystic kidney disease (CyKD) is a predominantly familial disease in which gene discovery has been led by family-based and candidate gene studies, an approach that is susceptible to ascertainment and other biases.

Methods

Using whole genome sequencing data from 1,209 cases and 26,096 ancestry-matched controls participating in the 100,000 Genomes Project, we adopted hypothesis-free approaches to generate quantitative estimates of disease risk for each genetic contributor to CyKD, across genes, variant types and allelic frequencies.

Results

In 82.3% of cases, a qualifying potentially disease-causing rare variant in an established gene was found. There was an enrichment of rare coding, splicing, and structural variants in known CyKD genes, with novel statistically significant gene-based signals in *COL4A3* and (monoallelic) *PKHD1*. Quantification of disease risk for each gene (with replication in the separate UK BioBank study) revealed substantially lower risk associated with genes more recently associated with autosomal dominant polycystic kidney disease, with odds ratios for some below what might usually be regarded as necessary for classical Mendelian inheritance. Meta-analysis of common variants did not reveal significant associations but suggested this category of variation contributes 3-9% to the heritability of CyKD across European ancestries.

Conclusion

By providing unbiased quantification of risk effects per gene, this research suggests that not all rare variant genetic contributors to CyKD are equally likely to manifest as a Mendelian trait in families. This information may inform genetic testing and counselling in the clinic.

Keywords: genomics, cystic kidney disease, renal, ADPKD, WGS

Introduction

Cystic kidney disease (CyKD) is a term for any disease that causes multiple fluid filled cysts within the kidney (excluding cystic degeneration of chronically diseased/failed kidneys). The term Polycystic Kidney Disease (PKD) generally refers to CyKD in which the kidneys become enlarged and replaced by cysts and is almost always inherited as a Mendelian trait (either autosomal dominant or autosomal recessive). 90% of CyKD is caused by Autosomal Dominant Polycystic Kidney Disease (ADPKD) which accounts for approximately 10% of all patients receiving kidney replacement therapy for kidney failure (KF) in the United Kingdom (UK) (1). It is estimated that between 1 in 800-1,000 of the population has ADPKD (2) making this the commonest monogenic cause of life-shortening disease worldwide.

Rare monoallelic variants in two genes, *PKD1* and *PKD2,* cause the majority of ADPKD, whilst biallelic variants in *PKHD1* cause the majority of ARPKD. *PKD1* accounts for ~80% of ADPKD diagnoses whilst *PKD2* accounts for ~15%(2). *PKD1* encodes Polycystin-1 (PC1), a large multi-domain glycoprotein whilst *PKD2* encodes Polycstin-2 (PC2) which is a nonspecific cation channel that interacts with PC1. Both proteins are found in primary cilia and play a role in mechanotransduction, transferring external information to the cell. Whilst the function of PC1, PC2 and the PC1-2 complex is still not fully understood, it is increasingly accepted that these proteins prevent cystogenesis via a dose dependent mechanism (3). *PKHD1* encodes the fibrocystin protein and acts via a common polycystin pathway to cause cysts in ARPKD usually

manifesting at a younger age than ADPKD and associated with extra-renal features including liver fibrosis (4).

There has been increasing recognition of the genetic heterogeneity of CyKD including contributions of genes other than *PKD1/PKD2*. Among the 10-15% of patients suspected to have ADPKD who have no explaining variants in either *PKD1* or *PKD2,* next generation sequencing (NGS) has identified several other genes in which variants cause ADPKD including *DNAJB11* (5)*, GANAB* (6)*, ALG9* (7) and more recently *IFT140* (8) and *ALG5* (9). Presently, a clinical diagnosis of polycystic kidney disease implies a high risk of kidney failure but pathogenic variants in these more recently described genes seem to carry a lower penetrance, manifesting as sporadic or non-Mendelian disease in individuals or families, with a less severe phenotype. To date, estimates of penetrance and outcomes with these rarer genetic variants are based predominantly on case series and reports of individual families, increasing the risk of ascertainment bias because more severely affected families are more likely to have come to light. Providing accurate data about the adverse consequences of these rarer variants is essential to prognostication, informing life and reproductive choices for patients and family members.

Obtaining a molecular diagnosis for CyKD can also inform choice of therapy (10,11). Genetic testing for *PKD1* variants is challenging owing to its high GC content, numerous repetitive regions, and the presence of six pseudogenes that share 97% of their sequence with *PKD1* (12,13). This has traditionally necessitated the use of Sanger sequencing of a long-range PCR-amplification product (13), a method that is technically difficult and expensive. Whole genome sequencing (WGS) has been shown to offer high quality *PKD1* sequencing (14). WGS provides more uniform coverage and

avoids capture bias, resulting in increased sensitivity to detect structural variants (SVs) and rare single nucleotide variants (SNVs) compared with exome (or capture-based) sequencing platforms, even within regions targeted by the latter (15,16). In addition, WGS benefits from economies of scale and the associated costs have dropped dramatically over the past decade (17) . However, few studies have looked at the utility of WGS in CyKD on a large scale (18).

Large-scale WGS datasets facilitate the assessment of the association of variants across the allelic spectrum with a certain phenotype in an unbiased genome-wide manner. This is especially useful in the study of the genomics of rare diseases, where approaches reliant on the sequencing of candidate gene panels or single genes of interest lead to discovery bias. In this study we perform genome-wide association analyses using WGS data from 1,209 CyKD cases and 26,096 ancestry matched controls. We perform variant-level, gene-level, pathway-level, and time-to-event association analyses including rare, common, and structural variants in a diverse-ancestry population providing the largest unbiased assessment of CyKD to date. We identify strong associations in known (*PKD1, PKD2, DNAJB11, IFT140)* and novel genes (*COL4A3,* monoallelic *PKHD1)* across various variant types. We leverage this data to conduct the largest common variant GWAS of CyKD to date, showing no significant associations, and use this to define the common variant heritability of CyKD as between 3-9%. This approach provides a comprehensive, unbiased framework for large-scale WGS analysis which can be utilized to gain insights into the molecular mechanisms underlying other rare diseases going forward.

Results

A full study workflow with high level results can be found in Figure 1.

The CyKD cohort demographics.

The cystic kidney disease cohort consisted of 1,558 participants recruited to the 100,000 Genomes Project (100KGP) of which 1,294 were probands (full breakdown of recruited pedigrees in Supplementary Table 1a). The demographic information and top five most frequent HPO codes of probands are shown in Table 1.

The clinically validated arm of the 100KGP gave a molecular diagnosis to 53% of the CyKD cohort.

Of these probands, 1,290 had a genetic outcome from the 100KGP clinical pipeline: 640 (52.93%) were solved, 34 (2.81%) were partially solved, 79 (6.54%) had missing data and 537 (44.42%) were unsolved. The top three molecular diagnoses were: *PKD1* truncating (340[26%]), *PKD2* truncating (122[9.5%]) , and *PKD1* non-truncating (118[9.1%]) variants. The full breakdown of solved cases and the types of variants can be found in Supplementary Table 2 (three 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 a gene recorded despite being listed as solved).

Outcome data shows those with pathogenic *PKD1* variants have the worst renal prognosis followed by those without a diagnosis.

Of the 1,290 cases, 578 (44.8%) had data regarding kidney function in the form of Human Phenotype Ontology (HPO) [PMID: 37953324] or Hospital Episode Statistics (HES) codes of whom 398 (68.9%) had reached KF (full breakdown in Supplementary Tables 1b and S1c). Survival

analysis of these cohorts followed what is known about the renal prognosis of variants in CyKD (Figure 2).

Combining the clinical and research pipelines identified a potential explaining genetic variant in the majority of CyKD cases.

Of the 1,209 ancestry matched CyKD cases, 994 (82%) had a potentially explaining monogenic or single structural variant identified when combining the clinical pipeline results with those who had variants identified via collapsing gene-based analyses (Supplementary Table S3). These are discussed in more detail below. Of note the research cohort consists of a subset of cases with ancestry matched controls (n=1,209).

Unbiased rare variant analysis highlights *IFT140* and *COL4A3* as important genes involved in CyKD.

Rare variant analysis of the ancestry matched cohort of 1,209 cases and 26,096 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 ($P=3.52x10^{-7}$, OR 1.07, 95% CI 0.95-1.21), and COL4A3 ($P=1.26x10^{-6}$, OR=3.02, 95% CI 2.10-4.22,* Figure 3a). There was no evidence of genomic inflation (lambda<1, see QQ-plot in Supplementary Figure 3b).

Removing cases solved by 100KGP and patients that had a bioinformatically ascertained potentially disease-causing variant in a known cystic gene left 308 cases (at the time of analysis *IFT140* was not a known CyKD gene). Repeating the rare variant analysis under the "missense+" mask in this group showed a significant enrichment of cases with variants in *IFT140 ($P=1.26x10^{-}$*

[16], OR=5.57, 95% CI 3.63-8.21) and *COL4A3 (P=6.83x10^{-7}, OR=4.93 95% CI 2.77-8.11)* compared with 26,096 controls (Figure 3b).

Using REVEL, a different annotation method for missense variants in the total and unsolved CyKD cohort, led to largely similar results but with smaller association values. In the total cohort; *PKD1 (P=1.63x10^{-35}, OR=22.41, 95% CI 14.57-34.49, PKD2 (P=3.84x10^{-14}, OR=21.87 95% CI 10.47-45.71)* and *COL4A3 (P=8.84x10^{-6}, OR=3.42 95% CI 2.04-5.47)* remained significant with the loss of the *DNAJB11* signal. Of note *PKHD1 (P=8.12x10^{-5}, OR=3.56 95% CI 1.62-7.02)* and *COL4A4 (P=7.74x10^{-4}, OR=3.52 95% CI 1.80-6.42)* increased their significance. In the unsolved CyKD cohort *COL4A3 (P=1.09x10^{-6} OR=6.50 95% CI 3.01-12.50)* remained significant (Supplementary tables S6 and S8) .

Of note other known genes causing CyKD such as *GANAB* and *ALG8* were not found to be enriched at a population level in this cohort. Analysis of a combined unsolved CyKD and polycystic liver disease cohort (n=359) did not reveal any additional associated genes (Supplementary tables S10 and 14).

IFT140 and *COL4A3* variants in unsolved individuals most likely represent their primary diagnosis.

Out of the 308 unsolved cases, 27 (8.8%) had a qualifying variant in *IFT140* under the "missense+" mask. Of the 27 cases, all were heterozygous for the qualifying variants. None of the variants individually reached genome-wide significance.

Analysis of SVs intersecting with *IFT140* revealed two additional cases (0.65%) with heterozygous exon-crossing SVs from the 308 unsolved cystic disease cases: one patient had a 3.6kb deletion spanning exon 16 and 17 and another patient had two different inversions (12.6kb and 8.1kb) .

Four exon crossing SVs (three deletions and one tandem duplication) were seen in four different controls (0.015%). There was enrichment of *IFT140* SVs in cases versus controls (*P*= 0.0032) although this was not significant at genome-wide level. None of the 27 initial cases with *IFT140* SNVs had detectable CNVs affecting *IFT140* compared to three CNVs seen in 26,096 controls (*P*= 1).

There were no plausible second variants within *IFT140* or other known cystic kidney genes in any of these individuals. A full variant and phenotypic breakdown of the *IFT140* cases can be found in Supplementary Table 4a.

Amongst the 15 unsolved CyKD patients with qualifying variants in *COL4A3* under the "missense+" mask, 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. There was no plausible second variant in known cystic kidney genes that could explain the phenotype. Of note four of the *COL4A3* patients had liver cysts. We reanalysed these four patients searching for known genetic causes of polycystic liver disease, but none were found. A full variant and phenotypic breakdown of the *COL4A3* cases can be found in Supplementary Table 4b.

Analysis of loss-of-function (protein truncating) variants identifies monoallelic defects of *PKHD1* in unsolved CyKD

Collapsing rare variants that had a high confidence call for loss-of-function under the "LoF" mask (i.e. analysis restricted to protein length-altering variants, excluding all missense variants)

revealed significant enrichment of cases for *PKD2 ($P=3.05x10^{-147}$, OR=130.85, 95% CI 83.66-215.37), PKD1 ($P=1.29x10^{-117}$,OR=36.01, 95% CI 30.52-42.23), IFT140 ($P=3.00x10^{-25}$, OR=14.03, 95%CI 7.91-24.45), DNAJB11 ($P=1.84x10^{-12}$, OR=1.07, 95% CI 0.95-1.21) and PKHD1 ($P=2.98x10^{-08}$, OR=4.07 95%CI 2.24-6.88)* (Figure 4).

Removing cases with qualifying variants in *IFT140* and *COL4A3* left 266 unsolved cases in whom rare variant testing did not reveal any additional significant associations (Supplementary Figure 3).

In all presented analyses, the patients were heterozygous for their qualifying variants, except in *DNAJB11* where 59 of the 369 cases that had qualifying variants within the "missense+" mask were homozygous.

There were 61 predicted LoF variants in *PKHD1* that made up the association signal in the LoF mask analysis of the whole CyKD cohort. These were seen in 50 cases of which 22 were solved, two were partially solved, 24 were unsolved and two were unascertainable. All 50 cases were heterozygous for the variant that made up the signal.

Of the 22 solved cases, three patients had a diagnosis of ARPKD secondary to biallelic *PKHD1* variants, and 19 had a diagnosis of ADPKD due to variants in *PKD1* or *PKD2*. In the two 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 also had a computationally predicted high impact non-truncating variant in *PKD1*, and one (in addition to the *PKHD1* variant) had a predicted high impact non-truncating *PKD2* variant.

In the remaining 18 cases with a single heterozygous *PKHD1* LoF variant there were no SNVs or SVs that would imply compound heterozygosity (and a diagnosis of ARPKD), or potentially pathogenic variants in any other gene associated with CyKD. Two patients had a second *PKHD1* variant with CADD >20 in *PKHD1,* but both had been deemed "likely benign" by Clinvar (Clinvar ID: 1187104 and 102305).

In total, 634 (2.4%) of the 26,096 controls carried qualifying monoallelic *PKHD1* LoF variants. When compared to the 18 (6.7%) out of 266 unsolved cases with no clear molecular diagnosis there is a significant enrichment of *PKHD1* variants in the unexplained CyKD cohort ($P=5.85x10^{-6}$, OR=2.92, 95% CI 1.69-4.76). Three of the 18 monoallelic *PKHD1* cases had reached KF at a median age of 42 years. There was no statistical difference between the rates of liver cysts between monoallelic *PKHD1* cohort and the general CyKD cohort (*P=0.31*). The full demographic details of the *PKHD1* cohort can be found in Supplementary table 4c.

Non-coding collapsing analysis of the no variant detected (NVD) cohort revealed enrichment of splice site variants in *PKD1* and *PKD2*

Removing the cases with qualifying *IFT140*, *COL4A3* and monoallelic *PKHD1* variants led to no further enrichment in the NVD cohort under the "missense+" or "LoF" gene collapsing tests. However, in the remaining 248 cases versus 26,096 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= $4.22x10^{-8}$ OR=398.51 95% CI 39.10-16384, DL P=$6.32x10^{-6}$ OR=no variants in controls)* and for DL in *PKD2 (P=$5.97x10^{-10}$ OR=no variants in controls) (*Figure 5).

There was no 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.

Rare variant analysis by primary variant does not reveal contribution of variants in other genes.

Using the primary variant, the CyKD cohort was divided into cases with *PKD1* and *PKD2* truncating and non-truncating variants, respectively. Excluding the primary gene in each cohort did not identify significant enrichment of any additional genes.

A full list of the summary statistics which includes the variants that make up each association can be found in Supplementary Tables 5-26.

Structural variants in *PKD1, PKD2* and the 17q12 loci play an important role in cystic kidney disease.

Exome wide gene-based SV analysis was performed in all CyKD cases and ancestry matched controls. Across all combined types of SV (CNVs, deletions, duplications, inversions) there was significant enrichment in *PKD1 ($P=2.02x10^{-14}$, OR=2.52 95% CI 1.69-3.63), PKD2 ($P=7.48x10^{-12}$, OR=3.51, 95% CI 1.74-6.37)* and genes within 17q12 locus including *HNF1B ($P=8.81x10^{-9}$, OR=7.11, 95% CI 3.41-13.66).* Of note, two genes within proximity of *PKD2* also reached genome-wide significance: *SPARCL1 ($P=5.76x10^{-7}$)* and *HSD17B11 ($P=8.69x10^{-6}$)* but these were made up of large CNVs that also encompassed *PKD2* (Figure 6).

The *PKD1* signal was driven by small deletions <10kb (median size 1.14kb, IQR 2.60) (*$P=2.17x10^{-22}$, 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 *($P=4.12x10^{-8}$, OR=8.70, 95% CI 3.72-18.80),* the signal was driven by deletions >10kb

(median size in *PKD2* 405kb, IQR 1273kb; 17q12 1550kb, IQR 94kb) with no other loci reaching genome-wide significance. No genes reached genome wide significance for duplications.

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

Of the 11 patients with 17q12 loci CNVs in the CyKD cohort, one patient had a *PKD1* non-truncating SNV and two had *PKD1* truncating SNVs that met the criteria for being likely disease-causing. One patient had a known *HNF1B* 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=248), there was significant enrichment for large (>10kb) deletions at the 17q12 loci ($P=9.21x10^{-9}$, *OR=24.04 94% CI 8.00-60.71)* which were detected in seven probands.

Of the seven 17q12 patients, the median age was 13.5 years, significantly lower than the total cystic disease cohort (*P<0.05).* None of the patients had reached kidney failure or had HPO or HES codes pertaining to diabetes; a full breakdown of phenotypic profile can be found in Supplementary Table 27 and all summary statistics from the SV analysis can be found in Supplementary Table 28-29.

More recently described genes in CyKD are less penetrant than *PKD1* and *PKD2.*

There was marked variation in the proportion of individuals with each gene/variant type that were documented to have CyKD, with the figures broadly comparable between 100KGP and UKBB (Table 2).

SeqGWAS of CyKD reveals no robust common variant associations.

A seqGWAS of 1,209 CyKD cases and 26,096 ancestry-matched controls using 10,377,275 variants with a MAF>0.1% (Supplementary Figure 4) revealed only a single variant reaching genome-wide statistical significance on chromosome 8, chr8:92259567:A:C (*P=1.38x10^{-8}, OR 0.72, MAF 0.23).* There was no evidence of genomic inflation (lambda 0.99). To confirm/refute this surprising finding we meta-analysed this dataset with those from the UK, Japanese and FinnGen biobanks. In the non-100KGP datasets there was evidence of association at several loci, most notably a stop gain in *PKHD1* in the FinnGen cohort (see Supplementary Figures 5-6 for individual study Manhattan plots) but the chr8:92259567:A:C signal was not replicated and likely to be a false positive. Overall in the combined analysis of 2,923 CyKD cases and 900,824 controls across 6,641,351 variants there were no genome-wide significant associations (Supplementary Figure 7).

Subgroup analysis by primary disease-causing variant type did not reveal any genome-wide significant loci (see Supplementary Figure 8 and 9).

The proportion of heritability attributable to common variants was between 3%-9%

Within the 100KGP CyKD cohort the proportion of phenotypic variance (h^2) explained by additive common and low-frequency variation among individuals of European ancestry was 9.0% (SE 7.6%). Using the summary statistics from the combined FinnGen/UKBB CyKD GWAS the

estimated heritability was 3.0% (SE 9.7%). The large standard errors reflect low power to detect heritability within this cohort.

Time to event analysis did not reveal any trans-acting genetic modifiers of severity.

Within the pre-ancestry matched CyKD cohort, 398 of the 1,288 probands had reached KF (30.9%) with a median age of 52 years (IQR 16). Time to event genetic association analysis did not reveal any genome-wide significant associations – either in the total cohort or stratified by primary gene or variant type (see Supplementary figure 10).

Discussion

Of the 1,209 ancestry matched CyKD patients, 994 (82%) had a qualifying potentially pathogenic SNV or SV identified through a combination of clinical grade and unbiased research analyses of biobank-scale WGS data.

The high diagnostic yield of WGS to investigate CyKD has led to this technology being made available to patients presenting with CyKD in the UK via the National Health Service's Genomic Medicine Service (19) (though it must be noted that as yet a proportion of these variants do not necessarily meet ACMG criteria for issuing a clinically actionable molecular diagnosis).

The data presented also clarifies the underlying genetic architecture of cystic kidney disease. They point strongly to the conclusion that cystic kidney disease is extensively driven by monogenic mechanisms via rare variants of multiple different types, with a small contribution from common variants.

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*).

Secondly, we provide robust statistical evidence that *COL4A3 is* associated with cystic kidney disease. Smaller studies have hinted at this association (20) and sequencing of unexplained KF patients in an American cohort, showed a significant proportion of unexplained cystic cases were attributed to the *COL4A* family of genes (21). Using collapsing burden testing (as opposed to SKAT-O) another group has also observed enrichment of rare variants in *COL4A3* in CyKD in the 100KGP (22). In the UK Biobank, *COL4A4* is the most strongly associated *COL4A* gene with a

cystic kidney disease phenotype ($P=5.85x10^{-4}$) and it is likely the play of chance, i.e. the rare variant distribution in our cohort versus UKBB, that accounts for the observation of association with *COL4A3* rather than *COL4A4* (or both genes): it is likely that a larger study would have the power to detect associations with both genes. Of note, four patients in the *COL4A3* cohort had liver cysts and despite no potentially pathogenic variants being found in genes associated with polycystic liver disease (*PRKCSH*, *SEC63*, *LRP5*) we cannot completely rule out the presence of a second variant contributing to this phenotype.

Thirdly, our attempts to find common variants that might contribute to the CyKD phenotype by meta-analysing more than 2,000 cases with nearly one million controls did not reveal any significant associations. In fact, in the Finnish population that has undergone significant genetic bottlenecks causing positive selection for certain recessive variants there is an enrichment of a known pathogenic *PKHD1* variant at an allelic frequency that borders the "rare" variant mask (rs137852949, MAF in Finnish population = $7.48x10^{-3}$, MAF in non-Finnish European population = $3.24x10^{-4}$) but meets inclusion in FinnGen. This variant has been implicated as a heterozygous cause of polycystic liver disease (23) and its enrichment in our cohort as a heterozygous entity provides evidence for the role of *PKHD1* as a monoallelic cause of cystic kidney disease. Of note, two of the monoallelic *PKHD1* cases also exhibited hepatic fibrosis which would be consistent with ARPKD, however, interrogation of their genomes did not reveal a plausible second variant. Potential explanations for this include: that these cases had an unannotated second variant (perhaps hypomorphic rather than truly pathogenic); they had somatic mosaicism for a second pathogenic *PKHD1* allele; they had an alternative cause for hepatic fibrosis; or that monoallelic *PKHD1* variants can manifest with this phenotype. Long term follow-up or more detailed

acquisition of phenotypic data from patients with monoallelic *PKHD1* variants may help to answer this in the future.

Our analysis of common variant heritability suggests that 3%-9% of CyKD may be explained by low penetrance common variation, though our efforts to ascertain this were underpowered resulting in large standard errors. The small magnitude of this contribution explains why individual genome-wide significant variants were not detected – the small effect size means an additional 440 cases would be required to detect a heritability signal with >80% power and that substantially larger studies would be needed to detect the loci contributing to this risk. (A full power calculation can be found in Supplementary Figure 9). Our efforts to use age of KF per driving molecular diagnosis within CyKD is an attempt to try and unpick common variant contributions to disease severity and quantitatively define genetic modifiers, a key question faced by those researching CyKD, but power was not sufficient to detect any significant signals. This represents the largest systematic analysis of whether oligogenic or polygenic mechanisms are important in the aetiology of CyKD and with a rapidly increasing number of CyKD patients undergoing WGS as part of routine clinical care in the UK and establishment of a National Genomic Research Library, our study sets up an analytical pathway to address this in the future with ever-increasing power.

Similar to our recent work on urinary stone disease highlighting the importance of intermediate effect rare variants as a risk factor (24), as well as work describing a low-frequency *UMOD* variant (present in 0.1% of European ancestry individuals) that confers an intermediate risk of KF (25), we show that CyKD represents another disease enriched for such variation. As shown in Table 2, the odds ratio for a CyKD diagnosis for each gene across both the 100KGP cohort and the UK Biobank is highly variable. This highlights that even highly deleterious variants in some of

the more recently reported CyKD genes are likely to have such a low penetrance that they may seldom exhibit Mendelian patterns of inheritance and may be perhaps regarded as intermediate risk factors for developing CyKD rather than 'pathogenic' variants in the Mendelian disease sense. 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 or reproductive interventions for these disorders.

Finally, our findings are replicated in the UK Biobank with the top gene associations with cystic kidney disease being *PKD1 ($P=9.83x10^{-63}$), PKD2 ($P=1.64x10^{-60}$)* and *IFT140 ($P=4.52x10^{-15}$)* in a European ancestry cohort of 531 patients and 239,516 controls (26). We calculate that the UK Biobank CyKD cohort is powered to detect genes that explain ≥8% of the total phenotypic variance, meaning genes associated with smaller effect sizes are unlikely to be identified.

The larger odds ratios observed for *PKD2* compared to *PKD1* is notable. We believe ascertainment bias of patients with *PKD1* variants in the 100KGP might explain this as patients with *PKD1* variants tend to present earlier to healthcare services with hypertension and other sequalae. In an unrelated analysis of patients with severe early onset hypertension in 100KGP we found *PKD1* to be the most strongly associated gene on collapsing analysis with many of the cases having been solved (27). This suggests that many CyKD-*PKD1* patients have a clinical diagnosis early in life and were therefore not entered into the 100KGP (100KGP recruitment criteria dissuaded recruitment of typical CyKD patients) meaning there are fewer *PKD1* patients (56% of those with potentially pathogenic variants) compared with the prior literature. Among 655 French patients with CyKD and a molecular diagnosis reported for clinical purposes, 80% had *PKD1* variants (28) and consistent with similar data from patients undergoing genetic testing for

CyKD in a clinical setting in the UK (unpublished RaDaR data: 438/550 (80%) patients have *PKD1* variants) . This suggests that the 100KGP and UK Biobank cohorts are probably depleted for people with *PKD1*-associated disease compared with people undergoing genetic testing for clinical reasons. It is also possible, however, that clinical sequencing cohorts may be susceptible to ascertainment bias, where patients with more severe or earlier onset disease are more likely to have a molecular test. Finally it is important to note that these odds ratios represent the odds of being ascertained for CyKD and should not automatically be seen as a marker for disease severity or age at kidney failure.

Using WGS we have also undertaken the first systematic assessment of the structural and non-coding variant contribution to CyKD. These contribute to unsolved cases highlighting the power WGS has in identifying sites previously untested by traditional sequencing technologies. Whilst splice site variants have been implicated in individual families with unexplained CyKD (29,30) this analysis gives quantitative statistical evidence at a population level that suggests these sites should be scrutinized in clinical analysis of *PKD1* and *PKD2*, something that is increasingly being recognized (31). Equally, utilizing methodology similar to the gnomAD SV working group (32) we find, as they did, that SVs play a larger role in the variant landscape than previously thought. Whilst SVs in CyKD genes have been implicated in small numbers, this cohort level analysis attributes at least 3.35% (40/1209) of the cystic disease burden to SVs, a similar proportion to the recently described *IFT140* gene at a population level. These discoveries are made possible by WGS which, unlike older methods, such as multiplex ligation-dependent probe amplification (MLPA) or arrays, enable accurate calling of many different types of SVs genome-wide. These findings should help inform decisions about the sensitivity of short-read WGS and other potential

sequencing approaches, such as RNA-sequencing or long read DNA (33) sequencing, in a clinical setting.

In the remaining unsolved cases we did not find any further enrichment at a variant, gene, pathway, or SV level. Given the findings above, one explanation is that we 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 as a consequence of non-monogenic developmental disorders, somatic mosaicism that WGS would be unable to detect, or undocumented environmental exposures that we have not accounted for such as diet, lithium exposure (34) or smoking that are known to affect CyKD risk or phenotype and may account for some of the missing heritability. Irrespective, this work gives an estimate of the cohort size (an additional 2000 cases using the assumptions in the power section of the methods) needed to power future studies to discover additional monogenic causes of CyKD using unbiased genome-wide approaches. As more patients with CyKD are sequenced as part of their routine healthcare in the UK it is possible that this threshold will be passed, and further monogenic causes will be discovered using this type of methodology. Coupled with developments in analytical techniques, identification of variants across the allelic frequency and disease risk spectrum may further extend understanding of the biological basis of cystic kidney disease.

This study has several other limitations. From a phenotype perspective we are reliant on age at KF as determined by hospital coding systems as our only marker of outcome, and were unable to access granular phenotypic and imaging data for our cohort, limiting our ability to apply several prognostic tools (11,35) that would have aided in further stratification as well as potentially

improving power by reducing the chance of misclassification of cases as controls. Secondly, whilst this study represents the largest WGS study in CyKD to date, we were underpowered to detect common variant signals in our seqGWAS with an OR<3, this has limited our ability to find tractable signals in all forms of GWAS as well as in the heritability analysis. The heritability estimates also relied on pre-computed taggings from the UKBB potentially introducing biases influenced by assumptions about genetic architecture and population structure. Furthermore, while WGS allows for more accurate SV calling over traditional microarray, this process is highly dependent on the algorithm used for calling with the possibility of false positives. Ideally long-range sequencing and independent validation would allow for more complete SV detection. Equally, we were unable to functionally characterize the splice variants given the lack of RNA sequencing, meaning our conclusions rest on the enrichment of such variants in cases compared with controls and any clinical actionability for participants in the 100KGP would be subject to cDNA confirmation on a case-by-case basis.

In summary, this study provides the most comprehensive WGS analysis of CyKD to date highlighting the contributions to disease risk of different types of variants across the allelic spectrum. Whilst some of our novel findings (*COL4A3* and *PKHD1*) will need additional validation both functionally and statistically, these findings can be used to inform genomic sequencing and counselling strategies offered to patients. This study also provides a blueprint for the unbiased analyses of other rare diseases using WGS on a biobank scale.

Methods

Sex as a biological variable

In all presented analyses sex, as determined by X and Y genotypes, was used as a covariate in all models generated.

The 100,000 Genomes Project

The 100,000 Genomes Project (100KGP) is one of the largest disease-based sequencing initiatives in the world in which WGS data from large numbers of NHS patients with rare diseases and cancer, and their relatives, have been generated (36,37). Key strengths of this dataset with respect to the study of rare diseases are that all germline samples are processed and analysed using a shared pipeline and that sequencing data is available for many individuals without the phenotype under study, drawn from the same population. This allows for robust control of technical artefacts, allele frequency and variant burden in the population, in contrast to previous sequencing studies.

Recruitment to the 100KGP is via a network of 13 NHS Genomic Medicine Centres (GMCs) and includes collection of phenotype data hierarchically encoded using Human Phenotype Ontology (HPO) codes (38), facilitating computerized analysis of clinical features. CyKD patients were recruited to the project if they met the following criteria:

- >5 cysts affecting one or both kidneys with at least 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;
- and/or:
    - features suggestive of classical ADPKD who had not undergone prior genetic testing of *PKD1* and *PKD2.*

Participants were excluded if they suffered from kidney failure (KF) due to identified (non-cystic) disease, if they had multicystic dysplastic kidney(s) or if they had a prior genetic diagnosis for their condition. We used the Genomics England dataset (v15) (39), which contains WGS data, details of clinical phenotypes encoded using Human Phenotype Ontology (HPO) terms and structed data automatically extracted from National Health Service (NHS) hospital records, collected for more than 90,259 cancer and rare disease patients (see Data availability) as well as their unaffected relatives to generate the cohorts. The study workflow and a full description of the cohort creation can be found in Supplementary Methods 1-4. After quality control, relatedness filtering and ancestry matching (Supplementary Methods 1-4 and Supplementary Figure 2), we were left with 1,209 cases and 26,096 controls for analysis.

Potentially pathogenic variants were ascertained through a combination of the clinical arm of the 100KGP and bioinformatically as detailed below to create subgroups stratified by enrichment for primary molecular cause of CyKD. We performed all single-variant, gene-burden and structural variant analysis in the total cohort and each molecular subgroup (except the "other genes" group). We 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, see Supplementary Figure 9). We

performed European only analysis in the unexplained CyKD cohort to highlight the advantages of ancestry matching (supplementary methods 4 and supplementary tables 9 and 13).

Single-variant seqGWAS

Whole-genome single-variant association analysis (seqGWAS) was carried out using the R package SAIGE (40) (version 0.42.1) which uses a generalised linear mixed model (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: minor allele count (MAC) ≥20, missingness <1%, Hardy-Weinberg equilibrium (HWE) $p>10^{-6}$, and differential missingness $p>10^{-5}$. When case-control ratios are unbalanced, as in our study, type 1 error rates are inflated because the asymptotic assumptions of logistic regression are invalidated. SAIGE employs a saddle point approximation (41) to calibrate score test statistics and obtain more accurate p-values than the normal distribution.

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 (42) was used to create Manhattan and Q-Q (quantile-quantile) plots. 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.

Metanalysis of GWAS data

A metanalysis of cystic kidney disease GWAS using summary statistics from our analysis, a combined UK/Japanese Biobank (UKBB/JBB) analysis of 220 phenotypes including polycystic kidney disease (19,093,042 variants) and Finngen (version 8) analysis of cystic kidney disease (43) (19,441,692 variants) was performed using METAL [version 2011-03-025](44). The summary statistics from the UKBB/JBB analyses were lifted over from build 37 to 38 using the UCSC liftover tool(45). A full breakdown of the biobank phenotypes can be found in Supplementary Methods 5. Between the three data sets 8,217,458 variants 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 ($P<1\times10^{-5}$) and those in which the minimum/maximum allele frequencies differed by >0.05 were excluded leaving 6,641,352 variants across 2,923 cases and 900,824 controls. The genomic inflation factor (lambda), calculated based on the 50th percentile, was 1.01 indicating no significant population stratification.

Single-variant seqGWAS time to event analysis

Genetic Analysis of Time-to-Event phenotypes (GATE) (46) was used to conduct a time to event (TTE) analysis utilizing the Cox proportional hazard model that accounts for heavily censored phenotypes and low frequency variants. 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 (full list of codes used in supplementary table 32) that

would highlight whether a patient had reached kidney failure (KF) as well as codes pertaining to their stage of chronic kidney disease (CKD) from stage 1-5 (Supplementary Table 1). The age of KF, as determined by the earliest occurrence of a clinical code for KF, was used as the end point in the TTE analysis and those who were yet to reach KF were censored. The same genomic and phenotype data as per the single variant seqGWAS was used to conduct the TTE GWAS.

Heritability estimation using common variants.

Narrow sense heritability (h^2), the contribution of phenotypic variation from additive genetic factors was estimated using two methods: GCTA-LDMS (47) and LDAK-SumHer (48). GCTA-LDMS was applied to the 100KGP WGS data using a European subset of the total ancestry-matched CyKD cohort (full details in Supplementary Methods 6).

Summary statistics from the cystic kidney disease analysis of the combined European ancestry Finngen and UK BioBank (780 FinnGen cases and 424 UKBB cases with 375,708 FinnGen controls and 417,905 UKBB controls) were used with LDAK-SumHer to calculate heritability under the BLD-LDAK model using the pre-computed taggings (a record of the relative expected heritability tagged by various predictors), calculated from the UKBB.

The observed heritability was then liability adjusted to account for the population prevalence of CyKD relative to its representation in the 100KGP (49). In this analysis a CyKD prevalence of 0.001 was used to transform the observed heritability to a liability threshold model.

Rare variant collapsing analysis

Prior to collapsing analysis, variants were filtered based on different criteria called masks. The masks used for this analysis were a rare, damaging missense mark ("missense+"), a high confidence loss of function mask ("LoF"), an intronic mask ("intronic"), a splice site mask, a 3-prime untranslated region mask (3'-UTR) and a 5-prime UTR mask (5'-UTR). For the total CyKD cohort and the unsolved NVD cohorts we also ran a rare exome variant ensemble learner (REVEL) (50) derived mask to investigate missense signals in more detail. Full details of the mask parameters and quality control can be found in Supplementary Methods 7.

We applied the "missense" and "LoF" masks to the total 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 above looking for previously undetected gene signals. Association testing was performed using the Scalable and Accurate Implementation of Generalized mixed model (SAIGE-GENE) (v0.42.1) (51) to ascertain whether rare coding variation was enriched in cases on a per-gene basis exome-wide. Sex and the top ten principal components were included as fixed effects when fitting the null model. Full details about the use of SAIGE can be found in Supplementary Methods 8.

Subgroup analysis stratified by primary variant and depleting analysis

Patients who have their phenotype "solved" by the clinical multi-disciplinary team (MDT) had a report issued with the details of the molecular diagnosis. Depending on the diagnosis these

patients were placed into different cohorts: *PKD1*-truncating (*PKD1*-T), PKD1-non truncating (*PKD1*-NT), *PKD2*-truncating (*PKD2*-T), *PKD2* non-truncating (*PKD2*-NT), "other gene" (encompassing other genes in the panel) and no variant detected (NVD). In the patients with NVD we 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 (52) and placing them in the relevant cohort. The filtering was performed using bcftools and filter-VEP. For each subsequent round of analysis (across SNV and SVs) if a gene or SV was found to be significantly enriched in cases, we identified the cases that contained qualifying variants and removed them from the NVD cohort and re-analysed the cohort, eventually leaving 184 cases with no clear genetic cause of disease identified.

We performed all single-variant, gene-burden, and structural variant analysis in each molecular subgroup (bar the "other genes" group). We 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, see supplementary figure 9).

Pathway analysis

For the cohort of patients that had no molecular diagnosis, the summary statistics from their rare variant SKAT-O analysis with SAIGE was analysed using the Gene set analysis Association using Sparse Signals method (GAUSS) (53) with default settings. The summary statistics were analysed using the canonical (3759 pathways) and hallmark (50 pathways) curated gene set

pathways from the Gene Set Enrichment Analysis (GSEA) group (54). The results of these analyses can be found in Supplementary Table 30-31.

Exome-wide Structural variant analysis

Structural variants were called from WGS using the Genomics England pipeline that incorporates CANVAS (55) to detect copy number (>10kb) and MANTA (56) to identify SVs greater than 50bp. 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 (150bp). 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.

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.

For each sample BEDTools(57) was used to extract SVs that intersected at least one exon by a minimum of 1 bp. 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 the gnomAD SVs(32) with allele frequency >1% and or a dataset of common (AF>0.1%) SVs generated from 12,243 cancer patients recruited to 100KGP. SVs were then merged using SURVIVOR (58) allowing a maximum distance of 300bp between pairwise breakpoints and allele frequencies calculated using BCFtools (59).

After removal of overlapping common variants, 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.

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.

Power

PAGEANT (60) is a power calculation tool for rare variant collapsing tests that uses the underlying distribution of gene size and MAF of variants from the ExAC dataset (61). Under the assumption that 80% of variants collapsed per gene in the SAIGE-GENE analysis were causal, we

calculated that we had >80% power to detect a gene signal that accounted for >4% of the variance of the phenotype with an exome-wide significance threshold of P=2.5x10^{-6} in the total case control cohort.

For single variant analysis, power was calculated using the R package genpwr (62) under an additive model using the conventional genome-wide significance threshold of P<5x10^{-8}. With this sample size at an allele frequency of 1%, single variant association testing was sufficiently powered (>80%) to detect alleles with an odds ratio (OR) > 3. Supplementary figure 11 details the results of the power calculations for the CyKD GWAS in more detail.

For heritability analysis we used the GCTA-GREML power calculator (63) which revealed we had a 54% chance of detecting 5% heritability within the 100KGP cohort. This is likely to be even lower in the summary statistics methods applied to the combined UKBB/FinnGen metanalysis.

Statistics

For single-variant association analyses (including time to event analyses) a significance threshold of P<5x10^{-8} was applied for genome-wide statistical significance. For gene-based rare variant analyses, a Bonferroni-corrected significance threshold of P<2.58x10^{-6} (0.05/19,364 genes) was used. 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 two-sided Fisher's test.

For pathway analysis a two-sided Fisher's exact test was utilized to compare cases to controls with a Bonferroni-corrected significance threshold of P<0.05/(n=pathways tested).

For structural variant (SV) analysis, a two-sided Fisher's exact test was utilized to compare cases and controls under a dominant inheritance model, with a Bonferroni-corrected significance threshold of $P<2.5x10^{-6}$ 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)

When comparing demographic and phenotypic characteristics between cases and controls, Student's t-tests and ANOVA were applied as appropriate. All t-tests were two-tailed unless otherwise specified. Two-way ANOVA was used when examining interactions between multiple factors.

The statistical significance for all tests was determined using a P value threshold of 0.05, unless specified otherwise for specific analyses. All statistical analyses were performed in R.

Study Approval

Ethical approval for the 100,000 Genomes Project was granted by the Research Ethics Committee for East of England – Cambridge South (REC Ref: 14/EE/1112). Participants provided written informed consent for the use of their genetic and clinical data.

Data Availability

All collapsing gene and pathway analyses can be found in the Supplementary Tables.

All seqGWAS summary statistics can be found at: https://zenodo.org/records/10613736

Details of the aggregated dataset used for the analysis can be found at: https://re-docs.genomicsengland.co.uk/aggv2/

Genomic and phenotype data from participants recruited to the 100,000 Genomes Project can be accessed by application to Genomics England Ltd (https://www.genomicsengland.co.uk/about-gecip/joining-research-community/).

Code Availability

Code used for the analyses in this paper can be found at: https://github.com/oalavijeh/cykd_paper.

Details of the ancestry derivation methods used by Genomics England can be found at: https://re-docs.genomicsengland.co.uk/gen_sim/

Details of the rare variant workflow can be found at: https://re-docs.genomicsengland.co.uk/avt/

Details of the common variant GWAS workflow can be found at: https://re-docs.genomicsengland.co.uk/gwas/

References

1. Evans K, Pyart R, Steenkamp R, Whitlock T, Stannard C, Gair R, et al. UK renal registry 20th annual report: introduction. karger.com [Internet]. Available from: https://www.karger.com/Article/Abstract/490958

2. Lanktree MB, Haghighi A, Guiard E, Iliuta IA, Song X, Harris PC, et al. Prevalence estimates of polycystic kidney and liver disease by population sequencing. J Am Soc Nephrol. 2018 Oct 1;29(10):2593–600.

3. Kim DY, Park JH. Genetic Mechanisms of ADPKD. Adv Exp Med Biol. 2016 Oct 1;933:13–22.

4. Ong ACM, Harris PC. A polycystin-centric view of cyst formation and disease: the polycystins revisited. Kidney Int. 2015 Oct;88(4):699–710.

5. Cornec-Le Gall E, Olson RJ, Besse W, Heyer CM, Gainullin VG, Smith JM, et al. Monoallelic Mutations to DNAJB11 Cause Atypical Autosomal-Dominant Polycystic Kidney Disease. Am J Hum Genet. 2018 May 5;102(5):832.

6. Porath B, Gainullin VG, Cornec-Le Gall E, Dillinger EK, Heyer CM, Hopp K, et al. Mutations in GANAB, Encoding the Glucosidase IIα Subunit, Cause Autosomal-Dominant Polycystic Kidney and Liver Disease. Am J Hum Genet. 2016 Jun 2;98(6):1193–207.

7. Besse W, Chang AR, Luo JZ, Triffo WJ, Moore BS, Gulati A, et al. ALG9 Mutation Carriers Develop Kidney and Liver Cysts. J Am Soc Nephrol. 2019;30(11):2091–102.

8. Senum SR, Li Y (sabrina) M, Benson KA, Joli G, Olinger E, Lavu S, et al. Monoallelic IFT140 pathogenic variants are an important cause of the autosomal dominant polycystic kidney-spectrum phenotype. Am J Hum Genet. 2022 Jan 6;109(1):136–56.

9. Lemoine H, Raud L, Foulquier F, Sayer JA, Lambert B, Olinger E, et al. Monoallelic pathogenic ALG5 variants cause atypical polycystic kidney disease and interstitial fibrosis. Am J Hum Genet. 2022 Aug 4;109(8):1484–99.

10. Gansevoort RT, Arici M, … TB-. ND, 2016 U. Recommendations for the use of tolvaptan in autosomal dominant polycystic kidney disease: a position statement on behalf of the ERA-EDTA Working Groups on. academic.oup.com [Internet]. Available from: https://academic.oup.com/ndt/article-abstract/31/3/337/2460159

11. Cornec-Le Gall E, Audrézet M-P, Rousseau A, Hourmant M, Renaudineau E, Charasse C, et al. The PROPKD score: a new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. Am Soc Nephrol. 2016;27:942–51.

12. Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, et al. Identification of gene mutations in autosomal dominant polycystic kidney disease through targeted resequencing. J Am Soc Nephrol. 2012 May 1;23(5):915–33.

13. Audrézet M-P, Audrézet A, Cornec-Le Gall E, Chen J-M, Redon S, Qú Eréer´eré I, et al. Autosomal dominant polycystic kidney disease: Comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. Wiley Online Library. 2012 Aug;33(8):1239–50.

14. Mallawaarachchi AC, Hort Y, Cowley MJ, Mccabe MJ, Minoche A, Dinger ME, et al. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur J Hum Genet. 2016 May 11;24(11):1584–90.

15. Biesecker LG, Green RC. Diagnostic Clinical Genome and Exome Sequencing. N Engl J Med. 2014 Jun 19;370(25):2418–25.

16. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020 Jun 24;583(7814):96–102.

17. The Cost of Sequencing a Human Genome [Internet]. Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

18. Mallawaarachchi AC, Lundie B, Hort Y, Schonrock N, Senum SR, Gayevskiy V, et al. Genomic diagnostics in polycystic kidney disease: an assessment of real-world use of whole-genome sequencing. European Journal of Human Genetics 2021 29:5. 2021 Jan 12;29(5):760–70.

19. England NHS. NHS Genomics Medicine Service Testing Directory [Internet]. [cited 2023 May 10]. Available from: https://www.england.nhs.uk/publication/national-genomic-test-directories/

20. Gulati A, Sevillano AM, Praga M, Gutierrez E, Alba I, Dahl NK, et al. Collagen IV Gene Mutations in Adults With Bilateral Renal Cysts and CKD. Kidney International Reports. 2020 Jan 1;5(1):103.

21. Groopman EE, Marasa M, Cameron-Christie S, Petrovski S, Aggarwal VS, Milo-Rasouly H, et al. Diagnostic Utility of Exome Sequencing for Kidney Disease. N Engl J Med. 2019 Jan 10;380(2):142–51.

22. Cipriani V, Vestito L, Magavern EF, Jacobsen JO, Arno G, Behr ER, et al. Rare disease gene association discovery from burden analysis of the 100,000 Genomes Project data. medRxiv [Internet]. 2023 Dec 21; Available from: http://dx.doi.org/10.1101/2023.12.20.23300294

23. Besse W, Dong K, Choi J, Punia S, Fedeles SV, Choi M, et al. Isolated polycystic liver disease genes define effectors of polycystin-1 function. J Clin Invest. 2017 May 1;127(5):1772–85.

24. Sadeghi-Alavijeh O, Chan MM, Moochhala SH, Genomics England Research Consortium, Howles S, Gale DP, et al. Rare variants in in the sodium-dependent phosphate transporter gene SLC34A3 explain missing heritability of urinary stone disease. Kidney Int [Internet]. 2023 Jul 4; Available from: http://dx.doi.org/10.1016/j.kint.2023.06.019

25. Olinger E, Schaeffer C, Kidd K, Elhassan EAE, Cheng Y, Dufour I, et al. An intermediate-effect size variant in *UMOD* confers risk for chronic kidney disease. Proc Natl Acad Sci U S A. 2022 Aug 16;119(33):e2114734119.

26. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021 Aug 10;597(7877):527–32.

27. Abstracts from the 2022 Annual Scientific Meeting of the British and Irish Hypertension Society (BIHS). J Hum Hypertens. 2022 Sep 5;36(1):1–22.

28. Cornec-Le Gall E, Audrézet M-P, Chen J-M, Hourmant M, Morin M-P, Perrichot R, et al. Type of PKD1 mutation influences renal outcome in ADPKD. J Am Soc Nephrol. 2013 May;24(6):1006–13.

29. Claverie-Martin F, Gonzalez-Paredes FJ, Ramos-Trujillo E. Splicing defects caused by exonic mutations in PKD1 as a new mechanism of pathogenesis in autosomal dominant polycystic kidney disease. RNA Biol. 2015;12(4):369–74.

30. Wang K, Zhao X, Chan S, Cil O, He N, Song X, et al. Evidence for pathogenicity of atypical splice mutations in autosomal dominant polycystic kidney disease. Clin J Am Soc Nephrol. 2009 Feb;4(2):442–9.

31. Hort Y, Sullivan P, Wedd L, Fowles L, Stevanovski I, Deveson I, et al. Atypical splicing variants in PKD1 explain most undiagnosed typical familial ADPKD. NPJ Genom Med. 2023 Jul 7;8(1):16.

32. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020 May 27;581(7809):444–51.

33. Borràs DM, Vossen RHAM, Liem M, Buermans HPJ, Dauwerse H, van Heusden D, et al. Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing. Hum Mutat. 2017 Jul;38(7):870–9.

34. Grünfeld J-P, Rossier BC. Lithium nephrotoxicity revisited. Nat Rev Nephrol. 2009 May;5(5):270–6.

35. Irazabal MV, Rangel LJ, Bergstralh EJ, Osborn SL, Harmon AJ, Sundsbak JL, et al. Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials. J Am Soc Nephrol. 2015 Jan;26(1):160–72.

36. 100000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868–80.

37. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, et al. National Genomic Research Library [Internet]. figshare; 2020. Available from: https://figshare.com/articles/dataset/GenomicEnglandProtocol_pdf/4530893/7

38. Köhler S, Gargano M, Research NM-. …. A, 2021 U. The human phenotype ontology in 2021. academic.oup.com [Internet]. Available from: https://academic.oup.com/nar/article-abstract/49/D1/D1207/6017351

39. Genomics England. The National Genomics Research and Healthcare Knowledgebase v5. 2019;

40. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet. 2020 Jun 1;52(6):634.

41. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am J Hum Genet. 2017 Jul 6;101(1):37–49.

42. D. Turner S. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw. 2018 May 19;3(25):731.

43. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner K, et al. FinnGen: Unique genetic insights from combining isolated population and national health register data [Internet]. bioRxiv. 2022. Available from: https://www.medrxiv.org/content/10.1101/2022.03.03.22271360v1.abstract

44. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010 Sep 1;26(17):2190–1.

45. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res [Internet]. 2006;34(Database issue). Available from: https://pubmed.ncbi.nlm.nih.gov/16381938/

46. Dey R, Zhou W, Kiiskinen T, Havulinna A, Elliott A, Karjalainen J, et al. Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. Nat Commun. 2022 Sep 16;13(1):5437.

47. Yang J, Lee SH, Wray NR, Goddard ME, Visscher PM. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. Proc Natl Acad Sci U S A. 2016 Aug 9;113(32):E4579-80.

48. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using summary statistics. Nat Genet. 2020 Apr;52(4):458–62.

49. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011 Mar 11;88(3):294–305.

50. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 6;99(4):877–85.

51. Zhou W, Bi W, Zhao Z, Dey KK, Jagadeesh KA, Karczewski KJ, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat Genet. 2022 Sep 22;54(10):1466–9.

52. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019 Nov;51(11):1560–5.

53. Dutta D, VandeHaar P, Fritsche LG, Zöllner S, Boehnke M, Scott LJ, et al. A powerful subset-based method identifies gene set associations and improves interpretation in UK Biobank. Am J Hum Genet. 2021 Apr 1;108(4):669–81.

54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.

55. Roller E, Ivakhno S, Lee S, Royce T, Bioinformatics ST-., 2016 U. Canvas: versatile and scalable detection of copy number variants. academic.oup.com [Internet]. Available from: https://academic.oup.com/bioinformatics/article-abstract/32/15/2375/1743834

56. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220–2.

57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

58. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017 Jan 24;8:14061.

59. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021 Feb 16;10(2). Available from: https://pubmed.ncbi.nlm.nih.gov/33590861/

60. Derkach A, Zhang H, Chatterjee N. Power Analysis for Genetic Association Test (PAGEANT) provides insights to challenges for rare variant association studies. Bioinformatics. 2018 May 1;34(9):1506–13.

61. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 17;536(7616):285–91.

62. Moore CM, Jacobson SA, Fingerlin TE. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. Hum Hered. 2019;84(6):256–71.

63. Visscher PM, Hemani G, Vinkhuyzen AAE, Chen G-B, Lee SH, Wray NR, et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genet. 2014 Apr;10(4):e1004269.

Acknowledgments

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The authors gratefully acknowledge the participation of the patients and their families recruited to the 100,000 Genomes Project.

OSA is funded by an MRC Clinical Research Training Fellowship (MR/S021329/1). MYC is funded by a Kidney Research UK Clinical Research Fellowship (TF_004_20161125). DPG is supported by the St Peter's Trust for Kidney, Bladder and Prostate Research.

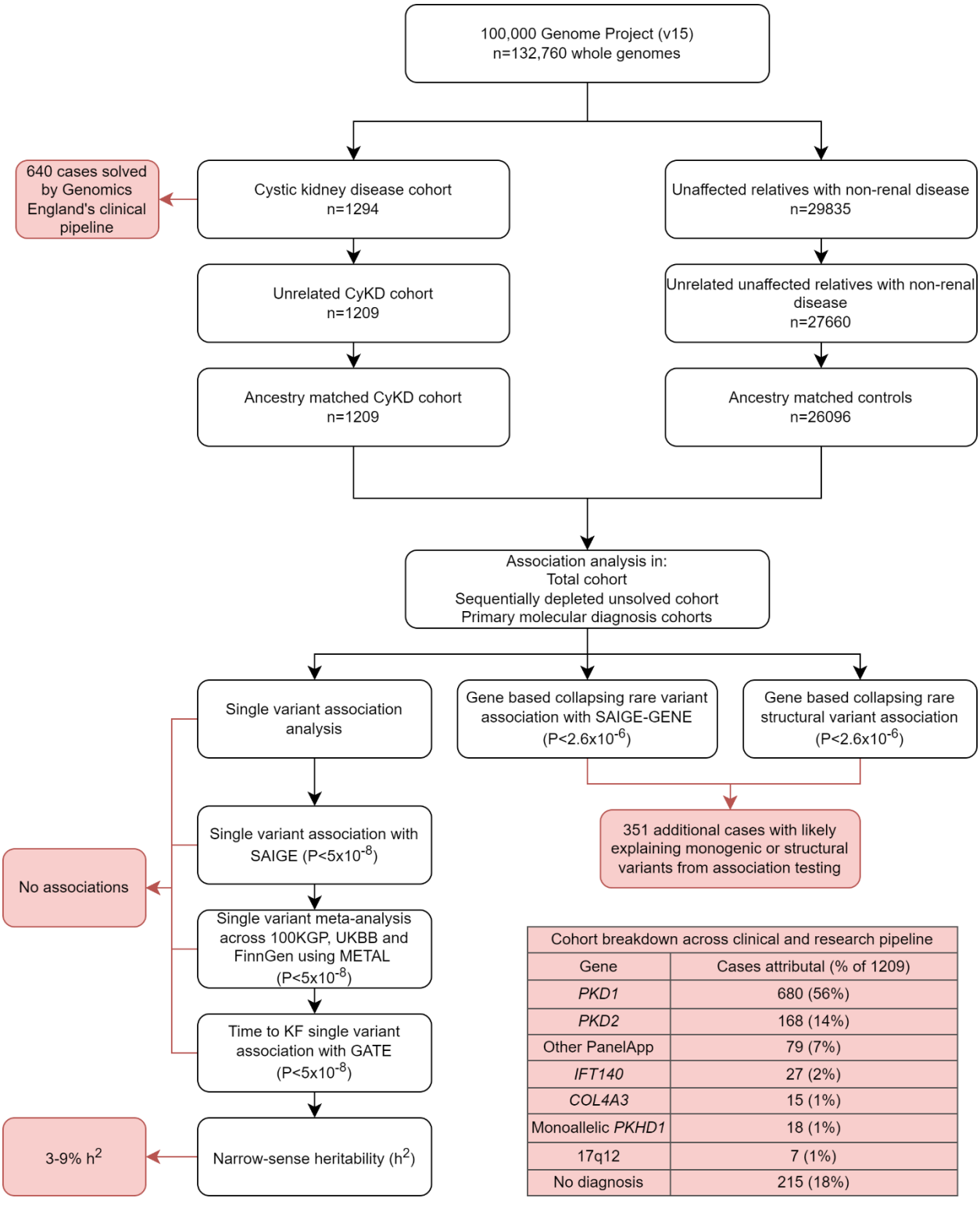
Author Contributions

DPG conceived the study. OSA conducted the analyses and wrote the manuscript with extensive input from MYC, DPG and APL. GD aided with analyses; CV provided scripting support. AS, AK and AH provided bioinformatic assistance from the Genomics England side. HS, DB, RS, APL, and DPG provided tutelage and guidance throughout.

Competing interests

The authors declare no competing interests.

Figures



Cohort breakdown across clinical and research pipeline	
Gene	Cases attributal (% of 1209)
PKD1	680 (56%)
PKD2	168 (14%)
Other PanelApp	79 (7%)
IFT140	27 (2%)
COL4A3	15 (1%)
Monoallelic *PKHD1*	18 (1%)
17q12	7 (1%)
No diagnosis	215 (18%)

Figure 1 – Study workflow detailing high level methods and results

Flow diagram depicting the methods and high-level results from the study. 100KGP -100,00 genome project, UKBB - UK Biobank, GATE – Genetic Analysis of Time-to-Event.

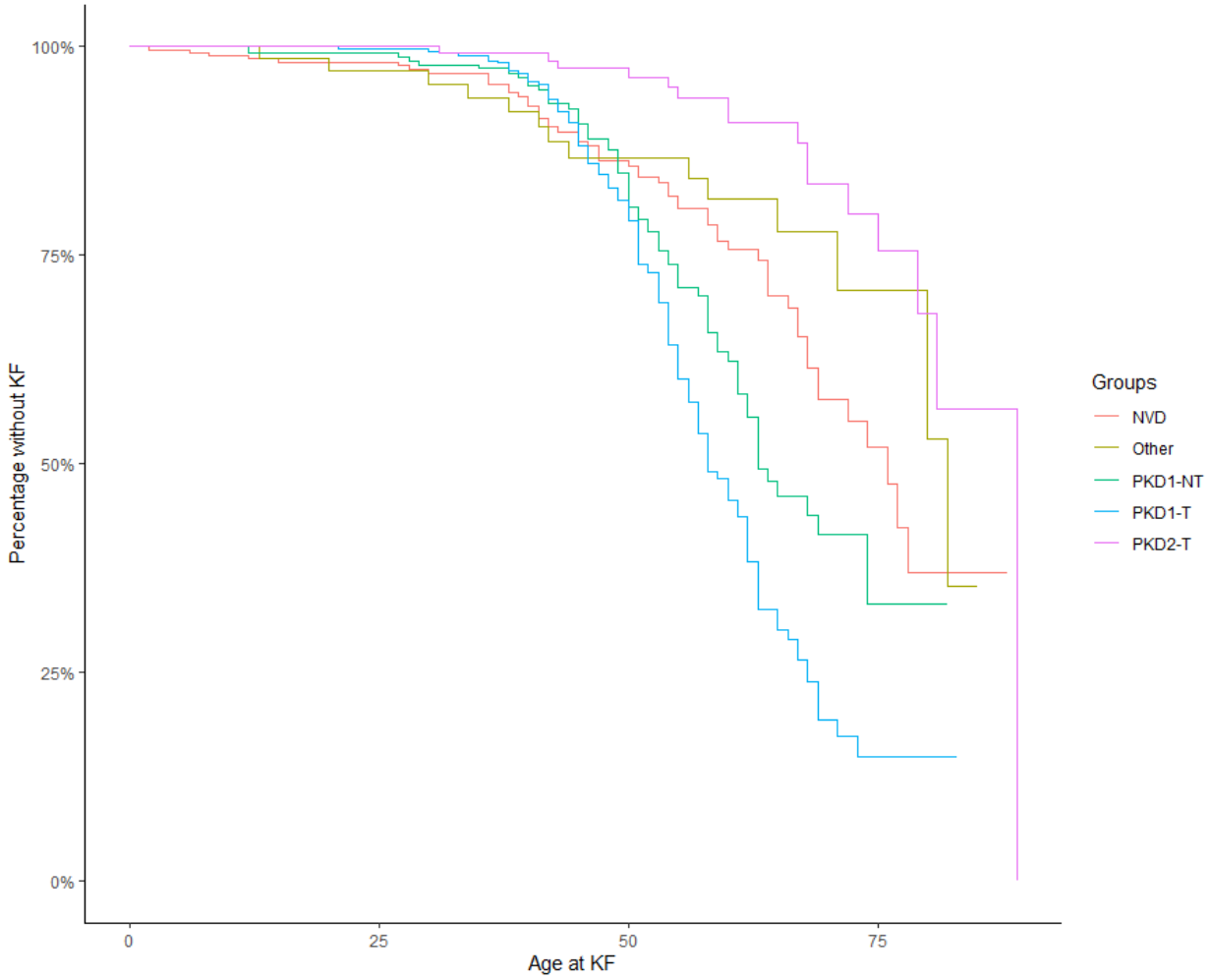


Figure 2 - Outcome data shows those with pathogenic *PKD1* variants have the worst renal prognosis followed by those without a diagnosis.

Kaplan-Meier graph of kidney failure by variant type. 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

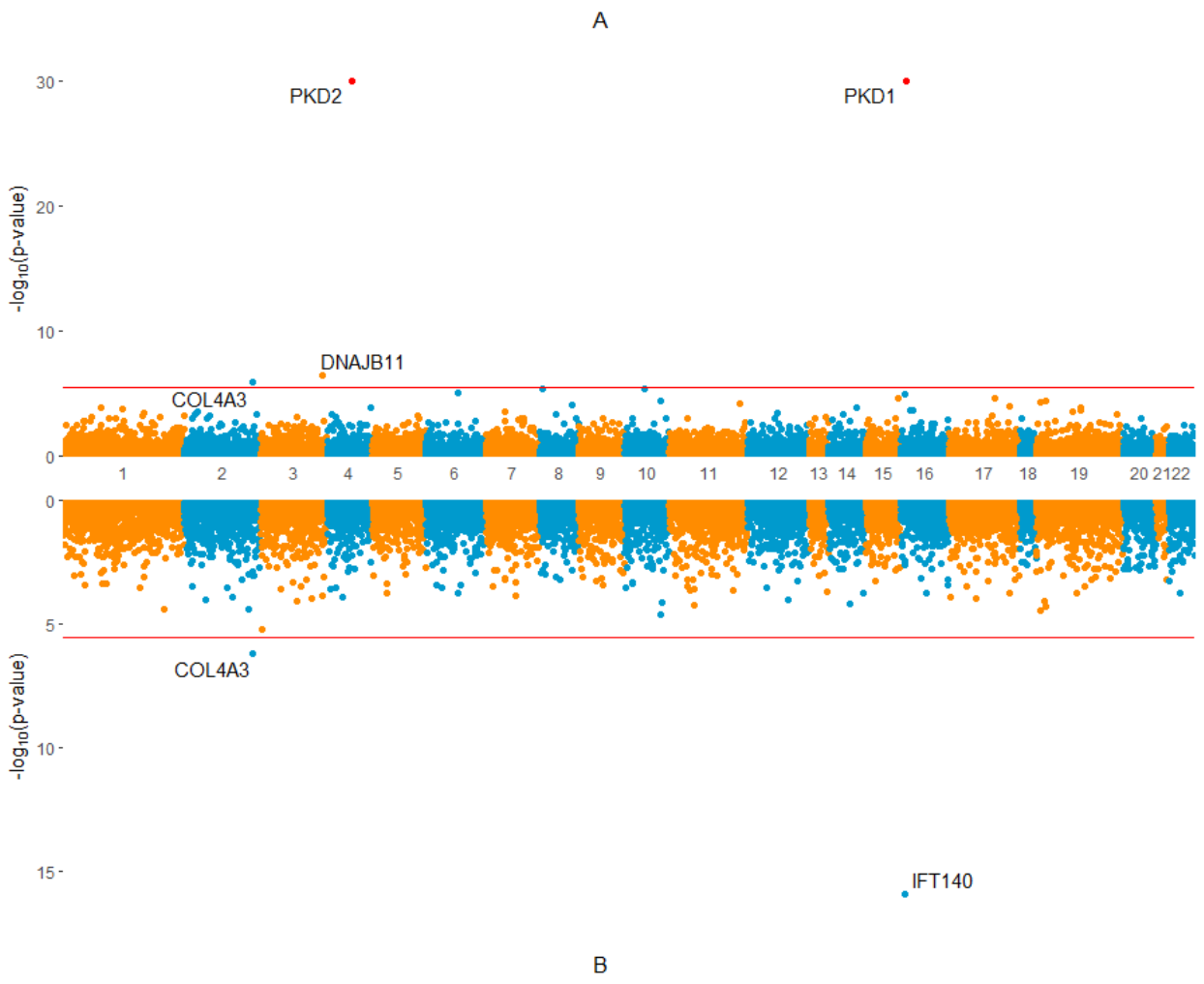


Figure 3 - Unbiased rare variant analysis highlights *IFT140* and *COL4A3* as important genes involved in CyKD

Gene-based Miami plots of the SAIGE-GENE "missense+" analyses. Each point is a gene, made up of variants that qualified under their respective mask. A. Missense+ analysis of the total ancestry matched cohort of 1,209 cases and 26,096 showing a significant enrichment of cases for PKD1, PKD2, DNAJB11 and COL4A3. PKD1 and PKD2 are highlighted as the plot is capped at an association of $1x10^{-30}$, actual associations: PKD1 ($P=1.17x10^{-309}$), PKD2 ($P=1.96x10^{-150}$). B. Missense+ analysis of the depleted cohort of 308 cases versus 26,096 controls showing enrichment of cases for IFT140 and COL4A3. The horizontal line indicates the threshold for exome-wide significance.

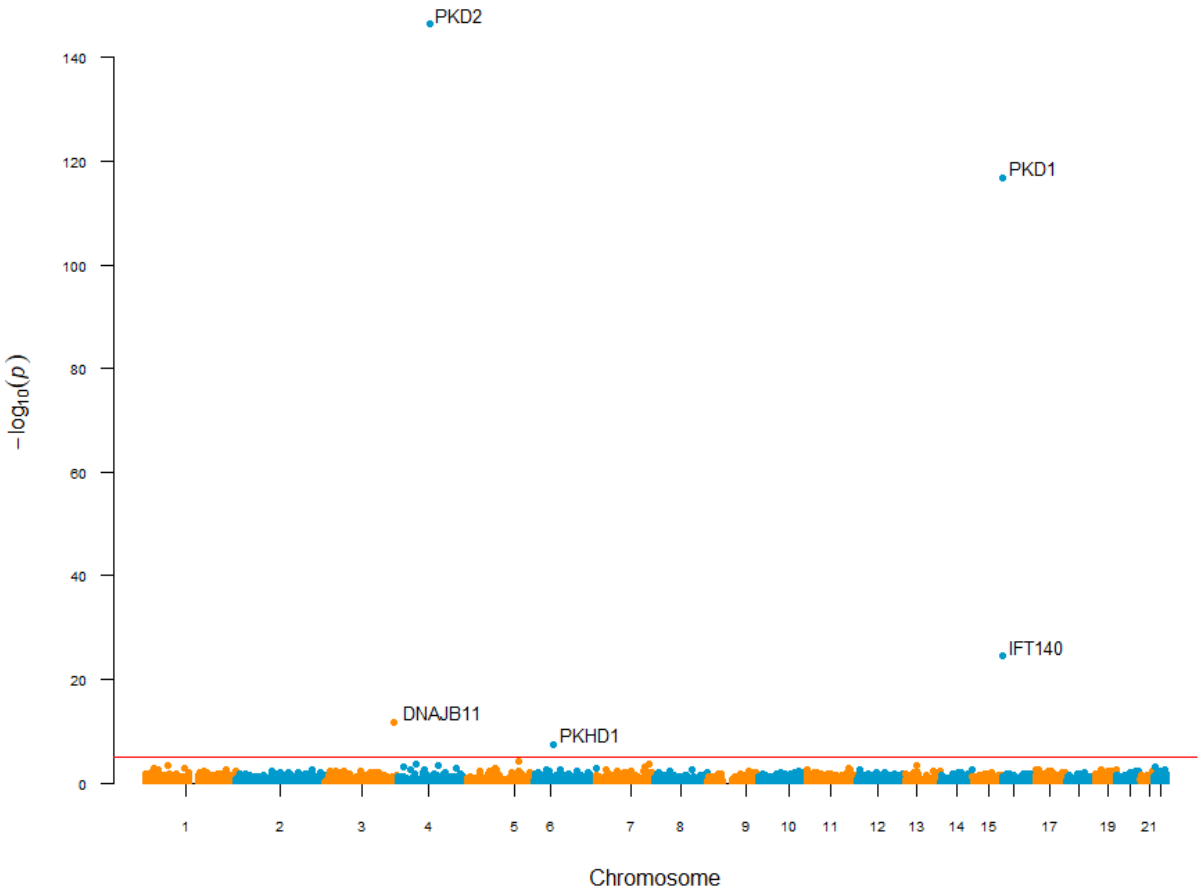


Figure 4 - Analysis of loss-of-function (protein truncating) variants identify monoallelic defects of *PKHD1* in unsolved CyKD.

Gene-based Manhattan plot of the SAIGE-GENE "loss-of-function" analysis of the total ancestry matched cohort of 1,209 cases and 26,096 showing a significant enrichment of cases for PKD1, PKD2, DNAJB11, IFT140 and PKHD1. Each point is a gene, made up of variants that qualified under their respective mask. The red line indicates an exome-wide significance level of P=2.5x10-6.

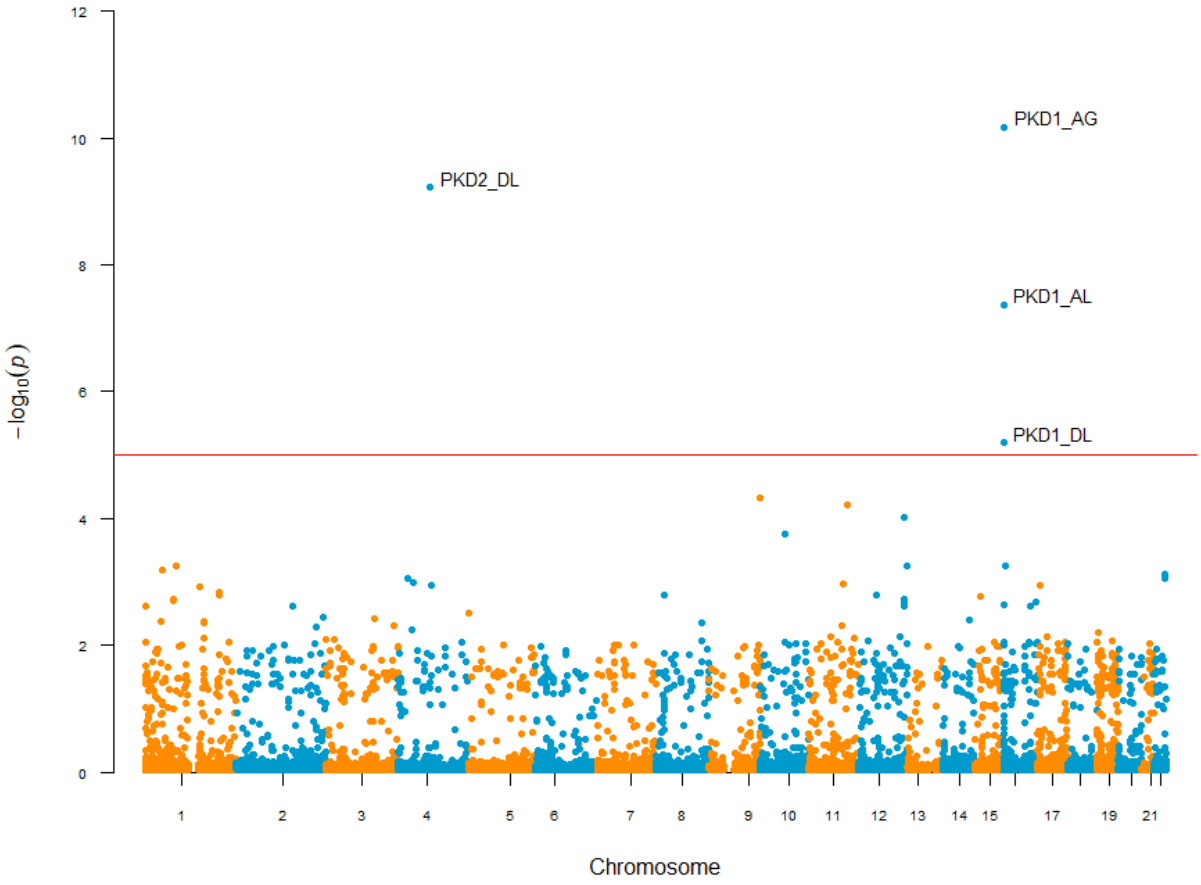


Figure 5 - Non-coding collapsing analysis of the no variant detected (NVD) cohort revealed enrichment of splice site variants in *PKD1* and *PKD2*

Gene-based Manhattan plot of SAIGE-GENE splicing analysis. Each point is a gene representing the significance of the association with cystic kidney disease in 248 cases versus 26096 controls, made up of variants that are highly likely (SpliceAI score >0.8) to impact splicing. The horizontal line indicates the threshold for genome wide significance (genome or exome? Add

the threshold used). There was significant enrichment in acceptor gain (AG), acceptor loss (AL) and donor loss (DL) variants for PKD1 (AG P= 6.70x10-11, AL P= 4.22x10-08, DL P=6.32x10^{-6}) and for DL in PKD2 (P=5.97x10^{-10}).

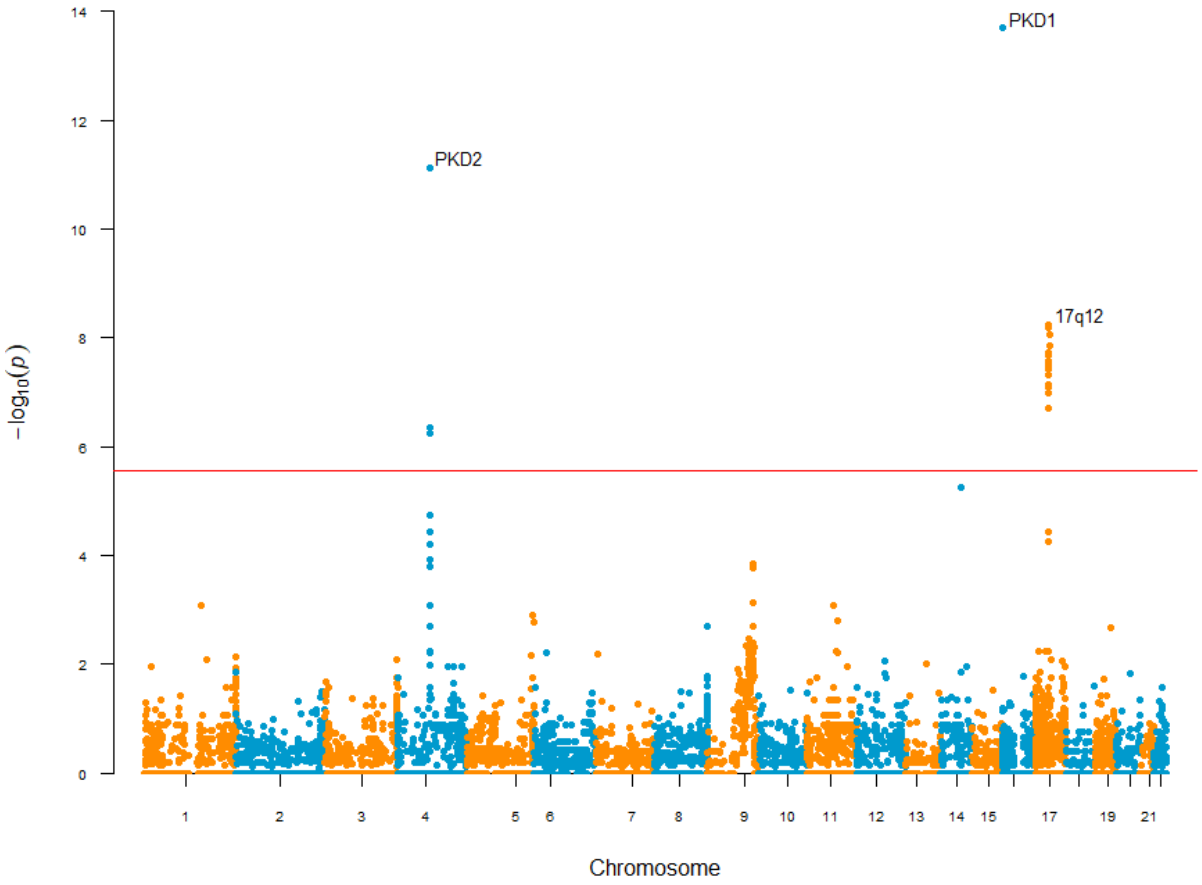


Figure 6 – Structural variants in *PKD1*, *PKD2* and the 17q12 loci play an important role in cystic kidney disease.

Gene-based Manhattan plot. Each point is a gene representing the significance of the association with cystic kidney disease in 1,209 cases and 26,096 controls, made up of rare (in-house MAF<0.01), exon crossing SV/CNVs that have been called by MANTA/CANVAS. Common SV/CNVs (MAF>0.1%) seen in gnomAD or the 100KGP cancer cohort were excluded. The horizontal line indicates the threshold for genome-wide significance. The significant associations were PKD1 ($P=2.02x10^{-14}$), PKD2($P=7.48x10^{-12}$) and the 17q12 locus ($P=8.81x10^{-9}$).

Tables

Table 1- Demographic and phenotypic breakdown of the recruited cystic kidney disease probands and controls. Ancestry is inferred using whole genome sequencing (as described in the methods). HPO, Human Phenotype Ontology.

Demographics	Case (N=1294)	Control (N=26096)
Female	669 (51.75%)	14,557 (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
Inferred Ancestry		
European	960 (74.19%)	20255 (77.62%)
African	65 (5.02%)	553 (2.12%)
American	4 (0.31%)	42 (0.16%)
South Asian	90 (6.96%)	2982 (11.43%)
East Asian	15 (1.16%)	158 (0.61%)
Admixed	160 (12.36%)	2106 (8.07%)
HPO phenotypes		
Multiple renal cysts	1,085 (83.85%)	NA
Hypertension	697 (53.86%)	NA
Enlarged Kidney	513 (39.64%)	NA
Hepatic cysts	383 (29.60%)	NA
Haematuria	162 (12.52%)	NA

Table 2 – Odds ratio of developing CyKD in the 100KGP (n=741) and the UKBB (n=825) in each different gene. OR for UKBB taken from the AstraZeneca analysis [ref] using the model closest to our analysis.

***Association not statistically significant in association analysis, included here as they are clinically reportable genes for this phenotype. 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.*

Gene	OR of developing CyKD (100KGP)	OR of developing CyKD (UKBB)
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 truncating	5.40 (0.11-54.56)**	Not seen
GANAB non-truncating	1.63 (0.79-3.02)**	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)**