Supplemental Online Content

Chang AR, Moore BS, Luo JZ, et al. Exome sequencing of a clinical population for autosomal dominant polycystic kidney disease. *JAMA*. Published December 27, 2022. doi:10.1001/jama.2022.22847

eAppendix. Supplementary Methods

eTable 2. Coding Definition for Comorbidities in Table 1

eTable 3. Variants in Patients With Other Kidney Diseases or ADPKD Family History Alone

eTable 8. Typical ADPKD Phenotype Cases With "Other" (Not *PKD1* or *PKD2*) Gene Variants

eFigure 1. Typical ADPKD Imaging Phenotype

eFigure 2. Mild ADPKD Imaging Phenotype

eFigure 3. Atypical ADPKD Imaging Phenotype

eFigure 4. Flow Diagram for Cohorts Described in the Study

eFigure 5. Sample Pedigrees and Images of Carriers of *IFT140* Variants

eFigure 6. Sample Pedigrees and Images of Carriers of Novel *PKD1* Variants

(For eTables 1, 4, 5, 6, and 7, please see Supplement 1 [Excel document].)

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix. Supplementary Methods

Data Sharing Statement

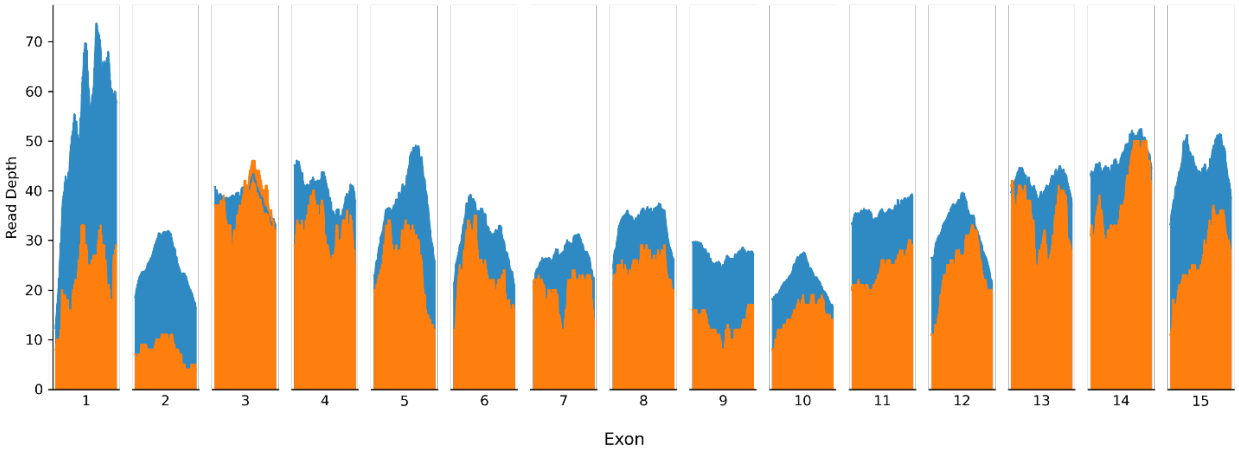
The data supporting the findings of this study are available within the article and Supplementary Data files listed above. Additional information for reproducing the results described in the article is available upon reasonable request and subject to a data use agreement. Pertinent data will be deposited into public repositories, (e.g., ClinVar) upon publication.

Exome sequencing

Exome sequencing was performed in collaboration with Regeneron Genetics Center as previously described (28). A modified version of the xGEN probe from Integrated DNA Technologies (IDT) was used for target sequence capture (29). Sequencing was performed by paired end 75bp reads on either an Illumina HiSeq2500 or NovaSeq. Coverage depth was sufficient to provide more than 90% coverage of the targeted bases in for 99% of samples. Alignments and variant calling were based on GRCh38 human genome reference sequence.

Copy number variants for *PKD1*, *PKD2*, and atypical ADPKD genes were estimated using lattice-aligned mixture models (CLAMMS) software(30). In brief, a mixture model is fit for each exome capture region to characterize the expected distribution of coverage for a copy number state. A hidden Markov model is used to call CNVs for each individual using normalized coverage values. As experimentally validating CNV calls in our setting is difficult, a number of measures were taken to ensure that only high-quality calls were included in the analysis. First, CLAMMS computes quality metrics based on probabilities from the hidden Markov model. These metrics were used to filter for high-confidence CNV calls. To further validate these CNVs, we examined the compressed Reference-oriented Alignment Map (CRAM) files. CRAM files show alignment

of exome sequencing reads to the reference exome (similar to binary alignment maps [BAM] files) and contain read-depth as well as reference and alternate allele counts. Briefly, one should expect a shallower average read depth for the regions deleted as part of the CNV. Using average read depth in these regions we find lower read depth for the samples identified as CNV deletion (see exons 1-2 for PKD2 in the example shown below from one of the patients).



The example provided above shows sequencing depth across 15 exons in *PKD2* with the reference panel in blue and CNV deletion carrier in orange. The carrier's deletion spans the first two exons of *PKD2*, and a corresponding drop in sequencing depth is observed, orange peaks are lower compared to the blue reference peaks, supporting the validity of the CNV call.

Nonsense or PTVs (protein truncating variants) are defined in this study as variants that cause a start-loss, frameshift, or early termination/stop-gain of the encoded protein. Variants in *PKD1, PKD2, ALG8, ALG9, DNAJB11, GANAB, HNF1B, IFT40, LRP5, PKHD1,* and *SEC61* were annotated using the Ensembl Variant Effect Predictor with Ensembl 90 definitions (31). We included PTV variants that had GQ score >20, AD_ALT>3, AD_ALT/AD_REF ≥0.5, and MAF<0.01 in the MyCode population. Missense variants in *PKD1* and *PKD2* were included using the same criteria, but only if previously reported in the PKDB as Likely Pathogenic, or Likely Hypomorphic. We cross-referenced all variants with ClinVar interpretation and stars rating

system (32). *PKD1* and *PKD2* were additionally cross-referenced with the PKDB (33) as well as VarSome (https://doi.org/10.1093/bioinformatics/bty897), a database that considers case-control, computational, functional, and family segregation data to provide ACMG-guided classification of variants' pathogenicity(32). For patients with clinical diagnosis of ADPKD, we also cataloged any accompanying rare (MAF <0.001) missense variants in *PKD1 and PKD2* and rare pLOF or missense variants in 11 other genes that have been described in patients with cystic kidney disease (*ALG8, ALG9, DNAJB11, GANAB, HNF1B, IFT40, LRP5, PKHD1, PRKCSH, SEC61B, SEC63*).

Genetic Ancestry

Briefly, principal components were calculated for HapMap3 samples using high-quality common SNPs, and each sample of the study population was projected onto those principal components (Staples et al. Am J Human Genet. 2018; PMID 29727688). A kernel density estimator was trained on the HapMap3 principal components for five ancestral classes: Admixed American, African, East Asian, European, and South Asian. The likelihood of belonging to each ancestral class was calculated for each sample, and the most likely ancestral class was assigned. If multiple or no classes reached a high likelihood, the sample was assigned as Unknown

Phenotyping using EHR

We used data from the electronic health record (EHR) to determine whether participants had been diagnosed with PKD using 9th and 10th International Classification Diseases (ICD) diagnosis codes (Q61.2, Q61.3, 753.13, 753.12). Additional diagnoses examined included other cystic kidney diseases (Q61.5, Q61.8, Q61.9, 753.10), acquired kidney cyst (N28.1, 593.2), congenital kidney cyst (Q61.00, Q61.01, Q61.02, 753.11, 753.19), or liver cystic disease (Q44.6, 573.8); a composite outcome of "any kidney/liver cyst diagnosis" included all the ICD codes.

Chart Reviews

To enhance genotype-phenotype analyses, additional chart review was performed on 1) patients with ICD diagnosis codes for PKD; 2) patients who had *PKD1* or *PKD2* variants that were PTVs, or missense variants previously described as LP or likely hypomorphic in PKDB; 3) family members of patients with chart-confirmed PKD who also had available exome sequencing data. Chart review including review of imaging was done by at least 1 nephrologist with focus on kidney and liver imaging, nephrolithiasis, cerebral aneurysms, history of dialysis, transplant, family history of ADPKD or cerebral aneurysms, and clinical genetic testing. Additionally, blinded review of imaging was done by at least 1 radiologist with additional review by a senior radiologist with expertise in abdominal imaging for questionable cases, and cases were discussed until consensus was achieved.

Participants who had a PKD ICD code but whose clinical diagnosis/phenotype was more consistent with autosomal recessive polycystic kidney disease (ARPKD), tuberous sclerosis complex (TSC), or congenital abnormalities of kidney and urinary tract (CAKUT) were examined separately from the ADPKD cohort.

Statistical Analyses

Genotype-based approach. First, we examined the prevalence of participants with rare PTV variants (MAF<0.01), or large deletions detected by CNV analysis in *PKD1*, *PKD2*, and any of the 11 other cystic disease-associated genes (*ALG8, ALG9, DNAJB11, GANAB, HNF1B, IFT40, LRP6, PKHD1, PRKCSH, SEC61B, SEC63B)*. We also examined the proportion of participants with rare missense mutations in *PKD1* or *PKD2* listed in the PKDB listed as likely pathogenic, or likely hypomorphic. We then evaluated whether carriers of PTVs or large

deletions in each of the 11 other cystic disease-associated genes were at increased risk of ICD-diagnosed PKD, liver cystic disease, or any kidney/liver cystic disease. Logistic regression was used to assess the association of PTVs with ADPKD determined by ICD code. The severely imbalanced nature of the rare genotypic data may result in bias in the maximum likelihood estimations. Additionally, the strong predictive power of variables such as the truncation of *PKD1* or *PKD2* for diagnosis of ADPKD could result in the problem of separation. For this reason, Firth logistic regression was used which reduces the bias by penalizing the likelihood function by Jeffrey's invariant prior. The regression was adjusted for age, sex, year of first outpatient encounter at Geisinger, and genetically-determined ancestry. First and second-degree relatives were removed in this analysis to avoid false positives due to presence of related individuals. For each gene, the logistic regression coefficient represents the expected change in log-odds of the outcome occurring if a variant is present in the gene. A larger coefficient may be interpreted as a larger increase in the relative risk of the outcome.

Phenotype-based approach. After chart review was performed to confirm PKD on individuals with at least one code for PKD in the EHR, we examined exome sequencing data for PTVs and CNV deletions in *PKD1/PKD2*, reported missense variants in the PKDB, rare and novel *PKD1/PKD2* missense/in frame deletion variants, and rare variants in the 11 other putative cystic genes. Previously reported variants were then grouped based on the PKDB classifications. After identifying first- and second-degree relatives of each participant with ADPKD using PRIMUS(37) and EHR-documented family history, we reconstructed pedigrees of patients with ADPKD. We then reviewed charts of family members of ADPKD patients to evaluate evidence for co-segregation of the variants in question with disease.

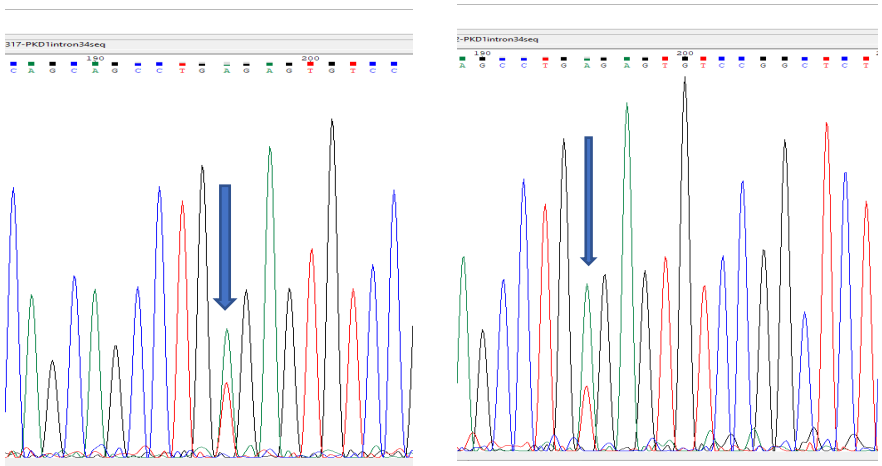
Steps to verify diagnostic accuracy

Limited Sanger sequencing and bioinformatic approach to validate assignment of variants to PKD1 and its pseudo genes

Six pseudogenes on chromosome 16, *PKD1P1-PKD1P6*, replicate with high homology exons 1-33 in *PKD1*. Presence of these pseudogenes can confound interpretation of sequencing data as the sequence captured and amplified from the pseudogenes may be misaligned to *PKD1* and alternatively sequence captured from *PKD1* maybe incorrectly misaligned to one of the pseudogenes. We therefore took additional bioinformatic steps to address this potential issue. The main purpose of this exercise is to identify potential false negative variants that may have been misaligned to one of the pseudogenes in the population exome analysis. We manually forced aligned all reads from *PKD1* and pseudogenes *PKD1P1-PKD1P6* to the *PKD1* sequence and assigned variants to these forced aligned sequences which yielded 39 subject samples with 20 potential variants. While exons 1 to 33 in *PKD1* and its pseudogenes are 98.4% identical (31), there are some unique loci in almost every exon. We assumed that those unique loci were reads that belonged to *PKD1* and not indicative of a variant in a pseudogene read. Therefore, if a read included both a unique *PKD1* site and a rare variant, we classified that read as a real *PKD1* variant. Alternatively, if a read included a unique *PKD1* site but not a variant or a unique pseudogene locus with a variant, we attributed that read as a pseudogene . This process allowed us to verify that 34 participants did not have variants in *PKD1* mis-attributed to a pseudogene. The additional 5 participants' DNA was used for long range PCR followed by Sanger sequencing. All Sanger sequences samples matched the *PKD1* reference sequence. This indicates that the reads assigned to *PKD1* are truly *PKD1* reads and not from the pseudogenes.

We also performed confirmatory sanger sequencing in 4 patients with a suspected *PKD1* PTV and no ADPKD diagnosis. Each sample was amplified using primers within unique intronic regions to amplify the PKD1 site of interest.

Two splice site variants were verified. Sanger sequencing below are the tracing for each sample showing presence of the heterozygous variants at the site indicated by blue arrow.



A third, different splice site variant was a false positive and was removed from the PTV group analysis.

The fourth variant was not a frameshift but an in-frame deletion which had not passed initial quality control but is now classified as an in-frame deletion with no phenotype

Use of available clinical genetic testing

We also examined data from a convenience sample of 17 MyCode participants with cystic kidney disease or ESKD who had undergone clinical genetic testing and had either a pathogenic or VUS in a cystic gene. There was 100% concordance in this convenience sample with variants in cystic genes for all 17 participants detected using MyCode exome sequencing data (10 *PKD1* variants, 1 *HNF1B* variant, 1 *ALG8* variants, 2 *PKHD1* variants, 1 *TSC2*, 2 *PKHD1* partial deletions, 2 large *PKD1* deletions, 3 *PKD2* variants).

eTable 2. Coding Definition for Comorbidities in Table 1

*In some cases we used more stringent predefined algorithms to define cases and controls in EHR

Comorbidities	Codes
ADPKD	Q61.2, Q61.3, 753.13, 753.12
Any kidney/liver cyst	ADPKD (Q61.2, Q61.3, 753.13, 753.12), cystic kidney diseases (Q61.5, Q61.8, Q61.9, 753.10), congenital kidney cyst (Q61.00, Q61.01, Q61.02, 753.11, 753.19), liver cystic disease (Q44.6, 573.8).
Nephrolithiasis	592.0, 592.1, 592.9, 594.0, 594.1, 594.2, 594.8, 594.9, 274.11, N20
ESKD	Dialysis CPT codes 36818-36821, 36831-36833, 90940, 90951-90966, 90967-90970, 90989, 90993, G0257, G9231, S9339, 36147, 90918-90921, 90925, G0308-G027, G0392, G0393
	Dialysis ICD codes 39.27, 39.42, 39.53, 39.54, 585.6, V45.11, V45.12, V56.1, V56.2, V56.31, V56.32, V56.8, V45.1, N18.6, Z91.15, N18.5+Z99.2
	Transplant CPT codes 00868, 50340, 50360, 50365, 50380, S2065, G9231
	Transplant ICD codes 00.91, 00.92, 00.93, 55.53, 55.69, V42.0, 0TY****, Z94.0
Cerebral aneurysm	437.3, I67.1
Cardiac valvular abnormalities	395.0, 746.3, 396.2, and 424.1, 394.0, 424.0, 746.02, 424.3, 746.09, 397.0, I35.0, I35.2, I06.0, I06.2, Q23.0,I35.1, I06.1, Q23.1, I34.2, I05.0, I05.2, Q23.2,I34.0, I05.1, Q23.3,I37.0, I37.2, Q22.0, Q22.1, Q24.3, Q22.2, I37.1, I36.0,I36.2, I07.0, I07.2, Q22.4, I36.1, I07.1
Cystic liver disease	ICD9: 573.8
	ICD10: Q44.6

Hypertension	At least two of the following:
	• At least two records of hypertensive medication
	• At least 3 outpatient high BP readings (Systolic >= 140 or Diastolic >= 90) each over 7 days apart
	• Diagnosis:
	  ○ Diagnosed on patient's problem list
	  ○ At least two ICD9/ICD10 codes within two years

Dyslipidemia	ICD-9: 272
	ICD-10: E78
Diabetes mellitus	At least two of the following: • Diagnosed on patient's problem list

	• Active antidiabetic medication
	• >50% of outpatient labs abnormal in last 2 years
	  ○ Fasting glucose >= 126 mg/dL
	  ○ Non-fasting glucose >= 200 mg/dL
	  ○ HbA1c >= 6.5

Cerebrovascular disease	ICD-9: 362.34, 430.x - 438.x
	ICD-10: G45.x, G46.x, H34.0, I60.x - I69.x
Heart failure	At least one of the following: ➢ Active heart failure diagnosis on patient's problem list ➢ ICD code for diagnosis of heart failure and prescription of heart failure medication • ICD9: 402.01, 402.11, 402.91, 428.0, 420.1, 428.2*, 428.3*, 428.4*, 428.9 • ICD10: I11.0, I13.0, I13.2, I50.1, I50.2*, I50.3*, I50.4*, I50.9 • Heart failure medication: Furosemide, Lasix, Bumetanide, Bumex, Torsemide, Demadex, Ethacrynic acid, Edecrin, Metolazone, Zaroxolyn
Cancer	ICD-9: 140.x - 172.x, 174.x - 195.8, 200.x - 208.x, 238.6, 196.x - 199.x
	ICD-10: C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C81.x - C85.x, C88.x, C90.x - C97.x, C77.x - C80.x
Coronary artery disease	ICD-9: 410, 411.8, 414
	ICD-10: I21 - I25
	ICD-10: I21 - I25

eTable 3. Variants in Patients With Other Kidney Diseases or ADPKD Family History Alone

Study_ID	Sex	Age	Phenotype	Chr:Pos:Ref:Alt	Gene	HGVSp	GHS MAF	gnomAD MAF	Mayo PKDB Call (PKD1/PKD2) ClinVar Call (PKHD1/TSC2
ADPKD265	M	35-40	ARPKD	6:52058349:G:A	PKHD1	Arg496Ter			Pathogenic
ADPKD165	M	45-50	ARPKD						
ADPKD170	M	30-35	ARPKD	6:52024750:A:G	PKHD1	Ile1687Thr			Pathogenic/Likely pathogenic
ADPKD114661	M	35-40	ARPKD	DEL:6:51619080:51619520 6:51659907:G:A	PKHD1	CNV_del Gln3407Ter			Pathogenic
ADPKD333	F	35-40	ARPKD	DEL:6:51619080:51619520 6:51659907:G:A	PKHD1	CNV_del Gln3407Ter			Pathogenic
ADPKD152817	M	20-25	CAKUT	DEL:4:86354529:86354529	PKD2	CNV_del			
ADPKD279	F	25-30	CAKUT	4:78318845:G:A	FRAS1	Ala666Thr	8.64E-06	8.93E-06	
				4:78477959:A:G	FRAS1	Ile2666Val	5.47E-05	1.42E-04	
				16:51141732:T:C	SALL1	Ser164Gly	5.19E-05	0	
ADPKD53641	M	20-25	CAKUT	6:52076302:T:C	PKHD1	Gln141Arg			
ADPKD271	F	55-60	TSC	16:2056705:C:T	TSC2	Pro237Leu			Conflicting
ADPKD71641	F	15-Oct	TSC	16:2080179:C:T	TSC2	Arg1138Ter			Pathogenic
ADPKD328	F	35-40	Fam hx of ADPKD only	4:88038501:CGTAA:C	PKD2	splice donor variant			not in Mayo PKDB Pathogenic/Likely pathogenic
ADPKD139089	F	15-20	Fam hx of ADPKD only	16:2092573:A:G	PKD1	Trp3726Arg			Highly Likely Pathogenic
ADPKD147228	F	10-May	Fam hx of ADPKD only	16:2091121:C:T	PKD1	Trp3922Ter			Definitely Pathogenic
ADPKD157323	F	25-30	Fam hx of ADPKD only	16:2092573:A:G	PKD1	Trp3726Arg			Highly Likely Pathogenic
ADPKD343	F	40-45	Fam hx of ADPKD only						
ADPKD171010	M	40-45	Multiple liver & renal cysts	4:88043387:C:T	PKD2	Arg417Ter			Definitely Pathogenic

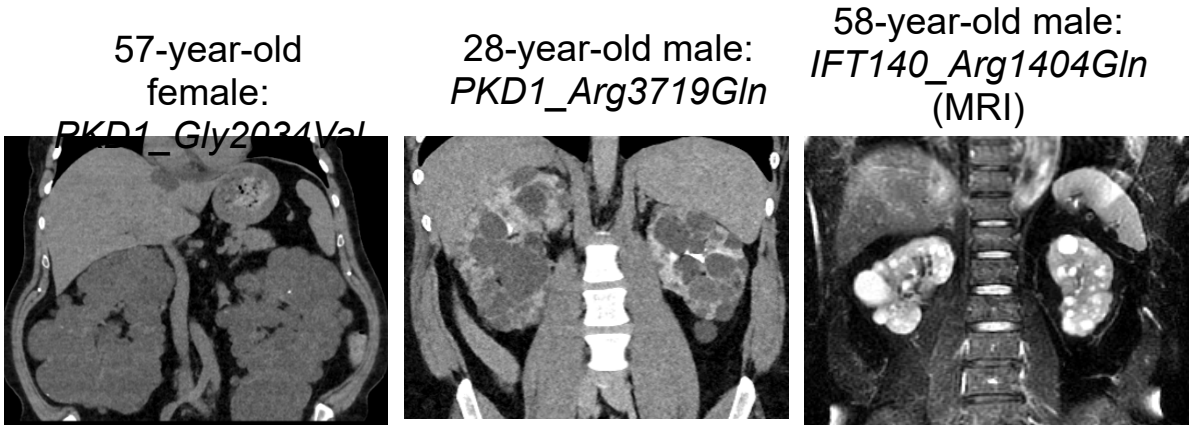
Abbreviations: ADPKD (autosomal dominant polycystic kidney disease), ARPKD (autosomal recessive polycystic kidney disease), CAKUT (congenital anomalies of kidney and urinary tract), tuberous sclerosis complex (TSC), Mayo PKDB (polycystic kidney database)

eTable 8. Typical ADPKD Phenotype Cases With "Other" (Not *PKD1* or *PKD2*) Gene Variants

Gene	Variant	Category
IFT140	Trp459Ter	Truncating
IFT140	Val653Ter	Truncating
IFT140	Met444Val	Missense_VUS
IFT140	Val851Met	Missense_VUS
HNF1B	Arg181Gln	Missense_VUS
HNF1B	Ala373Pro	Missense_VUS
ALG9	Arg112Cys	Missense_VUS
PKHD1	Gln3770His	Missense_VUS

eFigure 1. Typical ADPKD Imaging Phenotype

Example abdominal CTs/MRI for typical ADPKD: bilateral and diffuse distribution of cysts, with moderate or severe replacement of kidney tissue by cysts, where all cysts contribute similarly to total kidney volume

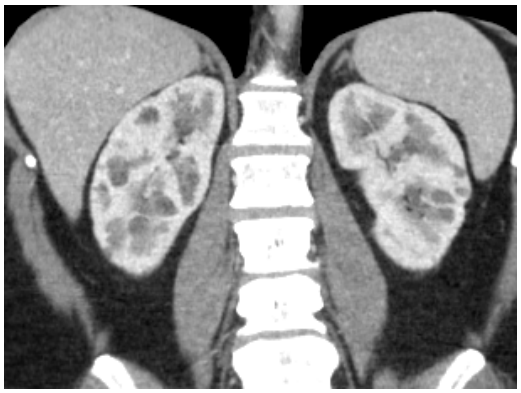


Abbreviations: CT (computed tomography), MRI (magnetic resonance imaging), ADPKD (autosomal dominant polycystic kidney disease)

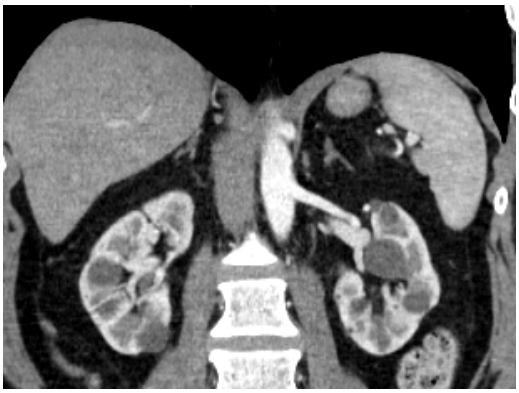
eFigure 2. Mild ADPKD Imaging Phenotype

Example abdominal CTs for mild ADPKD: bilateral/diffuse replacement of kidney tissue by cysts

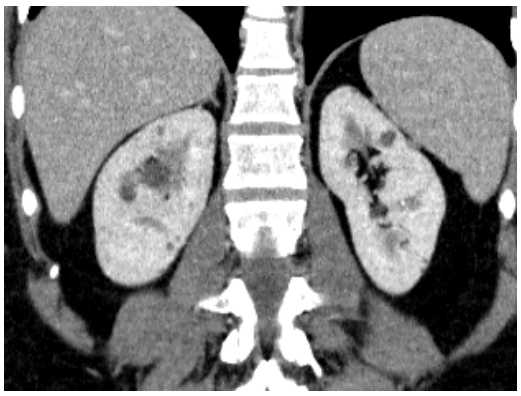
48-year-old female:
PKD1_Asp1165Gly



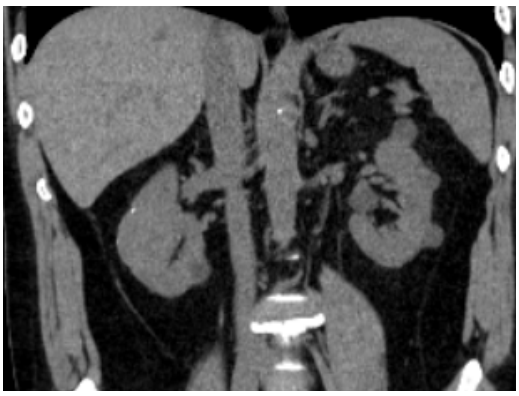
37-year-old male:
PKD2_ Gln300Ter and
PKD2 _Thr1237Met



28-year-old female:
PKD2_Arg786GlyfsTer25



53-year-old male:
GANAB_Asp647Val

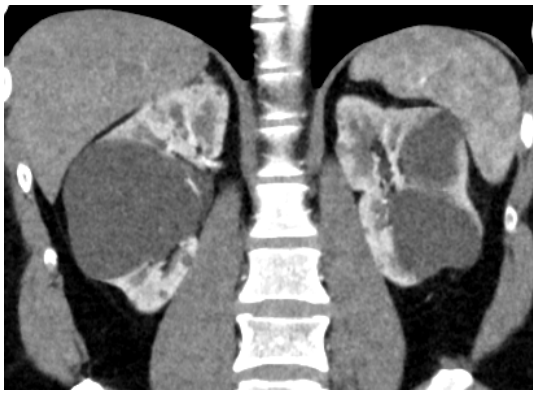


Abbreviations: CT (computed tomography), ADPKD (autosomal dominant polycystic kidney disease)

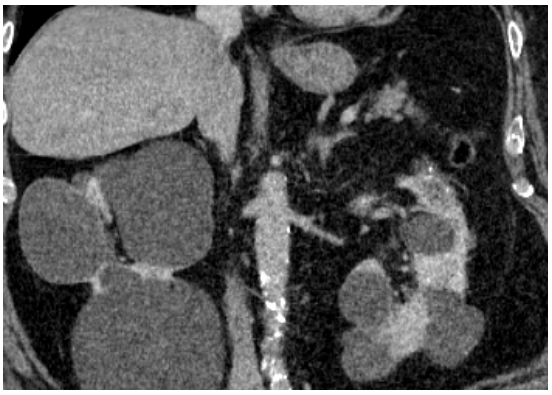
eFigure 3. Atypical ADPKD Imaging Phenotype

Example abdominal CTs for atypical ADPKD: per Mayo clinic imaging classification (see Irazabal MV, et al. J Am Soc Nephrol. 2015;26(1):160-72)

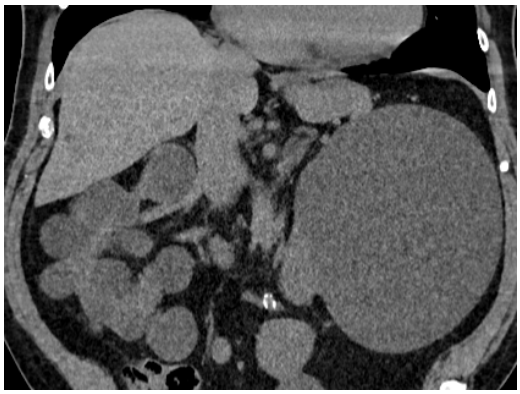
40-year-old male:
PKD1 _Gly2034Val



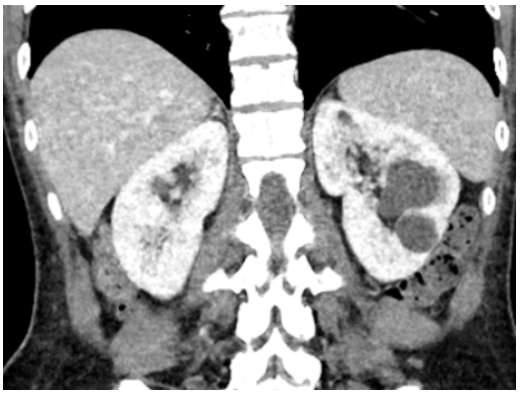
75-year-old male:
IFT140 Trp653Ter



71-year-old male:
PKD1 Ala3889Ser

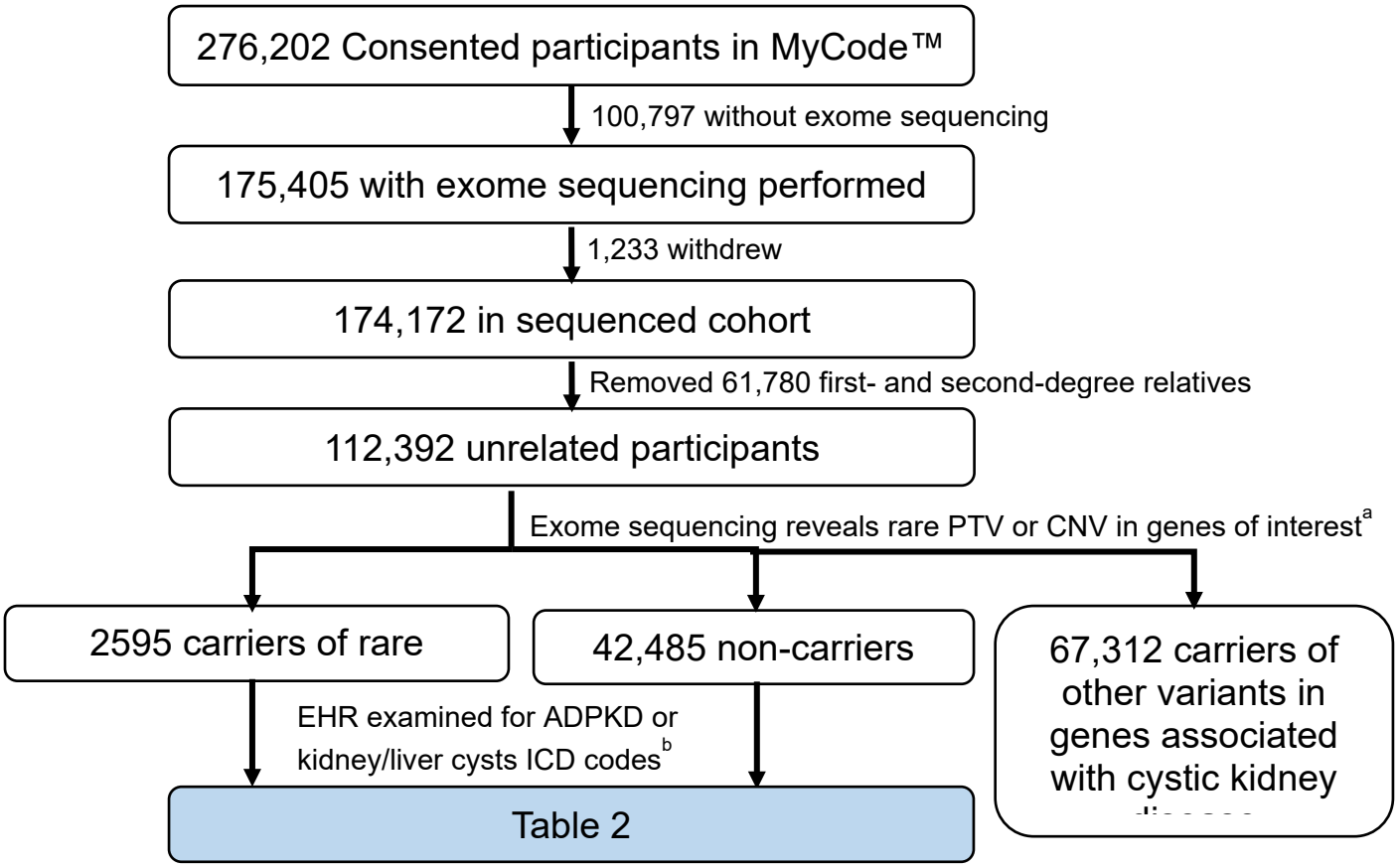


32-year-old female:
GANAB Asp647Val



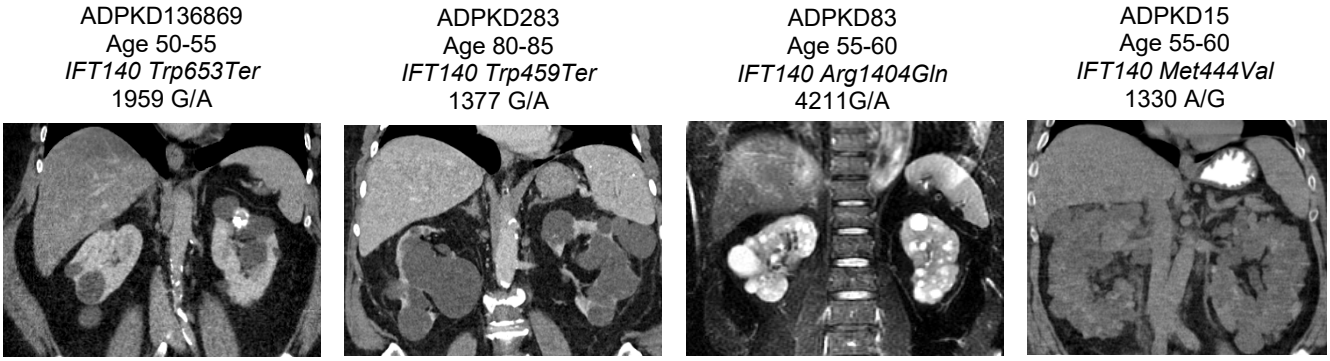
Abbreviations: CT (computed tomography), ADPKD (autosomal dominant polycystic kidney

eFigure 4. Flow Diagram for Cohorts Described in the Study



a. Genes of interest are genes related to cystic kidney or liver disease and include: *PKD1, PKD2, ALG8, ALG9, DNAJB11, GANAB, HNF1B, IFT140, LRP5, PKHD1, PRKCSH* and *SEC63.*
b. ADPKD and kidney/liver cysts ICD diagnosis as defined in Figure 1 and Table 2
c. Abbreviations: ADPKD (autosomal dominant polycystic kidney disease), CNV (copy number variant), EHR(electronic health record), PTV (protein truncating variant)

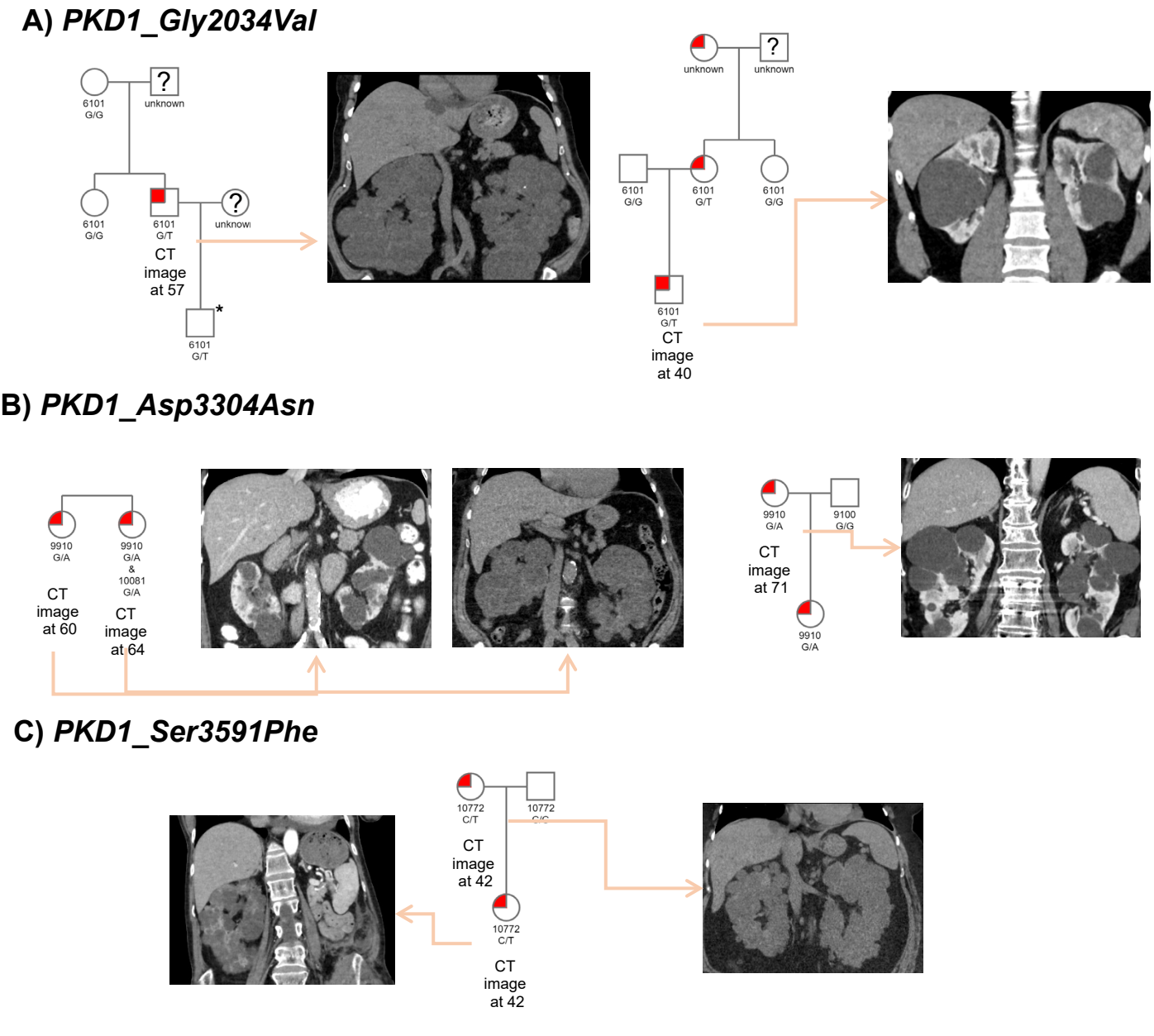
eFigure 5. Sample Pedigrees and Images of Carriers of *IFT140* Variants



Abbreviations: CT (computed tomography)

eFigure 6. Sample Pedigrees and Images of Carriers of Novel *PKD1* Variants

A) Pedigrees and imaging of carriers of novel *PKD1 Gly2034Val* missense variant. Three out of four carriers have ADPKD, while the one without current evidence of ADPKD is <25 years old (*). Non-carrier relatives in either family do not have an ADPKD phenotype.
B) Pedigree and available CT imaging for carriers of *PKD1 Asp3304Asn.* In total, 6/10 have imaging evidence of ADPKD, 3 are shown. Of the 4 carriers with no evidence, 3 do not have abdominal imaging and one is under age 40.
C) Pedigree and imaging for carriers of *PKD1Ser3591Phe, of whom* 3 of 3 have ADPKD. While non-carrier family members do not have ADPKD.



Abbreviations: CT (computed tomography), ADPKD (autosomal dominant polycystic kidney disease)