scientific reports

OPEN

A genotype-to-phenotype approach suggests under-reporting of single nucleotide variants in nephrocystin-1 (*NPHP1*) related disease (UK 100,000 Genomes Project)

Gary Leggatt[1,3,5✉], Guo Cheng[1], Sumit Narain[1], Luis Briseño-Roa[2], Jean-Philippe Annereau[2], The Genomics England Research Consortium*, Christine Gast[1,3], Rodney D. Gilbert[1,4] & Sarah Ennis[1]

Autosomal recessive whole gene deletions of nephrocystin-1 (*NPHP1*) result in abnormal structure and function of the primary cilia. These deletions can result in a tubulointerstitial kidney disease known as nephronophthisis and retinal (Senior–Løken syndrome) and neurological (Joubert syndrome) diseases. Nephronophthisis is a common cause of end-stage kidney disease (ESKD) in children and up to 1% of adult onset ESKD. Single nucleotide variants (SNVs) and small insertions and deletions (Indels) have been less well characterised. We used a gene pathogenicity scoring system (GenePy) and a genotype-to-phenotype approach on individuals recruited to the UK Genomics England (GEL) 100,000 Genomes Project (100kGP) (n = 78,050). This approach identified all participants with *NPHP1*-related diseases reported by NHS Genomics Medical Centres and an additional eight participants. Extreme *NPHP1* gene scores, often underpinned by clear recessive inheritance, were observed in patients from diverse recruitment categories, including cancer, suggesting the possibility of a more widespread disease than previously appreciated. In total, ten participants had homozygous CNV deletions with eight homozygous or compound heterozygous with SNVs. Our data also reveals strong in-silico evidence that approximately 44% of *NPHP1* related disease may be due to SNVs with AlphaFold structural modelling evidence for a significant impact on protein structure. This study suggests historical under-reporting of SNVS in *NPHP1* related diseases compared with CNVs.

Nephrocystin-1 (*NPHP1*) was first localised to chromosome 2q13 before a sizeable homozygous deletion was found in this same region[1]. Saunier et al. and Hildebrandt et al. identified *NPHP1* within the minimum deletion interval along with another gene, *MALL*[2,3]. They both noted additional single nucleotide variants (SNVs) in *NPHP1* but not in *MALL* in patients with a heterozygous deletion. Saunier et al. later demonstrated that the *NPHP1* gene was flanked by two inverted 358 kilobase (kb) low copy repeats (LCRs) and that these both included a smaller LCR of 45 kb, containing the patient's deletion breakpoints[4].

Human chromosome two has arisen from the fusion of two ancestral great ape chromosomes, and the telomere-telomere fusion point is located at 2q13. Telomeres have several repeat sequences, which may explain the presence of these LCRs[5]. LCRs are found in 5–15% of the human genome and provide a point where nonallelic

[1]University of Southampton, Duthie Building (MP 808), Southampton General Hospital, Tremona Road Shirley, Southampton SO16 6YD, UK. [2]Medetia, Imagine Institute for Genetic Diseases, 24 Boulevard du Montparnasse, 75015 Paris, France. [3]Wessex Kidney Centre, Portsmouth Hospitals University NHS Trust, Southwick Hill Road, Cosham, Portsmouth PO6 3LY, UK. [4]Southampton Children's Hospital, Southampton General Hospital, Tremona Road Shirley, Southampton SO16 6YD, UK. [5]University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Tremona Road Shirley, Southampton SO16 6YD, UK. *A list of authors and their affiliations appears at the end of the paper. ✉email: G.Leggatt@soton.ac.uk

homologous recombination (NAHR) mediated cross-over can occur[5,6]. This unequal cross-over results in structural variation, including copy number variants (CNVs) such as duplications, deletions or inversions[5].

NPHP1 is localised to the connecting cilia in photoreceptor cells and to the base of the primary renal cilium (known as the transition zone) of renal epithelial cells of the collecting ducts[7,8]. Primary cilia are near-ubiquitous sensory organelles that protrude from the cell surface[9]. Diseases associated with *NPHP1* are therefore known as ciliopathies[10].

Homozygous deletion of *NPHP1* is the commonest cause of nephronophthisis (20–25% of cases)[11–13]. Nephronophthisis, which means 'vanishing nephron' in Greek, is a tubulointerstitial kidney disease classically divided into three categories, based on the age of presentation: infantile, juvenile, and, less commonly, adolescent[14]. Ultrasound examination reveals small to normal-sized kidneys, with increased echogenicity due to fibrosis and loss of corticomedullary differentiation. Small corticomedullary cysts may develop later in the disease course due to loss of normal tissue[14]. The infantile form differs most significantly from the other forms with *in-utero* onset, oligohydramnios and ESKD in the first two years of life but with enlarged, rather than normal or shrunken kidneys and more widespread cyst development[15]. Juvenile and adolescent nephronophthisis may be associated with polyuria and polydipsia preceding the development of chronic kidney disease as well as growth retardation[14].

In addition to nephronophthisis, 15% of patients have retinal dystrophy (Senior–Løken syndrome) or neurological (Joubert syndrome) phenotypes[16]. The retinal phenotype may be mild[17]. Joubert syndrome (JS) is characterised by cerebellar ataxia, mental retardation, hypotonia, and neonatal respiratory dysregulation[18]. The cerebellar malformation can be seen on MRI scans as the so-called 'molar tooth sign' (MTS). The malformations are less severe in patients with *NPHP1* mutations with milder JS than with other gene causes[18–20].

The homozygous deletion of *NPHP1* is fully penetrant and, therefore, pathognomonic of the nephronophthisis phenotype[11,13]. It is well established that nephronophthisis is a leading cause of ESKD in children[14]. However, more recent data from Snoek et al. identified homozygous deletions in *NPHP1* in 0.9% of adult onset ESKD[21]. One individual first reached ESKD as late as 61 years old. 69% of patients had a prior diagnosis other than nephronophthisis, tubulointerstitial or cystic kidney disease[21]. Although reported in association with nephronophthisis, small insertions and deletions (indels) and single nucleotide variants (SNVs) are less well characterised in *NPHP1*[22].

While single variant association testing has successfully identified common disease variants, it remains underpowered for rare variants even with high case numbers[23]. We, therefore, applied our gene-level scoring system, GenePy[24]. GenePy is a software that transforms variant level data into gene-level data. GenePy generates a score of whole gene pathogenicity for each person by incorporating in silico deleteriousness metrics, population allele frequency and observed zygosity for each variant[24]. The effect of multiple variants within a gene is combined into a single cumulative gene pathogenicity score for each individual. GenePy scores are continuous but do not follow a normal distribution. Higher GenePy scores are intuitive at a base level as higher scores represent greater pathogenic mutational burden due to rare and deleterious mutations. GenePy scores can be used as the basis to prioritise large numbers of candidate variants.

The UK Genomics England (GEL) 100,000 Genomes Project (100kGP), which completed recruitment in 2018, aimed to sequence the genomes of patients with cancer and rare disease and linked this data with digital clinical records. The aims were to improve clinical diagnosis, tailor therapies, enable new scientific discoveries and link research with a clinically integrated genomic medicine service in the UK National Health Service (NHS)[25].

Using whole-genome sequencing and longitudinal clinical data from the 100kGP project; we aimed to identify pathogenic variation in *NPHP1* and associated phenotypes using a genotype-to-phenotype approach. We generated GenePy scores for *NPHP1* for all individuals recruited to the 100kGP and further incorporated copy number variation (CNV) data. We also interrogated the frequency of CNVs and SNVs in individuals recruited to 100kGP and assessed the evidence for SNVs contributing to the phenotype. We applied a gene first approach or genotype-to-phenotype approach by considering the genotypes of all individuals recruited to the 100kGP irrespective of their disease. We then undertook reverse phenotyping to relate the phenotypic spectrum associated with differing mutations.

Methods

Ethical.

The 100,000 genomes project was approved by the National Research Ethics Service Research Ethics Committee for East of England—Cambridge South Research Ethics Committee. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s) as part of the original study.

Recruitment.

Access to the GEL dataset for the project (research registry ID 109) was approved by the renal GEL clinical interpretation partnership (GeCIP), and all analyses were conducted within the GEL research environment. Patients with GEL-defined eligibility criteria for either rare disease or cancer consented to germline whole-genome sequencing linked to clinical phenotype data, including longitudinal electronic health records. Specific Human Phenotype Ontology (HPO) data were collected at the time of recruitment depending on the participant's recruitment category[26]. Additional longitudinal phenotype data included the hospital episode statistics (HES) database containing details of all admission, emergency and outpatient appointments at NHS hospitals in England recorded as ICD-10 codes[27]. The majority of participants provided blood samples, although a small number provided saliva for germline DNA extraction. In addition to germline DNA, some participants recruited to the cancer arm also provided somatic tumour cell DNA.

Phenotype data. Phenotype data was extracted from LabKey Main programme interim data release v11.0[28]. Human Phenotype Ontology (HPO) terms were provided by recruiting Genomic Medicine Centres (GMC) according to the GEL Protocol[29].

Computational pipeline. Genomic data was sourced from an aggregate multi-sample variant call format (VCF; aggV2) main release version 11 (17/12/20). All samples were sequenced with 150 bp paired-end reads in a single lane of an Illumina HiSeq X instrument. Raw sequencing output in the form of a binary base call (BCL) file was processed on the Illumina North Star Version 4 Whole Genome Sequencing Workflow (NSV4, v. 2.6.53.23), which comprises the iSAAC Aligner (v. 03.16.02.19) and Starling Small Variant Caller (v.2.4.7). CNVs were called by the Genomics England pipeline using Illumina canvas software[30]. Samples were aligned to the NCBI Genome Reference Consortium Human Build 38 assembly with decoy sequences. Single sample gVCFs were aggregated using Illumina gVCF genotyper (v.2019.02.26).

A multisource gVCF was extracted from the aggregated gVCF from which the *NPHP1* gene locus (defined using NCBI GRCh38/hg38 (chr2:110,123,335–110,205,062)) was selected using bcftools 1.6[31,32].

Variants meeting minimum quality standards (GQ > 20, mean GQ of > 35, DP > 10 and maximum missingness of 70%) were identified using vcftools 0.1.16 and retained for downstream analysis (n = 12,780)[33].

GenePy. In order to generate whole gene pathogenicity scores, we applied the GenePy algorithm v.1.3 for both coding and non-coding variants[24,34]. The GenePy score (S) for a given gene (g) and individual (h) is the sum of the effect of all variants (k) where each biallelic variant locus (i) in a gene is weighted according to its predicted deleteriousness (using CADD(Di), zygosity and global allele frequency (gnomAD v3.0)) (f).

$$S_{gh} = -\sum_{i=1}^{k} D_i \log_{10}(f_{i1} \cdot f_{i2})$$

The VCF was then annotated using ANNOVAR 1.0 for gnomAD_version 3.0 allele frequency data (f) and the RefGene database of genes[35]. Combined Annotation Dependent Depletion (CADD) v1.6 was used to annotate for deleteriousness (D)[36]. We limited variants to those with CADD PHRED scores greater than 15 (n = 214). This cut-off was chosen as it represents the median value for all possible canonical splice site changes and non-synonymous variants in CADD.

Copy number variation (CNV). The GenePy algorithm is streamlined to include indels and SNVs represented in VCF files but does not systematically incorporate CNV data. Therefore, a GenePy score for CNV whole gene deletions of *NPHP1* was integrated with the whole gene score by arbitrarily assigning (D) as the maximal CADD 1.6 score observed for a stop-gain variant within the *NPHP1* gene and allele frequency using gnomAD structural variant v2 ($f = 1.798 \times 10^{-3}$).

Structural modelling. The mutations identified were modelled both in the Missense 3D Database and using Maestro Suite v13.1 and BioLuminate 4.6 Release 2022-1 (Maestro & BioLuminate, Schrödinger, LLC, New York, NY, 2021)[37–41]. Mutations were identified to be pathogenic following the criteria: (1) the substitution replaces a buried charged residue with an uncharged residue, (2) substitution disrupts all side-chain/side-chain H-bond(s) and/or side-chain/main-chain bond(s) H-bonds, (3) substitution breaks all side-chain/side-chain H-bond(s) and/or side-chain/main-chain H-bond(s) formed by the wild type which was buried, (4) substitution leads to an expansion or contraction of the cavity volume of ≥ 70 Å^3, (5) the substitution breaks a salt bridge formed by wild type which was buried. The maximum N–O bond length is 5.0 Å.

Results

The genotype-to-phenotype approach incorporating SNVs, indels and CNVs into ranked *NPHP1* GenePy scores identified renal, retinal, or neurological phenotypes amongst the top-ranked individuals (see Table 1). Age at onset of disease is likely to be overestimated, as HES data was only available from the date of recruitment onwards. In total 26 participants were identified with biallelic recessive genotypes consistent with monogenic *NPHP1*-related disease. Our approach identified eighteen patients with a renal or retinal phenotype and a genotype constant with the monogenic autosomal recessive disease. The NHS genomic medicine centres have previously reported eight of these. In total ten homozygous CNV whole-gene deletions of *NPHP1* were discovered. Homozygous or compound heterozygous SNVs consistent with monogenic disease were identified in eight patients with renal, retinal, or neurological phenotypes (see Fig. 1).

Assessment of pathogenicity of variants. All SNVs and indels had allele frequencies less than or equal to 4×10^{-3} according to gnomAD v3.0 and CADD 1.6 PHRED scores between 15 and 47 (see Table 2). Assessment of pathogenicity of variants was undertaken following the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology guidelines for the interpretation of sequence variants[42]. Phasing of GEL whole genome sequencing is not available, and confirmation in participants with possible compound heterozygous mutations required analysis of parent genomes where available. However, the compound heterozygous missense mutations must be *in-trans* with heterozygous CNV deletions, whereas homozygous variants do not require phase information. Assessment of parental genomes for segregation of candidate variants was possible in three cases. Participant nine was possibly compound heterozygous for R668C and S666C; however, both variants were present in the father, confirming presence *in-cis*. Participant thirty-three was

Patient number	Consanguinity	Kidney Phenotype (age)	Retinal Phenotype (age)	Neurological Phenotype (age)	Normalised Specific Disease	Type	Sex	Predicted Ancestry	Genotype 1	Genotype 2	GEL Pipeline Gene
1	Yes	11 to 15			Unexplained kidney failure in young people	Proband	M	AFR	Stop gained	Stop gained	*NPHP1*
2	Yes	21 to 25			Unexplained kidney failure in young people	Relative	M	AFR	Stop gained	Stop gained	*NPHP1*
3	No				Intellectual disability	Relative	F	EUR	CNV deletion	Missense	
4	No	31 to 35	46 to 50	36 to 40	Rod-cone dystrophy	Proband	F	EUR	CNV deletion	Missense	*NPHP1*
5	No	36 to 40	41 to 45		Rod-cone dystrophy	Relative	F	EUR	CNV deletion	Missense	*NPHP1*
6	Yes				NA	Cancer	F	EUR	CNV deletion	Missense	
7	No		56 to 60		Rod-cone dystrophy	Proband	M	SAS	Compound het Missense*	Compound het Missense*	
8				6 to 10	Intellectual disability	Proband	M	EUR	Compound het Missense*	Compound het Missense*	
9	No				Intellectual disability	Relative	M		Missense variants *in-cis*	Missense variants *in-cis*	
10	Yes		26 to 30		Rod-cone dystrophy	Proband	F	SAS	Missense	Missense	*PDE6B*
11	No		0 to 1		Rod-cone dystrophy	Proband	F		Stop gained		
12	Yes				Rod-cone dystrophy	Relative	M	AFR	Stop gained		
13	Yes				Undiagnosed metabolic disorders Epilepsy plus other features	Relative	F	SAS	Compound het intronic*	Compound het intronic*	*NSUN2*
14	Yes			0 to 1	Intellectual disability	Proband	M	EUR	Compound het Missense*	Compound het Missense*	*TAFFAZIN*
15	Yes	11 to 15			Congenital Anomaly of the Kidneys and Urinary Tract (CAKUT)	Proband	M	SAS	Missense	Missense	
16					IUGR and IGF abnormalities	Relative	M	SAS	Compound het Missense*	Compound het Missense* (benign)	
17	No			26 to 30	Exceptionally young adult onset cancer	Proband	F	EUR	Compound het intronic*	Compound het intronic*	
18	Yes	41 to 45			NA	Cancer	M	SAS	Compound het Missense*	Compound het Missense* (benign)	
19				6 to 10	Undiagnosed metabolic disorders	Proband	M	AFR	Stop gained	Intronic variant	
20	No	31 to 35			Unexplained kidney failure in young people	Proband	M		CNV deletion	CNV deletion	*NPHP1*
21	No	16 to 20			Primary ciliary dyskinesia	Proband	F	SAS	CNV deletion	CNV deletion	
22	Yes				Unexplained kidney failure in young people	Proband	M	SAS	CNV deletion	CNV deletion	*NPHP1*
23	Yes	11 to 15			Renal tract calcification (or Nephrolithiasis or nephrocalcinosis)	Proband	F	EUR	CNV deletion	CNV deletion	
24	No	31 to 35	46 to 50		Rod-cone dystrophy	Proband	M		CNV deletion	CNV deletion	
25	No	21 to 25			Unexplained kidney failure in young people	Proband	M		CNV deletion	CNV deletion	
26	No	26 to 30	41 to 45		Inherited macular dystrophy	Proband	M	SAS	CNV deletion	CNV deletion	solved
27			21 to 25		Rod-cone dystrophy	Proband	F	EUR	CNV deletion	CNV deletion	
28	Yes	6 to 10			Congenital Anomaly of the Kidneys and Urinary Tract (CAKUT)	Proband	F	EUR	CNV deletion	CNV deletion	*NPHP1*
29		51 to 55	51 to 55		NA	Cancer	F	EUR	CNV deletion	CNV deletion	
30	No				Paediatric motor neuronopathies	Relative	F	EUR	Inframe deletion	Intronic variant	
31	No	56 to 60		56 to 60	Hereditary ataxia	Proband	M		Compound het Missense*	Compound het Missense*	

Continued

32	No				Hereditary ataxia	Relative	F		Compound het Missense*	Compound het Missense*	
33	Yes	11 to 15		16 to 20	Mitochondrial disorders	Proband	M	EUR	Compound het Missense*	Compound het Missense*	
34	No				Mitochondrial disorders	Relative	F		Compound het Missense*	Compound het Missense*	
35	No				Amyotrophic lateral sclerosis or motor neuron disease	Relative	F		Compound het Missense*	Compound het Missense*	
36	No				NA	Cancer	M	EAS	Compound het Missense*	Compound het Missense*	
37					Ultra-rare undescribed monogenic disorders	Relative	F	SAS	Missense		
38	No	76 to 80			Familial Thoracic Aortic Aneurysm Disease	Relative	F	EUR	Stop gained		
39	Yes				Intellectual disability	Relative	M	EAS	Missense*	Intronic variant*	
40	Yes				NA	Cancer	F		Missense	Intronic variant	
41					NA	Cancer	M	AFR	CNV deletion	Intronic variant	
42	No				NA	Cancer	F	AFR	CNV deletion	Intronic variant	
43	No				Intellectual disability	Proband	M	EUR	Missense*	Intronic variant*	
44	No				Holoprosencephaly	Relative	M	SAS	Missense*	Intronic variant*	*MPDZ*
45	No	0 to 1			Congenital Anomaly of the Kidneys and Urinary Tract (CAKUT)	Proband	M	AFR	Missense*	Intronic variant*	
46	No		41 to 45		Rod-cone dystrophy	Proband	M		Stop gained		*EYS*
47				46 to 50	Charcot-Marie-Tooth disease	Proband	M	EUR	Compound het Missense*	Compound het Missense*	
48	Yes			6 to 10	Intellectual disability	Proband	M	EUR	Stop gained		
49	No			21 to 25	Epilepsy plus other features	Proband	M	EUR	Stop gained		
50					NA	Cancer	M	EUR	Stop gained		

Table 1. Top 50 GenePy scores from 78,050 participants across the Genomics England recruitments, including cancer and rare disease. Individuals in dark blue (n = 18) have both a genotype and phenotype consistent with *NPHP1* related disease. Participants highlighted in light blue (n = 8) have a genotype consistent with *NPHP1* related disease but no phenotype recorded in the Genomics England research environment. This may either be because of inaccurate phenotyping, documentation or because they are yet to present to a healthcare professional or because of incompletely penetrant disease. *Currently unable to confirm phase. Genomics England did not permit publication of more detailed phenotype data due to confidentiality concerns.

possibly compound heterozygous for R545K and R444C; however, both variants were present in the participant's mother, confirming presence *in-cis*. Compound heterozygous variants in participant fourteen (R639I and Y78H) were confirmed to segregate with one variant in each parent. No variants were demonstrated to be de novo. No variants were predicted to affect splicing, including many intronic variants with moderate to high CADD scores. A list of these variants, including heterozygous pathogenic variants and those that failed to segregate or were predicted benign in ClinVar, are included in supplemental data (see Supplementary Table 1).

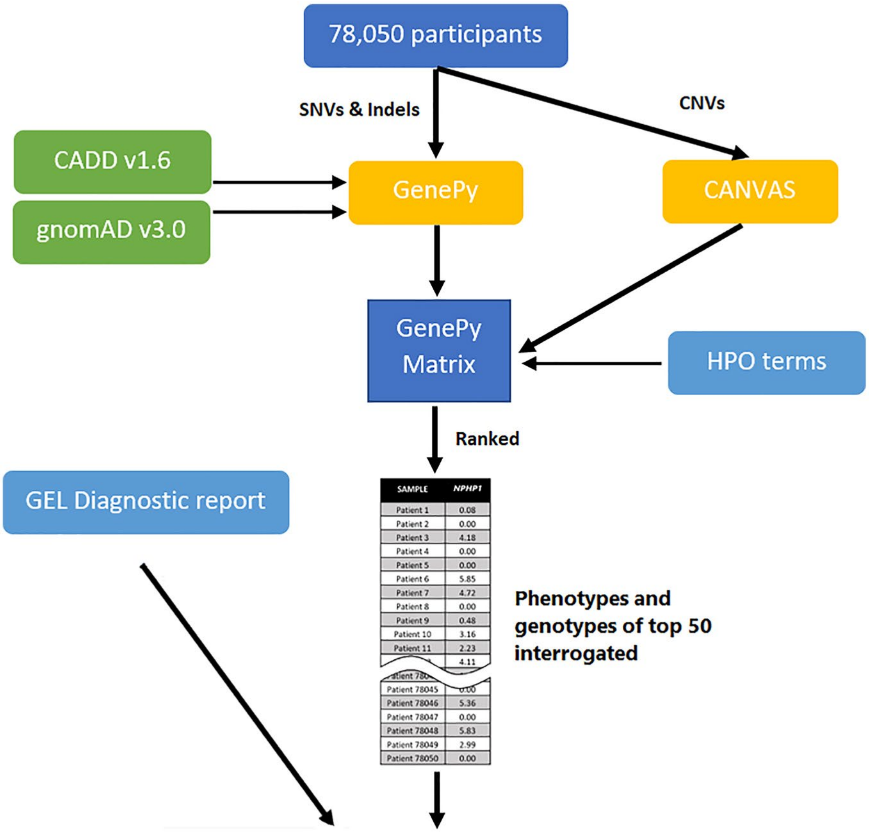
Copy number variant (CNV) analyses.

CNV analysis revealed 424/78,050 participants from 344 families (0.54%) with CNV deletions identified across cancer and rare disease cohorts. The allele frequency of CNV deletions was similar in cancer and rare disease cohorts and twice the rate described in gnomAD SV v2.1 (see Table 3). The latter may not be surprising given exome-based sequencing is more challenging for calling CNVs.

Ten homozygous CNV deletions from ten different families were joint ranked twenty in order of descending GenePy score. Six of these were not previously reported by the GEL genomic medicine centres. Eight participants with homozygous CNV deletions had renal failure with either ESKD (n = 6) or CKD stage 4 (n = 2). Age at first recorded renal phenotype using either HES ICD-10 data or recruitment HPO terms ranged from 6 to 52 years. Four participants had retinal dystrophy with age at first recorded retinal phenotype ranging from 25 to 54. One participant had no phenotype data available but had been recruited due to 'unexplained renal failure in the young', which required onset of ESKD before the age of 50. One participant was recruited into the cancer cohort for malignant melanoma and had ESKD with tubulointerstitial nephritis and hypertensive renal disease by age 52 and hereditary retinal dystrophy by age 54. All ten patients with CNV deletions were deemed pathogenic by ACMG guideline standards due to the predicted null variant nature of the gene deletion giving very strong evidence of pathogenicity.

Single nucleotide variants (SNVs).

Eight participants had a phenotype consistent with renal or retinal disease and homozygous or compound heterozygous SNVs. These included a homozygous stop-gain mutation (two siblings from one family), heterozygous CNV deletions with a missense mutation (two siblings from one family), two different homozygous missense mutations (two individuals from two families), and two different



	Individuals with a biallelic genotype consistent with *NPHP1*-related disease (n=26)	
	Renal or retinal phenotype	No renal or retinal phenotype
Homozygous CNV	10 (4)	0
Heterozygous CNV & missense	2 (2)	2
Stop gained	2 (2)	0
Homozygous missense	2 (0)	0
Compound heterozygous missense*	2 (0)	6
Total	18 (8)	8

Figure 1. Bioinformatic workflow with numbers of participants with biallelic genotypes for monogenic *NPHP1*-related disease. Comparison of genotype-to-phenotype approach (participants with and without penetrant renal or retinal phenotypes) and those reported by the GEL clinical pipeline are in brackets. This approach was fully sensitive to pick up individuals as solved by the GEL clinical interpretation pipeline at the time of writing.

possible compound heterozygous missense mutations (two individuals from two families). The participants with homozygous stop gained and homozygous missense mutations and three of the ten with homozygous CNV deletions were known to have consanguinity. Of these eight the two siblings with homozygous stop-gained variants could be reported as pathogenic according to ACMG guidelines due to the null variant. The other SNVs were rare (4.00E−04) or novel missense variants with strong in-silico evidence of pathogenicity with CADD scores of between 18 and 28.9.

An additional eight participants from eight families had recessive genotypes consistent with monogenic *NPHP1*-related disease but without a documented phenotype. These included six with possible compound heterozygous missense mutations and two compound heterozygous CNV deletions *in-trans* with a missense mutation. One participant without a documented renal or retinal phenotype was recruited to the cancer cohort. The

Chr:Position	Ref	Alt	Consequence	Amino acid	gnomAD frequency	CADD phred	Clinvar	ACMG	Patient Number																									
									1	2	3	4	5	6	7	8	10	14	15	17	20	21	22	23	24	25	26	27	28	29	31	32	35	47
whole gene			del				P	P			▨ t	▨ t	▨ t	▨ t							■	■	■	■	■	■	■	■	■	■				
2:110123943	G	A	ns	R684W	2.79E-05	27.0	VUS	VUS				▨ t	▨ t																					
2:110123988	G	A	ns	R669C	1.40E-05	20.9	VUS	VUS																										
2:110123993	G	C	ns	S667C		24.8		VUS																										
2:110125647	C	A	ns	R640I		29.7		VUS										▨ t																
2:110125677	G	A	ns	S630L	2.00E-04	25.7	VUS	VUS																										▨
2:110131762	A	G	ns	M576T		19.3		VUS								▨																		
2:110131780	T	C	ns	E570G		28.7		VUS			▨ t																							
2:110143602	C	T	ns	R546K	4.00E-04	18.0	CI	VUS												▨											▨	▨	▨	
2:110148020	G	A	ns	R445C	2.79E-05	26.1	VUS	VUS																							▨	▨	▨	
2:110150198	C	T	sg	W437*		47.0		P	■	■																								
2:110163096	T	C	ns	I327V	4.89E-05	15.6	VUS	VUS						▨ t																				
2:110164702	T	C	ns	K253E	1.40E-05	18.2		VUS											■															
2:110168547	C	T	ns	E177K	1.40E-05	26.0		VUS								▨																		
2:110178520	A	G	ns	Y78H	1.20E-03	25.7	CI	VUS										▨ t																▨
2:110179656	C	G	ns	E58Q		26.3		VUS							▨																			
2:110201427	T	C	ns	Y46C		25.0	VUS	VUS												▨														
2:110204926	G	T	ns	R15S	4.88E-05	28.9	CI	VUS							▨		■																	
Phenotype			Renal						▨	▨		▨	▨						▨		▨	▨	▨	▨	▨	▨	▨		▨	▨	▨			
			Retinal									▨	▨		▨		▨								▨		▨	▨		▨				
			Neurological									▨				▨		▨		▨											▨			▨

Table 2. Eighteen pathogenic or assumed pathogenic variants from 26 individuals consistent with autosomal recessive NPHP1-related disease. Dark blue background indicates homozygous variants, and pale blue indicates heterozygous variants; *t* in-trans. *Chr:Position* chromosome and position of 5′ base of variant in *Homo sapiens* (human) genome assembly GRCh38, *Ref* reference allele, *Alt* alternative allele, *Consequence* Consequence of the variant (*del* CNV whole gene deletion of NPHP1, *sg* stop gained variant, *ns* non-synonymous), *Amino acid* amino acid change, *gnomAD frequency* gnomAD genome v3.0 allele frequency (all populations), *CADD phred* Combined Annotation Dependent Depletion score (phred scale) version 1.6, *Clinvar* Clinical significance according to Clinvar (*P* pathogenic, *VUS* variant of uncertain significance, *CI* conflicting interpretations of pathogenicity), *ACMG* assessment of pathogenicity after application of American College of Medical Genetics and Genomics and the Association for Molecular Pathology guidelines for the interpretation of sequence variant (*P* pathogenic, *VUS* variant of uncertain significance).

	Rare disease participants in aggregated gVCF	Cancer participants with germline DNA in aggregated gVCF	gnomAD v2.1 SV
Heterozygous CNV (total number of individuals)	331	83	39
Homozygous CNV (total number of individuals)	9	1	0
Total allele count	349	85	39
Number of individuals	62,898	15,152	21,694
Allele frequency	5.55×10^{-3}	5.61×10^{-3}	1.80×10^{-3}

Table 3. Heterozygous and homozygous *NPHP1* CNV deletions and allele frequencies for rare disease, all cancer germline compared with gnomAD v2.1 SV.

other was the relative of a participant recruited with intellectual disability and no phenotype data available for renal or retinal disease.

Many heterozygous CNV deletions (n = 414) were discovered, including six in the top 50 GenePy rankings. Two heterozygous CNV deletions were identified in participants recruited to the cancer cohort, but no renal or retinal phenotype was recorded. However, the deletions were in-trans with an intronic variant with a CADD 1.6 PHRED score of 15.56 and found in an active enhancer mark (H3K27ac). The remaining heterozygous CNVS were joint ranked 593 and not found with any additional predicted pathogenic variants by CADD 1.6.

AlphaFold structural modelling of variants. To explore if any the pathogenic or assumed pathogenic variants (described in Table 2) could have an impact in the structural integrity of NPHP1, we mapped

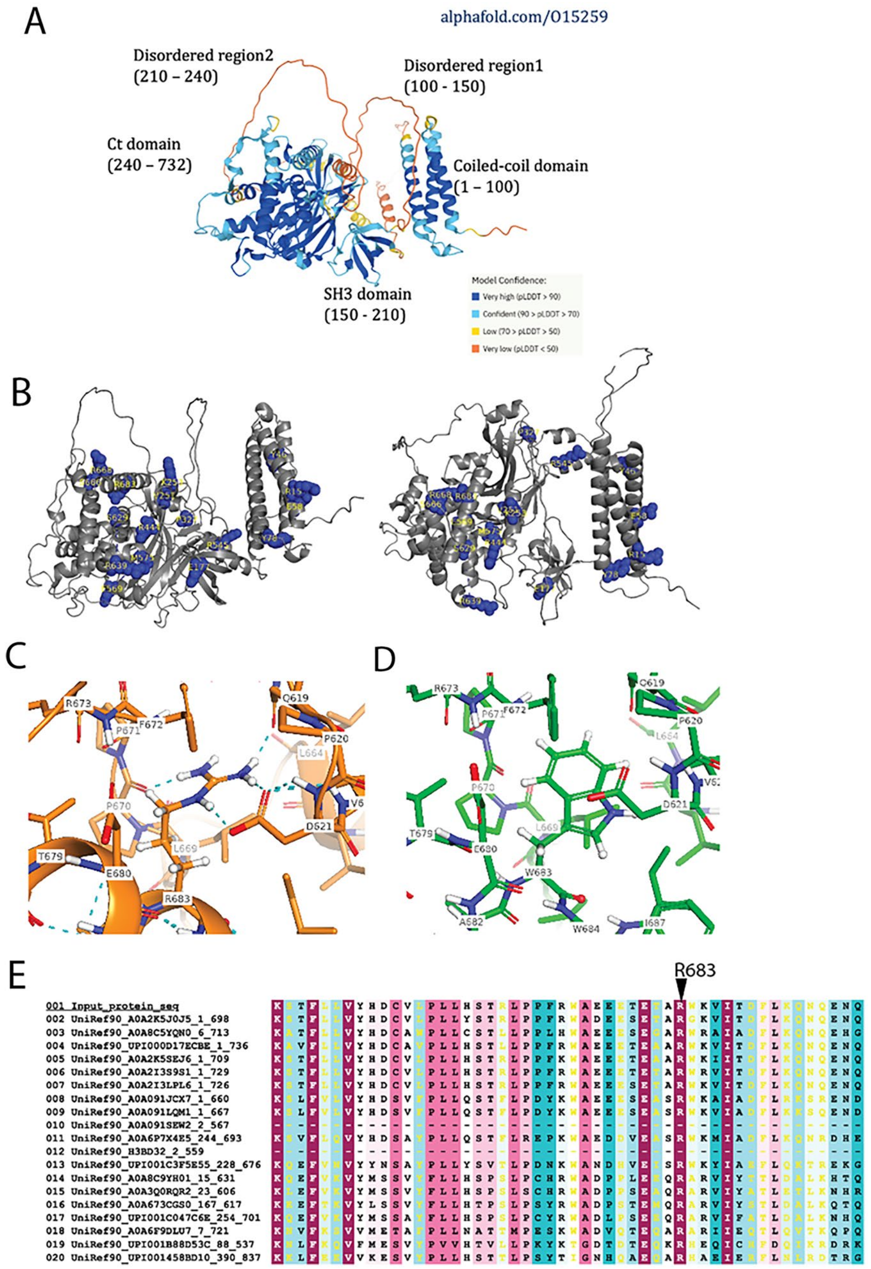

A
alphafold.com/O15259
Disordered region2 (210 – 240)
Disordered region1 (100 - 150)
Ct domain (240 – 732)
Coiled-coil domain (1 – 100)
SH3 domain (150 - 210)
Model Confidence:
Very high (pLDDT > 90)
Confident (90 > pLDDT > 70)
Low (70 > pLDDT > 50)
Very low (pLDDT < 50)
B
C
R673
P671
F672
Q619
P620
L664
P670
D621
L669
T679
E680
R683
D
V62
W683
A682
W684
I687
E
R683
001 Input_protein_seq
002 UniRef90_A0A2K5J0J5_1_698
003 UniRef90_A0A8C5YQN0_6_713
004 UniRef90_UPI000D17ECBE_1_736
005 UniRef90_A0A2K5SEJ6_1_709
006 UniRef90_A0A2I3S9S1_1_729
007 UniRef90_A0A2I3LPL6_1_726
008 UniRef90_A0A091JCX7_1_660
009 UniRef90_A0A091LQM1_1_667
010 UniRef90_A0A091SEW2_2_567
011 UniRef90_A0A6P7X4E5_244_693
012 UniRef90_H3BD32_2_559
013 UniRef90_UPI001C3F5E55_228_676
014 UniRef90_A0A8C9YH01_15_631
015 UniRef90_A0A3Q0RQR2_23_606
016 UniRef90_A0A673CGS0_167_617
017 UniRef90_UPI001C047C6E_254_701
018 UniRef90_A0A6F9DLU7_7_721
019 UniRef90_UPI001B88D53C_88_537
020 UniRef90_UPI001458BD10_390_837

Figure 2. Alphafold modelling of NPHP1 variants. (**A**) Model of the NPHP1 polypeptide as predicted by AlphaFold and manually annotated by the main regions predicted. (**B**) Mapping of mutations identified by GenePy within the NPHP1 structural model. (**C**) Residue R683 and its predicted side-chain entourage in NPHP1 model. (**D**) Residue W683 and its predicted side-chain entourage in NPHP1-R683W mutant model. In (**C**) and (**D**) H-bonds are depicted as cyan dotted lines. (**E**) Multiple sequence alignment showing conservation of R683 in NPHP1 orthologues (mainly vertebrates); the colour coding ranges according to the level of phylogenetic conservation, from variable (cyan) and average (white) to conserved (maroon).

the most conspicuous changes into the structural model of NPHP1 (UNIPROT#O15259) recent released by AlphaFold[43,44]. In short, the AlphaFold algorithm is an AI system based on the analysis of contact maps of co-evolving residues extracted from multiple sequence alignments (MSAs) of sequences, which feed two neural networks: one trained on Protein Data Bank (PDB) structures to predict interatomic angles and distances; another, trained to score the geometry and structural accuracy. This model identifies five different structural regions within the NPHP1 polypeptide chain (see Fig. 2): an N-terminal domain enriched by coil-coiled helical structures (residues 1 to 100, approximately), a first disordered region 1 (residues 100 to 150), a SH3 domain (r.150 to 210), a second disordered region (210 to 240), and a globular C-terminal domain rich in both beta-sheets and alpha helices (240 to 732). The quality of the prediction in generally consistently high in most of these regions, with the exclusion the two disordered regions. Experimental structural data is available for the N-terminal coil-

coiled domain, like that of the BAG domain protein family, known to mediate the anti-apoptotic functions, and the SH3 domains thought to be involved in cell adhesion[45,46].

The structural model of NPHP1 allowed assessment of the emplacement and possible impact on the structure of the mutations identified using both a missense predictor and Maestro Suite v13.1 and BioLuminate 4.6 Release 2022-1 (Maestro & BioLuminate, Schrödinger, LLC, New York, NY, 2021)[37–41]. Firstly, all the mutations were found located in predicted domain-containing regions (Supplementary Fig. 1B and Supplementary Table 2). Secondly, most mutations were predicted not to have a significant impact on the structure—except for M575T and R683W; however, for M575T a significant change in the cavity is predicted given the residues size reduction (not shown). Also, for R684W on the other side, the prediction on change include: this substitution replaces a buried charged residue with an uncharged residue (TRP), disrupts multiple H-bond with the residues P670, D621, Q619, and disrupts a salt bridge formed by NE atom of R683 and OD1 atom of D621. Along with a contraction of 90.504 $\mathring{A}^3$ (Supplementary Fig. 1C,D; Supplementary Table 3). Finally, we confirmed that R683 was phylogenetically conserved (Supplementary Fig. 1E).

Discussion

Eighteen participants were identified with genotypes consistent with a diagnosis of recessive *NPHP1*-related disease and renal or retinal phenotypes. Of these, ten participants with homozygous *NPHP1* deletions could be considered pathogenic according to ACMG guidelines. Eight patients had homozygous and compound heterozygous SNVs, including two heterozygous variants of uncertain significance (VUS) in-trans with CNV deletions. Of these, only the two with homozygous stop gain mutations can be classified as pathogenic according to ACMG criteria.

Evidence for pathogenicity.

The other SNVs were rare (4.00E−04) or novel missense variants with strong in-silico evidence of pathogenicity. The CADD PHRED scores were between 18 and 28.9. At the risk of losing some causative variants, we chose a cut off CADD score of 15 and above. This cut-off was chosen as it represents the median value for all possible canonical splice site changes and non-synonymous variants in CADD. Variants with CADD PHRED scores over 15 are predicted to be in the top 0.5% of most deleterious substitutions that can occur in the human genome. However, as is true of missense variants in many other genes they cannot be officially classified as clinically pathogenic without expensive and time consuming in-vitro or in vivo functional assays. This poses a wider problem with the bottle neck to clinical interpretation now moving from high through put genomics to a lack of high throughput functional assays to confirm pathogenicity. Given these variants are rare or even private to a specific family, the cost of developing 'well-established in vitro or in vivo functional studies' is likely to be prohibitive and result in a loss of translational benefit to patients and their families. Where functional experimental evidence is available it supports roles of some protein domains in apoptosis and cell adhesion[45,46]. AlphaFold structural modelling of these SNVs additionally predicted significant impact on protein structure.

Identification of participants who may yet to develop disease.

An additional eight participants had a recessive SNV genotype consistent with monogenic *NPHP1*-related disease but no documented renal or retinal phenotype. One was recruited with cancer, and three were 'unaffected' relatives. Generally, phenotype data are scarce for participants recruited to cancer and for rare disease participants other than the proband. The GEL recruitment process for cancer participants does not routinely require documentation of kidney function, even though this would invariably have been tested in all patients undergoing treatment for their cancer. 'Unaffected' relatives with no phenotype recorded at recruitment and no longitudinal ICD-10 HES data could be due to either not being clinically evaluated, having subclinical disease, or having a clinical disease that is not documented, that may present later or is not fully penetrant.

When assessing compound heterozygosity, the difficulty is that some variants may lie on the same chromosome. This cannot currently be assessed with short-read data such as that used by Genomics England. We confirmed the presence in-trans in one participant through parental assessment. Their young age may explain the lack of phenotype data for probands with compound heterozygous variants. They may yet develop significant disease in their lifetime (two were under 15 years old). Array-based data on homozygous *NPHP1* CNV deletions from Snoek et al. suggests that individuals may not reach end-stage kidney disease until the seventh decade[21]. Longitudinal follow up of these participants using HES data from GEL will be required to clarify penetrance and confirm other cases.

Benefits and limitations of gene first approach using GenePy.

The large number of participants in the 100kG project offered a unique opportunity to identify the full spectrum of pathogenic variation in *NPHP1* and related diseases. We therefore applied a more objective, agnostic genotype-to-phenotype approach. A significant benefit over a case–control approach is that it precludes the need to exclude participants from analyses based on predicted ancestry or relatedness. This is particularly important given the under-representation of individuals from non-European ancestries in the genetic literature. As with all approaches that implement short read data, this approach is subject to the same limitations. This includes reduced sensitivity to detect variants in highly variable regions and absence of phase representation. A lack of phase data is particularly problematic in adult-onset disease where parent DNA may not be available to determine zygosity, therefore limiting the ability to assess compound heterozygous variants. As with all large-scale data, GenePy scoring is dependent upon data quality and the elimination of systematic bias and technical artefacts. However, the genomics England data used was of high integrity, with all samples sequenced with 150 bp paired-end reads in a single lane of an Illumina HiSeq X.

Potential for diagnostic uplift. Our data support the benefit of a gene first approach that allows the identification of genotypes and associated phenotypes unbiased with respect to their phenotype. This approach can be applied to any gene and includes participants who may yet to develop disease. Notably, the genotype–phenotype approach did not miss any individuals reported as solved by the GEL clinical interpretation pipeline at the time of writing. According to GMC exit questionnaire data, nine participants from seven families were reported by GEL using the clinical interpretation pipeline. However, one was reported for another retinal gene (*PDE6B*) other than *NPHP1*. This highlights the possibility that some participants have more than one monogenic molecular diagnosis. In one whole exome study of 7374 patients, 101 (4.9%) had diagnoses that involved more than one locus, and associated phenotypes could be distinct or overlapping[47]. All patients with more than one genetic diagnosis were known to be consanguineous.

The contribution of SNVs to *NPHP1*-related disease. Homozygous CNV deletions of *NPHP1* are a well-defined cause of kidney failure. However, the contribution of SNVs and indels to kidney and retinal disease is less well characterised. This study suggests that pathogenic *NPHP1* variants are more common than previously estimated and that SNVs may account for up to 44% of diagnoses of *NPHP1*-related monogenic disease. This may represent a conservative estimate as participants with no recorded renal or retinal phenotype (due to young age or lack of identification) may yet demonstrate these phenotypes.

Conclusion

The gene first (genotype-to-phenotype) approach combined with a gene pathogenicity score (GenePy) enabled the identification of the full molecular genetic spectrum of nephrocystin-1 (*NPHP1*) in the UK 100,000 genomes project irrespective of relatedness or ancestry. Supportive data from AlphaFold structural modelling suggests that up to 44% of diagnoses of *NPHP1*-related disease may be caused by SNVs in the addition to the well described CNV deletions. Increasing availability of phased genomic data from long read sequencing is expected to further the power of this approach and allow confident assessment of compound heterozygous variants. Longitudinal follow up will provide data on the ability to predict future disease states in high GenePy scoring individuals.

Data availability

Full data are available in the Genomic England Secure Research Environment. Access is controlled to protect the privacy and confidentiality of participants in the Genomics England 100,000 Genomes Project and to comply with the consent given by participants for use of their healthcare and genomic data. Access to full data is permitted to researchers after registration with a Genomics England Clinical Interpretation Partnership (GeCIP) (https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/) and by contacting the corresponding author upon reasonable request.

Received: 26 August 2022; Accepted: 23 March 2023
Published online: 09 June 2023

References

1. Konrad, M. *et al.* Large homozygous deletions of the 2q13 region are a major cause of juvenile nephronophthisis. *Hum. Mol. Genet.* **5**, 367–371 (1996).
2. Hildebrandt, F. *et al.* A novel gene encoding an SH3 domain protein is mutated in nephronophthisis type 1. *Nat. Genet.* **17**, 149–153 (1997).
3. Saunier, S. A novel gene that encodes a protein with a putative src homology 3 domain is a candidate gene for familial juvenile nephronophthisis. *Hum. Mol. Genet.* **6**, 2317–2323 (1997).
4. Saunier, S. *et al.* Characterization of the NPHP1 locus: Mutational mechanism involved in deletions in familial juvenile nephronophthisis. *Am. J. Hum. Genet.* **66**, 778–789 (2000).
5. Yuan, B. *et al.* Comparative genomic analyses of the human NPHP1 locus reveal complex genomic architecture and its regional evolution in primates. *PLoS Genet.* **11**, e1005686 (2015).
6. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
7. Fliegauf, M. *et al.* Nephrocystin specifically localizes to the transition zone of renal and respiratory cilia and photoreceptor connecting cilia. *J. Am. Soc. Nephrol.* **17**, 2424–2433 (2006).
8. Eley, L., Moochhala, S. H., Simms, R., Hildebrandt, F. & Sayer, J. A. Nephrocystin-1 interacts directly with Ack1 and is expressed in human collecting duct. *Biochem. Biophys. Res. Commun.* **371**, 877–882 (2008).
9. Reiter, J. F. & Leroux, M. R. Genes and molecular pathways underpinning ciliopathies. *Nat. Rev. Mol. Cell Biol.* **18**, 533–547. https://doi.org/10.1038/nrm.2017.60 (2017).
10. Wolf, M. & Hildebrandt, F. Nephronophthisis. *Pediatr. Nephrol.* **26**, 181–194. https://doi.org/10.1007/s00467-010-1585-z (2011).
11. Halbritter, J. *et al.* Identification of 99 novel mutations in a worldwide cohort of 1,056 patients with a nephronophthisis-related ciliopathy. *Hum. Genet.* **132**, 865–884 (2013).
12. Hildebrandt, F., Attanasio, M. & Otto, E. Nephronophthisis: Disease mechanisms of a ciliopathy. *J. Am. Soc. Nephrol.* **20**, 23–35. https://doi.org/10.1681/ASN.2008050456 (2009).
13. Braun, D. A. & Hildebrandt, F. Ciliopathies. *Cold Spring Harb. Perspect. Biol.* **9**, 028191 (2017).
14. Hildebrandt, F. & Zhou, W. Nephronophthisis-associated ciliopathies. *J. Am. Soc. Nephrol.* **18**, 1855–1871. https://doi.org/10.1681/ASN.2006121344 (2007).
15. Hildebrandt, F. *et al.* Molecular genetic identification of families with juvenile nephronophthisis type 1: Rate of progression to renal failure. *Kidney Int.* **51**, 261–269 (1997).
16. Caridi, G. *et al.* Renal-retinal syndromes: Association of retinal anomalies and recessive nephronophthisis in patients with homozygous deletion of the NPH1 locus. *Am. J. Kidney Dis.* **32**, 1059–1062 (1998).
17. Birtel, J. *et al.* NPHP1 gene-associated nephronophthisis is associated with an occult retinopathy. *Kidney Int.* **100**, 1092–1100 (2021).
18. Parisi, M. A. *et al.* The NPHP1 gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome. *Am. J. Hum. Genet.* **75**, 82–91 (2004).

19. Caridi, G. *et al.* Nephronophthisis type 1 deletion syndrome with neurological symptoms: Prevalence and significance of the association. *Kidney Int.* **70**, 1342–1347 (2006).
20. Castori, M. *et al.* NPHP1 gene deletion is a rare cause of Joubert syndrome related disorders. *J. Med. Genet.* **42**, e9 (2005).
21. Snoek, R. *et al.* NPHP1 (Nephrocystin-1) gene deletions cause adult-onset ESRD. *J. Am. Soc. Nephrol.* **29**, 1772–1779 (2018).
22. Wang, Y., Chen, F., Wang, J., Zhao, Y. & Liu, F. Two novel homozygous mutations in NPHP1 lead to late onset end-stage renal disease: A case report of an adult nephronophthisis in a Chinese intermarriage family. *BMC Nephrol.* **20**, 1–5 (2019).
23. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785 (2010).
24. Mossotto, E. *et al.* GenePy—A score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinform.* **20**, 254 (2019).
25. Genomics England. *The National Genomics Research and Healthcare Knowledgebase v5*. https://doi.org/10.6084/m9.figshare.4530893.v5 (2019).
26. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
27. World Health Organization. *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelins. ICD-10 Classification of Mentall and Behavioural Disorders* (1992).
28. Nelson, E. K. *et al.* LabKey Server: An open source platform for scientific data integration, analysis and collaboration. *BMC Bioinform.* **12**, 71 (2011).
29. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
30. Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics* **32**, 2375–2377 (2016).
31. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
32. NPHP1 nephrocystin 1 [Homo sapiens (human)]—Gene—NCBI. https://www.ncbi.nlm.nih.gov/gene/4867
33. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
34. GitHub—UoS-HGIG/GenePy-1.3: Latest GenePy version compatible with both GRCh38 and GRCh37 (hg19). https://github.com/UoS-HGIG/GenePy-1.3
35. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
36. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
37. Khanna, T., Hanna, G., Sternberg, M. J. E. & David, A. Missense3D-DB web catalogue: An atom-based analysis and repository of 4M human protein-coding genetic variants. *Hum. Genet.* **140**, 805–812 (2021).
38. Ittisoponpisan, S. *et al.* Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?. *J. Mol. Biol.* **431**, 2197–2212 (2019).
39. Zhu, K. *et al.* Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins Struct. Funct. Bioinform.* **82**, 1646–1655 (2014).
40. Salam, N. K., Adzhigirey, M., Sherman, W., Pearlman, D. A. & Thirumalai, D. Structure-based approach to the prediction of disulfide bonds in proteins. *Protein Eng. Des. Sel.* **27**, 365–374 (2014).
41. Beard, H., Cholleti, A., Pearlman, D., Sherman, W. & Loving, K. A. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLoS One* **8**, e82849 (2013).
42. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
43. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
44. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
45. Mannella, V. *et al.* The N-terminal domain of NPHP1 folds into a monomeric left-handed antiparallel three-stranded coiled coil with anti-apoptotic function. *ACS Chem. Biol.* **14**, 1845–1854 (2019).
46. Le Maire, A. *et al.* Solution NMR structure of the SH3 domain of human nephrocystin and analysis of a mutation-causing juvenile nephronophthisis. *Proteins Struct. Funct. Bioinform.* **59**, 347–355 (2005).
47. Posey, J. E. *et al.* Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* **376**, 21–31 (2017).

Acknowledgements

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project 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 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. David Rinaldo for his helpful advice with the modelling using Maestro & BioLuminate suite (Schrodinger, NY). Analyses of copy number variants in *NPHP1* were initially performed by Dr Guillermo del Angel, Alexion Pharmaceuticals Inc. This study was supported by the National Institute for Health Research (NIHR) Southampton Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions

Analysis and interpretation of the whole genome sequencing data were performed principally by G.L., G.C., S.N. and S.E. G.L. prepared Tables 1, 2, and 3 and Supplementary Table 1. Expertise regarding proteomics was provided by L.B.R. and J.P.A. L.B.R. prepared Supplementary Fig. 1 and Supplementary Tables 2 and 3. Clinical perspective was provided by G.L., C.G. and R.D.G. The manuscript was written by G.L. and L.B.R., with input from all other authors.

Competing interests

JPA and LBR are shareholders at Medetia Pharmaceuticals; JPA and LBR are authors in the patent application WO2109/075369A1, currently on PCT National examination phase. There are no other conflicts of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-32169-4.

Correspondence and requests for materials should be addressed to G.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Genomics England Research Consortium

J. C. Ambrose[6], P. Arumugam[6], R. Bevers[6], M. Bleda[6], F. Boardman-Pretty[6,7], C. R. Boustred[6], H. Brittain[6], M. A. Brown[6], M. J. Caulfield[6,7], G. C. Chan[6], A. Giess[6], J. N. Griffin[8], A. Hamblin[6], S. Henderson[6,7], T. J. P. Hubbard[6], R. Jackson[6], L. J. Jones[6,7], D. Kasperaviciute[6,7], M. Kayikci[6], A. Kousathanas[6], L. Lahnstein[6], A. Lakey[6], S. E. A. Leigh[6], I. U. S. Leong[6], F. J. Lopez[6], F. Maleady-Crowe[6], M. McEntagart[6], F. Minneci[6], J. Mitchell[6], L. Moutsianas[6,7], M. Mueller[6,7], N. Murugaesu[6], A. C. Need[6,7], P. O'Donovan[6], C. A. Odhams[6], C. Patch[6,7], D. Perez-Gil[6], M. B. Pereira[6], J. Pullinger[6], T. Rahim[6], A. Rendon[6], T. Rogers[6], K. Savage[6], K. Sawant[6], R. H. Scott[6], A. Siddiq[6], A. Sieghart[6], S. C. Smith[6], A. Sosinsky[6,7], A. Stuckey[6], M. Tanguy[6], A. L. Taylor Tavares[6], E. R. A. Thomas[6,7], S. R. Thompson[6], A. Tucci[6,7], M. J. Welland[6], E. Williams[6], K. Witkowska[6,7], S. M. Wood[6,7] & M. Zarowiecki[6]

[6]Genomics England, London, UK. [7]William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK. [8]University of East Anglia, Norwich, UK.