


Exome analysis links kidney malformations to developmental disorders and reveals causal genes

Received: 16 November 2024

Accepted: 18 July 2025

Published online: 07 August 2025

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Congenital anomalies of the kidneys and urinary tract (CAKUT) are developmental disorders that commonly cause pediatric chronic kidney disease and mortality. We examine here rare coding variants in 248 CAKUT trios and 1742 singleton CAKUT cases and compare them to 22,258 controls. Diagnostic and candidate diagnostic variants are detected in 14.1% of cases. We find a significant enrichment of rare damaging variants in constrained genes expressed during kidney development and in genes associated with other developmental disorders, suggesting phenotype expansion. Consistent with these data, 18% of CAKUT patients with diagnostic variants have neurodevelopmental or cardiac phenotypes. We identify 40 candidate genes, including *CELSR1*, *SSBP2*, *XPO1*, *NR6A1*, and *ARID3A*. Two are confirmed as CAKUT genes: *ARID3A* and *NR6A1*. This study suggests that many yet-unidentified syndromes would be discoverable with larger cohorts and cross-phenotype analysis, leading to clarification of the genetic and phenotypic spectrum of developmental disorders.

Congenital anomalies of the kidney and urinary tract (CAKUT) are diagnosed in 0.5% of live births^{1,2} and account for 50% of pediatric kidney failure. CAKUT includes a spectrum of developmental defects affecting the kidney number, size, morphology, and position, and the lower urinary tract (outflow abnormalities)³. Among CAKUT patients, those with anomalies affecting kidney size and morphology (hypoplastic kidneys, dysplastic kidneys, multicystic dysplastic kidneys) and number (kidney agenesis) have a high risk of progression to kidney failure³. Identification of a genetic cause for CAKUT can impact prognosis and clinical management^{4,5}. Although multiple causative genes for CAKUT have been identified, surveys based on modest-sized cohorts indicate that they only account for 5–20% of cases⁶. The two genes accounting for the highest proportion of genetic diagnoses of CAKUT are *HNF1B* and *PAX2*^{7,8}. Despite the high diagnostic yield, genetic testing is not uniformly implemented in pediatric practice.

While CAKUT most often occurs in isolation, it can also manifest in conjunction with developmental defects in other organs. For example, 30% of infants with congenital heart disease also have urinary tract defects^{9,10}. Shared pathogenesis of CAKUT and neurodevelopmental disorders, such as autism spectrum disorder and intellectual disability

(ID), has also been suggested^{6,11}. For example, Bardet-Biedl, CHARGE, Cornelia de Lange syndrome, and Smith-Magenis syndromes are associated with CAKUT, congenital heart disease, and autism spectrum disorder. Similarly, structural variants such as the 22q11.2 deletion (DiGeorge syndrome), 7q11.23 deletion (Williams syndrome), 16p11.2 deletions/duplications, and 17q12 deletion are associated with all these disorders¹². However, the possibility of shared underlying genetic mechanisms and pathways between CAKUT and other developmental disorders has not been systematically examined.

The majority of CAKUT is detected in patients without a known family history of disease, suggesting the possibility of de novo variants or recessive disease. The analysis of de novo variants has been a successful approach to gene discovery for developmental disorders^{13–16}. Another successful approach is gene-burden case-control association studies, enabling the identification of genes for schizophrenia, amyotrophic lateral sclerosis^{17,18}, retinal diseases¹⁹, epilepsy²⁰, and CAKUT²¹. Both approaches take advantage of large population databases to restrict the analysis and test enrichment for exceedingly rare, predicted deleterious variants, which are more likely to be associated with developmental anomalies like CAKUT. We hypothesized that similar analyses could identify additional CAKUT-causing genes.

✉ e-mail: Hila.MiloRasouly@columbia.edu; ag2239@cumc.columbia.edu

In this work, we show that the rate of diagnostic and candidate diagnostic variants in cases with CAKUT is 14.1%. Using de novo and gene-burden case-control association analysis, we identify 40 candidate genes, and confirm two of them as causal CAKUT genes: *ARID3A*

and *NR6A1*. This study provides evidence for the shared genetics of CAKUT with other developmental disorders.

Results

High genetic heterogeneity of kidney anomalies

The study included a total of 1990 unrelated probands with CAKUT ascertained based on presence of kidney defects including unilateral or bilateral renal agenesis, hypoplastic kidney disease, dysplastic kidney disease or multicystic dysplastic kidney disease (Table 1) and 518 parents (507 unaffected and 11 affected), as well as 22,258 controls (Supplementary Table 1 and Supplementary Fig. 1). For the diagnostic analysis, genes associated with multiple Human Phenotype Ontology (HPO) terms, or multiple independent associations with renal agenesis, hypoplastic kidney disease, dysplastic kidney disease or multicystic dysplastic kidney disease in the literature were defined as “known genes” ($N=208$, Supplementary Data 1 and 2), whereas genes with single reports, or limited evidence for causality in the literature were defined as “probable genes” ($N=297$, Supplementary Data 1 and 2). Exome sequencing and microarray analysis identified diagnostic variants in 205 (10.3%) probands across 56 known genes and 25 known structural variants (Fig. 1a). Of those, 12 probands had dual diagnoses. In addition, 34 (1.7%) probands harbored candidate diagnostic variants, defined as borderline variants of uncertain significance in known genes (“VUS-high”)²², and 41 (2.1%) probands with P/LP variants in genes with emerging associations with renal agenesis, hypoplastic kidney disease, dysplastic kidney disease or multicystic dysplastic kidney disease (“probable genes”, Fig. 1a and Supplementary Data 3 and 4). We did not report variants of unknown significance, unless they were VUS-high. Of the 133 probands with a P/LP variant in a known gene, 52 (39%) had variants in *HNFIβ*, *PAX2*, *EYAI*, *DSTYK* or *GREB1L*, and 31 (23.3%) were singleton diagnoses (Fig. 1b). Of the 72 probands with diagnostic structural variants, 29 (40.3%) had the 17q12 (RCAD) deletion and 8 (11.1%) had single gene deletions identified by exome sequencing (Fig. 1c).

Autosomal dominant disorders accounted for the majority of diagnoses ($n=180$, 87.8%), followed by autosomal recessive disorders ($n=20$, 9.8%) and X-linked disorders ($n=5$, 2.4%). Amongst the 180 probands with dominant diagnoses, family information was available for 25; of those, 14 (56%) had de novo variants, 6 (24%) inherited the variant from a parent with CAKUT, and 5 (20%) inherited the variant from a parent without reported CAKUT.

Overall, we observed a higher diagnostic rate in individuals with bilateral CAKUT (20% vs 7% in those with unilateral CAKUT, chi-square p -value = 1.8×10^{-7} , Fig. 1d), in females (11% vs 9% in males, chi-square p -value = 3.8×10^{-3}), and in those with extra-renal anomalies (12% vs 9%, chi-square p -value = 0.02). We also observed a higher diagnostic rate depending on the kidney phenotype (13% in probands with cystic dysplastic kidney diseases vs 7% in probands with renal agenesis and 12% in probands with renal hypodysplasia, chi-square p -value = 1.9×10^{-3}). We only observed a trend for increased diagnostic rate based on reported family history of kidney disease (13% vs 9%, p -value = 0.06), but it was not statistically significant. There were no differences in diagnostic rate between singletons and trios (10% diagnostic rate in both).

Enrichment for de novo variants in constrained genes in trios with kidney anomalies

A total of 370 de novo variants were identified in 248 CAKUT trios, and the number of de novo variants per trio ranged from 0 to 12. The distribution of de novo variants per trio was consistent with prior reports (Supplementary Fig. 2). While there was no enrichment of de novo synonymous variants, we observed a 1.85-fold enrichment of de novo loss-of-function variants (LoF, FDR q -value = 3.42×10^{-4} , Table 2), and a 1.36-fold enrichment of de novo missense (FDR q -value = 1.54×10^{-5}) in

Table 1 | Characteristics of the cases included in the study

		N	Proportion
Total		1990	
Sex	Female	814	41%
	Male	1176	59%
Participants' enrollment location or origin	Poland	842	42%
	Italy	536	27%
	Macedonia	159	8%
	Columbia University Irving Medical Center	132	7%
	Chronic kidney disease in children study (CKiD)	112	6%
	Deciphering developmental disorders (DDD)	97	5%
	Netherlands	52	3%
	Croatia	37	2%
	Other	23	1%
Genetic ancestry (PCA-based)	Europe	1736	87%
	South America (Latino/Hispanic)	70	4%
	Africa	54	3%
	Asia	17	1%
	Admixed	97	5%
Kidney phenotype	Hypoplastic or dysplastic kidney disease	816	41%
	Kidney agenesis	741	37%
	Multi-cystic dysplastic kidney disease	433	22%
Laterality	Bilateral	149	7%
	Unilateral	1144	57%
	Unknown	697	35%
Family history of kidney disease	Yes	324	16%
	None reported ^a	1666	84%
Additional lower urinary tract phenotypes	Including vesicoureteral reflux (VUR), ureter, bladder, and urethra anomalies	439	22%
Extra-renal phenotypes (not mutually exclusive)	Congenital heart disease	186	9%
	ID and developmental delay	119	6%
	Genital	147	7%
	Gastro-intestinal	120	6%
	Facial dysmorphism	118	6%
	Eye	63	3%
	Ear (including hearing impairment)	50	3%
	Endocrine	66	3%
	None reported ^a	1078	54%
Study cohort	Trios	248	12%
	Singletons (gene-burden analysis)	1742	88%
	Singletons (single variant analysis)	1566	79%

PCA principal component analysis.

^aExtra-renal presentations and family history of kidney disease were not consistently collected in different cohorts.

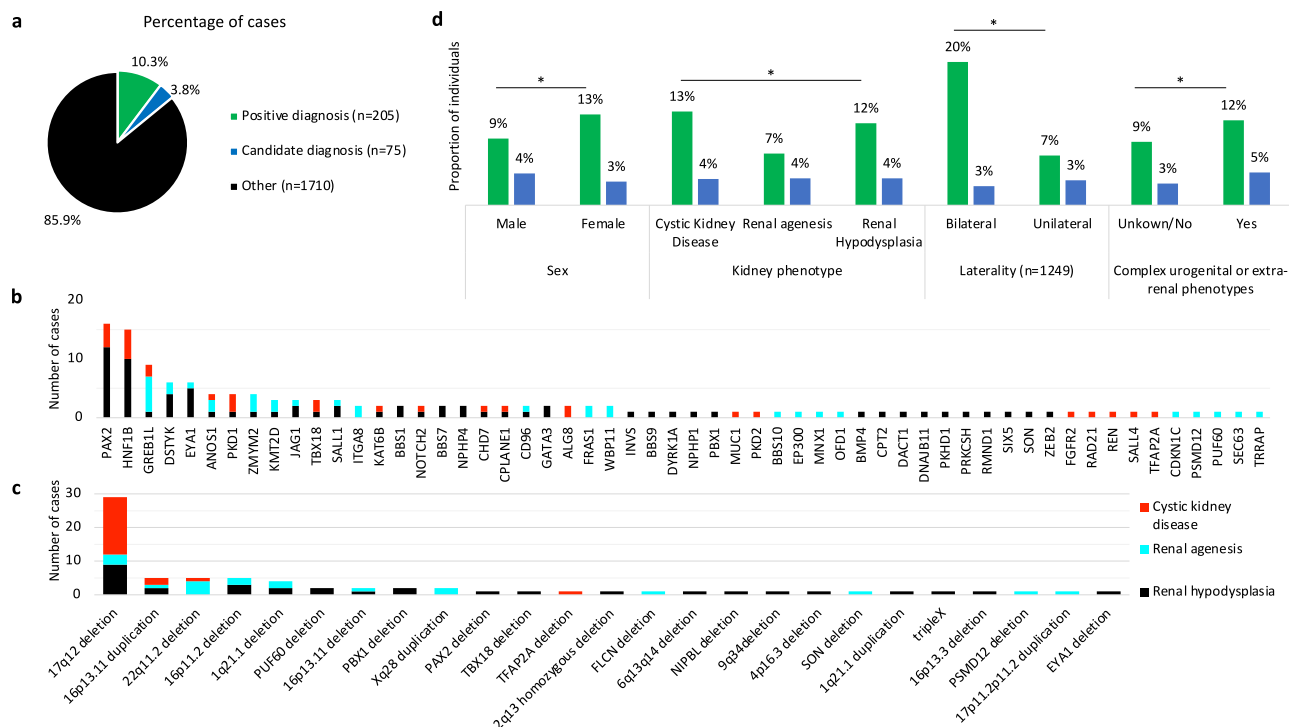


Fig. 1 | Diagnostic analysis in 1990 individuals. **a** Percentages of cases with diagnostic findings (green) and a candidate diagnosis (dark blue: VUS-high in a known gene, P/LP in a candidate gene or candidate structural variant) and those without a diagnostic or candidate variant (black). **b** Number of cases with an LP/P variant diagnostic or candidate finding (cases with two variants were counted once, with the variant most likely associated with their form of CAKUT) per gene and CAKUT (red: cystic kidney disease, black: renal hypoplasia and light blue: renal agenesis). **c** number of cases with a diagnostic or candidate structural variant per genomic area and CAKUT (red: cystic kidney disease, black: renal hypoplasia and light blue: renal agenesis). **d** Proportion of cases with diagnostic/

candidate findings based on clinical characteristics (green: diagnostic findings and dark blue: candidate diagnosis). Diagnostic rate comparisons were performed using two-sided chi-square tests: (i) Bilateral vs. unilateral CAKUT: 20% vs. 7% (Chi-square p -value = 1.8×10^{-7} ; $n = 1249$). (ii) Females vs. males: 11% vs. 9% (Chi-square p -value = 3.8×10^{-3} ; $n = 1710$). (iii) Extra-renal anomalies vs. isolated CAKUT: 12% vs. 9% (Chi-square p -value = 0.02; $n = 1710$). (iv) Kidney phenotype: 13% (cystic dysplastic), 7% (agenesis), 12% (hypodysplasia) (Chi-square p -value = 1.9×10^{-3} ; $n = 1710$). Asterisks (*) indicate statistical significance. No adjustments for multiple comparisons were applied.

cases compared to expectations. Of the 291 non-synonymous de novo variants, only eight were in genes known to be associated with dominant forms of CAKUT (*HNF1B*, *KMT6B*, *PAX2*, *PSMD12*, *SON*, *NIPBL*, *PUF60*, *GREB1L*). The enrichment for de novo variants was more pronounced for constrained genes ($pLi > 0.9$, $LOEUF < 0.35$, $misZ > 3.09$), particularly those that are highly expressed during kidney development. When restricting the analysis to constrained genes highly expressed in human nephron-progenitor cells, we observed a 7.64-fold enrichment for LoF variants (FDR q -value = 8.52×10^{-5}), and a 3.78-fold enrichment for missense variants (FDR q -value = 1.35×10^{-3}). Similarly, we observed a significant enrichment of LoF and missense variants in constrained genes highly expressed during early mouse kidney development (Table 2).

Enrichment of rare variants in constrained genes by gene-based burden analysis

We also performed gene-based burden analysis in 1742 unrelated probands with CAKUT (not included in the trio analysis) and 22,258 genetically matched controls. Under the LoF and missense model, only *PAX2*, known to be associated with CAKUT, reached exome-wide significance (p -value $< 10^{-6}$, Fig. 2a and Supplementary Data 5–9). *HNF1B* and *PAX2*, both known to be associated with CAKUT, reached exome-wide significance under the LoF model (Supplementary Fig. 3), while no gene reached exome-wide significance under the missense model (Supplementary Fig. 4). A total of 36 cases had two or more qualifying variants in known CAKUT genes, including 18 with a heterozygote qualifying variant in a gene associated with an autosomal dominant form of CAKUT, and of those 17 had a diagnostic variant (including one with two diagnostic variants and one with a diagnostic and a candidate

variant). All remaining qualifying variants were heterozygote variants in genes associated with recessive diseases.

When we restricted the genome-wide analysis to constrained genes ($pLi > 0.9$, $LOEUF < 0.35$), we observed a significant enrichment of LoF variants (odds ratio $OR_{(LoF)} = 1.48$; FDR q -value = 4.03×10^{-9}) and missense variants ($OR = 1.31$; FDR q -value = 3.91×10^{-6}) (Fig. 3a and Supplementary Table 2). The enrichment was slightly more pronounced when we restricted the analysis to constrained genes that are highly expressed in human nephron-progenitor cells ($OR_{(LoF)} = 1.70$, FDR q -value = 1.12×10^{-4} ; $OR_{(Mis)} = 1.31$, FDR q -value = 2.70×10^{-3}). We also noted that significantly more cases than controls harbored three or more LoF or missense qualifying variants in constrained genes (12% of cases vs 7% of controls, p -value = 1.2×10^{-28} , Fig. 3b and Supplementary Table 3). We observed a nominal association between the number of LoF or missense qualifying variants and the presence of extra-renal anomalies (13% of cases with extra-renal anomalies vs 10% of cases with unknown or no extra-renal anomalies had 3 or more qualifying variants, p -value = 3.03×10^{-2} , Fig. 3c).

Enrichment for rare damaging variants in genes associated with ID/autism spectrum disorder in trio and case-control analyses

We searched for rare damaging variants in genes associated with ID, autism spectrum disorder, and congenital heart disease. We detected an excess of de novo LoF variants in constrained genes associated with ID and/or autism spectrum disorder both in the trios (4.11-fold enrichment, FDR q -value = 3.35×10^{-4} , Table 2) and in the case-control dataset (LoF: $OR_{(LoF)} = 1.63$, FDR q -value = 9.18×10^{-7} ; $OR_{(Mis)} = 1.32$, FDR q -value = 6.44×10^{-8} , Supplementary Table 2).

Table 2 | De novo enrichment analysis in 248 trios

Model	Variant type	# Genes	Observed		Expected		Enrichment	p-value	FDR q-value
			n	Rate	n	Rate			
Genome-wide	Synonymous	18,931	78	0.31	81.6	0.33	0.95	0.67	0.77
	Missenses	18,931	246	0.99	180.5	0.73	1.36	2.18×10^{-6}	1.54×10^{-5} **
	LoF	18,931	45	0.18	24.3	0.10	1.85	1.14×10^{-4}	3.42×10^{-4} *
	LoF and missense	18,931	291	1.17	204.9	0.83	1.42	8.76×10^{-9}	3.94×10^{-7} ***
Constrained genes	Synonymous	500	3	0.01	5	0.02	0.60	0.87	0.88
	Missenses	951	30	0.12	16.6	0.07	1.8	2.01×10^{-3}	6.03×10^{-3}
	LoF	2850	18	0.07	5.7	0.02	3.14	3.25×10^{-5}	9.75×10^{-4} *
	LoF and missense	3039	93	0.38	50.6	0.20	1.84	5.97×10^{-8}	1.01×10^{-6} ***
Human nephron-progenitor cells (18 weeks) and CG	Synonymous	662	3	0.01	3	0.01	1.00	0.58	0.76
	Missenses	237	11	0.04	2.9	0.01	3.78	2.26×10^{-4}	1.35×10^{-3}
	LoF	631	8	0.03	1	4.03×10^{-3}	7.64	1.42×10^{-5}	8.52×10^{-5} **
	LoF and Missenses	662	27	0.11	8.2	0.03	3.3	1.63×10^{-7}	1.39×10^{-6} ***
Mouse embryonic kidney (E15.5) and CG	Synonymous	491	1	4.03×10^{-3}	2.1	8.47×10^{-3}	0.48	0.88	0.88
	Missenses	201	7	0.03	2.4	9.68×10^{-3}	2.96	0.01	0.02
	LoF	457	6	0.02	0.7	2.82×10^{-3}	8.74	8.11×10^{-5}	1.62×10^{-4} *
	LoF and Missenses	491	17	0.07	5.7	0.02	3.01	8.74×10^{-5}	2.97×10^{-4} *
ID and/or autism spectrum disorder and CG	Synonymous	1043	6	0.02	8.5	0.03	0.71	0.85	0.88
	Missenses	461	14	0.06	9.3	0.04	1.5	0.09	0.09
	LoF	962	10	0.04	2.4	0.01	4.11	2.23×10^{-4}	3.35×10^{-4} *
	LoF and Missenses	1043	38	0.15	21.6	0.09	1.76	8.87×10^{-4}	2.15×10^{-3}
Congenital heart defects and CG	Synonymous	166	0	0	2.2	8.87×10^{-3}	0	1	1
	Missenses	67	5	0.02	1.5	0.01	3.24	0.02	0.03
	LoF	153	1	4.03×10^{-3}	0.4	1.61×10^{-3}	2.43	0.34	0.34
	LoF and missense	166	9	0.04	3.7	0.02	2.42	0.01	0.03
Immune and CG	Synonymous	217	1	4.03×10^{-3}	1.2	4.84×10^{-3}	0.81	0.71	0.86
	Missenses	96	5	0.02	1.5	6.05×10^{-3}	3.44	0.02	0.03
	LoF	195	2	8.06×10^{-3}	0.4	1.61×10^{-3}	5.18	0.06	0.07
	LoF and missense	217	8	0.03	3.3	0.01	2.45	0.02	0.03

Constrained: for LoF: $pLi > 0.9$ and $oe_lof_up_ < 0.35$; for missenses: $misZ > 3.09$, for LoF and missenses, $pLi > 0.9$ and $oe_lof_up_ < 0.35$ and/or $misZ > 3.09$.

CG constrained genes, LoF loss of function variants (stop-gained, stop-lost, start-lost, splice-site, and frameshift variants).

***FDR q-value $< 10^{-5}$; **FDR q-value $< 10^{-4}$; *FDR q-value $< 10^{-3}$.

There was no significant enrichment in genes encoding the innate immune system, used as a negative control. Combining the trio and case-control analyses we confirmed the significant enrichment of variants in constrained genes (FDR q -value_(LoF) = 3.87×10^{-12} , FDR q -value_(miss) = 1.11×10^{-7} , Table 3), genes highly expressed in human nephron-progenitor cells (FDR q -value_(LoF) = 6.30×10^{-8} , FDR q -value_(miss) = 8.48×10^{-5}), genes associated with ID and autism spectrum disorder (FDR q -value_(LoF) = 4.68×10^{-9} , FDR q -value_(miss) = 5.51×10^{-8}), and genes associated with congenital heart disease (FDR q -value_(LoF) = 1.9×10^{-2} , FDR q -value_(miss) = 3.02×10^{-3}).

These findings were further supported by an analysis of clinical phenotypes. Of the 177 probands with diagnostic variants, 60 had variants in genes associated with CAKUT and developmental delay, and despite the paucity of clinical information, 11 (18%) of those had reported ID or developmental delay. Similarly, 50 probands had variants in genes associated with CAKUT and congenital heart disease, and 8 (16%) of them had reported congenital heart disease (Supplementary Table 3).

Two causal genes, *NR6A1* and *ARID3A*

The trio and burden analyses identified 40 candidate genes for CAKUT prioritized based on predicted deleteriousness, mutational constraint,

known association with ID, autism spectrum disorder, or congenital heart disease, enrichment in gene-burden analysis under the LoF model (OR > 10) or de novo status (Supplementary Data 10 and 11). One of the candidate genes, *CELSR1*, has been previously suggested as a candidate gene for the CAKUT presentation in individuals with Phelan-McDermid syndrome²³. We identified 7 cases with cystic kidney diseases and 9 controls with LoF variants in *CELSR1* ($pLi=1$). In the genes *SSBP2*, *XPO1*, *MAP4K4*, *NR6A1*, and *ARID3A*, at least two cases were identified with LoF variants, while no LoF variants were identified in controls (Supplementary Table 5 and Supplementary Data 10 and 11). *NR6A1* (nuclear receptor subfamily 6 group A member 1) and *ARID3A* (AT-rich interaction domain 3A), with four and three LoF variants in cases, were both confirmed as causal disease genes by the identification of additional independent probands with CAKUT and related phenotypes.

NR6A1. Review of sequencing data from Columbia CAKUT cases not included in the discovery cohort identified 6 additional cases with *NR6A1* predicted deleterious variants (predicted splice variants in two cases and predicted deleterious missense variants with REVEL > 0.8 in four cases, Table 4 and Fig. 4). One of these cases harbors an R92W variant that segregates in additional affected relatives (Fig. 4b). Structural analysis predicts R92W is highly deleterious because it would

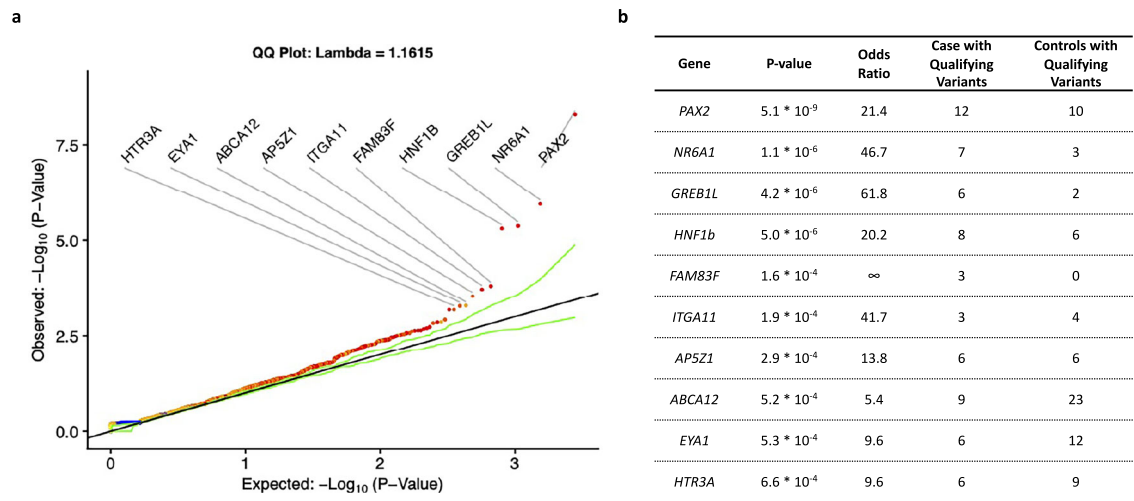


Fig. 2 | Gene-burden and gene-set enrichment analyses. **a** The gene-burden analysis was performed by extracting the number of cases and controls with and without a qualifying variant (QV) per gene. The exact two-sided Cochran–Mantel–Haenszel (CMH) test was used to test for enrichment of qualifying variants in cases vs controls, while controlling for cluster. Quantile–Quantile

probability plot of the p -values generated by the gene-burden analysis focused on qualifying variants (LoF and predicted deleterious missense). Red indicates case-enriched p -values. The green lines represent the 95% confidence interval. Empiric confidence interval distributions created by permutations ($n = 1000$). **b** List of the ten most significant genes.

abrogate the formation of high affinity salt bridges with the DNA phosphate backbone and would cause clashes within the DNA binding domain (Fig. 4c, d). The Columbia CAKUT proband with the R92W variant also has an eye coloboma but no other variants in genes causing eye defects (Fig. 4b). Through collaboration with the NIH, we identified two unrelated individuals ascertained for developmental eye phenotypes who also had CAKUT and harbored pathogenic *NR6A1* variants (a large deletion and the R92W variant, see related manuscript by Neelathi et al.²⁴). Thus, the R92W variant was identified in three unrelated probands with CAKUT and eye abnormalities in the Columbia and NIH cohorts. We also identified an additional *NR6A1* LoF variant in an individual with CAKUT through collaboration with the University Medical Center Utrecht. In summary, we identified *NR6A1* predicted deleterious variants in 10 cases in the Columbia cohort (including 6 LoF and 1 de novo; and 2 with CAKUT and eye anomalies), and 3 additional cases in external cohorts (including 2 LoF variants; 2 with CAKUT and eye anomalies and 1 with eye anomalies), identifying *NR6A1* as a gene associated with kidney, eye, and other congenital anomalies.

ARID3A. In addition to the three LoF variants in our discovery cohort (Table 5), a review of the Columbia CAKUT biobank identified two additional CAKUT cases with *ARID3A* LoF variants. Through collaboration with Boston Children’s Hospital, we identified two additional individuals with LoF in *ARID3A*, both confirmed as de novo variants. One individual displays vesicoureteral reflux (VUR) Grade 5, Solitary Kidney, Posterior Urethral Valves, Megaureter, Hydronephrosis, and chronic kidney disease stage 3. The other individual displays Megaureter, hydronephrosis, VSD, Atrial septal defect, and Hypoplasia of facial muscles. Altogether, we identified seven predicted LoF in *ARID3A* in patients with kidney and genitourinary defects. The finding of seven independent LoF variants, including two de novo variants, identifies *ARID3A* as a haploinsufficient CAKUT gene. A phenotypic review indicated that patients with *ARID3A* LoF variants have multiple genitourinary defects, and a few had extra-renal abnormalities. Pathway analysis demonstrated that *ARID3A* and *NR6A1* are co-expressed with *SALL4*, a gene known to cause CAKUT (Supplementary Fig. 5).

Low-frequency risk variants

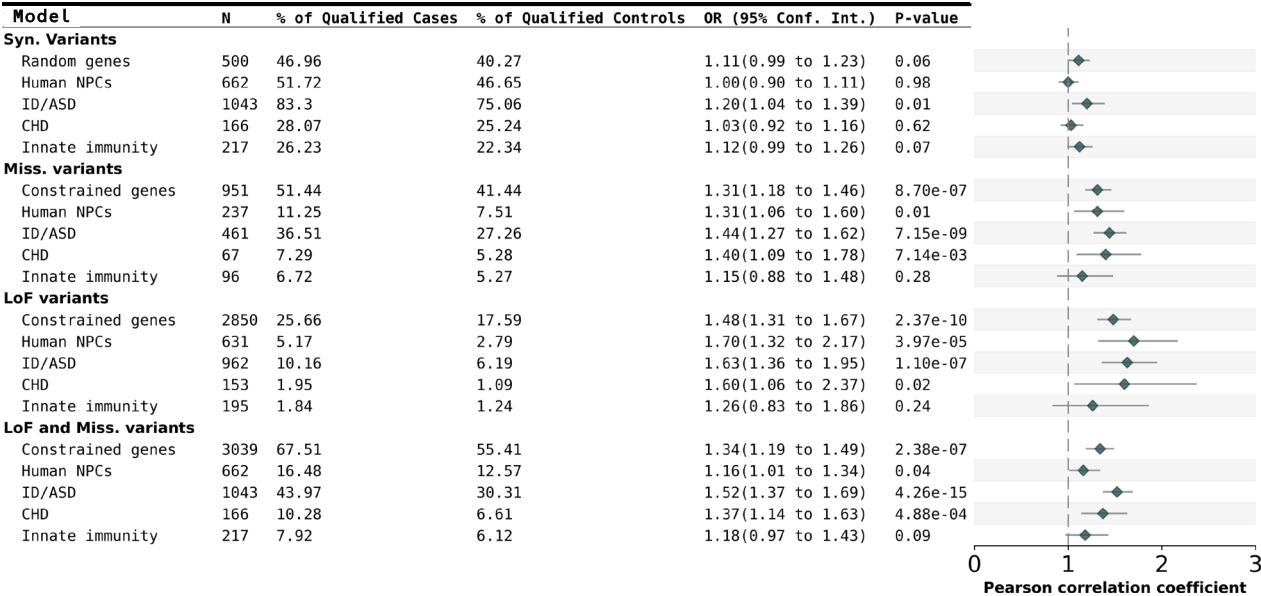
We also performed an exome-wide single variant association study for low-frequency risk variants in genes known to be associated with

CAKUT (Supplementary Table 2). We identified four suggestive associations in genes known to cause autosomal dominant or recessive CAKUT (*DSTYK*, *KMT2D*, *PKHD1*, and *SDCCAG8*), including the low-frequency *DSTYK* variant with a prior association with CAKUT (Supplementary Table 4)²⁵.

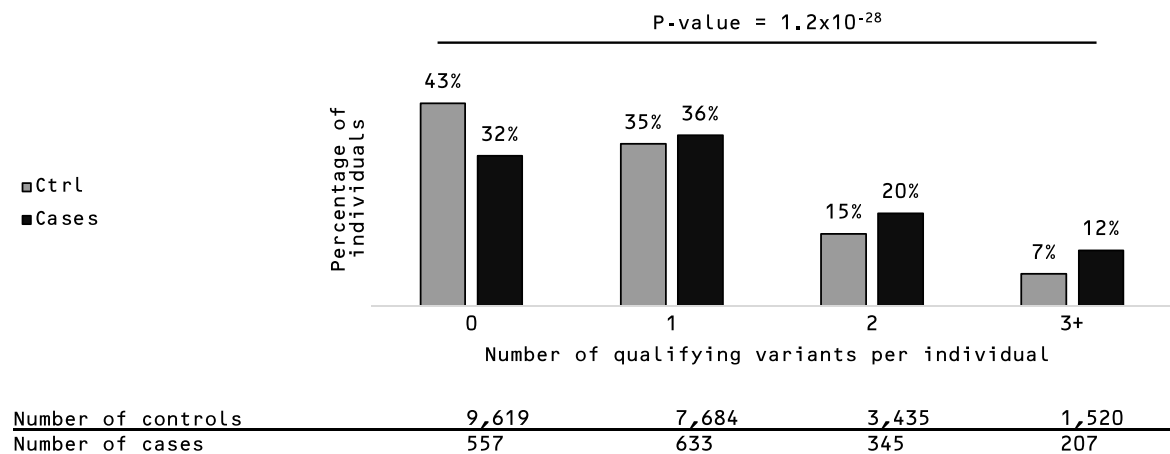
Discussion

In this cohort of 1990 individuals with CAKUT, the diagnostic, trio, and case-control analyses point to a wide number of genes for CAKUT. We identified diagnostic variants in 57 genes and 25 structural variants. Combined with the candidate diagnostic variants (54 genes and four structural variants), the overall diagnostic rate of 14.1% in this cohort falls in the previously reported range of 2–27% for CAKUT^{26,27}. *HNF1B* and *PAX2* accounted for a larger fraction of diagnoses and therefore reached genome-wide significance in the case-control analysis, while other genes known to be associated with CAKUT each contributed a small fraction of cases, highlighting the long tail in the distribution of rare genetic diseases that cause CAKUT. The trio and case-control analyses also demonstrated a significant excess of ultrarare deleterious variants, mostly in constrained genes expressed during early kidney development that are not known to be associated with CAKUT, suggesting many yet-to-be-identified genetic causes that would require larger cohorts for discovery. Moreover, CAKUT patients harbored a significant burden of deleterious variants in ID or autism spectrum disorder genes, suggesting shared pathogenesis. The candidate diagnoses also strengthened the previously reported association of CAKUT for 35 known genes and three duplications associated with congenital heart disease and ID or autism spectrum disorder syndromes. These data are consistent with our prior studies showing significant overlap between CNV disorders underlying CAKUT, congenital heart disease, ID, and autism spectrum disorder, and that CAKUT patients with pathogenic CNVs often have undetected or subclinical neurodevelopmental phenotypes^{4,6,28,29}. In this cohort, we similarly found that the CAKUT patients with cardiac or neurodevelopmental comorbidities were more likely to have a known genetic syndrome and to carry more qualifying variants than those without extra-renal manifestations. Altogether, these findings indicate that CAKUT, congenital heart disease, ID, and autism spectrum disorder share underlying genetic mechanisms and pathways and point to the potential for phenotypic expansion for known and yet-to-be-identified developmental

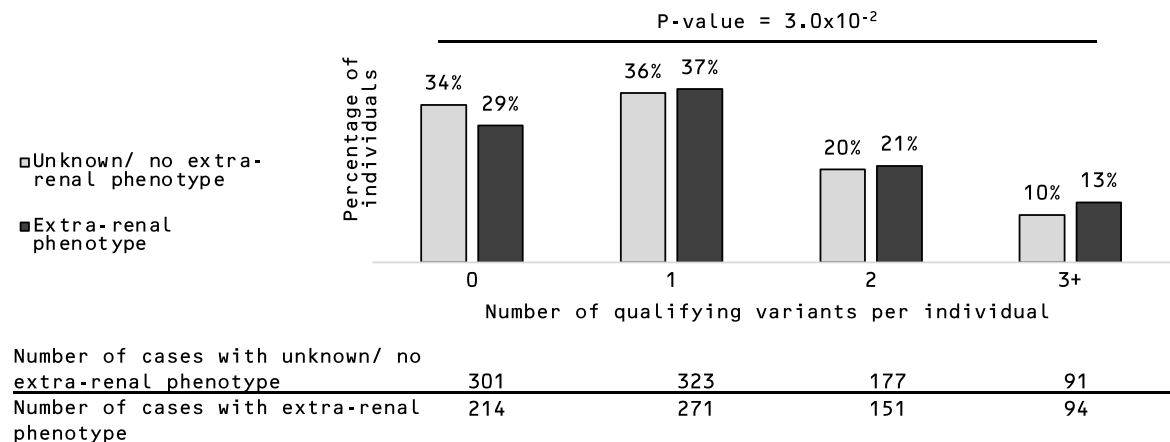
a



b



c



disorders. Hence, a cross-phenotype analysis of developmental disorders may identify genetic disorders and potential modifiers that determine the spectrum and severity of defects.

The reasons for the variable severity of clinical phenotypes in CAKUT patients are unknown; genetic, environmental, or developmental factors have been invoked³⁰. We found that 5% CAKUT cases harbored two or more ultrarare deleterious variants in constrained genes, and these individuals had a higher likelihood of presenting with

multiple developmental phenotypes. This finding is supported by our prior study, where we similarly found that CAKUT patients with extrarenal phenotypes often harbor multiple rare, gene-disrupting CNVs⁶. In addition to a higher burden of ultrarare variants, we identified several low-frequency risk variants in *DSTYK*, *KMT2D*, *PKHD1*, and *SDCCAG8*, which may modify the severity of disease, consistent with prior studies demonstrating that common variants also influence the risk of developmental phenotypes. Taken together, these data suggest

Fig. 3 | Gene-set enrichment analyses and distribution of the number of qualifying variants per individual. **a** Forest plot depicting the gene-set analysis using six gene sets based on three gene burden analyses. **b** Number of cases (black) and controls (Ctrl: gray) with 0, 1, 2, or at least 3 qualifying variants (LoF and predicted deleterious missense variants in constrained genes). Statistical significance of the difference between cases and controls was assessed using a two-sided Wilcoxon rank sum test. A significant difference was observed (OR = 1.1, p -value = 1.2×10^{-28}). **c** Number of cases with extra-renal phenotypes (dark gray) and cases without known extra-renal phenotypes (light gray) with 0, 1, 2, or at least 3 qualifying variants (LoF and predicted deleterious missense variants in constrained genes).

Statistical significance of the difference between cases with and without known extra-renal phenotypes was assessed using a two-sided Wilcoxon rank sum test (OR = 1.1; p -value = 3.03×10^{-2}). No adjustment for multiple comparisons was applied. CHD congenital heart disorder, ID/ASD intellectual delay and autism spectrum disorder, LoF loss-of-function variants, mis missense variants, NPCs nephron progenitor cells, OR odds ratio, syn synonymous variants. Predicted deleterious missense variants: subRVI domain score percentile < 50% and predicted damaging based on at least 2 of the following criteria: (1) REVEL > 0.5; (2) PrimateAI > 0.8; (3) AlphaMissense score > 0.6; and (4) CADD score > 25. Constrained genes had $pLI \geq 0.9$ and $LoEF < 0.35$ and/or $misZ > 3.09$.

Table 3 | Meta-analysis of gene burden and de novo gene sets analysis

Gene-set	# of genes	Model	Case-control analysis p -value	De novo analysis p -value	Fisher p -value	FDR q -value
Constrained genes	500	Syn	0.06	0.87	0.21	0.24
Constrained genes	2850	LoF	2.37×10^{-10}	3.25×10^{-5}	2.58×10^{-13}	3.87×10^{-12}
	952	Mis.	8.70×10^{-7}	2.01×10^{-3}	3.80×10^{-8}	1.11×10^{-7}
Human NPCs (18 weeks)	663	Syn	0.98	0.58	0.89	0.9
	632	LoF	3.97×10^{-5}	1.42×10^{-5}	1.26×10^{-8}	6.30×10^{-8}
	238	Mis.	1.08×10^{-2}	2.26×10^{-4}	3.39×10^{-5}	8.48×10^{-5}
ID or autism spectrum disorder	1043	Syn	0.01	0.85	0.05	0.075
	962	LoF	1.10×10^{-7}	2.23×10^{-4}	6.24×10^{-10}	4.68×10^{-9}
	461	Mis.	7.15×10^{-9}	0.09	1.47×10^{-8}	5.51×10^{-8}
Congenital heart disease	286	Syn	0.6	1	0.9	0.9
	267	LoF	0.02	0.03	0.01	0.019
	97	Mis.	0.007	0.02	1.41×10^{-3}	3.02×10^{-3}
Immune	217	Syn	0.07	0.71	0.20	0.24
	195	LoF	0.24	0.06	0.08	0.11
	96	Mis.	0.28	0.02	0.04	0.067

Constrained genes only (Constrained: ^afor LoF: $pLI > 0.9$ and $oe_{lof_up_} < 0.35$, ^bfor missenses: $misZ > 3.09$). LoF loss of function variants (stop-gained, stop-lost, start-lost, splice-site, and frameshift variants), Mis. missense variants, Syn synonymous variants, NPCs top decile expression in nephron progenitor cells at 18 weeks of gestation, constrained genes only, ASD genes associated with autism spectrum disorder extracted from the AutismDB, constrained genes only, CHD genes associated with congenital heart disease, constrained genes only.

oligogenic or polygenic determination for some CAKUT cases, and variation in genetic background as a potential contributor to phenotypic severity.

We identified two causal genes for CAKUT, *NR6A1* and *ARID3A*. We found multiple probands and families with *NR6A1* mutations presenting with kidney and eye malformations and a few other developmental phenotypes. *NR6A1* encodes the protein germ cell nuclear factor (GCNF), also known as RTR (Retinoid Receptor-Related Testis-Associated Receptor). *Nr6a1* is expressed at 15 days post coitum (dpc) in mice in the ureteric tip and the cap mesenchyme (TS19-TS28), and in the early tubule and renal interstitium³¹. *Nr6a1*^{-/-} embryos have abnormal optic vesicle morphology, neural tube closure, and additional developmental anomalies leading to their demise at 10.5 dpc. Because kidneys develop after 12 dpc, conditional inactivation in cells important for kidney development will be required to study the function of *Nr6a1* on kidney development. The role of *NR6A1* variants in causing eye and kidney defects was confirmed by the identification of independent cases ascertained for eye anomalies²⁴. The incomplete penetrance and expressivity identified in carriers of *NR6A1* variants is a common phenomenon in autosomal dominant diseases³².

We identified a total of 5 individuals with unilateral renal agenesis and two individuals with hydronephrosis with rare predicted deleterious LoF variants in *ARID3A*, and two of those variants occurred de novo. Interestingly, a review of 13 cases with a 19p13.3 microdeletion that encompasses *ARID3A* reported 3 cases (23%) with unilateral renal agenesis³³. *ARID3A* encodes a DNA-binding protein highly expressed in E14.5 murine kidneys³¹. Inactivation of *Arid3a* leads to embryonic

lethality at E12.5 for most mice due to impaired hematopoiesis, but the few surviving *Arid3a* null mice display abnormal cell proliferation and structural abnormalities in their kidneys^{34,35}. In *Xenopus*, *ARID3A* binds to regeneration signal-response enhancers, reduces the levels of the constitutive heterochromatin marker histone three lysine nine trimethylation, promotes cell cycle progression, and causes the outgrowth of nephric tubules³⁶. In addition, an *Arid3a*^{-/-} kidney mouse cell line, KKPS5, was reported to spontaneously develop into multicellular nephron-like structures in vitro, which can form mouse nephron structures if engrafted into immunocompromised medaka mesonephros³⁷. Interestingly, we found that *ARID3A* is co-expressed with *NR6A1* and *SALL4* and is predicted to function in a common transcriptional module³⁸⁻⁴⁰. Taken together, these data suggest that *ARID3A* plays a regulatory role in nephrogenesis, potentially by balancing cellular proliferation. Hence, other co-expressed genes in the *ARID3A* transcriptional module may be candidates for kidney developmental disorders. Further functional studies in cellular models, organoids, or animal models could decipher the role of *ARID3A* during kidney development and help identify additional candidate genes.

Our study has multiple clinical and research implications that chart a path forward for genetic studies of CAKUT. The clinical implications of a genetic diagnosis are numerous, including providing a definitive diagnosis, prognostic information, decision support for transplantation decisions, and identification of affected family members²⁷. The diagnostic rate observed in this cohort supports the Kidney Disease: Improving Global Outcomes (KDIGO) and the National Kidney Foundation (NKF) recommendation that genetic

Table 4 | Predicted Loss-of-function variants in NR6A1

ID	Gene	Variant (genomic location Hg19)	c.DNA	Protein change	TraP	Inheritance	Kidney anomaly	Extra-renal anomalies
Ind 1	NR6A1	9-127533299-C-G	c.100G > C	p.Gly34Arg (predicted p.?)	0.848	Het, inherited from a healthy father	Multicystic dysplastic kidney (MCDK)	Unk
Ind 2	NR6A1	9-127316735-C-T	c.245G > A	p.Arg82Gln(predicted p.?)	0.514	Het, unknown	Renal agenesis	Seizures, no spine anomalies
Ind 3	NR6A1	9-127316655-G-A	c.337C > T	p.Gln113Ter	0.049	Het, unknown	Renal agenesis, unilateral, VUR, congenital	No signs of heart defects, no ASD, no hearing impairment
Ind 4	NR6A1	9-127302354-G-T	c.554C > A	p.Ser185Ter	0.526	Het, unknown	CAKUT	Unk
Ind 5	NR6A1	9-127300561-C-A	c.631G > T	p.Glu211Ter	0.564	Het, unknown	Renal agenesis, unilateral	Eyes: anisometropia, heart: patent ductus arteriosus and uterus agenesis, hirsutism
Ind 6	NR6A1	9-127287097-C-T	c.1254G > A	p.Trp418Ter	0.082	Het, de novo	Crossed ectopia of the kidney, left kidney with reduced function	Bronchial asthma, growth within normal limits
Ind 7	NR6A1	9-127285074-T-C	c.1352-2A > G	p.?	0.611	Het, unknown	Renal agenesis, bilateral	Deceased at birth

All variants are absent in gnomAD v3.
Het heterozygote.

testing should be offered to individuals with CAKUT^{41,42}. Our data highlight the significant genetic heterogeneity of CAKUT and expand the list of potential diagnostic genes for this disorder. The genetic overlap with other developmental disorders suggests that a diagnostic workup of CAKUT patients should include a broader gene list, and CAKUT should be considered as an expansion phenotype for genetic disorders associated with cardiac, eye, or brain developmental disorders. In addition, CAKUT patients should be screened for extrarenal developmental phenotypes that may have been missed during initial evaluation, particularly neurodevelopmental phenotypes, which may become evident in later years. Conversely, CAKUT genes should be considered in the workup for other developmental phenotypes. The determinants of variable clinical expression of kidney malformations suggest a role for genetic modifiers or redundant developmental programs that influence the ultimate clinical outcome. While the diagnostic and gene burden analysis nominated several candidate genes such as *CELSR1*, *SSBP2*, *XPO1*, or *MAPK4*, the excess number of rare variants in conserved genes detected in the trio and case-control analyses also points to the existence of many additional CAKUT genes that will require larger cohorts for discovery. Assuming an odds ratio of ~10, about 21,000 individuals with CAKUT would be needed for 80% statistical power to detect a gene with exome-wide significance⁴³, and increasing the number of controls would only reduce the number of cases need to reach this power to about 20,000, pointing to the importance of collaborative efforts like the GeneMatcher initiative⁴⁴. Finally, the genes for kidney malformations may also point to environmental factors that can disrupt developmental programs. For example, several genes implicated in CAKUT, such as *RET* or *GREB1L*, interact with retinoic acid signaling, which is important for organogenesis^{45,46}. Similarly, disruption of the gut microbiota in obese mice may influence *ARID3A* binding to its targets by altering specific metabolites⁴⁷.

Study limitations include incomplete clinical information about the spectrum of developmental phenotypes and a lack of genetic ancestral diversity. In addition, although this is the largest genetic study of CAKUT patients to date, a still larger sample size will be needed to comprehensively assess the spectrum of genes and variants contributing to CAKUT. The study's reliance on exome sequencing may still have limitations in detecting certain types of genetic abnormalities, such as deep intronic or other non-coding mutations, that are relevant to the diagnostic analysis. Utilizing whole genome sequencing or long-read sequencing in future studies could enhance diagnostic capabilities by identifying additional genetic variants not captured by exome sequencing or CNV analysis. Detailed phenotyping, consistent collection of family history, and larger cohorts would allow a significant increase in both diagnostic yield and identification of disease-causing genes that may have smaller effect sizes or act under a recessive model. Collaborative studies combining patients with CAKUT, congenital heart disease, ID, and autism spectrum disorder will also help overcome the challenge associated with genetic heterogeneity and limited penetrance, and also enable long-term assessment of functional status, clinical complications, and reproductive outcomes.

Methods

Written informed consent was obtained from all participants recruited for this study or, where applicable, from their legal guardians prior to participation in this study. For the DDD cohort, written informed consent specific to research on undiagnosed developmental disorders was obtained from all families, as approved by the UK and Irish Research Ethics Committees. For the control cohort, a more general form of written informed consent was secured, which covered participation in research studies of this type, including the use of individual sequencing data for genetic research. The study complies with all relevant regulations. Approval for human research was obtained from

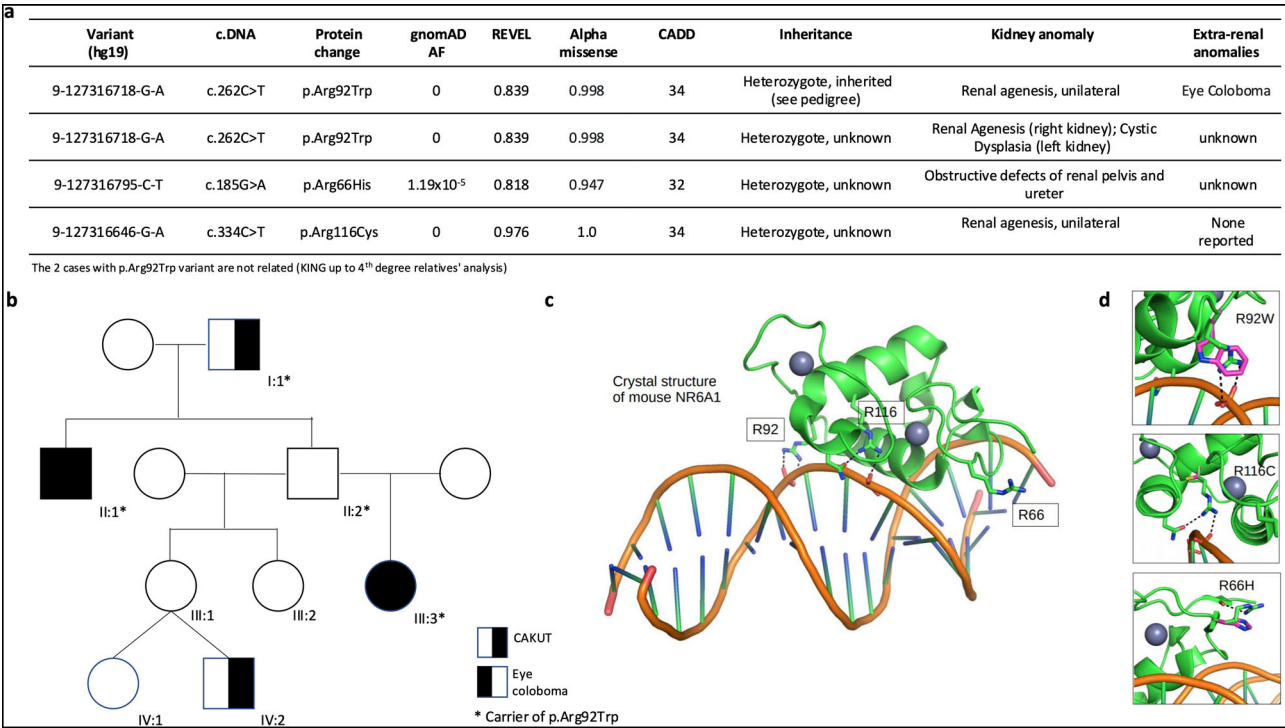


Fig. 4 | Missense variants in *NR6A1*. **a** Four independent Columbia CAKUT cases with predicted deleterious missense variants in *NR6A1*. **b** Family carrying the p.Arg92Trp (R92W) variant in *NR6A1*. (I:1 pelvic and small kidney, II:1 renal agenesis and Eye coloboma, III: 3 renal agenesis and eye coloboma, and IV: 2 renal agenesis and unknown genetic status). **c** Crystal structure of mouse *NR6A1* and location of

the three amino acids in which missense variants were identified in the four cases. **d** Modeling the potential impact of two of the three missense variants on the protein structure (R92W and R116C are modeled on the *NR6A1* crystal structure PDB ID: 5KRB), the potential structural consequences of R66H could not be determined from the published *NR6A1* structure.

the Institutional Review Board of Columbia University Medical Center, as well as the Ethics Review Boards at collaborating institutions (Poznan University of Medical Sciences, Poland; University of Brescia and Spedali Civili of Brescia, Italy; Hospital University of Padova, Italy; University of Messina, Italy; IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy; ARNAS Brotzu Hospital, Cagliari, Italy; IRCCS G. Gaslini, Genoa, Italy; University of Milan, Italy; University of Calabria, Italy; Parma University Hospital, Parma, Italy; University of Torino, Torino, Italy; Università degli Studi della Campania “Luigi Vanvitelli”, Naples, Italy; Ospedale Maggiore Policlinico, Milan, Italy; University of Split, Split, Croatia; University Children’s Hospital, Skopje, North Macedonia). The animal research was carried out in accordance with the Columbia University Animal Care Protocol (protocol number: AC-AABL5584).

Participants

The diagnosis of CAKUT was made by nephrologists or urologists based on pertinent imaging data such as renal ultrasound, MRI, or CT scans. The study cohort was composed of 1781 individuals and 176 parents recruited for genetic studies of kidney disease at Columbia University and collaborating institutions from Poland, Italy, Macedonia, the Netherlands, Croatia, and other countries (Table 1). In addition, the study includes 112 CAKUT probands from the Chronic Kidney Disease in Children (CKiD) study⁴⁸ and 277 samples from the Deciphering Developmental Delay (DDD) study¹⁶, including 90 trios of probands with CAKUT and parents without CAKUT and seven probands whose parents have CAKUT. Probands included 858 cases with hypoplastic or dysplastic kidney disease, 735 cases with kidney agenesis, and 397 cases with multicystic dysplastic kidney disease (Table 1). The majority of cases were of European genetic ancestry based on Principal Component Analyses (PCA, 87%) and were not reported to have extra-renal phenotypes (71%). Of the 1990 probands, 248 were

analyzed as trios (no family history, Supplementary Fig. 1). Family relationships were confirmed using KING software⁴⁹. The sequence data from 1742 unrelated (KING V1.4: up to third-degree, kinship coefficient > 0.0884) probands with CAKUT who were not included in the trio analysis were compared to data from 22,258 unrelated controls matched based on genetic ancestry (PCA-based, Supplementary Table 1). The controls were enrolled in studies unrelated to CAKUT or as controls and were available through the Columbia Institute for Genomics Medicine. All consented to have anonymized sequence data available for secondary genetic analysis and were called and annotated using the same bioinformatic pipeline as cases. Controls were defined as either individuals who do not have CAKUT based on ultrasound (2814 individuals with biopsy-documented IgA nephropathy or C3 glomerulopathy), individuals with disorders unrelated to kidney disease (e.g., amyotrophic lateral sclerosis, *n* = 10,114), or individuals enrolled as controls or healthy family members to diverse studies (*n* = 9330).

Exome sequencing

Exome sequencing was performed on 2290 samples (1893 probands and 316 parents) through the Yale Center for Mendelian Genomics, the New York Genome Center, and the Columbia Institute for Genomics Medicine, using different exome kits (Supplementary Table 5) at a minimum average depth of 30× to a maximum of 250× as described in refs. 21,50. All samples had a minimum of 90% coverage at 10× across the Consensus Coding Sequence (CCDS) regions. In addition, the CRAM files from the DDD study (Agilent V5 exome kit) were downloaded and converted into FASTQ files¹⁶. All probands, parents, and controls’ sequence data were processed using the same bioinformatic pipeline. FASTQ data were processed and aligned to hg19/GRCh37 using DRAGEN and GATK haplotyper caller, and variants were called using the Genome Analysis Toolkit (GATK) v3.6 best practices.

Table 5 | Predicted Loss-of-function variants in ARID3A

ID	Gene	Variant (genomic location Hg19)	c.DNA	Protein change	TraP	Inheritance	Kidney anomaly	Extra-renal anomalies
Ind. 8	ARID3A	19-926060-G-C	c.-268 + 1G > C	p.?	0.903	Het, inherited from mother (unknown clinical status)	Congenital hydronephrosis	Hutch diverticulum
Ind. 9	ARID3A	19-929604-C-T	c.76 C > T	p.Gln26Ter	0.089	Het, unknown	Renal agenesis, unspecified	Highly functioning autism, emotional hypersensitivity
Ind. 10	ARID3A	19-929838-C-T	c.310 C > T	p.Arg104Ter	0.828	Het, de novo	Renal agenesis, posterior urethral valves, VUR Grade 5, megaureter, hydronephrosis, CKD3	Unknown
Ind. 11	ARID3A	19-964432-G-C	c.950 + 1G > C	p.?	0.939	Het, unknown	Renal agenesis, unilateral, VUR, congenital, left kidney: dysplasia	Dyscalculia
Ind. 12	ARID3A	19-964831-A-T	c.951-2A > T	p.?	0.529	Het, unknown	Renal agenesis, unilateral, ectopic kidney (includes pelvic kidney)	Unknown
Ind. 13	ARID3A	19-964918-C-T	c.1036C > T	p.Arg346Ter	0.015	Het, unknown	Renal agenesis, unilateral	Normal development
Ind. 14	ARID3A	19-966630-ACCCT-G	c.1259-1263delACCCCT	p.His420fs	-	Het, de novo	Megaureter, hydronephrosis	VSD, atrial septal defect, hypoplasia of facial muscles

All variants are absent in gnomAD v3.
Het heterozygote.

Analyses were performed using the in-house Analysis Tool for Annotated Variants for retrieving, annotating, and filtering variants in large cohorts⁵¹.

Diagnostic analysis

We used a combination of automated and manual curation to generate a gene list for diagnostic analysis. HPO terms associated with CAKUT were utilized to capture a comprehensive list of relevant genes⁵². Genes associated with a single HPO term underwent a manual curation of the literature to determine gene-phenotype associations. The list was also annotated for the presence of extra-renal manifestations (Supplementary Data 1 and 2).

The single-nucleotide variants and small insertions and deletions were classified according to the American College of Medical Genetics and Genomics (ACMG) guidelines for clinical sequence interpretation⁵³. Structural variants were identified with both microarray data and exome data. To identify copy-number variants with microarray data, 1342 samples (67.4%) were genotyped on various Illumina arrays (Supplementary Fig. 1 and Supplementary Table 5). Raw data was first processed with Affymetrix Power Tools and the PennCNV-Affy protocol or with Illumina GenomeStudio v2011, to obtain probe-level logR-ratio and b allele frequency values. Raw intensity data were processed in GenomeStudio v2011 (Illumina). PennCNV software was used to determine CNV calls. PennCNV and PLINK software were used for quality control^{6,28}. To identify structural variants in exome data, we used XHMM⁵⁴. As XHMM does not readily differentiate between homozygous and heterozygous deletions, we only analyzed deletions in genes known to be associated with dominant forms of CAKUT.

When available, parental inheritance data were used to determine the variants' phase and de novo status. Variants were classified as pathogenic (P), likely pathogenic (LP), or variants of uncertain significance highly suspicious to become pathogenic because of phenotype association (VUS-high)²². Only variants that matched the inheritance pattern of the associated condition were considered. P/LP variants and full gene deletions in genes known to cause CAKUT were considered diagnostic. In the probable genes, only P/LP variants that matched the inheritance pattern of the associated condition are reported.

De novo variants' identification

Variants were annotated using Ensembl canonical transcripts. In cases with multiple transcripts, the most deleterious effect was included. De novo variants were defined as variants present in the proband (Read Depth ≥ 10 , Alternate allele depth ≥ 3 , $QUAL_{snv} \geq 50$, $QUAL_{indel} \geq 300$, $GQ \geq 20$, proportion of missing alleles less than 20%, Alternate read percentage 20-80%). The stringency of these quality filters was tested using inherited variants in the same trios, demonstrating that they removed less than 1% of inherited variants. Finally, only rare variants based on the Genome Aggregation Database⁵⁵ (gnomAD v2.1.1 global allele frequency (AF) $\leq 10^{-5}$ and Columbia Institute for Genomics Medicine AF $\leq 10^{-4}$) absent in the parents of the trios analyzed were retained. All variants were then manually curated using the Integrative Genomics Viewer interactive software tool⁵⁶.

De novo enrichment analysis

Trio analysis was restricted to functional de novo variants: LoF variants (LoF: canonical splice-site, stop-gain, stop-loss, start-lost, and frame-shift insertions and deletions), missense variants, and protein-altering (LoF and Missense) variants. To estimate the probability of a de novo variant with each one of those effects, we used denovolyzer⁵⁷ and the updated mutation table from denovoWEST⁵⁸. The *p*-value generated by denovolyzer was obtained from a Poisson test. For the genome-wide analysis, a Bonferroni corrected *p*-value threshold at $\alpha = 0.05$ is used, i.e., 1.3×10^{-6} as recommended by the denovolyzer developers⁵⁷.

Predefined gene-sets

Our primary analysis focused on comparing the burden of de novo variants or ultra-rare variants in two sets of mutation-intolerant genes as predicted by gnomAD v2.1.1: one set of genes with a high probability of being loss-of-function (LoF) intolerant (pLI score > 0.9, LOEUF > 0.35; $n = 2851$ genes), and one set of genes intolerant to missense variation (missense Z score > 3.09, $n = 952$ genes). We also analyzed a subset of those two gene sets.

The first gene-set includes genes highly expressed in nephron-progenitor cells (NPCs) in 18-week human embryonic kidneys⁵⁹ ($n = 610$ genes for the LoF intolerant set and $n = 228$ genes for the missense-intolerant set). The NPC gene set used expression data derived from single-cell RNA sequencing data by Hochane et al.⁵⁹. We specifically selected genes in the top decile of expression in NPCs compared to other cell types at this developmental stage.

The second gene-set includes genes highly expressed (top decile) during early murine kidney development at E15.5 ($n = 434$ genes for the LoF intolerant set, and $n = 189$ genes for the missense-intolerant set). The genes highly expressed during early murine kidney development were identified using bulk RNAseq assays. Four embryos from mixed C57BL/6 Tac mice were harvested at 15.5 days post-conception (dpc). Microdissected kidneys were submerged in RNAlater (Qiagen), then homogenized using a micropestle (Avantor #211-2100). Total RNA was purified by RNeasy mini kit (Qiagen) according to the manufacturer's instructions. Quantification and quality control were done by Bioanalyzer (now Agilent) followed by Illumina TruSeq library preparation and sequencing with Illumina HiSeq2500. Briefly, poly-A pull-down was used to enrich mRNAs from total RNA samples. RNA extraction was performed using the TruSeq RNA prep kit. 30 million single-end 100-bps reads were obtained per sample (i.e., both kidneys of one embryo). RTA (Illumina) was used for base calling, and bcl2fastq (version 1.8.4) for converting BCL to fastq format and adapter trimming. Reads were mapped to the reference genome UCSC/mm9 with Tophat version 2.1.0 with settings: 4 mismatches and 10 maximum multiple hits. The relative abundance of genes and splice isoforms was estimated with Cufflinks (version 2.0.2) with default settings to obtain FPKM values for each gene.

We also created a set of genes associated with ID⁶⁰ or autism spectrum disorder⁶¹ and subdivided the set into 962 genes intolerant to LoF variants and 461 genes intolerant to missense variants.

In addition, we analyzed a gene-set comprising 525 genes associated with congenital heart disease, compiled using existing clinical targeted panels offered to patients with congenital heart disease. The list was subdivided into a set of 154 genes intolerant to LoF variants and a set of 68 genes intolerant to missense variants.

As a negative control, we analyzed a Reactome gene-set comprising 1124 genes associated with the innate immune system⁶². We then subdivided the set into 195 genes intolerant to LoF variants and 96 genes intolerant to missense variants.

The p -values were corrected based on the number of gene-sets and models included in the analysis (10), so the Bonferroni corrected p -value = 0.001. To adjust for the multiple testing, we calculated the FDR using the FDR function in the fuzzySim R package⁶³. All statistical analyses were performed in R version 4.3.1. The list of genes in each gene set can be found online (<https://doi.org/10.5281/zenodo.15312278>).

Gene-burden analysis

We performed a gene-burden genome-wide search for enrichment of “qualifying variants” in protein-coding genes in 1742 CAKUT compared to 22,258 controls⁶⁴. We restricted the analyses to variants within CCDS regions or the 2 bp canonical splice sites. Furthermore, we only considered variants that fulfilled all of the following QC criteria: $\geq 10\times$ coverage of the site, quality score (QUAL) ≥ 50 , genotype quality score (GQ) ≥ 20 , quality by depth score (QD) ≥ 5 , mapping quality score

(MQ) ≥ 40 , read position rank sum score (RPRS) ≥ -3 , mapping quality rank sum score (MQRS) ≥ -10 , Fisher's strand bias score (FS) ≤ 60 (SNVs) or ≤ 200 (indels), strand odds ratio (SOR) ≤ 3 (SNVs) or ≤ 10 (indels), GATK variant quality score recalibration filter “PASS”, “VQSRTTrancheSNP90.00to99.00”, or “VQSRTTrancheSNP99.00to99.90”, alternate allele fraction for heterozygous calls ≥ 0.3 . To control the differences in coverage, only variants covered in at least 70% of the case-control cohort and with a maximal 7% difference in coverage between cases and controls were included. To control for population stratification, we split the cohort into clusters using principal component analysis (PCA) on a set of predefined variants to capture population structure⁶⁵. To identify clusters based on genetic ancestry, we used the Louvain method of community detection on the first six principal components (PCs)⁶⁴. To assess the quality of the clusters (Supplementary Table 6), we performed further dimensionality reduction using the uniform manifold approximation and projection (UMAP). We performed Fisher's exact test to find associations between genes with qualifying variants and the phenotype. A quantile–quantile (QQ) plot was generated to evaluate the resulting p -values⁶⁴. We then meta-analyzed the results of all the ten clusters and generated a combined p -value and odds ratio using the Cochran–Mantel–Haenszel test⁶⁶. To qualify, the variants' allele frequency (AF) had to be less than 1×10^{-5} in the gnomAD v2.1.1 popmax allele frequency and less than 1×10^{-4} in the Columbia Institute for Genomics Medicine biobank. This AF was chosen as all diagnostic variants identified had a gnomAD v2.1.1 MAF $< 1 \times 10^{-5}$. Only LoF variants with high confidence based on LOFTEE⁶⁵ and identified in regions with low regional allele frequency⁶⁷ were “qualified”. Missense variants located in constrained domains based on a subRV⁶⁸ domain score percentile $< 50\%$ and predicted damaging based on at least three of the following criteria: (1) REVEL⁶⁹ > 0.5 ; (2) PrimateAI⁷⁰ > 0.8 ; (3) AlphaMissense⁷¹ score > 0.6 ; and (4) CADD⁷² score > 25 were “qualified”. Three analyses were performed, one for both LoF and qualified missense variants, one for LoF only, and one for qualified missense only. After Bonferroni correction for the ~18,000 genes in the genome ($\alpha = 0.05$), genome-wide significance was identified as p -value $< 2.8 \times 10^{-6}$. To confirm that this correction was adequate, we performed a case-control analysis of synonymous variants only and identified the lowest p -value = 5.8×10^{-5} . The number of qualifying LoF variants per sample in constrained genes was calculated to compare the distribution in cases and controls.

Candidate genes analysis

To prioritize candidate genes amongst the genes with de novo variants, we identified genes not known to be associated with CAKUT but highly expressed in nephron-progenitor cells and in E15.5 murine kidneys, or with LoF de novo variants in genes associated with ID or autism spectrum disorder, and/or congenital heart disease. Amongst the genes identified through the gene-burden analysis under the LoF model, constrained genes (pLI > 0.5) with OR > 10, and at least two cases with qualifying variants and p -value < 0.07 were prioritized. In those candidate genes, rare variants were identified in all cases and controls, regardless of their quality, as both the de novo and case control quality filters are extremely stringent. Colleagues with CAKUT cohorts at Boston Children's Hospital, University Medical Center Utrecht, and others were contacted to query for additional variants in those genes. To determine the potential impact of rare variants on splicing, we added to the tools described above for the case-control analysis the TraP (Transcript-inferred Pathogenicity) score⁷³. Candidate variants in the cases were then Sanger-validated.

Low-frequency variants analysis

To identify genetic risk factors for CAKUT, we performed a single variant analysis of low-frequency variants (defined as gnomAD allele frequency $\leq 1\%$). To reduce the risk of genetic stratification, we only included unrelated cases ($n = 1566$) and matched controls ($n = 13,031$) with European genetic ancestry. We tested a total of 1497 low-

frequency variants in genes known to be associated with CAKUT, identified in at least ten individuals and covered in at least 90% of the cases and controls, gnomAD filter “PASS”, and with a maximum difference of 0.01 gnomAD v2.1.1 AF. Variants passed the same criteria as described in the case-control analysis, and the maximum allowed difference in coverage between cases and controls was 5%. Variant-level *p*-values were generated using Fisher’s exact two-sided test. As we ran a single model, after Bonferroni correction based on the number of variants, we identified *p*-value $\leq 3.3 \times 10^{-5}$ for genome-wide significant associations and a cutoff of OR > 1.5 and *p*-value $\leq 10^{-3}$ for suggestive associations. To determine the potential deleteriousness of low-frequency variants, we added to the tools described above for the case-control analysis the AlphaMissense score⁷¹.

Pathway analysis

To assess whether the candidate genes identified in this study co-express with any of the known CAKUT genes, we used COXPRESdb⁷⁴. We first used the “CoExSearch” option to identify the top 300 genes co-expressed with each candidate gene. We then extracted from those 300 genes the known CAKUT genes. We combined the lists obtained for each candidate gene and used the “NetworkDrawer” option using the automatic platform, Cytoscape, “draw PPIs”, and either the “add a few genes” or “add many genes”.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The exome sequencing and microarray data for the cases generated in this study have been deposited in dbGaP (phs001749.v2.p1) [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001749]. Controlled access is required to protect participant privacy and comply with consent. The consent does not restrict access to any particular disease or research type. Data are available upon submission of a Data Access Request via dbGaP, with approval contingent on a Data Use Agreement prohibiting commercial use or re-identification. Requests are reviewed within four weeks. Sequencing data for 2660 IgAN samples used as controls were not generated for this study but obtained from other Columbia University investigators and have been deposited in dbGaP under accession number phs002480.v5.p4. [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/molecular.cgi?study_id=phs002480.v5.p4]. Controlled access is required to protect participant privacy and comply with consent. The consent does not restrict access to any particular disease or research type. Data are available upon submission of a Data Access Request via dbGaP, with approval contingent on a Data Use Agreement prohibiting commercial use or re-identification. Requests are reviewed within four weeks. DDD data was not generated for this study, it was obtained from the European Genome-phenome Archive (EGA) under accession code EGAS00001000775. EGA access requires approval by their Data Access Committee and a DUA. Data for diagnostic variants identified from the study have been submitted to the NCBI ClinVar database [SCV006081115-SCV006081242]. The E15.5 WT mouse RNAseq data used in this study are available in the GEO database under accession code GSE300198. Gene lists are uploaded to <https://doi.org/10.5281/zenodo.15312278>. The exome data for the rest of the controls were not generated for this study but obtained from other Columbia University investigators. They are available from investigators upon request by contacting the corresponding author, Dr. Ali Gharavi, at ag2239@cumc.columbia.edu. The data will be shared upon establishment of a Data Use Agreement (DUA). Requests are reviewed within four weeks. Raw data for all graphs are reported in the Supplementary Information.

References

1. CDC. Centers for Disease Control and Prevention (CDC): Congenital Heart Defects (CHDs). <https://www.cdc.gov/ncbddd/heartdefects/data.html> (2024).
2. World Health Organization. Congenital Disorders https://www.who.int/health-topics/congenital-anomalies#tab=tab_1 (2025).
3. Murugapoopathy, V. & Gupta, I. R. A primer on congenital anomalies of the kidneys and urinary tracts (CAKUT). *CJASN* **15**, 723–731 (2020).
4. Verbitsky, M. et al. Genomic disorders and neurocognitive impairment in pediatric CKD. *J. Am. Soc. Nephrol.* **28**, 2303–2309 (2017).
5. Alharbi, S. A., Alshenqiti, A. M., Asiri, A. H., Alqarni, M. A. & Alqah-tani, S. A. The role of genetic testing in pediatric renal diseases: diagnostic, prognostic, and social implications. *Cureus* **15**, e44490 (2023).
6. Verbitsky, M. et al. The copy number variation landscape of congenital anomalies of the kidney and urinary tract. *Nat. Genet.* **51**, 117–127 (2019).
7. Thomas, R. et al. HNF1B and PAX2 mutations are a common cause of renal hypodysplasia in the CKiD cohort. *Pediatr. Nephrol.* **26**, 897–903 (2011).
8. Weber, S. et al. Prevalence of mutations in renal developmental genes in children with renal hypodysplasia: results of the ESCAPE study. *J. Am. Soc. Nephrol.* **17**, 2864–2870 (2006).
9. Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T. & Correa, A. Prevalence of congenital heart defects in metropolitan Atlanta, 1998–2005. *J. Pediatr.* **153**, 807–813 (2008).
10. San Agustin, J. T. et al. Genetic link between renal birth defects and congenital heart disease. *Nat. Commun.* **7**, 11103 (2016).
11. Clothier, J. & Absoud, M. Autism spectrum disorder and kidney disease. *Pediatr. Nephrol.* **36**, 2987–2995 (2021).
12. Moreno-De-Luca, D. et al. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am. J. Hum. Genet.* **87**, 618–630 (2010).
13. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
14. Qi, H. et al. De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLoS Genet.* **14**, e1007822 (2018).
15. Wang, T. et al. Integrated gene analyses of de novo variants from 46,612 trios with autism and developmental disorders. *Proc. Natl. Acad. Sci. USA* **119**, e2203491119 (2022).
16. Deciphering Developmental Disorders, S Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
17. Cirulli, E. T. et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436–1441 (2015).
18. Gregory, J. M., Fagegaltier, D., Phatnani, H. & Harms, M. B. Genetics of amyotrophic lateral sclerosis. *Curr. Genet Med Rep.* **8**, 121–131 (2020).
19. Eade, K. et al. Serine biosynthesis defect due to haploinsufficiency of PHGDH causes retinal disease. *Nat. Metab.* **3**, 366–377 (2021).
20. Zhu, X. et al. A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLoS Genet.* **13**, e1007104 (2017).
21. Sanna-Cherchi, S. et al. Exome-wide association study identifies GREB1L mutations in congenital kidney malformations. *Am. J. Hum. Genet.* **101**, 789–802 (2017).
22. Rehm, H. L. et al. The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. *Genet Med* **25**, 100947 (2023).

23. McCoy, M. D. et al. Genetics of kidney disorders in Phelan-McDermid syndrome: evidence from 357 registry participants. *Pediatr. Nephrol.* **39**, 749–760 (2024).
24. Neelathi, U. M. et al. Variants in NR6A1 cause a novel oculo vertebral renal syndrome. *Nat. Commun.* **16**, 6111 (2025).
25. Martino, J. et al. Mouse and human studies support DSTYK loss of function as a low-penetrance and variable expressivity risk factor for congenital urinary tract anomalies. *Genet. Med.* **25**, 100983 (2023).
26. Kolvenbach, C. M., Shril, S. & Hildebrandt, F. The genetics and pathogenesis of CAKUT. *Nat. Rev. Nephrol.* **19**, 709–720 (2023).
27. Claus, L. R., Snoek, R., Knoers, N. V. A. M. & Van Erde, A. M. Review of genetic testing in kidney disease patients: diagnostic yield of single nucleotide variants and copy number variations evaluated across and within kidney phenotype groups. *Am. J. Med. Genet. Pt C.* **190**, 358–376 (2022).
28. Verbitsky, M. et al. Genomic imbalances in pediatric patients with chronic kidney disease. *J. Clin. Invest.* **125**, 2171–2178 (2015).
29. Verbitsky, M. et al. Genomic disorders in CKD across the Lifespan. *J. Am. Soc. Nephrol.* **34**, 607–618 (2023).
30. Walawender, L., Becknell, B. & Matsell, D. G. Congenital anomalies of the kidney and urinary tract: defining risk factors of disease progression and determinants of outcomes. *Pediatr. Nephrol.* **38**, 3963–3973 (2023).
31. Harding, S. D. et al. The GUDMAP database—an online resource for genitourinary research. *Development* **138**, 2845–2853 (2011).
32. Zschocke, J., Byers, P. H. & Wilkie, A. O. M. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat. Rev. Genet.* **24**, 442–463 (2023).
33. Peddibhotla, S. et al. Expanding the genotype-phenotype correlation in subtelomeric 19p13.3 microdeletions using high resolution clinical chromosomal microarray analysis. *Am. J. Med. Genet.* **161**, 2953–2963 (2013).
34. Popowski, M., Lee, B.-K., Rhee, C., Iyer, V. R. & Tucker, H. O. Arid3a regulates mesoderm differentiation in mouse embryonic stem cells. *J. Stem Cell Ther. Transpl.* **1**, 52–62 (2017).
35. Webb, C. F. et al. The ARID family transcription factor bright is required for both hematopoietic stem cell and B lineage development. *Mol. Cell Biol.* **31**, 1041–1053 (2011).
36. Suzuki, N., Hirano, K., Ogino, H. & Ochi, H. Arid3a regulates nephric tubule regeneration via evolutionarily conserved regeneration signal-response enhancers. *eLife* **8**, e43186 (2019).
37. Webb, C. F. et al. A developmentally plastic adult mouse kidney cell line spontaneously generates multiple adult kidney structures. *Biochem. Biophys. Res. Commun.* **463**, 1334–1340 (2015).
38. Slattery, M. L., Pellatt, D. F., Mullany, L. E., Wolff, R. K. & Herrick, J. S. Gene expression in colon cancer: a focus on tumor site and molecular phenotype. *Genes Chromosomes Cancer* **54**, 527–541 (2015).
39. Stelloo, S. et al. Deciphering lineage specification during early embryogenesis in mouse gastruloids using multilayered proteomics. *Cell Stem Cell* **31**, 1072–1090.e8 (2024).
40. Shang, Z. et al. Single-cell RNA-seq reveals dynamic transcriptome profiling in human early neural differentiation. *GigaScience* **7**, gij117 (2018).
41. KDIGO Conference Participants. Genetics in chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int.* **101**, 1126–1141 (2022).
42. Franceschini, N. et al. Advancing genetic testing in kidney diseases: report From a National Kidney Foundation Working Group. *Am. J. Kidney Dis.* <https://doi.org/10.1053/j.ajkd.2024.05.010> (2024).
43. Johnson, J. L. & Abecasis, G. R. GAS POWER CALCULATOR: web-based power calculator for genetic association studies. Preprint at <https://doi.org/10.1101/164343> (2017).
44. Bruel, A.-L. et al. 2.5 years' experience of GeneMatcher data-sharing: a powerful tool for identifying new genes responsible for rare diseases. *Genet. Med.* **21**, 1657–1661 (2019).
45. Brophy, P. D. et al. A gene implicated in activation of retinoic acid receptor targets is a novel renal agenesis gene in humans. *Genetics* **207**, 215–228 (2017).
46. Batourina, E. et al. Vitamin A controls epithelial/mesenchymal interactions through Ret expression. *Nat. Genet.* **27**, 74–78 (2001).
47. Mishra, S. P. et al. A mechanism by which gut microbiota elevates permeability and inflammation in obese/diabetic mice and human gut. *Gut* **72**, 1848–1865 (2023).
48. Furth, S. L. et al. Design and methods of the chronic kidney disease in children (CKiD) prospective cohort study. *Clin. J. Am. Soc. Nephrol.* **1**, 1006–1015 (2006).
49. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
50. Groopman, E. E. et al. Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
51. Ren, Z. et al. ATAV: a comprehensive platform for population-scale genomic analyses. *BMC Bioinforma.* **22**, 149 (2021).
52. Gargano, M. A. et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
53. 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).
54. Fromer, M. & Purcell, S. M. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet.* **81**, 7.23.1–21 (2014).
55. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 v(2020).
56. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinforma.* **14**, 178–192 (2013).
57. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 1–15 (2015).
58. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
59. Hochane, M. et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLoS Biol.* **17**, e3000152 (2019).
60. PanelApp. *Intellectual Disability—Microarray and Sequencing (Version 5.204)*. <https://panelapp.genomicsengland.co.uk/panels/285/> (2025).
61. Basu, S. N., Kollu, R. & Banerjee-Basu, S. AutDB: a gene reference resource for autism research. *Nucleic Acids Res.* **37**, D832–D836 (2009).
62. REACTOME. *Innate Immune System (R-HSA-168249)*. <https://www.reactome.org/content/detail/R-HSA-168249> (2005).
63. Barbosa, A. M. fuzzySim: applying fuzzy logic to binary similarity indices in ecology. *Methods Ecol. Evol.* **6**, 853–858 (2015).
64. Povysil, G. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
65. Cameron-Christie, S. et al. Exome-based rare-variant analyses in CKD. *J. Am. Soc. Nephrol.* **30**, 1109–1122 (2019).

66. Povysil, G. et al. Assessing the role of rare genetic variation in patients with heart failure. *JAMA Cardiol.* **6**, 379–386 (2021).
67. Krishna Murthy, S. B. et al. Assisting the analysis of insertions and deletions using regional allele frequencies. *Funct. Integr. Genomics* **24**, 104 (2024).
68. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9 (2016).
69. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
70. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
71. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
72. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).
73. Gelfman, S. et al. Annotating pathogenic non-coding variants in genic regions. *Nat. Commun.* **8**, 236 (2017).
74. Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic Acids Res.* **51**, D80–D87 (2023).

Acknowledgements

Donald E. Wesson Research Fellowship from the ASN Foundation for Kidney Research (HMR). Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the National Institutes of Health (NIH): Genetics of Human Renal Hypodysplasia (5R01DK080099-13, AGG), George M O'Brien Center (5U54DK104309-10, AGG), Advancing equitable implementation of genomic medicine in nephrology (5K01DK132495-02, HMR), and R56 DK138164, P20 DK116191, R01 DK103184, R01 DK115574, SSC. Research reported in this publication was also supported by the National Institute of General Medical Sciences of the NIH (1S10OD030363-01A1 (SMM) and the Partner in Dutch Kidney Foundation ArtDECO consortium (DKF 20OC002, AvE and IL). The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003]. This study makes use of DECIPHER (<http://www.deciphergenomics.org>), which is funded by Wellcome [grant number WT223718/Z/21/Z]. See Nature PMID: 25533962 or <https://www.ddduk.org/> for full acknowledgment. Source: <https://ega-archive.org/studies/EGAS00001000775>. The authors thank the following individuals or groups for contributing control samples: A Signature Program of CURE, Anna Mae Diehl, B. Grinton, C. Hulette, Croasdaile Village, D. Lancet, E. Davis, E. Nading, Farfel, G. Cavalleri, Heidi White, I. Scheffer, J. Burke, J. McEvoy, J. Samuels, K. Whisenhunt, Kenneth Schmader, Mamata Yanamadala, N. Delanty, S. Sisodiya, Shelley McDonald, A. Need, A. Poduri, C. Chen, C. Depondt, Carol Woods, Chia-siang Chen, Cynthia Moylan, D. Levy, Duke University Health System Nonalcoholic Fatty Liver Disease Research Database and Specimen Repository, E. Pras, Epilepsy Genetics Initiative, G. Nestadt, K. Welsh-Bomer, M. Gennarelli, M. Hauser, M. Sum, M. Walker, Maher, UCB funding, Manal Abdelmalek, N. Katsanis, "National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology

(CHAVI) (U19-AI067854), National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery (UM1-AI100645)", National Institute on Aging (R01AG037212, P01AG007232). ALSO include: data collection and sharing for the WHICAP project (used as controls in this analysis) was supported by the Washington Heights-Inwood Columbia Aging Project (WHICAP, P01AG07232, R01AG037212, RF1AG054023) funded by the National Institute on Aging (NIA) and by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873. We acknowledge the WHICAP study participants and the WHICAP research and support staff for their contributions to this study, R. Buckley R. Wapner, S. Berkovic, S. Delaney, S. Palmer, S. Schuman, T. Young, the ALS Sequencing Consortium; the Washington University Neuromuscular Genetics Project, the Epi4K Consortium and Epilepsy Phenome/Genome Project, The Murdock study community registry and biorepository Pro0001196; Kristen Newby, V. Shashi, V. Shashi, Undiagnosed Diseases Network, Y. Wang.

Author contributions

A.A.G., S.S.-C., and H.M.R. conceptualized the research goals and aims. A.A.G., K.S.B., and H.M.R. designed the analyses. D.A.F., D.B.G., G.P., and I.L.L. developed the methodology. J.K., N.T., O.H., S.A.B., S.Y., and K.S.B. performed the programming and implemented the computer code and supporting algorithms. A.B., A.K., J.M., M.V., N.E., N.V., S.R.M., S.S., T.L., K.S.B., and H.M.R. performed the analyses. J.Z., S.Y., and K.S.B. performed the data curation. K.S.B. and H.M.R. synthesize study data. A.B., K.S.B., and H.M.R. created the figures. A.J.D., F.L., M.M., and V.K. managed and coordinated participants' enrollment. A.A., A.Bad., A.C.C., A.G., A.L.B., A.M., A.M.K., A.R., A.S., A.S.B., A.M.v.E., B.H.K., C.I., C.L.S., D.B., D.C., D.D., D.J.C., D.M., D.S., E.F., E.G., F.H., F.Lin, F.S., F.T., G.B.A., G.K., G.J., G.L.M., G.M., G.M.G., G.Mo., G.Z., I.G., I.L., I.P., J.R., J.w., J.Z., K.K., K.M., K.O.S., K.Z., L.B., L.G., M.B., M.K.B.K., M.M., M.M.-W., M.M.M.H., M.R., M.S., M.Sz., M.T., M.Z., N.S.U., O.B., O.m.B., P.A.C., P.E., P.M., P.S., P.Z., R.J.C., R.R., R.W., S.A., S.G., S.K., S.M., S.N., S.P., T.H., V.C., V.G., V.J.L., V.T., X.M., and Y.C. provided the patients' samples. H.M.R. wrote the first draft of the manuscript. A.A.G., S.S.C., and K.S.B. made critical revisions to the manuscript and provided critical comments. A.A.G. supervised the work. A.A.G., D.B.G., and R.P.L. secured the funding for this work.

Competing interests

A.G.G. has served on advisory boards for Natera through a service agreement with Columbia University. A.G.G. has served on advisory boards for Actio Biosciences, Novartis, Vera, Vertex, and Travere. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62319-3>.

Correspondence and requests for materials should be addressed to Hila Milo Rasouly or Ali G. Gharavi.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Hila Milo Rasouly ^{1,2,66}✉, Sarath Babu Krishna Murthy^{2,66}, Natalie Vena^{1,2}, Gundula Povysil^{3,4}, Andrew Beenken¹, Miguel Verbitsky ¹, Shirlee Shril⁵, Iris Lekkerkerker⁶, Sandy Yang², Atlas Khan ¹, David Fasel², Janewit Wongboonsin ^{7,8}, Jeremiah Martino^{1,9}, Juntao Ke¹, Naama Elefant^{2,3}, Nikita Tomar², Ofek Harnof ², Sergey Kisselev¹, Shiraz Bheda², Sivan Reytan-Miron¹, Tze Y. Lim^{1,10}, Anna Jamry-Dziurla¹¹, Francesca Lugani¹², Jun Y. Zhang ¹, Maddalena Marasa ^{1,13}, Victoria Kolupaeva^{1,2}, Emily E. Groopman^{1,14}, Gina Jin ¹, Iman Ghavami¹, Kelsey O. Stevens¹, Arielle C. Coughlin^{1,15}, Byum Hee Kil¹, Debanjana Chatterjee¹, Drew Bradbury¹, Jason Zheng¹, Karla Mehl¹, Maria Morban¹, Rachel Reingold^{1,16}, Stacy Piva¹, Xueru Mu¹, Adele Mittrori ^{1,17}, Agnieszka Szmigielska ¹⁸, Aleksandra Gliwińska¹⁹, Andrea Ranghino^{20,21}, Andrew S. Bomback¹, Andrzej Badenski ¹⁹, Anna Latos-Bielenska¹¹, Valentina Capone²², Anna Materna-Kiryluk¹¹, Antonio Amoroso ^{23,24}, Claudia Izzi²⁵, Claudio La Scola^{26,27}, David Jonathan Cohen¹, Domenico Santoro ²⁸, Dorota Drozd ²⁹, Enrico Fiaccadori³⁰, Fangming Lin ³¹, Francesco Scolari³², Francesco Tondolo³³, Gaetano La Manna ^{33,34}, Gerald B. Appel¹, Gian Marco Ghiggeri ¹², Gianluigi Zaza³⁵, Giovanni Montini ^{22,36}, Giuseppe Masnata ³⁷, Grażyna Krzemien¹⁸, Isabella Pisani²⁹, Jai Radhakrishnan¹, Katarzyna Zachwieja²⁸, Loreto Gesualdo¹⁷, Luigi Biancone^{38,39}, Davide Meneghesso⁴⁰, Malgorzata Mizerska-Wasiak¹⁸, Marcin Tkaczyk⁴¹, Marcin Zaniew⁴², Maria K. Borszewska-Kornacka⁴³, Maria Szczepanska ¹⁹, Marijan Saraga⁴⁴, Maya K. Rao¹, Monica Bodria⁴⁵, Monika Miklaszewska²⁸, Natalie S. Uy⁴⁶, Olga Baraldi^{33,47}, Omar Bjanid¹⁹, Pasquale Esposito^{48,49}, Pasquale Zamboli⁵⁰, Pierluigi Marzuillo ⁵¹, Pietro A. Canetta ¹, Przemyslaw Sikora⁵², Rik Westland ⁵³, Russell J. Crew¹, Shumyle Alam⁵⁴, Stefano Guarino⁵¹, Susanna Negrisolo^{55,56}, Thomas Hays ⁵⁷, Shrikant Mane⁵⁸, Valeria Grandinetti³², Velibor Tasic ⁵⁹, Vladimir J. Lozanovski ^{59,60}, Yasar Caliskan ⁶¹, David Goldstein ^{3,62}, Richard P. Lifton⁶³, Iuliana Ionita-Laza^{64,65}, Krzysztof Kiryluk ¹, Albertien M. van Eerde ⁶, Friedhelm Hildebrandt ⁵, Simone Sanna-Cherchi ¹ & Ali G. Gharavi ^{1,2}✉

¹Division of Nephrology, Department of Medicine, Columbia University Medical Center, New York, NY, USA. ²Center for Precision Medicine and Genomics, Department of Medicine, Columbia University Medical Center, New York, NY, USA. ³Genomic & Bioinformatics Analysis Resource (GenBAR), Columbia University Medical Center, New York, NY, USA. ⁴Waypoint Bio, New York, NY, USA. ⁵Division of Nephrology, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ⁶Department of Genetics, University Medical Center Utrecht, Utrecht, Netherlands. ⁷Renal Division, Department of Internal Medicine, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. ⁸Division of Renal Medicine, Brigham and Women's Hospital, Boston, MA, USA. ⁹Dyson College of Arts and Sciences, Pace University, Pleasantville, NY, USA. ¹⁰Unit of Genomic Variability and Complex Diseases, Department of Medical Sciences, University of Turin, Turin, Italy. ¹¹Department of Medical Genetics, Poznan University of Medical Sciences, Poznan, Poland. ¹²Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Istituto Giannina Gaslini, Genoa, Italy. ¹³Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Clinical Research Center for Rare Diseases Aldo e Cele Daccò, Bergamo, Italy. ¹⁴Children's National Hospital, Washington, D.C., USA. ¹⁵Icahn School of Medicine at Mount Sinai, New York, NY, USA. ¹⁶Department of Dermatology, Weill Cornell Medicine, New York, NY, USA. ¹⁷Division of Nephrology, Dialysis and Transplantation Unit, Department of Precision and Regenerative Medicine and Ionian Area (DiMePre-J), University of Bari Aldo Moro, Bari, Italy. ¹⁸Department of Pediatrics and Nephrology, Medical University of Warsaw, Warsaw, Poland. ¹⁹Department of Pediatrics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland. ²⁰Nephrology, Dialysis and Renal Transplantation Unit, Azienda Ospedaliera Universitaria delle Marche, Ancona, Italy. ²¹Università Politecnica delle Marche, Ancona, Italy. ²²Division of Pediatric Nephrology, Dialysis and Transplant Unit, Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, Milan, Italy. ²³Department of Medical Sciences, University of Turin, Turin, Italy. ²⁴Non-coding RNAs and RNA-based Therapeutics, Italian Institute of Technology, CMP3VdA, Aosta, Italy. ²⁵Clinical Genetics Unit, Department of Molecular and Translational Medicine, University of Brescia and Spedali Civili, Brescia, Italy. ²⁶Pediatric Nephrology, Pediatric Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy. ²⁷Paediatric Palliative Care, Community Paediatrics Unit Rimini Riccione, Department of Women's Health, Childhood and Adolescence Rimini. AUSL of Romagna, Rimini, Italy. ²⁸Division of Nephrology, University of Messina, Messina, Italy. ²⁹Department of Pediatric Nephrology and Hypertension, Faculty of

Medicine, Jagiellonian University Medical College, Krakow, Poland. ³⁰Nephrology Unit, Parma University Hospital, Parma, Italy. ³¹Division of Pediatric Nephrology, Department of Pediatrics, Columbia University Medical Center, New York, NY, USA. ³²Division of Nephrology, Spedali Civili and University, Brescia, Italy. ³³Nephrology, Dialysis and Kidney Transplant Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy. ³⁴Department of Medical and Surgical Sciences (DIMEC), Alma Mater Studiorum University of Bologna, Bologna, Italy. ³⁵Nephrology, Dialysis and Transplantation Unit, Department of Pharmacy, Health and Nutritional Sciences, University of Calabria, Arcavacata, Italy. ³⁶Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy. ³⁷Department of Complex Diseases and Pediatric Nephrology, ARNAS Brotzu Hospital, Cagliari, Italy. ³⁸Division of Nephrology, Dialysis and Transplantation, Department of Medical Sciences, University of Torino, Torino, Italy. ³⁹AOU Città della Salute e della Scienza Hospital, Torino, Italy. ⁴⁰Division of Nephrology, Hospital University of Padova, Padova, Italy. ⁴¹Department of Pediatrics, Immunology and Nephrology, Polish Mother's Memorial Hospital Research Institute, Lodz, Poland. ⁴²Department of Pediatrics, University of Zielona Góra, Zielona Góra, Poland. ⁴³Department of Pediatrics, School of Medicine with the Division of Dentistry in Zabrze, Medical University of Silesia in Katowice, Katowice, Poland. ⁴⁴University of Split, School of Medicine, Split, Croatia. ⁴⁵Primary Care Unit, Distretto Sud-Est, Parma, Italy. ⁴⁶Division of Nephrology, Department of Pediatrics, Weill Cornell Medicine, New York, NY, USA. ⁴⁷Nephrology and Dialysis Unit, Santa Maria delle Croci Hospital-Ravenna, AUSL Della Romagna, Ravenna, Italy. ⁴⁸Department of Internal Medicine and Medical Specialties (DIMI), University of Genoa, Genoa, Italy. ⁴⁹Unit of Nephrology, Dialysis and Transplantation, IRCCS Ospedale Policlinico San Martino, Genoa, Italy. ⁵⁰Division of Nephrology, AORN San Giuseppe Moscati, Avellino, Italy. ⁵¹Department of Woman, Child and of General and Specialized Surgery, Università degli Studi della Campania "Luigi Vanvitelli", Naples, Italy. ⁵²Department of Pediatric Nephrology, Medical University of Lublin, Lublin, Poland. ⁵³Department of Pediatric Nephrology, Emma Children's Hospital—Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ⁵⁴Division of Urology, El Paso Children Hospital, El Paso, TX, USA. ⁵⁵Department of Women's and Children's Health, University of Padova, Padua, Italy. ⁵⁶Pediatric Research Institute "IRP Città della Speranza", Padua, Italy. ⁵⁷Division of Neonatology, Department of Pediatrics, Columbia University Medical Center, New York, NY, USA. ⁵⁸Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. ⁵⁹University Children's Hospital, Faculty of Medicine, Skopje, North Macedonia. ⁶⁰Department of General, Visceral, and Transplant Surgery, University Medical Center of the Johannes Gutenberg University, Mainz, Germany. ⁶¹Division of Nephrology, Saint Louis University, St. Louis, MO, USA. ⁶²Actio Biosciences, San Diego, CA, USA. ⁶³Rockefeller University, New York, NY, USA. ⁶⁴Department of Biostatistics, Columbia University, New York, NY, USA. ⁶⁵Department of Statistics, Lund University, Lund, Sweden. ⁶⁶These authors contributed equally: Hila Milo Rasouly, Sarath Babu Krishna Murthy.

✉ e-mail: Hila.MiloRasouly@columbia.edu; ag2239@cumc.columbia.edu