LETTER TO JMG

Comprehensive genomic analysis of *PKHD1* mutations in ARPKD cohorts

A M Sharp, L M Messiaen, G Page, C Antignac, M-C Gubler, L F Onuchic, S Somlo, G G Germino, L M Guay-Woodford

J Med Genet 2005;**42**:336–349. doi: 10.1136/jmg.2004.024489

Autosomal recessive polycystic kidney disease (ARPKD; MIM 263200) is an important childhood nephropathy, occurring in 1 in 20 000 live births.[1] The clinical phenotype is dominated by dilatation of the renal collecting ducts, biliary dysgenesis, and portal tract fibrosis. Affected children often present in utero with enlarged, echogenic kidneys, as well as oligohydramnios secondary to poor urine output. Approximately 30% of affected neonates die shortly after birth as a result of severe pulmonary hypoplasia and secondary respiratory insufficiency. Those who survive the perinatal period express widely variable disease phenotypes with systemic hypertension, renal insufficiency, and portal hypertension due to portal tract fibrosis as the most common clinical features.[2]

Linkage analysis indicates that mutations in a single locus on chromosome 6p12 are responsible for all typical forms of ARPKD.[3 4] Two groups working independently have identified *PKHD1* (MIM 606702) as the locus responsible for ARPKD and have demonstrated that this novel gene is among the largest in the human genome, extending over at least 470 kb and including a minimum of 86 exons.[5 6] Both *PKHD1* and its mouse orthologue (*Pkhd1*) encode a complex and extensive array of splice variants, with most abundant transcriptional expression in fetal and adult kidney and weaker expression in other tissues including liver and pancreas.[5 7] The longest *PKHD1* transcript includes 67 exons with an open reading frame (ORF) composed of 66 exons that encode a 4074 amino acid protein, polyductin/fibrocystin.[5 6] The full length protein is predicted to have several immunoglobulin-like, plexin, transcription factor (IPT) domains and multiple parallel beta-helix 1 (PbH1) repeats in its approximately 3860 amino acid extracellular amino terminus; a single transmembrane (TM) spanning domain; and a short, cytoplasmic carboxyl terminus with potential phosphorylation sites. Alternatively spliced transcripts are predicted to fall into two broad groups. The first subset, polyductin-M, is comprised of polypeptides that contain the single TM element but vary with respect to inclusion of the other predicted domains. The second subset, polyductin-S, lacks the TM domain and thus its members may be secreted.[5] The *PKHD1* gene products share structural features with hepatocyte growth factor receptor and plexins, members of a superfamily of proteins involved in regulation of cellular adhesion and repulsion as well as cell proliferation. In addition, recent studies have demonstrated that like other cystoproteins, polyductin/fibrocystin is expressed in the primary apical cilium.[6 8–12]

Based on the available data, ARPKD appears to result from partial or complete loss of polyductin/fibrocystin function. However, the mechanisms by which *PKHD1* mutations cause clinical disease phenotypes are not well understood. Gene based analyses have been complicated by the large gene size and reported mutation detection rates have ranged from 47% to 61%.[5 6 13–15] The limited mutation detection rates and the absence of mutational hot spots in *PKHD1* have confounded efforts to examine potential genotype-phenotype correlations. These methodological challenges must be overcome before such correlative analyses are revealing and gene based examination is robust enough for clinical diagnostic testing.

In the current study, we have refined the mutation detection strategies and evaluated all 86 predicted exons, including the 67 exons in the longest ORF transcript as well the 19 alternative exons. Our mutation detection rate of 82.7% is the best reported to date in an ethnically diverse population with a wide range of ARPKD associated phenotypes. We have examined potential correlations between disease phenotypes and specific mutational mechanisms and/or linear positions along *PKHD1*. Consistent with previous

Key points

- Mutations at a single locus, *PKHD1*, are responsible for all typical forms of autosomal recessive polycystic kidney disease (ARPKD). We have refined previously reported mutation detection strategies and evaluated all 86 predicted *PKHD1* exons, including the 67 exons in the longest open reading frame transcript as well the 19 alternative exons.
- We have rigorously examined the predicted pathogenicity of amino acid substitutions using the matrix criteria described by Miller and Kumar (Hum Mol Genet 2001;21:2319), as well as potential splice site alterations using the Splice Site Prediction by Neural Network (SSPNN) algorithm at http://www.fruitfly.org/seq_tools/splice.html.
- Our mutation detection rate of 82.7% is the best reported to date in an ethnically diverse population with a wide range of ARPKD associated phenotypes.
- We have re-categorised all reported mutations by numbering the exons according to their genomic order and providing the genomic nucleotide designation to clarify position. This re-compilation provides an essential platform for the robust interpretation of *PKHD1* variants identified in the course of prenatal testing or pre-implantation genetic diagnosis. We have submitted this data compilation to the *PKHD1* database maintained at http://www.humgen.rwth-aachen.de.

Abbreviations: ARPKD, autosomal recessive polycystic kidney disease; HNF-1, hepatocyte nuclear factor-1; ORF, open reading frame; SNP, single nucleotide polymorphism; SSPNN, Splice Site Prediction by Neural Network; TM, transmembrane

studies, we found that mutations are distributed along the *PKHD1* gene and patients carrying two potentially chain terminating mutations expressed the severe perinatal phenotype.[13–15] We have rigorously examined the predicted pathogenicity of amino acid substitutions and potential splice site alterations in both our dataset and all sequence variants reported to date. These systematic analyses, annotation of missense changes, and characterisation of mutations based on genomic position will provide an essential platform for the robust interpretation of *PKHD1* variants identified in the course of prenatal testing or pre-implantation genetic diagnosis.

METHODS

Patients and samples

A cohort of 59 unrelated families and individuals was ascertained from the databases at the University of Alabama at Birmingham (North American ARPKD Database and prenatal testing database at UAB Molecular Genomics Laboratory). ARPKD was diagnosed according to previously established criteria.[16] In addition, 16 affected fetuses from unrelated families with at least one previous ARPKD affected child were identified by haplotype analysis.[1] The pregnancies were terminated and ARPKD was confirmed by histopathological analyses.[17] Study subjects were ethnically diverse and represented the full spectrum of clinical presentations for ARPKD (table 1). Whole blood, chorionic villus samples, or products of conception were obtained from study subjects and family members under informed consent approved by the University of Alabama at Birmingham Institutional Review Board. Genomic DNA was prepared using standard protocols.[15] In addition, DNA samples were obtained from 100 anonymous, unrelated normal individuals.

PCR amplification

The 67 exons that compose the transcript (GenBank accession no. AF4800064) with the longest predicted ORF of *PKHD1* (GenBank accession no. AY129465) were amplified as a set of 80 amplicons. PCR primers were designed to amplify exon sequences, the adjacent splice sites, and 40–50 nucleotides of flanking sequence on each side as 200–400 bp products. When the exon size was greater than 400 bp, a series of overlapping primers was designed to limit the size of the amplicons. Exons 32, 59, 65, and 71 were amplified in seven, four, four, and two overlapping fragments, respectively. A number of the primers used in this study were described previously[15] and the remainder were designed using the Primer3 program (supplemental table 1; supplemental tables 1–4 are available from http://jmg.bmjjournals.com/supplemental). Several of these primers sets also amplified the predicted alternative exons 20a, 32a, 32b, 32c, 39a, 39b, 41a, 44a, 51a, 51b, 60a, 66a, and 71a[5] and we designed primers to amplify the remaining alternative exons 38, 38a, 39b, 41a, 62, 63, 64, and 71b (supplemental table 2).

Amplifications were performed in an MJ Research Dyad thermocycler using 100 ng genomic DNA in a total reaction volume of 50 µl, incorporating: 1×HotMasterMix (Eppendorf, Hamburg, Germany) and 400 pM of each forward and reverse primer (Integrated DNA Technologies, Coralville, IA, USA). The PCR reaction mixes were subjected to the following ''touchdown'' thermal cycling protocol: initial denaturation at 95℃ for 5 min; followed by 10 cycles at 95℃ for 30 s, 65℃ to 55℃ (−1℃ per cycle) for 30 s, 65℃ for 30 s; followed by 35 cycles at 95℃ for 30 s, 53℃ for 30 s, 65℃ for 30 s; followed by a final extension for 7 min at 65℃. The PCR reaction volume provided amplification product for denaturing high performance liquid chromatography (DHPLC) analysis as well as for sequencing templates.

Mutation detection

Mutation detection was performed by heteroduplex analysis using the Transgenomic WAVE DHPLC system (Transgenomic, Omaha, NE). For each amplicon, the optimal elution gradient for amplicon size and GC content as well as the optimal denaturing temperature were determined according to the WaveMaker version 4.1.44 system control software. Briefly, amplicons were prepared from three normal control samples and injected at five different temperatures: the determined optimal melting temperature, as well as temperatures ±0.5–1.5℃. Where necessary, GC clamps were included in primer sequences to normalise the melting profile of the amplicon and to permit higher injection temperatures. The DHPLC conditions used in these analyses are provided in supplemental table 1.

To enhance heteroduplex formation, PCR products were denatured at 95℃ for 5 min and allowed to gradually reanneal in the thermocycler block using 48 cycles of 1 min each with temperatures stepped down from 93.5℃ to 21.5℃ (−1.5℃ per cycle). Samples were then placed into the DHPLC autosampler and held at 10℃ as 5–8 µl aliquots were injected onto the column at the appropriate analytic temperature. The mobile phase consisted of a mixture of buffer A (0.1 M triethylammonium acetate (TEAA)), and buffer B (25% acetonitrile in 0.1 M TEAA) as per the manufacturer's instructions. Each fragment was eluted with a linear acetonitrile gradient at a flow rate of 0.9 ml/min. A normal control was processed with each batch of patient amplicons, and when possible, amplicons harbouring known mutations were included in the DHPLC analysis. For patients in whom no pathogenic variants were detected by DHPLC, their DNA samples were mixed with wildtype control DNA templates prior to re-amplification and DHPLC analysis. This ''sample mixing'' enhanced the DHPLC based detection of homozygous sequence variants in the patient DNAs.

Samples exhibiting altered chromatogram patterns or retention times with respect to normal controls were subjected to direct sequence analysis. PCR products were purified using the PCR Product Presequencing Kit (USB, Cleveland, OH, USA) and sequenced in both directions with BigDye Deoxy Terminator cycle sequencing on an ABI 3100 DNA sequencer (Applied Biosystems, Foster City, CA, USA). The primers used for DHPLC were also used as sequencing primers. Sequences were aligned and analysed using SeqMan II version 5.05 (DNASTAR, Madison, WI). All putative mutations were tested by segregation analysis when family material was available. To assess whether missense changes represented potential pathogenic mutations or benign polymorphisms, DNA samples from 100 unrelated normal control subjects were examined using DHPLC and sequence analysis.

The sequence variants were described using the nomenclature reported by Dunnen and Antonarakis[18] and the Human Genome Variation Society website (HGVS; http://www.hgvs.org/, accessed May 31, 2004). Nomenclature of *PKHD1* sequence variants was based upon the *PKHD1* mRNA sequence (GenBank accession no. AF4800064) with the A of the start codon designated as nucleotide 1. Exon and gDNA nucleotide numbers were reported from the *PKHD1* genomic DNA sequence (GenBank accession no. AY129465).

Statistical analyses

Genotype-phenotype comparisons were performed in those patients for whom two putative mutations were identified using a contingency table and a χ^2 significance test. We compared allelic and genotypic frequencies of chain terminating, splice site variants, and amino acid substitution mutations among our patients with and without perinatal

Table 1 *PKHD1* variants in the study cohort

Presentation	Variant type	Patient	Ethnic origin	Exon	Variant	
					cDNA level	Protein level
Fetal	Truncation/truncation	F3	French	59	c.9689delA	p.D3230fs
				59	c.9689delA	p.D3230fs
		F6	French	5	c.353delG	p.S118fs
				41	c.6653_6660dup	p.T2221fs
		F7	French	36	c.5895dupA	p.L1966fs
				69	c.11524 C→T	p.R3842X
		F13	French	59	c.9689delA	p.D3230fs
				59	c.9689delA	p.D3230fs
		F15	French	5	c.370 C→T	p.R124X
				36	c.5895dupA	p.L1966fs
Fetal	Splice/splice	F16	French	IVS32	c.5237-1 G→A	
				60	c.9998 G→C	p.R3333T
Fetal	Truncation/missense	F1	French	22	c.2269 A→C	p.I757L
				24	c.2520_2526del7	p.Y841fs
		F2	French	6	c.410 C→G	p.P137R
				23	c.2341 C→T	p.R781X
		F11	French	3	c.107 C→T	p.T36M
				59	c.9689delA	p.D3230fs
		F14	French	3	c.107 C→T	p.T36M
				36	c.5895dupA	p.L1966fs
		112-1	Caucasian American	12	c.806 G→T	p.G269V
				32	c.3761_3762delCCinsG	p.A1254fs
Fetal	Truncation/splice	F4	French	IVS5	c.391-5 A→G	
				59	c.9689delA	p.D3230fs
		F5	French	IVS50	c.7912-1 G→C	
				59	c.9689delA	p.D3230fs
Fetal	Splice/missense	F12	French	IVS33	c.5381-2 A→C	
				34	c.5410 C→T	p.R1804C
Fetal	Missense/missense	F10	French	53	c.8206 T→G	p.W2736G
				59	c.9719 G→T	p.R3240L
Fetal	Truncation	F9	French	34	c.5582delT	p.F1861fs
		36-4	Caucasian American	15	c.1212 G→A	p.W404X
Fetal	Splicing	12-5	Caucasian American	IVS37	c.6121+3_6121+4insT	
		F8	French	IVS53	c.8302+3 G→C	
Perinatal	Truncation/truncation	111-1	Caucasian American	32	c.3761_3762delCCinsG	p.A1254fs
				32	c.3761_3762delCCinsG	p.A1254fs
		39-1	Hispanic American	59	c.9689delA	p.D3230fs
				59	c.9689delA	p.D3230fs
		48-1	Hispanic American	40	c.6383delT	p.L2128fs
				59	c.9689delA	p.D3230fs
Perinatal	Truncation/splice	2-5	Caucasian American	IVS13	c.976+2 T→A	
				32	c.3761_3762delCCinsG	p.A1254fs
		10-20	Hispanic American	29	c.3364 G→A	p.G1122S
				32	c.3761_3762delCCinsG	p.A1254fs
		21-1	Caucasian American	14	c.977 G→T	p.G326V
				19	c.1774 C→T	p.R592X
Perinatal	Splice/in-frame indel	2-18	Caucasian American	IVS40	c.6490+2 T→C	
				53	c.8246_8247delGGinsCC	p.W2749S
Perinatal	Truncation/missense	6-3	Caucasian American	22	c.2180 dupA	p.N727fs
				65	c.10444 C→T	p.R3482C
		89-1	Caucasian American	58	c.8870 T→C	p.I2957T
				65	c.10637delT	p.V3546fs
		84-1	Hispanic American	32	c.3848 C→A	p.S1283X
				59	c.9311 T→G	p.F3104C
		13-1	Caucasian American	36	c.5895dupA	p.L1966fs
				54	c.8425 G→A	p.G2809R
		36-6	Caucasian American	52	c.8114delG	p.G2705fs
				56	c.8581 A→G	p.S2861G
		43-1	Caucasian American	3	c.107 C→T	p.T36M
				11	c.711_714delAATG	p.M238fs
		9-1	Caucasian American	36	c.5895dupA	p.L1966fs
				58	c.8870 T→C	p.I2957T
		19-1	Caucasian American	32	c.4415-4418del4ins11	p.C1472fs
				59	c.9725 G→T	p.G3242V
Perinatal	Splice/missense	50-1	Caucasian American	IVS12	c.881-1 G→A	
				69	c.11611 T→C	p.W3871R
Perinatal	Missense/missense	7-14	Caucasian American	30	c.3367 G→A	p.G1123S
				56	c.8588 A→G	p.Y2863C
		9-12	Caucasian American	3	c.107 C→T	p.T36M
				55	c.8518 C→T	p.R2840C
		10-16	Caucasian American	3	c.107 C→T	p.T36M
				3	c.107 C→T	p.T36M
		33-1	Caucasian American	3	c.107 C→T	p.T36M
				65	c.10805 C→T	p.P3602L
				65	c.10926 G→A	p.M3642I
		49-1	Caucasian American	30	c.3367 G→A	p.G1123S
				32	c.4751 G→T	p.S1584I

Table 1 Continued

Presentation	Variant type	Patient	Ethnic origin	Exon	Variant	
					cDNA level	Protein level
Perinatal	Truncation	28-4	Caucasian American	23	c.2341 C→T	p.R781X
		29-1	Caucasian American	32	c.3761_3762delCCinsG	p.A1254fs
Perinatal	Missense	2-15	Caucasian American	32	c.5236 G→A	p.G1746S
		3-12	Hispanic American	16	c.1409 G→A	p.G470D
		110-1	Caucasian American	16	c.1342 G→C	p.G448R
Perinatal		42-1	Caucasian Canadian		No mutation; no mixing	
		5-7	Caucasian American		No mutation; no mixing	
Perinatal		10-24	Caucasian American		No mutation after mixing	
		45-1	Caucasian American		No mutation after mixing	
Older presentation	Truncation/missense	7-1	Caucasian American	32	c.3761_3762delCCinsG	p.A1254fs
				32	c.5160 C→A	p.D1720E
		8-6	Caucasian American	9	c.664 A→G	p.I222V
				58	c.8829dupC	p.I2944fs
		14-1	Caucasian American	9	c.664 A→G	p.I222V
				32	c.3761_3762delCCinsG	p.A1254fs
		10-1	Caucasian American	30	c.3407 A→G	p.Y1136C
				32	c.3761_3762delCCinsG	p.A1254fs
		29-5	Caucasian American	3	c.107 C→T	p.T36M
				32	c.3761_3762delCCinsG	p.A1254fs
Older presentation	Truncation/in-frame deletion	15-1	Caucasian American	15	c.1159_1161delAAT	p.N387del
				36	c.5825 A→G	p.D1942G
Older presentation	Splice/missense	SB131	Saudi Arabian	22	c.2279 G→A	p.R760H
				32	c.4870 C→T	p.R1624W
Older presentation	Missense/missense	3-1	Caucasian American	30	c.3367 G→A	p.G1123S
				30	c.3367 G→A	p.G1123S
		4-1	Caucasian American	3	c.107 C→T	p.T36M
				9	c.664 A→G	p.I222V
		5-1	Caucasian American	37	c.5912 G→A	p.G1971D
				37	c.5912 G→A	p.G1971D
		8-19	South African	20	c.1880 T→A	p.M627K
				20	c.1880 T→A	p.M627K
		8-24	South African	20	c.1880 T→A	p.M627K
				20	c.1880 T→A	p.M627K
		10-10	Egyptian American	3	c.107 C→T	p.T36M
					c.107 C→T	p.T36M
		37-1	Caucasian American	3	c.107 C→T	p.T36M
				34	c.5387 T→C	p.L1796S
		7-4	Caucasian American	3	c.107 C→T	p.T36M
				44	c.6992 T→A	p.I2331L
		92-1	Turkish	15	c.1180 G→A	p.A394T
				54	c.8315 T→C	p.L2772P
		SB213	Saudi Arabian	32	c.4870 C→T	p.R1624W
				32	c.4870 C→T	p.R1624W
		SB253	Saudi Arabian	32	c.4870 C→T	p.R1624W
				32	c.4870 C→T	p.R1624W
Older presentation	Truncation	46-2	Caucasian American	36	c.5895dupA	p.L1966fs
		32-1	Caucasian American	59	c.9689DelA	p.D3230fs
		10-28	Caucasian American	20	c.1854delA	p.G619fs
Older presentation	Missense	10-5	Caucasian American	55	c.8518 C→T	p.R2840C
Older presentation		3-5	Caucasian American		No mutation; no mixing	
Older presentation		SB137	Saudi Arabian		No mutation after mixing	
CHF	Splice/missense	72-1	Hispanic American	IVS35	c.5751+3 G→A	
				59	c.9107 T→G	p.V3036G
CHF	Missense/missense	74-1	French Canadian	56	c.8581 A→G	p.S2861G
				58	c.8870 T→C	p.I2957T
				65	c.10319 T→A	p.V3440D

Nomenclature of *PKHD1* sequence variants was based upon the *PKHD1* mRNA sequence (GenBank accession no. AF480064) with the A of the start codon designated as nucleotide 1. CHF, congenital hepatic fibrosis.

presentation. Analyses were conducted in Excel, SPSS, and SAS.

RESULTS

PKHD1 mutations

We performed a systematic DHPLC based analysis of *PKHD1* mutations in a cohort of 75 unrelated ARPKD patients. Samples were screened for all 67 exons comprising the transcript with the longest continuous *PKHD1* ORF as well as the 19 predicted alternative exons. We detected a total of 173 variants in the *PKHD1* gene, 92 of which have not been previously reported. The sequence variations included 10% frameshifts (50% novel), 3.3% nonsense alterations (100% novel), 8% splicing variants (77% novel), 41% amino acid

Table 2 Truncating variants

Variant type	Exon	Variant: cDNA nucleotide	Variant: gDNA nucleotide	Variant: cDNA level	Protein level	Stop codon	Variant detection		Reference*
Frameshift	5	353	7635	c.353delG	p.S118fs	152	1/150		1, 2
	11	711	18048	c.711_714delAATG	p.M238fs	244	1/150		1, 5
	20	1854	33424	c.1854delA	p.G619fs	621	1/150	Novel	1
	22	2180	37316	c.2180dupA	p.N727fs	798	1/150	Novel	1
	24	2520	41496	c.2520_2526del7	p.Y841fs	871	1/150	Novel	1
	30	3528	59384	c.3528dupC	p.S1177fs	1189	1/200	Novel	1
	32	3761	61523	c.3761_3762delCCinsG	p.A1254fs	1302	10/150 (9)		1, 2, 3, 4, 5, 6
	32	4415	62177	c.4415_4418del4ins TATTCCCCTCT	p.C1472fs	1507	1/150	Novel	1
	34	5582	70144	c.5582delT	p.F1861fs	1973	1/150	Novel	1
	36	5895	127730	c.5895dupA	p.L1966fs	1969	6/150 (6)		1, 2, 3, 4, 5, 6, 7
	40	6383	175707	c.6383delT	p.L2128fs	2128	1/150		1, 6
	41	6653	178301	c.6653_6660dup	p.T2221fs	2229	1/150	Novel	1
	52	8114	251150	c.8114delG	p.G2705fs	2715	1/150	Novel	1
	58	8829	334291	c.8829dupC	p.I2944fs	2949	1/150		1, 4, 5
	59	9689	339686	c.9689delA	p.D3230fs	3263	11/150 (8)		1 ,2, 6
	65	10637	428124	c.10637delT	p.V3546fs	3567	1/150		1, 7
Nonsense	5	370	7652	c.370 C→T	p.R124X	124	1/150	Novel	1
	15	1212	27623	c.1212 G→A	p.W404X	404	1/150	Novel	1
	19	1774	31923	c.1774 C→T	p.R592X	592	1/150	Novel	1
	23	2341	39014	c.2341 C→T	p.R781X	781	2/150	Novel	1
	32	3848	61610	c.3848 C→A	p.S1283X	1283	1/150	Novel	1
	69	11524	454907	c.11524 C→T	p.R3842X	3842	1/150	Novel	1
In-frame deletion	15	1159	27570	c.1159_1161delAAT	p.N387del	4073	1/150		1, 4

Nomenclature of *PKHD1* sequence variants was based upon the *PKHD1* mRNA sequence (GenBank accession no. AF4800064) with the A of the start codon designated as nucleotide 1. *PKHD1* exons and genomic DNA (gDNA) nucleotide numbers (GenBank accession no. AY129465) were included for clarification of variant location.
*Numbers indicate the following: [1]current study, [2]Bergmann *et al*,[20] [3]Bergmann *et al*,[13] [4]Furu *et al*,[15] [5]Onuchic *et al*,[5] [6]Rossetti *et al*,[14] [7]Ward *et al*.[6]

substitutions (56% novel), 1.7% in-frame insertion/deletion alterations (67% novel), 12% silent exon variants (32% novel), and 24% intronic alterations (60% novel).

The following criteria were applied to predict whether a sequence variant was pathogenic: (i) a potential chain terminating effect on the longest predicted polypeptide; (ii) disruption of a canonical splice site or creation of a novel site; (iii) substitution of an evolutionarily conserved amino acid; (iv) alteration in the polarity or charge of an amino acid; and (v) assessment of the variant frequency in 200 control chromosomes. In total, we identified 124 pathogenic sequence variants among 150 test chromosomes for an overall detection efficiency of 82.7% (table 1). Homozygous *PKHD1* mutations were detected by mixing in eight samples. In three additional patients, there was insufficient DNA for mixing analysis. Mutations were identified on both chromosomes in 56 individuals (74.7%) and one mutation was found in 14 samples (18.7%). No disease causing variants were detected in six patients (8%).

Our patient cohort included six patients previously reported in Furu *et al*[15]: AL40/21-1, AL39/2-18, AL7/10-16, AL41/92-1, AL31/72-1, and AL33/74-1. In each case we confirmed the previously identified mutation and identified a second putative pathogenic sequence variant. In addition, while Furu *et al* did not detect any sequence variant in AL3/5-1, mixing analysis revealed that the patient was homozygous for the p.G1971D substitution.

Among the putative mutant alleles, 22 were predicted to cause chain termination with 16 frameshift variants (deletions, insertions, and duplications) and six nonsense mutations (table 2). Of note, c.9689delA was found on 10 alleles in seven patients; three were homozygotes and four were heterozygotes. Because these patients had different ethnic origins, a founder effect is less likely and this deletional event may represent a mutational ''hot spot''. Two additional frameshift variants, for example, c.3761_3762delCCinsG and c.5895dupA, were detected in several unrelated individuals from ethnically distinct populations, also suggesting the possibility of mutational hot spots. Of note, we identified the frameshift mutation c.3528dupC only in a control DNA sample while assessing the frequency of another variant in 200 control chromosomes. Given that the carrier frequency for *PKHD1* mutations is approximately 1:70 (1.4%) in non-isolated populations[19] and given the diverse array of mutant alleles described to date, it has been suggested that the prevalence of any individual mutation would be low in the normal population.[20] Therefore, it has been proposed that the absence of a *PKHD1* variant in 400 control chromosomes strongly indicates that it is a pathogenic alteration.[20] Our detection of this frameshift mutation in one of 200 control chromosomes suggests that, although supportive, a variant frequency below 0.25% is neither a sufficient nor a necessary criterion for categorising a sequence change as a pathogenic alteration.

We analysed all single nucleotide changes for potential splice site effects using Splice Site Prediction by Neural Network (SSPNN; http://www.fruitfly.org/seq_tools/splice.html). Splice site scores predicted the theoretical impact upon donor and acceptor site strength and the probability that sequence variants created novel splice sites. We found this to be a very useful tool to predict the effect of certain nucleotide changes on splicing, especially as mRNA work to complement our data is not currently feasible. Sixteen variants were predicted to alter splicing (table 3) with alterations in splice site scores generally >50%. Six variants disrupted the canonical GT splice donor or AG splice acceptor and presumably caused exon skipping. For c.7912-1 G→C, disruption of the 100% conserved canonical splice acceptor site had minimal impact (0.37 to <0.3) on the splice site score. However, we note that reductions of predicted splice site strength as low as 0.03 (3%) have been demonstrated to cause exon skipping.[21] Four mutations occurred at less conserved positions of the 5′ or 3′ splice sites: c.391-5 A→G, c.5751+3 A→G, c.6121+3_6121 +4insT, c.8302+3 G→C.

Six missense changes were predicted to alter splicing, of which two, p.R760H, and p.G2809R, altered conserved residues. As we did not have access to kidney mRNA and *PKHD1* is expressed at low levels in peripheral blood lymphocytes, the predicted impact of the splice site alterations could not be tested directly. We note however that investigations at the transcript level are likely to be problematic given the complex transcriptional profile of *PKHD1*.[5]

Of the 71 amino acid substitutions, 30 replaced residues that were either conserved in *Mus musculus* polyductin (NM_153179) and/or predicted to be conserved in polyductin-like proteins from *Rattus norvegicus* (XM_236979, XM_236984), *Gallus gallus*, and *Macaca fascicularis* (table 4). In addition, we applied the matrix criteria described by Miller and Kumar[22] to assess the statistical probability of pathogenicity for each amino acid substitution. This matrix, developed using disease associated human genetic variation and interspecific comparisons as well as Graham's chemical difference matrix, defines the relative likelihood that a missense change represents a polymorphism versus a pathogenic alteration. In our cohort, 24 missense changes were predicted to be pathogenic, and an additional three potentially pathogenic substitutions were identified. The p.Y1136C substitution was previously reported as a putative pathogenic mutation.[15] In the current study, this substitution was detected in another patient as well as a control chromosome and was predicted to have a higher pathogenic potential than polymorphic probability. Therefore, as a conservative estimate, 46% of the amino acid substitutions in our cohort were predicted to be pathogenic. We categorised an additional 20 variants as ''unclassified'' because the change disrupted non-conserved residues or the predicted polymorphic potential was higher than the pathogenic probability. Eight of these sequence variants were also detected in one control chromosome.

The missense alteration, p.T36M, has been described in each *PKHD1* mutation study reported to date (reviewed in Bergmann *et al*[20]) and has been proposed to either represent a founder effect, as most of the patients were of Central European origin, or constitute a mutational ''hot spot'', perhaps due to methylation induced deamination.[23] In the current study, the p.T36M change was identified in 10 unrelated individuals. Unlike the previous studies, these individuals were of diverse ethnic origins including Scotch Irish, French, African American, and Egyptian, as well as Central European. Thus, our analysis provides circumstantial evidence favouring the possibility that p.T36M occurs due to a frequent mutational event.

A second missense change, p.R1624W, was detected in three Saudi patients on five of six chromosomes. This observation may suggest a common founder allele in the Saudi population. However, this allele was not detected in two previously reported Saudi patients[15] nor in a fourth Saudi patient (SB137) in the current study. Moreover, the pathogenic potential of this sequence variant is not clear. The missense change disrupts a non-conserved residue but is predicted to have a high pathogenic potential. We were unable to examine the frequency of this variant in control Saudi chromosomes.

Both of the South African patients were homozygous for the p.M627K substitution. These unrelated children were of Afrikaner origin, a population with a higher prevalence of ARPKD[24] and other recessively transmitted disorders.[25] We are currently examining whether this substitution represents a founder effect in this population, information that could streamline gene based diagnostic testing in at risk Afrikaner children.

Two amino acid substitutions merit further discussion. The missense changes, p.V419S and p.W2749S, were caused by in-frame insertion and deletion mutations (in-frame indel) (table 4). The p.W2749S substitution was identified in a single patient (2-18), whereas the p.V419S variant was detected only in one control chromosome. While these sequence variants disrupted conserved residues, the Miller and Kumar criteria, which apply only to single nucleotide changes, were not informative regarding potential pathogenicity. Therefore, we included the in-frame indel mutations among the unclassified variants. We recognise, however, that the genetic events leading to these variations are likely to be

Table 3 Splice variants

Exon	gDNA nucleotide	Variant	cDNA alteration	Variant effect	WT SS	VAR SS	Variant detection		Reference*
Intronic									
6	11224	c.391-5 A→G	c.391-5 A→G	Loss of acceptor site Novel acceptor site formation	A: 0.40 A: <0.40	A: <0.4 A: 0.70	1/150	Novel	1
13	22521	c.881-1 G→A	c.881-1 G→A	Loss of acceptor site	A: 0.91	A: <0.4	1/150	Novel	1
33	64627	c.5237-1 G→A	c.5237-1 G→A	Loss of acceptor site	A: 0.97	A: <0.4	1/150	Novel	1
34	69941	c.5381-2 A→C	c.5381-2 A→C	Loss of acceptor site	A: 0.4	A: <0.4	1/150	Novel	1
51	239642	c.7912-1 G→C	c.7912-1 G→C	Loss of acceptor site	A: 0.37	A: <0.3	1/150	Novel	1
13	22619	c.976+2 T→A	c.976+2 T→A	Loss of donor site	D: 1.00	D: <0.4	1/150	Novel	1
35	77266	c.5751+3 A→G	c.5751+3 A→G	Loss of donor site	D: 0.91	D: <0.4	1/150	Novel	1
37	153506	c.6121+3_6121 +4insT	c.6121+3_6121 +4insT	Loss of donor site	D: 0.99	D: <0.4	1/150	Novel	1
40	175816	c.6490+2 T→C	c.6490+2 T→C	Loss of donor site	D: 0.78	D: <0.4	1/150	Novel	1
53	256755	c.8302+3 G→C	c.8302+3 G→C	Loss of donor site	D: 0.87	D: 0.23	1/150	Novel	1
Amino acid substitution									
14	24912	p.G326V	c.977 G→T	Loss of acceptor site	A: 0.59	A: <0.4	1/150		1, 2
29	54542	p.G1122S	c.3364 G→A	Loss of donor site	D: 0.85	D: <0.4	1/150		1, 3
60	340892	p.R3333T	c.9998 G→C	Loss of donor site	D: 0.91	D: 0.21	1/150	Novel	1
22	37415	p.R760H	c.2279 G→A	Decreased strength of donor site	D: 0.99	D: 0.84	1/150		1, 3
54	296362	p.G2809R	c.8425 G→A	Increased strength of alternate donor site	A: 0.32	A: 0.46	1/150	Novel	1
65	428413	p.M3642I	c.10926 G→A	Novel acceptor site formation	A: <0.4	A: 0.72	1/150	Novel	1

A, splice acceptor site; D, splice donor site; SS, splice site score (SSNN-BDGP); VAR, variant; WT, wildtype.
*Numbers indicate the following: [1]current study, [2]Furu *et al*,[15] and [3]Onuchic *et al*.[5]

Table 4 Amino acid substitutions

Exon	cDNA nucleotide	gDNA nucleotide	Variant	cDNA alteration	Pathogenic/polymorphism probability*	Mouse AA	Rat AA	Other AA	Amino acid group substitution		Normal control frequency		Reference†
									From	To			
Pathogenic variants													
3	107	4371	p.T36M	c.107 C→T	Equal	T35	T	T (*G gallus*)	Polar	Non-polar	0/200		1, 2, 3, 4, 5, 6, 7
6	410	11258	p.P137R	c.410 C→G	Pathogenic, higher	P135	P	P (*G gallus*)	Non-polar	Basic	0/200	Novel	1
9	664	16563	p.I222V	c.664 A→G	Equal	I220	I		Non-polar	Non-polar	0/100		1, 6
12	783	21499	p.I261M	c.783 A→G	Equal probability	I259	I		Non-polar	Non-polar	0/200	Novel	1
12	806	21522	p.G269V	c.806 G→T	Pathogenic, higher	G267	G		Non-polar	Non-polar	0/200	Novel	1
15	1180	27591	p.A394T	c.1180 G→A	Equal	A392	A		Non-polar	Polar	0/200	Novel	1
16	1342	29079	p.G448R	c.1342 G→C	Pathogenic, higher	G446	G		Non-polar	Basic	0/200	Novel	1
16	1409	29146	p.G470D	c.1409 G→A	Equal	G468	A		Non-polar	Acidic	0/200	Novel	1
20	1880	33450	p.M627K	c.1880 T→A	Pathogenic, higher	M625	M		Non-polar	Basic	0/200	Novel	1
30	3407	59263	p.Y1136C	c.3407 A→G	Pathogenic, higher	Y1134	F		Polar	Non-polar	1/200		1, 2
32	4751	62513	p.S1584I	c.4751 G→T	Pathogenic, higher	S1580			Polar	Non-polar	0/200		1, 4
32	5236	62998	p.G1746S	c.5236 G→A	Equal	G1742	G		Non-polar	Basic	0/200	Novel	1
34	5387	69949	p.L1796S	c.5387 T→C	Pathogenic, higher	L1792	L		Non-polar	Polar	0/200	Novel	1
36	5825	127660	p.D1942G	c.5825 A→G	Pathogenic, higher	D1938	D		Acidic	Non-polar	0/200		1, 4, 6
37	5912	153294	p.G1971D	c.5912 G→A	Equal	G1967	G		Non-polar	Acidic	0/200		1, 4
54	8315	296252	p.L2772P	c.8315 T→C	Pathogenic, higher	L2766	L		Non-polar	Non-polar	0/200		1, 4
55	8518	311769	p.R2840C	c.8518 C→T	Pathogenic, higher	Y2834	R		Basic	Non-polar	0/200	Novel	1
56	8588	314857	p.Y2863C	c.8588 A→G	Pathogenic, higher	Y2857	Y		Polar	Non-polar	0/200		1, 4
58	8870	334332	p.I2957T	c.8870 T→C	Pathogenic, higher	I2951	I		Non-polar	Polar	0/200		1, 3, 4, 5, 6, 7
59	9107	339104	p.V3036G	c.9107 T→G	Pathogenic, higher	V3030	V		Non-polar	Non-polar	0/200		1, 4
59	9311	339308	p.F3104C	c.9311 T→G	Pathogenic, higher	F3098	F		Non-polar	Non-polar	0/200	Novel	1
59	9719	339716	p.R3240L	c.9719 G→T	Pathogenic, higher	R3233	R		Basic	Non-polar	0/200	Novel	1
59	9725	339722	p.G3242V	c.9725 G→T	Pathogenic, higher	G3235	G		Non-polar	Non-polar	0/200	Novel	1
60	9866	340760	p.S3289I	c.9866 G→T	Pathogenic, higher	S3282	S		Polar	Non-polar	0/200		1, 2
61	10036	343108	p.C3346R	c.10036 T→C	Equal	C3339	C		Non-polar	Basic	0/200		1, 6
65	10444	427931	p.R3482C	c.10444 C→T	Pathogenic, higher	R3475	R		Basic	Non-polar	0/100		1, 3
65	10805	428292	p.P3602L	c.10805 C→T	Pathogenic, higher	P3595			Non-polar	Non-polar	0/200	Novel	1
69	11611	454994	p.W3871R	c.11611 T→C	Pathogenic, higher	W3864		W (*M fascicularis*)	Non-polar	Basic	0/200	Novel	1
Pathogenic variants – splicing													
14	977	24912	p.G326V	c.977 G→T	Pathogenic, higher	G324	G		Non-polar	Non-polar	0/120		1, 4
22	2279	37415	p.R760H	c.2279 G→A	Equal	C758	C		Basic	Basic	0/200		1, 5
29	3364	54542	p.G1122S	c.3364 G→A	Polymorphism, higher	G1120	G		Non-polar	Polar	0/200		1, 5
54	8425	296362	p.G2809R	c.8425 G→A	Pathogenic, higher	G2803	G		Non-polar	Basic	0/200	Novel	1
60	9998	340892	p.R3333T	c.9998 G→C	Polymorphism, higher	S3326	G		Basic	Polar	0/200	Novel	1
65	10926	428413	p.M3642I	c.10926 G→A	Equal	A3635			Non-polar	Non-polar	0/200	Novel	1
Unclassified variants													
22	2269	37405	p.I757L	c.2269 A→C	Polymorphism, higher	P755	I		Non-polar	Non-polar	0/200		1, 6
30	3367	59223	p.G1123S	c.3367 G→A	Polymorphism, higher	G1121	K		Non-polar	Polar	0/200		1, 3, 4
32	4038	61801	p.V1347M	c.4038 G→A	Pathogenic, higher	M1343			Non-polar	Non-polar	1/200	Novel	1
32	4139	61901	p.V1380A	c.4139 T→C	Equal	V1376			Non-polar	Non-polar	1/200	Novel	1
32	4870	62632	p.R1624W	c.4870 C→T	Pathogenic, higher	Q1620			Basic	Non-polar	0/200		1, 5
32	5160	62922	p.D1720E	c.5160 C→A	Polymorphism, higher	D1716			Acidic	Acidic	0/200	Novel	1
34	5410	69972	p.R1804C	c.5410 C→T	Pathogenic, higher	H1800	H		Basic	Polar	0/200	Novel	1

Table 4 Continued

Exon	cDNA nucleotide	gDNA nucleotide	Variant	cDNA alteration	Pathogenic/polymorphism probability*	Mouse AA	Rat AA	Other AA	Amino acid group substitution		Normal control frequency		Reference†
									From	To			
40	6436	175760	p.T2146S	c.6436 A→T	Polymorphism, higher	T2141	T		Polar	Polar	0/200	Novel	1
43	6854	183616	p.G2285E	c.6854 G→A	Equal	G2280			Polar	Acidic	1/200	Novel	1
44	6992	184012	p.I2331K	c.6992 T→A	Pathogenic, higher	M2326			Non-polar	Basic	0/200		1, 7
53	8206	256656	p.W2736G	c.8206 T→G	Polymorphism, higher	W2730	W		Non-polar	Non-polar	0/200	Novel	1
55	8521	311772	p.M2841V	c.8521 A→G	Equal	F2835	F		Non-polar	Non-polar	0/200	Novel	1
55	8546	311797	p.G2849A	c.8546 G→C	Polymorphism, higher	G2843	G		Non-polar	Non-polar	0/200	Novel	1
56	8581	314850	p.S2861G	c.8581 A→G	Polymorphism, higher	S2855	S		Polar	Non-polar	0/200		1, 4
57	8737	332769	p.V2913M	c.8737 G→A	Equal	V2907	V		Non-polar	Non-polar	1/200	Novel	1
57	8788	332820	p.R2930W	c.8788 C→T	Pathogenic, higher	R2924	R		Basic	Non-polar	1/200	Novel	1
59	9215	339212	p.A3072V	c.9215 C→T	Equal	A3066	A		Non-polar	Non-polar	1/200		1, 2
59	9788	339785	p.V3263A	c.9788 T→C	Equal	V3256	V		Non-polar	Non-polar	0/200	Novel	1
65	10666	428153	p.R3556C	c.10666 C→T	Pathogenic, higher	H3549			Basic	Non-polar	1/200	Novel	1
65	10319	427806	p.V3440D	c.10319 T→A	Pathogenic, higher	T3433	L		Non-polar	Acidic	0/200		1, 4
70	11737	460568	p.R3913C	c.11737 C→T	Pathogenic, higher	P3906		C (*M fascicularis*)	Basic	Non-polar	0/200	Novel	1
16	1255	28992	p.V419S	c.1255_1256delGTinsTC	Not covered by Miller	V417	V		Non-polar	Polar	1/200	Novel	1
53	8246	256696	p.W2749S	c.8246_8247 delGG insCC	Polymorphism, higher	W2743	W		Non-polar	Non-polar	0/200	Novel	1
Polymorphisms													
19	1736	31885	p.T579M	c.1736 C→T	Equal	T577	T		Polar	Non-polar	3/200		1, 2
22	2278	37414	p.R760C	c.2278 C→T	Pathogenic, higher	C758	C		Basic	Non-polar	70/200		1, 4, 6, 7
24	2489	41465	p.N830S	c.2489 A→G	Polymorphism, higher	N829	N		Polar	Polar	10/128		1, 6
32	3785	61547	p.A1262V	c.3785 C→T	Equal	V1259	S		Non-polar	Non-polar	9/238		1, 6, 7
32	5125	62887	p.L1709F	c.5125 C→T	Equal	L1705	L		Non-polar	Non-polar	2/200		1, 4, 6
35	5608	77120	p.V1870L	c.5608 G→T	Equal	V1870	E		Non-polar	Non-polar	9/200		1, 4, 6
35	5725	77237	p.R1909W	c.5725 C→T	Pathogenic, higher	Q1905			Basic	Non-polar	5/200	Novel	1
49	7675	219692	p.V2559L	c.7675 G→C	Equal	T2553	I		Non-polar	Basic	1/200	Novel	1
51	7921	239652	p.T2641A	c.7921 A→G	Equal	T2635	T		Polar	Non-polar	2/200		1, 4
54	8308	296245	p.T2770A	c.8308 A→G	Equal	T2764	T		Polar	Non-polar	2/200	Novel	1
54	8345	296282	p.G2782A	c.8345 G→C	Polymorphism, higher	C2776	C		Non-polar	Non-polar	2/200	Novel	1
56	8606	314875	p.T2869K	c.8606 C→A	Polymorphism, higher	V2863	V		Polar	Basic	1/200		1, 4, 6
59	9415	339412	p.D3139Y	c.9415 G→T	Pathogenic, higher	N3132	N		Acidic	Polar	1/200		1, 4, 5, 6
59	9577	339574	p.V3193I	c.9577 G→A	Polymorphism, higher	V3186	G		Non-polar	Non-polar	3/200	Novel	1
65	10515	428002	p.S3505R	c.10515 C→A	Pathogenic, higher	S3498	R		Non-polar	Basic	5/128		1, 6, 7
69	11525	454908	p.R3842L	c.11525 G→T	Pathogenic, higher	R3835	E	R (*M fascicularis*)	Basic	Non-polar	3/200		1, 2
70	11696	460527	p.Q3899R	c.11696 A→G	Pathogenic, higher	Q3892		Q (*M fascicularis*)	Polar	Basic	87/200		1, 4, 6, 7
70	11714	460545	p.I3905N	c.11714 T→A	Pathogenic, higher	K3898		N (*M fascicularis*)	Non-polar	Polar	11/200		1, 4, 6
71	12143	468450	p.Q4048R	c.12143 A→G	Pathogenic, higher	S4036		Q (*M fascicularis*)	Polar	Basic	42/128		1, 4, 7

*Miller[22]; †numbers indicate the following: [1]current study, [2]Bergmann *et al*,[20] [3]Bergmann *et al*,[13] [4]Furu *et al*,[15] [5]Onuchic *et al*,[5] [6]Rossetti *et al*,[14] [7]Ward *et al*.[6]

more complex than single nucleotide substitution, thus perhaps increasing the likelihood that these variants are associated with pathogenic effects. Functional studies will be required to examine the pathogenic potential of each of these variants further.

We note that a maximum of two putative mutations were detected in all except patients 33-1 and 74-1, in whom three potentially pathogenic missense changes were identified (table 1). For 33-1, the p.T36M and p.P3602L substitutions have been described in other ARPKD patients and are predicted to have high pathogenic potential. However, the p.M3642I variant is predicted to create a relatively strong acceptor splice site. Functional studies will be required to assess the relative pathogenicity of these three missense variants. For 74-1, the p.I2957T and p.V3440D variants have been reported previously in ARPKD patients and were not found in control chromosomes. Our analysis indicated that p.V3440D involves a substitution of a non-conserved residue but has a higher probability of being a pathogenic alteration than a polymorphism.[22] Therefore, we would categorise this

missense change as an unclassified variant. In comparison, the p.S2861G variant has a higher probability of being a polymorphism than a pathogenic alteration, the substitution involves a highly conserved residue, and it was not detected in 200 control chromosomes. Therefore, the pathogenic potential of this variant in patients 36-6 and 74-1 cannot be excluded. Unfortunately, for both patients 33-1 and 74-1, maternal and paternal DNAs were not available to define the allelic inheritance of these sequence variants.

PKHD1 sequence polymorphisms

In addition to the putative pathogenic changes described above, we also detected 19 single nucleotide polymorphisms (SNPs) that cause amino acid substitutions (table 4). As expected, these missense changes occurred frequently in the normal control population. However, 50% (8/19) involved conserved residues and the Miller and Kumar matrix predicted that these eight variants had a higher pathogenic potential than polymorphic probability. Therefore, such criteria should be interpreted as a guide and pathogenic potential will ultimately need to be assessed by functional studies.

We also detected silent nucleotide changes in both exons and intronic sequences (table 5). Those variants that were common in the control population were designated as polymorphisms. In addition, a number of these variants were detected only in the affected cohort, but no clear pathogenic potential could be ascribed. In the absence of functional studies, we have categorised these variants as probable polymorphisms. While several of these intragenic SNPs have been reported previously (reviewed in Bergmann *et al*[20]), many are novel. Taken together, these SNPs can be used in future analyses to define the haplotypes for mutant chromosomes.

Finally, we identified a number of single nucleotide changes that would be predicted to involve alternative exons. However, without information defining those transcripts that contain alternative exons and their predicted reading frames, we were unable to interpret the potential pathogenic impact of these sequence variants.

Genotype-phenotype correlations

To examine possible genotype-phenotype correlations, we stratified our patient cohort into two groups: the severe group (perinatal) in which the affected child died in the perinatal period; and the less severe group (non-perinatal or older presentation) in which the child either survived the perinatal period or presented at an older age beyond the perinatal period. Haplotype based diagnosis initially identified all the fetal cases and in each family the index child had died in the perinatal period. Therefore, for the purposes of our analyses, we combined the fetal and perinatal cohorts.

We identified both putative mutations in 56 (74.7%) of the 75 patients examined: 36/49 (73.5%) perinatal and 20/26 (76.9%) non-perinatal (table 1). We focused our genotype-phenotype analysis on these 56 patients and characterised the mutations into three classes: (i) chain terminating groups, including nonsense and insertion/deletion with frame shifting; (ii) putative splice site variants; and (iii) amino acid substitution. We reasoned that chain terminating mutations were likely to be complete loss of function variants. Splice site variants could be associated with chain termination if skipping occurs as an out-of-frame event or with a hypomorphic allele if in-frame skipping removes only a number of AAs. Amino acid substitutions may be an admixture of hypomorphic alleles of varying degrees of functional loss.

First, we confirmed that the observed mutational combinations were in Hardy-Weinberg equilibrium, suggesting that there is no obvious bias in the identification of mutations in our patient cohort. Homozygous mutations were detected in 12 patients and the remaining patients were compound heterozygotes. Homozygous chain terminating mutations were identified in four patients with the perinatal phenotype, whereas homozygous amino acid substitutions were found in one perinatal patient and six non-perinatal cases. None of the patients were known to be related. The presence of a common amino acid substitution among the Saudi and Afrikaner patients may be due to population bottlenecks, founding effects, and/or inbreeding.

We then examined the phenotype correlations in patients with various combinations of chain terminating mutations, splice site variants, and amino acid substitutions (table 6). Consistent with previous reports,[5 13–15 26] all patients with combinations of predicted chain terminating and splice site variants (n = 13) were in the perinatal cohort. There was a significant difference in the mutational spectra between the perinatal phenotype and the non-perinatal subset, with variant pairings in the perinatal group comprised mainly of clear pathogenic mutations (chain terminating and splicing variants) or combinations of more pathogenic variants with amino acid variants ($p = 0.00003$ by χ^2 test, $df = 2$). The presence of at least one amino acid substitution was significantly associated with the non-perinatal subset ($p = 0.0029$ by χ^2 test with $df = 1$; odds ratio 1.81, 95% confidence interval 1.374 to 2.406). However, we cannot definitively conclude that at least one amino acid mutation is required for survival. As in previous studies, the T36M variant was detected frequently in our cohort (11 patients), but the distribution of this substitution was not significantly different between the phenotypes.

Lastly, we examined whether the position of the putative *PKHD1* mutation within the longest ORF correlated with disease phenotype. RT-PCR data from mouse kidney indicates that probes from the 5′ and 3′ ends of the longest transcript reveal different transcriptional profiles (GG Germino, unpublished). Therefore, we partitioned the *PKHD1* ORF into three "bins": exon 1–20 (c.-222−1964); exon 21–37/39–50 (c.1965–7911); and exon 51–61/65–71 (7912–12225). Based upon these three bins, there is no significant difference in the distribution of mutations along the *PKHD1* ORF by phenotype ($p = 0.182$, χ^2 test, $df = 2$). Exons 32, 59, and 65 have the highest absolute number of mutations, but these are also some of the largest exons. In contrast, exons 3, 5, and 9 are short exons (78, 109, and 65 bp, respectively) but have significantly higher mutational rates than the rest of the ORF exons ($p<0.05$, by binomial test). Sequence variants involving 3.67–4.61% of the nucleotides in these exons are predicted to have pathogenic effects. When all sequence variants are taken into account (pathogenic, unclassified, and polymorphic), exons 7 and 22 have significantly higher variant rates. The significance of this observation, for example, potential domain effects, is unclear and will require further study. Finally, we note that no mutations were detected in exons 28, 42, 45, 46, or 66. The pathogenic significance of this observation remains to be determined.

Analysis of all reported mutations

The *PKHD1* gene is predicted to contain a minimum of 86 exons, 71 non-overlapping exons that span the entire length of the gene and 15 alternative exons that use different splice sites. The longest ORF transcript contains 67 of the 71 non-overlapping exons. We designed primers for all 86 predicted exons[5] and defined DHPLC protocols to examine each exon (supplemental tables 1 and 2).

Previous mutational studies have either numbered *PKHD1* exons sequentially from 1 to 67 according to the longest ORF

Table 5 Silent and intronic variants

Exon	cDNA nucleotide	gDNA nucleotide	Variant	Normal control frequency		Reference*
Unclassified variant from affected individuals – silent						
3	129	4393	c.129 T→C	0/200	Novel	1
32	4035	61797	c.4035 C→A	0/200	Novel	1
44	6900	183920	c.6900C→T	0/200	Novel	1
49	7533	219550	c.7533 A→G	0/200		1, 2
59	9402	339399	c.9402 G→A	0/200	Novel	1
Unclassified variant from affected individuals – intronic						
12	IVS11	21483	c.779−12insT	0/200	Novel	1
19	IVS18	31811	c.1694−32 C→G	0/200	Novel	1
43	IVS43	183691	c.6865+63 A→C	0/200	Novel	1
50	IVS50	231739	c.7911+19 T→C	0/200	Novel	1
54	IVS53	296213	c.8303−27 T→C	0/200	Novel	1
69	IVS69	4550663	c.11665+18delT	0/200	Novel	1
Unclassified variant (<1.4% normal control chomosomes) – silent						
13	888	22529	c.888 A→T	1/200		1, 2
30	3411	59266	c.3411 G→A	1/200	Novel	1
30	3468	59324	c.3468 G→A	1/200	Novel	1
Unclassified variant (<1.4% normal control chomosomes) – intronic						
2	IVS1	2533	c.-84−22dupT	1/200	Novel	1
7	IVS6	13953	c.449−78 C→T	1/200	Novel	1
44	IVS43	183872	c.6866−14 T→C	1/200	Novel	1
44	IVS44	184059	c.6996+43 C→T	1/200	Novel	1
52	IVS52	251246	c.8173+37 G→A	1/200	Novel	1
53	IVS52	256614	c.8174−10dupT	1/200	Novel	1
56	IVS55	314769	c.8555−55 G→A	1/200		1, 2
Polymorphisms – silent						
4	214	5113	c.214C→T	19/128		1, 2, 4
4	234	5113	c.234C→T	8/128		1, 3, 4
15	1185	27596	c.1185 C→T	8/200		1, 4
17	1587	30667	c.1587 T→C	9/200		1, 2, 3, 4
21	2046	24402	c.2046 A→C	21/128		1, 2, 4
32	3756	61518	c.3756 G→C	2/238		1, 3, 4
32	4920	62682	c.4920 A→G	3/238		1, 2, 3, 4
32	5199	62961	c.5199 C→A	5/200		1, 2
36	5896	127731	c.5896C→T	11/200		1, 2
49	7587	219604	c.7587 A→G	88/200		1, 2, 3, 4
50	7764	231573	c.7764 A→G	60/200		1, 2, 3, 4
57	8673	332705	c.8673 C→G	3/200	Novel	1
59	9237	339234	c.9237 G→A	35/200		1, 2, 3, 4
65	10521	428008	c.10521 C→T	5/128		1, 2, 3, 4
67	11340	439524	c.11340 T→C	2/150		1, 2, 3, 4
Polymorphisms – intronic						
7	IVS7	14128	c.527+19 T→C	50/200		1, 2, 3
7	IVS7	14160	c.527+51 G→T	10/200		1, 2
8	IVS8	15524	c.602+67 A→G	72/200		1, 2
16	IVS15	28961	c.1234−10 T→A	9/200		1, 2
16	IVS15	28966	c.1234−5 C→T	3/200	Novel	1
22	IVS22	37428	c.2279+13 T→G	4/200		1, 3
32	IVS31	61359	c.3629−32 A→G	3/320		1, 2
32	IVS32	63012	c.5236+14 A→G	3/200		1, 2
32	IVS32	63029	c.5236+31delCTTT	2/200		1
35	IVS34	77071	c.5601−42 T→C	5/200	Novel	1
40	IVS39	175649	c.6333-8_6333-7delTT	7/200	Novel	1
49	IVS49	219783	c.7733+33 C→T	62/200	Novel	1
49	IVS49	219793	c.7733+43 G→A	2/200	Novel	1
50	IVS49	231538	c.7734−4 T→C	60/200		1, 2
52	IVS51	251128	c.8108−16 G→C	9/200	Novel	1
53	IVS52	256605	c.8174−19 A→T	3/200	Novel	1
53	IVS53	256764	c.8302+12 T→A	21/320		1, 2
54	IVS54	296448	c.8440+71 A→G	3/200	Novel	1
55	IVS54	311660	c.8441−32 G→C	19/200		1, 2
57	IVS56	332603	c.8643−72 C→T	44/200		1, 2
57	IVS56	332657	c.8643−24 A→G	19/200		1, 2
58	IVS57	334241	c.8798−19 A→C	15/200		1, 2
59	IVS59	339836	c.9829+10 T→G	1/200		1, 2
65	IVS65	428672	c.11174+11 A→G	3/200	Novel	1
68	IVS67	448586	c.11399−71 G→A	10/200	Novel	1
69	IVS69	455095	c.11665+47delA	5/200	Novel	1

*Numbers indicate the following: [1]current study, [2]Furu *et al*,[15] [3]Rossetti *et al*,[14] [4]Ward *et al*.[6]

Table 6 Mutation type and phenotype in *PKHD1*

Phenotype	Mutation type: numbers of alleles and proportions Chain terminating groups	Putative splice site variants	Amino acid substitutions	Total
Type 1	32 (0.44)	11 (0.15)	29 (0.40)*,**	72
Type 2	6 (0.15)	2 (0.05)	30 (0.80)	40

	Mutational genotype: numbers of individuals and proportions Splice/splice	Truncation/truncation	Truncation/splice	Truncation/missense	Splice/missense	Missense/missense	Totals
Type 1	1 (0.027)	7 (0.19)	5 (0.14)	13 (0.36)	4 (0.11)	6 (0.17)	36
Type 2	0 (0)	0 (0)	0 (0)	6 (0.3)	2 (0.10)	12 (0.60)	20
Total	1	7	5	19	6	18	56

*p = 0.00003, significant difference between the two phenotypes among the mutation types; **p = 0.003, amino acid substitution is significantly associated with perinatal survival.

transcript[6 13 14 26] or provided two numerical identifiers per exon, one using the sequential numbering system and the second using the genomic number.[5 15] Given the potential pathogenic importance of sequence variants in the alternative exons (this study), we have re-categorised all the mutations reported to date (including the current study) by numbering the exons according to their genomic order and providing the genomic nucleotide designation to clarify position (supplemental table 3). This re-compilation provides a robust template for future mutational reports, particularly those that examine the pathogenic potential of sequence variants observed in alternative exons. We have submitted this data compilation to the *PKHD1* database maintained at http://www.humgen.rwth-aachen.de.

We re-analysed all reported missense mutations using the SSPNN website to determine whether previously reported missense changes involving the longest ORF exons could disrupt predicted splice sites (supplemental table 4). The variants c.6865+4 A→T, c.657 C→T, and p.Q1917R have been reported in previous studies and were predicted to cause splice site alterations. Our analysis using the SSPNN scores is consistent with these predictions. In addition, these analyses indicate that p.S1664F, p.S1867N, p.I2303F, p.C2688F, and p.S2983L all have a very high probability of causing aberrant splicing in the *PKHD1* ORF. We also determined that two variants in normal controls, p.T2938M (1/200[6]) and p.R3107P (1/400[20]) were likely to alter splice sites. In contrast, while previous reports have classified the variants p.I222V,[6] p.A17V,[6] and c.5381−9 T→G[14] as putative splice site mutations, our SSPNN scores were not consistent with a significant splice site effect. The pathogenic potential of these missense changes will require further analysis with RNA templates.

We note that 30 variants detected in the studies reported to date would be predicted to have pathogenic effects on the longest ORF but may also affect alternate *PKHD1* transcripts, as they involve nine of the alternative exons (supplemental table 3). These 30 variants comprise six frameshift, three splice, four nonsense, and 17 missense variants. The frameshift and splice variants would be responsible for chain termination of transcripts encoded by the longest ORF as well as transcripts containing alternate exons. The nonsense and missense variants may have an effect on alternate transcripts depending on the reading frame. Of the missense variants predicted to occur in alternate transcripts, 10 are classified as pathogenic and seven are unclassified in the longest ORF. Two variants from the unclassified group are silent variations in the longest ORF, however, an altered reading frame may cause a more pathogenic amino acid substitution, or in the case of c.6975 C→T, one reading frame may lead to the formation of a stop codon and subsequent chain termination. Further investigations must clarify the complex transcriptional profile of *PKHD1* before the pathogenic potential of sequence variants involving alternative exons can be assessed.

Finally, we re-examined all the reported amino acid substitutions and assessed putative pathogenicity using the Miller and Kumar matrix criteria. Supplemental table 4 provides a compilation of all these missense mutations stratified according to their likely pathogenic effect. These data should provide a useful reference for laboratories performing gene based diagnostic testing.

DISCUSSION

The current study represents the first comprehensive genomic analysis of *PKHD1*, a novel gene that includes a minimum of 86 exons.[5 6] All previous mutational studies have focused on the 67 exons that comprise the longest ORF transcript and reported mutation detection rates that vary from 47–61% in phenotypically diverse cohorts[5 6 13–15] to 85% among patients with the severe perinatal phenotype.[26]

We have postulated that genomic analysis of all 86 predicted exons should identify putative pathogenic changes among all biologically relevant exons. This information thus should provide a rational starting point to investigate the role of alternative exons among the extensive array of *PKHD1* splice variants and perhaps, once the complex transcriptional profile of *PKHD1* is better understood, increase the effectiveness of mutation detection strategies.

Mutation detection rate

Our overall detection efficiency among the 67 exons in the longest ORF transcript was 82.7%, the highest reported to date in a phenotypically diverse ARPKD cohort. The detection rates were not statistically different among the three phenotypic groups examined (fetal, perinatal, older presentation). When combined with linkage data indicating that mutations in *PKHD1* are causative in all typical cases of ARPKD,[3 4] these data suggest that the previous low detection rate among children with older presentations was more likely due to methodological issues than potential genetic heterogeneity.

A total of 319 *PKHD1* variants have been reported to date, including the 173 variants from this report, 92 of which were novel. Our data have increased the *PKHD1* variant database content by 28.8%. Among these novel variants, 44/92 (47.8%) were classified as likely pathogenic changes (splice site alterations, chain termination, and pathogenic amino acid substitutions) and 33/92 (35.9%) were unclassified variants, consisting primarily of amino acid substitutions at

non-conserved residues. Four variants, p.D1720E, p.W2736G, p.W2749S (in-frame indel), and p.G2849A, involved substitutions of highly conserved amino acids and were not detected in 200 normal chromosomes. However, the Miller and Kumar matrix predicted these missense changes were likely to be polymorphisms and thus we categorised them as unclassified variants.

We did not detect any mutations in 8% of our cohort and we found both mutations in only 15 (78.9%) of the 19 fetal cases, despite the fact that ARPKD was confirmed in all by histopathological analysis. In the former case, these findings may suggest misdiagnosis, but several of these children had affected siblings with pathoanatomically proven ARPKD. The unidentified *PKHD1* alterations in these patients could involve mutational events that are not detectable with our exon based method, for example, intronic changes that generate cryptic splice sites or large rearrangements that cause one to multiple exon deletions. Alternatively, these patients may carry alterations in regulatory elements. For example, the proximal promoter of the mouse *Pkhd1* gene contains an evolutionarily conserved hepatocyte nuclear factor-1 (HNF-1) binding site and mutations of this site inhibit promoter activity.[27] The human *PKHD1* promoter region also contains a putative HNF-1 site. When this site was examined in 21 patients using a PCR based strategy, no sequence variants were identified (Guay-Woodford and Igarashi, unpublished data). However, exhaustive analysis of this human *PKHD1* promoter region has not been performed and alternative promoters for this transcriptionally complex gene are at least theoretically possible. Finally, missense changes involving alternative exons were identified in several patients. For example, the missense change g.405493A→G was detected in exon 64 in patient 32-1 (data not shown). While only one other putative mutation was identified for this patient, the pathogenic potential of the g.405493A→G variant cannot be assessed in the absence of more detailed information regarding the putative ORF of alternative exon containing transcripts.

Direct gene based testing

Given that ARPKD is an often devastating disease with typical presentation in the perinatal period, there is strong demand for gene based diagnostic testing. However, the complexity of the gene, the variability in mutation detection efficiency, and the high frequency of missense mutations have complicated the development of an efficient clinical test and the robust interpretation of detected sequence variants. To address these issues, we have refined the DHPLC screening protocol by optimising primer design and DHPLC analytic conditions. We have included a template mixing step when no mutation was detected in the first round of screening to optimise the identification of homozygous variants. Without mixing, the detection rate would have been 75.3%. In addition, we have incorporated the matrix criteria described by Miller and Kumar[22] and the SSPNN algorithm into our analysis of missense mutations to optimise predictions regarding the pathogenicity of amino acid substitutions and potential splice site effects. We have compiled our missense mutations with all missense changes reported to date and categorised them as pathogenic, unclassified, or polymorphic based on these criteria. This compilation should serve as an important reference source for laboratories performing gene based diagnostic testing.

For newly identified missense changes, assessment of the likely pathogenicity remains problematic. Bergmann *et al*[20] have proposed that the absence of a *PKHD1* variant among 400 control chromosomes (<0.25%) is sufficient to conclude that the change is pathogenic. However, we have detected variants from all mutational classes in 1 of 200 control chromosomes (0.5%). Moreover, a few variants with high pathogenic potential have been detected only in normal chromosomes to date.

Therefore, we suggest the following criteria to predict potential pathogenicity of novel mutations: (i) a putative chain terminating effect on the longest predicted polypeptide; (ii) disruption of a canonical splice site or creation of a novel site (SSPNN algorithm); (iii) alteration of an evolutionarily conserved amino acid in the context of the Miller and Kumar matrix; and (iv) detection in <0.5% of normal chromosomes. We propose categorising missense changes that meet one of the first three criteria as pathogenic; those that meet criteria (ii) or (iii) but occur in >0.5% of normal chromosomes as unclassified, that is pathogenicity cannot be excluded; and those found in >0.5% of normal chromosomes as polymorphisms, particularly if they do not meet any of the first three criteria.

Finally, we recognise that despite our stratification strategies, missense mutations may pose challenges for clinical diagnostic laboratories engaged in gene based testing, particularly when the data are sought to guide prenatal diagnosis or pre-implantation genetic diagnosis. Therefore, when possible in these cases, we recommend a combined approach using haplotype analysis to complement *PKHD1* mutation screening.

Genotype-phenotype correlations

In our previous study,[15] we used a binary assessment of clinical presentation based on whether or not the affected children survived the immediate perinatal period. This endpoint was chosen because it was reliably available for all our patient samples. We tested the hypothesis that differences in the clinical severity observed among the patients resulted, at least in part, from the nature of the germ-line mutations. These analyses indicated that the presence of two chain terminating mutations invariably resulted in perinatal lethality and survival past the immediate perinatal period required the presence of at least one amino acid substitution mutation. However, the converse did not apply, that is perinatal lethality could occur in presence of an apparent missense variant.

The current study focused on those 56 patients in whom we had identified two putative mutations. We applied the same phenotype stratification and examined correlations with various combinations of chain terminating mutations, splice site variants, and amino acid substitutions. In addition, we examined whether the linear position of putative *PKHD1* mutations correlated with disease phenotype. Our experimental design has advantages over previous analyses because we did not combine data from several studies, we did not have a high frequency of sequence variants attributable to founder effects, and we did not include a mixed group of individuals with one and two identified mutations. As a result, we lowered the risk of biased results due to differential discovery rates for different mutational types. In addition, because all patients were from a single study and genotype information is complete, we had a higher probability of more accurately assessing genotype-phenotype correlations. That said, as in previous studies, all our patients with combinations of predicted chain terminating and splice site variants were in the perinatal cohort. The presence of at least one amino acid substitution was significantly higher in the non-perinatal subset. However, even this study had limited power and we cannot definitively conclude that at least one amino acid mutation is required for survival. In contrast, our analysis did not confirm previous observations[15] suggesting a significant correlation between the less severe phenotype and the distribution of mutations along the *PKHD1* ORF. In fact, we conclude that neither the location nor the type of

amino acid substitution correlates with disease severity. Of specific note, our two patients with isolated congenital hepatic fibrosis had mutations that were not unique in either type or position.

Future directions for mutation testing

Our study demonstrates that DHPLC provides an efficient and economical screening approach for a gene of the size and complexity of *PKHD1*. However, we acknowledge that a PCR based methodology is subject to various limitations. For example, PCR will not detect genomic rearrangements involving deletion or duplications of a few kilobases. Two recently described techniques, multiplex amplification and probe hybridisation (MAPH) and multiplex ligation dependent probe amplification (MLPA), allow detection of such mid-size rearrangements by simultaneously screening for the loss or duplication of up to 40 target sequences.[28] Both methods rely on sequence specific probe hybridisation to genomic DNA, followed by amplification of the hybridised probe, and semi-quantitative analysis of the resulting PCR products. We speculate that a subset of ARPKD patients in whom $\leqslant 1$ sequence variant was detected may carry such mid-size deletions or duplications and we are examining these patients further. In addition, we note that PCR based mutation detection methods may be adversely affected by intronic variations that do not themselves have pathogenic potential but compromise the binding of PCR primers.[29] Given the frequency of intronic variants reported in *PKHD1* (approximately 17%), it is at least theoretically possible that non-amplification of a mutant allele due to primer binding site variants may contribute to a reduced mutation detection rate in patients with strong clinical evidence for ARPKD.

Finally, we have considered RNA based methodologies, such as RT-PCR[30] and protein truncation testing,[31] as complementary strategies to enhance the identification of pathogenic variants in *PKHD1*. These approaches have proven to be quite robust in mutational analyses of transcriptionally complex genes such as *NF1*, increasing the detection efficiency rate to 95%.[21] However, several factors complicate the application of these methodologies to *PKHD1* testing. First, unlike neurofibromatosis (*NF1*) and hereditary breast cancer (*BRCA1*), ARPKD is transmitted as a recessive trait and detection of both mutations is required for definitive diagnosis. Second, *PKHD1* is expressed at very low levels in peripheral blood lymphocytes (see Onuchic *et al*[5] and unpublished data) and neither kidney tissue nor renal epithelial cells are generally available for clinical testing. Third, in mouse *Pkhd1*, the complex array of splice variants appears to vary, at least in part, in an organ specific fashion (Guay-Woodford, unpublished data). If the same holds true for *PKHD1*, it will be difficult to assess the biological impact of variants detected in templates from non-phenotypically affected organs.

We propose that the next major advance in *PKHD1* mutation detection efficiency will be predicated on an exhaustive examination of the *PKHD1* complex transcriptional profile. Moreover, comparative analysis of mRNA processing will be required to determine whether pathogenic variants are represented among the array of transcripts in more clinically accessible tissues such as peripheral blood lymphocytes, amniocytes, and chorionic villus cells. However, in the interim, we believe that DHPLC provides an acceptable screening tool for the detection of *PKHD1* variants. Our data and that recently reported by Bergmann *et al*[26] provide an appropriate platform to begin offering gene based diagnostic testing in prenatal cases, patients with unusual clinical presentations, and pre-implantation assessment of early stage embryos.

ACKNOWLEDGEMENTS

The authors thank the patients and families who were involved in these studies.

ELECTRONIC DATABASE INFORMATION

The following URLs are mentioned in this study: Berkeley Drosophila Genome Project (BDGP), http://www.fruitfly.org/seq_tools/splice.html (splice site neural network NNSPLICE 0.9); Blast, http://www.ncbi.nlm.nih.gov/BLAST/; Entrez Nucleotide, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide; Human Genome Variation Society (HGVS), http://www.hgvs.org/; and Primer3, http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi. Supplemental tables are available from http://jmg.bmjjournals.com/supplemental.

Authors' affiliations

A M Sharp, L M Guay-Woodford, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA
L M Messiaen, Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, USA
G Page, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA
C Antignac, Department of Genetics, Hopital Necker-Enfants Malades, Paris, France
M-C Gubler, INSERM U574, Hopital Necker-Enfants Malades, Paris, France
L F Onuchic, Department of Medicine, University of São Paulo School of Medicine, São Paulo, Brazil
S Somlo, Departments of Internal Medicine and Genetics, Yale University School of Medicine, New Haven, CT, USA
G G Germino, Departments of Medicine and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD, USA

These investigations were supported by the National Institutes of Health (GGG), the Polycystic Kidney Disease Foundation (LG-W), and a Clinical Scientist Award in Translational Research from the Burroughs-Wellcome Foundation (LG-W).

Competing interests: none declared

Correspondence to: Dr Lisa M Guay-Woodford, Division of Genetic and Translational Medicine, University of Alabama at Birmingham, 740 Kaul Human Genetics Building, 720 20th Street South, Birmingham, AL, USA; lgw@uab.edu

Received 25 June 2004
Revised version received 2 August 2004

REFERENCES

1 **Zerres K**, Muecher G, Becker J, Steinkamm C, Rudnik-Schoneborn S, Heikkila P, Rapola J, Salonen R, Germino G, Onuchic L, Somlo S, Avner E, Harman L, Stockwin J, Guay-Woodford L. Prenatal diagnosis of autosomal recessive polycystic kidney disease (ARPKD): molecular genetics, clinical experience, and fetal morphology. *Am J Med Genet* 1998;**76**:137–44.
2 **Guay-Woodford L**, Bernstein J. Other cystic kidney diseases. In: Johnson R, Feehally J, eds. *Comprehensive clinical nephrology*. London: Mosby International, 2003:611–26.
3 **Zerres K**, Muecher G, Bachner L, Deschennes G, Eggermann T, Kaariainen H, Knapp M, Lenner T, Misselwitz J, Muhlendahl Kv, Neumann N, Pirson Y, Rudnik-Schoneborn S, Steinbicker V, Wirth B, Scharer K. Mapping of the gene for autosomal recessive polycystic kidney disease (ARPKD) to chromosome 6p21-cen. *Nat Genet* 1994;**7**:429–32.
4 **Guay-Woodford LM**, Muecher G, Hopkins SD, Avner ED, Germino GG, Guillot AP, Herrin J, Holleman R, Irons DA, Primack W. The severe perinatal form of autosomal recessive polycystic kidney disease maps to chromosome 6p21.1-p12: implications for genetic counseling. *Am J Hum Genet* 1995;**56**(5):1101–7.
5 **Onuchic L**, Furu L, Nagasawa Y, Hou X, Eggermann T, Ren Z, Bergmann C, Senderek J, Esquivel E, Zeltner R, Rudnik-Schöneborn S, Mrug M, Sweeney W, Avner E, Zerres K, Guay-Woodford L, Somlo S, Germino G. *PKHD1*, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple IPT domains and PbH1 repeats. *Am J Hum Genet* 2002;**70**:1305–17.
6 **Ward C**, Hogan M, Rossetti S, Walker D, Sneddon T, Wang X, Kubly V, Cunningham J, Bacallao R, Ishibashi M, Milliner D, Torres V, Harris P. The

gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat Genet* 2002;**30**:259–69.
7 **Nagasawa Y**, Matthiesen S, Onuchic L, Hou X, Bergmann C, Esquivel E, Senderek J, Ren Z, Zeltner R, Furu L, Avner E, Moser M, Somlo S, Guay-Woodford L, Buttner R, Zerres K, Germino G. Identification and characterization of Pkhd1, the mouse orthologue of the human ARPKD gene. *J Am Soc Nephrol* 2002;**13**:2246–58.
8 **Guay-Woodford L**. Murine models of polycystic kidney disease: molecular and therapeutic insights. *Am J Physiol Renal Physiol* 2003;**285**:F1034–49.
9 **Zhang M**, Mai W, Li C, Cho S, Hao C, Moeckel G, Zhao R, Kim I, Wang J, Xiong H, Wang H, Sato Y, Wu Y, Nakanuma Y, Lilova M, Pei Y, Harris R, Li S, Coffey R, Sun L, Wu D, Chen X, Breyer M, Zhao Z, McKanna J, Wu G. PKHD1 protein encoded by the gene for autosomal recessive polycystic kidney disease associates with basal bodies and primary cilia in renal epithelial cells. *Proc Natl Acad Sci U S A* 2004;**101**:2311–6.
10 **Wang S**, Luo Y, Wilson P, Witman G, Zhou J. The autosomal recessive polycystic kidney disease protein is localized to primary cilia, with concentration in the basal body area. *J Am Soc Nephrol* 2004;**15**:592–602.
11 **Menezes F**, Cai Y, Nagasawa Y, Silva A, Watkins M, Silva A, Somlo S, Guay-Woodford L, Germino G, Onuchic L. Polyductin, the PKHD1 gene product, comprises isoforms expressed in plasma membrane, primary cilium and cytoplasm. *Kidney Int* 2004;**66**:1345–55.
12 **Masyuk TV**, Huang BQ, Ward CJ, Masyuk AI, Yuan D, Splinter PL, Punyashthiti R, Ritman EL, Torres VE, Harris PC, LaRusso NF. Defects in cholangiocyte fibrocystin expression and ciliary structure in the PCK rat. *Gastroenterology* 2003;**125**(5):1303–10.
13 **Bergmann C**, Senderek J, Sedlacek B, Pegiazoglou I, Puglia P, Eggermann T, Rudnik-Schneborn S, Furu L, Onuchic L, Baca MD, Germino G, Guay-Woodford L, Somlo S, Moser M, Buttner R, Zerres K. Spectrum of mutations in the gene for autosomal recessive polycystic kidney disease (ARPKD/PKHD1). *J Am Soc Nephrol* 2003;**14**:76–89.
14 **Rossetti S**, Torra R, Coto E, Consugar M, Kubly V, Malaga S, Navarro M, El-Youssef M, Torres V, Harris P. A complete mutation screen of PKHD1 in autosomal-recessive polycystic kidney disease (ARPKD) pedigrees. *Kidney Int* 2003;**64**:391–403.
15 **Furu L**, Onuchic L, Gharavi A, Hou X, Esquivel E, Nagasawa Y, Bergmann C, Senderek J, Avner E, Zerres K, Germino G, Guay-Woodford L, Somlo S. Milder presentation of recessive polycystic kidney disease requires presence of amino acid substitution mutations. *J Am Soc Nephrol* 2003;**14**:2004–14.
16 **Guay-Woodford L**, Desmond R. Autosomal recessive polycystic kidney disease (ARPKD): the clinical experience in North America. *Pediatrics* 2003;**111**:1072–80.
17 **Gagnadoux M**, Attie T, Amiel J, Gigarel N, Bonnefont J, Munnich A, Gubler M, Antignac C. Prenatal diagnosis of autosomal recessive polycystic kidney disease. *Arch Pediatr* 2000;**7**:942–7.
18 **Dunnen J**, Antonarakis S. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 2002;**15**:7–12.
19 **Zerres K**, Becker J, Muecher G, Rudnik-Schoneborn S, Onuchic L, Germino G, Somlo S, Avner E, Harman L, Stockwin J, Guay-Woodford L. Haplotype-based prenatal diagnosis in autosomal recessive polycystic kidney disease (ARPKD). *Am J Med Genet* 1998;**76**:137–44.
20 **Bergmann C**, Senderek J, Kupper F, Schneider F, Dornia C, Windelen E, Eggermann T, Rudnik-Schoneborn S, Kirfel J, Furu L, Onuchic L, Rossetti S, Harris P, Somlo S, Guay-Woodford L, Germino G, Moser M, Buttner R, Zerres K. PKHD1 mutations in autosomal recessive polycystic kidney disease (ARPKD). *Hum Mutat* 2004;**23**:453–63.
21 **Messiaen L**, Callens T, Mortier G, Beysen D, Vandenbroucke I, Roy NV, Speleman F, Paepe AD. Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Hum Mutat* 2000;**15**:541–55.
22 **Miller M**, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 2001;**21**:2319–28.
23 **Cooper D**, Youssoufian H. The CpG dinucleotide and human genetic disease. *Hum Genet* 1988;**78**:151–5.
24 **Lombard E**, Kromberg J, Thomson P, Milner L, Biljon Iv, Jenkins T. Autosomal recessive polycystic kidney disease: evidence for high frequency of the gene in Afrikaans-speaking population. *S Afr Med J* 1989;**76**:321–3.
25 **Grove S**, Kromberg J. Bernard-Soulier syndrome in two Afrikaner families. *S Afr Med J* 1985;**29**:1050–2.
26 **Bergmann C**, Senderek J, Schneider F, Dornia C, Kupper F, Eggermann T, Rudnik-Schoneborn S, Kirfel J, Moser M, Buttner R, Zerres K. PKHD1 mutations in families requesting prenatal diagnosis for autosomal recessive polycystic kidney disease (ARPKD). *Hum Mutat* 2004;**23**:453–63.
27 **Hiesberger T**, Bai Y, Shao X, McNally B, Sinclair A, Tian X, Somlo S, Igarashi P. Mutation of hepatocyte nuclear factor-1beta inhibits Pkhd1 gene expression and produces renal cysts in mice. *J Clin Invest* 2004;**113**:814–25.
28 **Sellner L**, Taylor G. MLPA and MAPH: new techniques for detection of gene deletions. *Hum Mutat* 2004;**23**:413–19.
29 **Jeffrey G**, Chakrabarti S, Hegele R, Adams P. Polymorphism in intron 4 of HFE may cause overestimation of C282Y homozygote prevalence in haemochromatosis. *Nat Genet* 1999;**22**:325–6.
30 **Vandenbroucke I**, Vandesompele J, Paepe AD, Messiaen L. Quantification of NF1 transcripts reveals novel highly expressed splice variants. *FEBS Lett* 2002;**522**:71–6.
31 **Hogervorst F**, Cornelis R, Bout M, Vliet Mv, Oosterwijk J, Olmer R, Bakker B, Klijn J, Vasen H, Meijers-Heijboer H. Rapid detection of BRCA1 mutations by the protein truncation test. *Nat Genet* 1995;**10**:208–12.