Identification of Gene Mutations in Autosomal Dominant Polycystic Kidney Disease through Targeted Resequencing

Sandro Rossetti,* Katharina Hopp,† Robert A. Sikkink,‡ Jamie L. Sundsbak,* Yean Kit Lee,‡ Vickie Kubly,* Bruce W. Eckloff,‡ Christopher J. Ward,* Christopher G. Winearls,§ Vicente E. Torres,* and Peter C. Harris*

*Division of Nephrology and Hypertension, †Mayo Graduate School, and ‡Advanced Genomics Technology Center, Mayo Clinic, Rochester, Minnesota; and §Oxford Radcliffe Hospital and Jesus College, University of Oxford, Oxford, United Kingdom

ABSTRACT

Mutations in two large multi-exon genes, *PKD1* and *PKD2*, cause autosomal dominant polycystic kidney disease (ADPKD). The duplication of *PKD1* exons 1–32 as six pseudogenes on chromosome 16, the high level of allelic heterogeneity, and the cost of Sanger sequencing complicate mutation analysis, which can aid diagnostics of ADPKD. We developed and validated a strategy to analyze both the *PKD1* and *PKD2* genes using next-generation sequencing by pooling long-range PCR amplicons and multiplexing barcoded libraries. We used this approach to characterize a cohort of 230 patients with ADPKD. This process detected definitely and likely pathogenic variants in 115 (63%) of 183 patients with typical ADPKD. In addition, we identified atypical mutations, a gene conversion, and one missed mutation resulting from allele dropout, and we characterized the pattern of deep intronic variation for both genes. In summary, this strategy involving next-generation sequencing is a model for future genetic characterization of large ADPKD populations.

J Am Soc Nephrol 23: ●●●–●●●, 2012. doi: 10.1681/ASN.2011101032

Autosomal dominant polycystic kidney disease (ADPKD) is one of the most common inherited cystic kidney diseases, with an incidence of 1 in 400 to 1 in 1000.[1,2] ADPKD is characterized by the development and progressive enlargement of cysts in the kidneys and other organs, eventually leading to ESRD. ADPKD is caused by mutations at two genes, *PKD1* (16p13.3) and *PKD2* (4q21). *PKD1* mutations account for approximately 85% and *PKD2* mutations for approximately 15% of the cases in clinically well characterized cohorts.[3] PKD1 patients reach ESRD approximately 20 years earlier than PKD2 patients (approximately 54 versus 74 years).[4] *PKD1* and *PKD2* encode polycystin 1 and 2 (PC-1 and PC-2), respectively. PC-1 is a large, transmembrane protein that interacts with PC-2, a transient receptor potential channel that regulates intracellular calcium.[5] Both proteins localize to the kidney primary cilium,[5] and may act as a flow-dependent mechanosensor regulating the differentiation and proliferation of tubular epithelial cells.[5]

Within ADPKD populations, a high level of allelic heterogeneity is observed, with a total of 436 pathogenic *PKD1* and 115 pathogenic *PKD2* mutations reported to date, the majority of which are private to a single pedigree (ADPKD Database [PKDB], http://pkdb.mayo.edu).

Gene conversions (GCs) are unusual mutational events that cause the transfer of sequence variants from segmental duplications into the master gene,

Received October 27, 2011. Accepted January 5, 2012.

Published online ahead of print. Publication date available at www.jasn.org.

Correspondence: Dr. Sandro Rossetti, Division of Nephrology and Hypertension, Mayo Clinic, 200 First Street SW, Rochester MN 55905. Email: rossetti.sandro@mayo.edu

and have been proven to be disease associated.[6] GCs have been previously described in ADPKD[7,8] but their exact genomic origin and extent have not been characterized.

ADPKD is typically diagnosed by imaging such as ultrasonography, computed tomography, or magnetic nuclear resonance,[9,10] with age-related criteria established for ultrasonography.[9,11] However, a diagnosis determined by imaging may be uncertain, particularly in young individuals (aged <30 years).[11] In such cases, molecular diagnostics is useful to determine a definite diagnosis.[3] Molecular testing also plays a role in the evaluation of potential living related kidney donors with doubtful imaging data, in individuals with a negative family history, and in cases of early onset ADPKD.[12] Furthermore, mutation characterization of clinical trials cohorts[3] provides genetic stratification for the evaluation of such trials.[13]

The 5′ two-thirds of the *PKD1* gene (exons 1–32) is duplicated six times on chromosome 16 within six pseudogenes (*PKD1P1-P6*).[14,15] The *PKD1P1-P6* pseudogenes share a 97.7% sequence identity with the genuine *PKD1*, although they carry some large deletions compared with the genuine *PKD1*.[15,16] Rare sequence divergences have been used to develop *PKD1* locus-specific amplicons to analyze the duplicated portion of the gene for mutations.[17] The *PKD1* genomic complexity and the high allelic heterogeneity of both *PKD1* and *PKD2* make molecular diagnostics challenging.[3]

High-throughput next-generation DNA sequencing (NGS) technologies have recently been developed, the common feature of which is the utilization of massive parallel sequencing of DNA strands after random fragmentation to produce millions of reads. These are subsequently re-aligned for sequence variant calling.[18–20] The feasibility of utilizing NGS for limited genomic regions has arisen through multiplexing by the introduction of bar codes, unique 6-bp tags, which allow the individual identification of samples analyzed within the same lane.[21] Bar coding and multiplexing of PCR and long-range PCR (LR-PCR) amplicons from groups of patients have been effectively used to characterize genomic regions up to approximately 150 kb.[22–30] Exon enrichment or capture protocols have also been developed for the analysis of specific genomic intervals or the whole exome.[31,32] However, these are not effective in duplicated genomic regions (*e.g.*, the *PKD1* gene) because they would lead to concurrent capture of the six pseudogenes.

In this study, we utilized pooling and multiplexing of samples to validate NGS for the mutation analysis of the ADPKD genes in a cohort of 230 ADPKD patients. These results show the feasibility of high-throughput NGS for the genetic characterization of large ADPKD cohorts. Furthermore, the utilization of fewer PCR primers and the possibility of characterizing the entire genomic structure of the *PKD1* and *PKD2* genes will help in detecting and characterizing atypical mutations (deep intronic variants, GCs, and ones missed due to allele dropout).

RESULTS

Development of LR-PCR Amplicons for the *PKD1* and *PKD2* Genes and Proof of Principle Experiment for Pooling and Multiplexing Samples for NGS

Because of the duplication of the *PKD1* gene (which already requires LR-PCR amplicons for locus-specific amplification[17]) and the limited genomic size of both genes combined (118 kb), we extended the number of LR-PCR amplicons to cover all of the coding regions of both genes tested (76.2 kb; *PKD1*-eight amplicons; *PKD2*-six amplicons) (Figures 1A and 2 and Supplemental Table 1).

A proof of principle experiment (Figure 3A) was performed employing one Illumina flow cell using 16 previously Sanger-characterized ADPKD patients (carrying 281 known sequence variants) and four novel cases. We pooled two to eight samples in a single bar-coded library (lanes 1–4) to test the maximum number of samples that can be pooled in a single bar-coded library while still detecting all of the positive controls. A second part of this experiment analyzed multiplexing of two to four bar-coded libraries of four pooled samples each (up to 16 samples, lanes 5–7), to evaluate the relationship between read depth and detection of positive controls after multiplexing (Figure 3A). One single bar-coded library of four unknown samples was run to mimic the planned mutation discovery workflow (lane 8).

The 281 Sanger-verified control variants allowed a detailed analysis of read depth (number of reads per known variant), coverage (percentage of the regions of interest adequately covered), sensitivity (proportion of true positives), and precision rate in the exonic regions (proportion of correctly identified mutations) for each of the bar-coded libraries (Table 1). Very high read depth was obtained for all control variants, and although variation of up to 18-fold in read depth was found within the same library (pool of four samples, 397× to 72,497×), all regions of interest were adequately covered (Figure 2). Single nucleotide variants were efficiently detected, but two *PKD1* deletions (38 and 15 bp in length, respectively) were missed due to the short, 51-bp reads used, lowering the overall sensitivity. For the pooling test, a loss of sensitivity was observed when pooling eight samples (four false negatives and 12 false positives), whereas the pool of four and six samples performed similarly well (Table 1). Multiplexing at this level did not affect sensitivity or precision rate significantly.

Taken together, this proof of principle experiment suggested the following: a conservative approach of pooling four samples per bar-coded library was feasible, at least 12 such libraries could be run per lane with an expected read depth of approximately 100×, and longer reads were required for the detection of indels longer than 15 bp.

Mutation Analyses of a Large ADPKD Cohort by Bar-Coded and Multiplexed NGS

On the basis of the proof of principle experiment, we utilized the previously developed amplicons to characterize a cohort of 264 ADPKD samples (230 novel and 34 internal controls) as 66

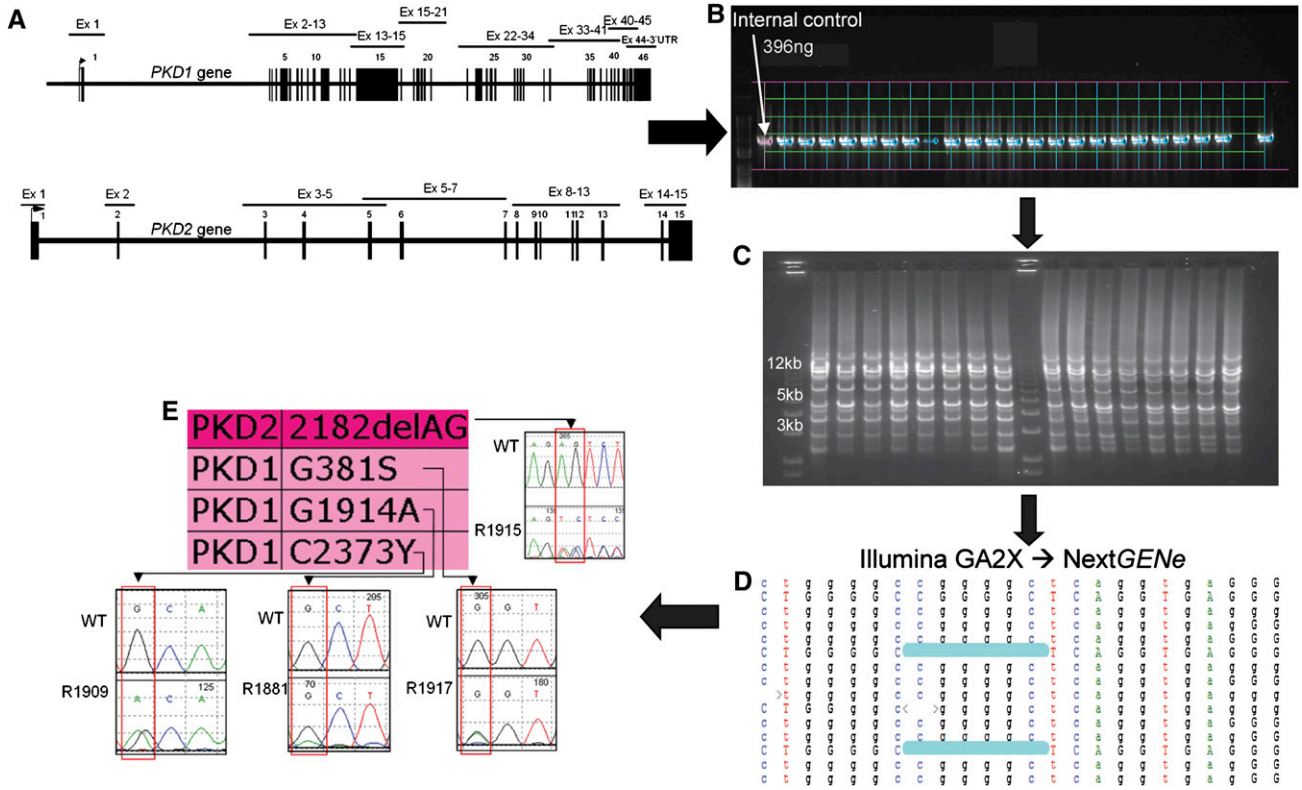

Figure 1. Schematic visualization of the NGS workflow used in this study. The workflow indicated by the arrows is as follows. (A) Amplicons were individually amplified by long-range PCR. The *PKD1* duplicated region (exons 1–33) was amplified as five locus-specific long-range amplicons (2.2–8.7 kb in size), and the same strategy was extended to the *PKD1* single-copy region (exons 33–46, three long-range amplicons 2.1–5.9 kb in size) and to the *PKD2* gene (six long-range amplicons, 1.2–13 kb in size), covering all coding regions and most intronic regions as a total of 76.2 kb. (B) Amplification was quality verified and normalized by gel densitometry to a sample of known concentration. Two microliters of each amplified product were run on a 0.8% agarose gel for quality check and quantification after fluorescent visualization. Lanes and bands were captured (green lines network), and the actual area of each amplified product intercepted (blue line above each band). Each band was quantified by comparison to a known control (rose line), and values transferred to a spreadsheet for the calculation of the appropriate volume to be used during amplicon assembly. (C) Amplicons were assembled equimolarly for each individual sample, and assembled samples were pooled for each indexed library. Assembled libraries were subsequently quality verified by gel electrophoresis. After assembly, 2 μl of the assembled material was fluorescently visualized on a 0.8% agarose gel to verify the approximate homogeneity of fluorescent intensity and presence of multiple bands corresponding approximately to the expected sizes. (D) Samples were sequenced on an Illumina GA2X instrument, and reads were exported as FASTAQ files, deconvoluted by bar code, and mined using the Next*GENe* software package. Mutation reports were exported for evaluation. Manual checks of called variants were performed by visualizing the Next*GENe* alignment as shown. (E) Variants were individually reconfirmed by sequencing the original four samples included in the corresponding library. In the example reported, the four samples included in this library were individually proven to carry the *PKD2* change c.2182_2183delAG and three *PKD1* amino acid substitutions (p.Gly381Ser, p.Gly1914Ala, and p.Cys2373Tyr, respectively) (indicated in the panel with the short designation 2182delAG, G381S, G1914A, and C2373Y because of space constraints).

libraries (Figures 3B and 4) using 101-bp reads. The 230 novel samples spanned the entire phenotypic spectrum of ADPKD and included severe and mild PKD cases, to mimic a "real-world" diagnostic setting. Patient samples were amplified separately and amplicons pooled equimolarly for libraries 1–66, whereas the DNA was pooled before amplification for libraries 67–74 (Figure 3B).

Data mining identified 2445 variants in the 230 novel cases. After quality filtering, 779 high-confidence variants were retained and 1666 low-confidence variants were removed (Concise Methods and Figure 4). The remaining 779 high-confidence variants were further filtered based on the likelihood of disease association (Figure 4). We individually confirmed by re-amplifying each of the four samples originally pooled in the same library and Sanger sequencing the 176 possible pathogenic variants and 58 likely neutral variants (Figure 4, Table 2, and Supplemental Tables 2 and 3). Of the 234 Sanger-verified variants, 213 (90%) were true positives and 21 (10%) false positives. Manual inspection of the 21 false positive variants revealed that they were due to misalignment

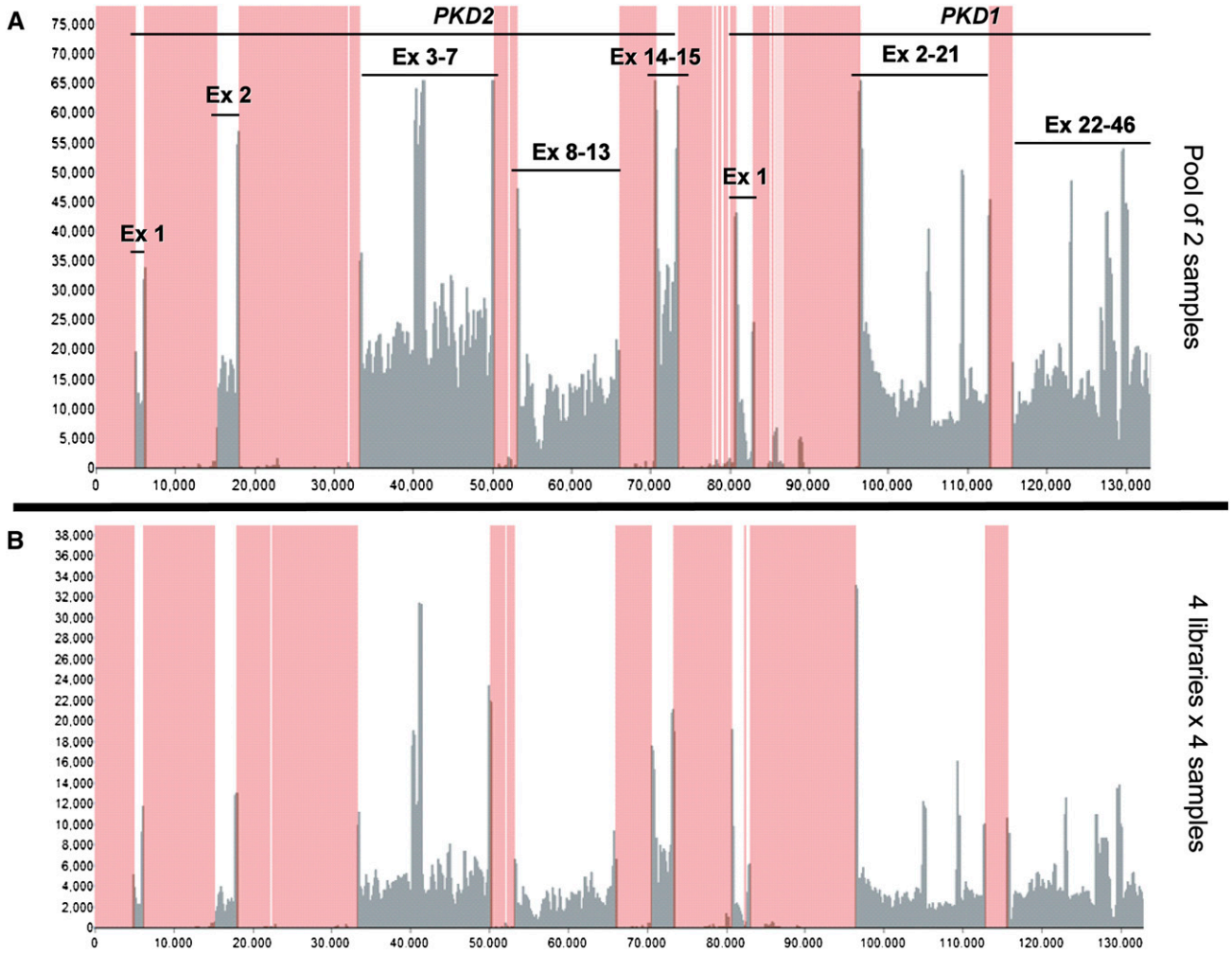

Figure 2. Read depth and coverage analysis of the proof of principle experiment. The diagram shows that all of the regions of interest (indicated by exon lines) were adequately covered, and compares the read depth obtained when pooling two (A) and multiplexing four libraries of four samples each (B) (lanes 1 and 7 in Figure 3A, respectively). This experiment confirmed that sufficient read depth was obtained when multiplexing (B) four libraries of four samples and suggested that sufficient read depth would be obtained even by multiplexing 12 such libraries per lane. The *x*-axis represents genomic interval, the *y*-axis represents number of reads, and the rose-colored areas are out of target regions. Ex, exon.

of reads during data mining, low quality coverage or, for *PKD1*, residual contamination from the pseudogenes (Table 3).

For the group of nonpathogenic variants that were not checked by Sanger sequencing (Supplemental Tables 2 and 3), data from the PKDB and/or "the NCBI Database of Single Nucleotide Polymorphisms (dbSNP)" high Next*GENe* score/read depth or detection in at least two different libraries (high-confidence variants) provided evidence that they were real variants.

The 155 possible pathogenic variants were classified either as definitely pathogenic or as variants of unknown clinical significance (VUCS). VUCS were further classified as highly likely pathogenic, likely pathogenic, indeterminate, likely hypomorphic, or likely neutral (see Concise Methods, Figure 4, and Table 2). Interestingly, these 155 true positive variants included deletions and insertions of up to approximately one-third of the total read length used in this experiment (Figure 5, A and B and Table 2).

Because of the pronounced phenotypic heterogeneity of the study cohort, we focused on the subset of 183 ADPKD probands with a diagnosis compatible with standard clinical and imaging criteria to fairly evaluate the overall detection rate of this experiment.[9,11] In this subset of samples, a detection rate of 115 of 183 (63%) was achieved (66 probands with definitively pathogenic variants, 35 with highly likely pathogenic variants, and 14 with likely pathogenic variants) (Figure 4 and Table 2). In the remaining 68 pedigrees, only indeterminate, likely hypomorphic, novel likely neutral variants (in 7 pedigrees) (Table 2), or synonymous and known polymorphisms (in 61 pedigrees, not shown) were found.

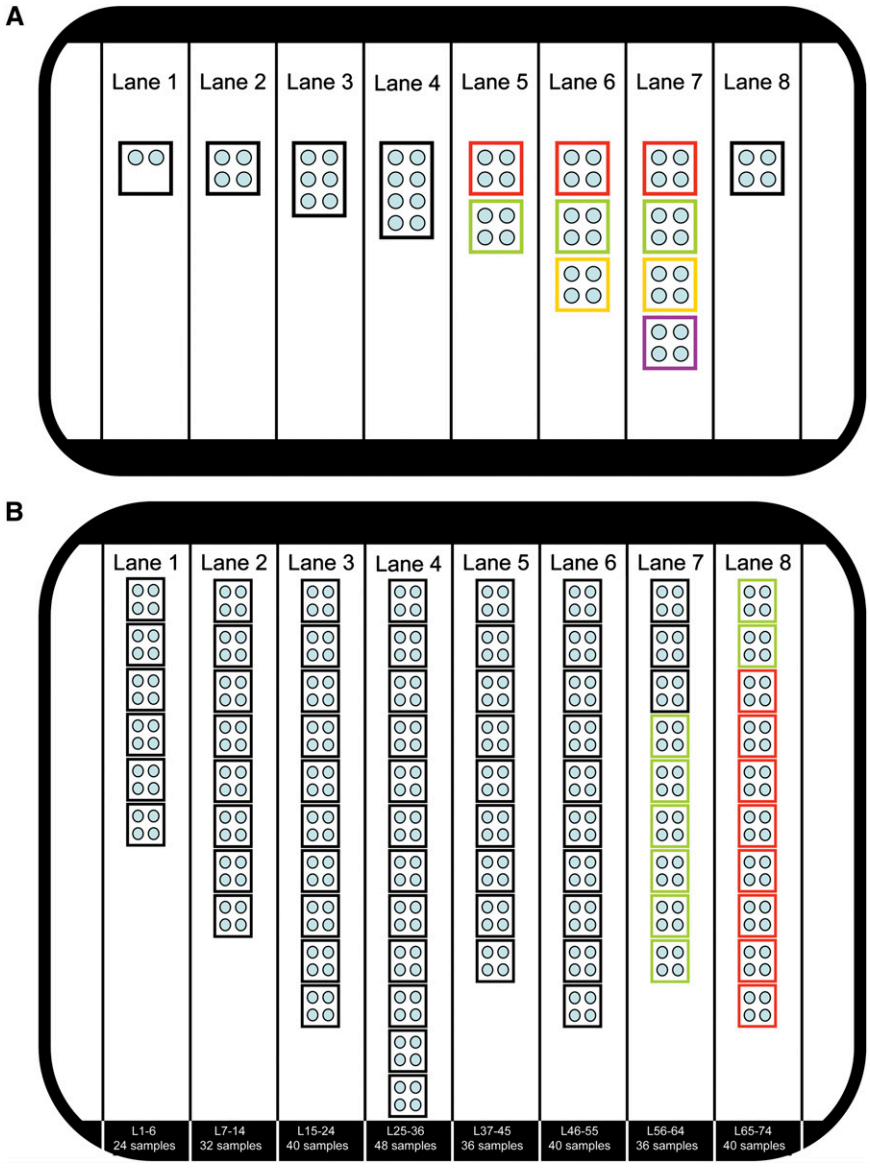

Figure 3. Layout of the two NGS experiments performed in this study. (A) Layout of the proof of principle experiment. The eight lanes of the Illumina flow cell used in the proof of principle experiment are shown: each individual library is shown as a box and each sample as a circle. This proof of principle experiment tested pooling (lanes 1–4) by combining between two and eight samples in the same library, and tested multiplexing (lanes 5–7) by running two to four libraries of four samples each (8, 12, and 16 samples, respectively). Colored boxes indicate the same library during multiplexing (lanes 5–7). Lane 8 tested the discovery workflow by running four unknown samples. All libraries were individually bar coded and paired-end sequenced as 51-bp reads, resulting in an average of 28.5 million paired reads per lane, 91% of which could be re-aligned (approximately 26 million) and generating 10.6 Gb of usable data for variant calling. (B) Layout of the experiment performed to characterize a cohort of 230 novel ADPKD samples. The eight lanes of the Illumina flow cell used in this experiment are shown, and the library and sample number run in each lane is indicated at the bottom. All libraries (boxes) were derived by pooling four samples (small circles). Libraries 1–66 (black and green boxes) were generated by pooling samples after individual PCR amplification, whereas libraries 67–74 (red boxes) were generated by pooling genomic DNA before amplification. Patients in libraries 67–74 (red) and 59–66 (green) are the same. All libraries were individually bar coded and paired end sequenced as 101-bp reads, resulting in an average of 43 million paired reads per lane, 81% of which could be re-aligned (approximately 35 million/lane) and generating 28.3 Gb of re-aligned data suitable for variant calling (much higher than the first experiment due to hardware upgrade of the Illumina instrument). Lane 4 sequenced the highest number of bar codes currently supported by Illumina (12 bar codes, 48 samples).

To further evaluate the sensitivity, specificity, and accuracy of this experiment, we utilized 34 samples that were concurrently characterized by Sanger sequencing during the timeframe of this experiment. In addition we resequenced 14 of the 68 unresolved pedigrees by Sanger sequencing, for a total of 48 samples for which both Sanger sequencing and NGS data were available (Tables 4 and 5). This comparison with the available Sanger data suggested 78% sensitivity, 100% specificity, and 60% accuracy in this experiment (Table 5). Manual inspection of the NGS alignment for the missing variants in the nine false negative pedigrees revealed that they were filtered out because of lack of coverage or too stringent threshold criteria used during data mining (Figure 5, C and D, respectively, and Table 4).

In two cases, NGS detected novel changes previously missed: a novel, highly likely pathogenic change (*PKD1*/p.Cys3081Arg) (Figure 6) and a likely neutral change (*PKD1*/p.Thr2250Met) in two Sanger mutation-negative samples (R1380 and R1930, respectively) (Table 4). Interestingly, manual inspection of the original Sanger chromatograms revealed that they were missed due to unequal amplification (allele dropout) during the original Sanger analysis, with almost complete loss of the mutant allele (p.Cys3081Arg) (Figure 6), as well as an operator-caused mistake with a variant lost at follow-up (p.Thr2250Met). Manual inspection of the binding sites of the original Sanger sequencing primers for p.Cys3081Arg revealed no polymorphisms as a possible cause of primer instability, which was also supported by repeated Sanger sequencing analysis (Figure 6).

Comparison of Pooling of Amplicons versus Pooling of DNA Strategy

Libraries 67–74 were matched to libraries 59–66 in patient content, but genomic

Table 1. Results of the proof of principle experiment

Strategy	Control Variants[a]	Average Read Depth[b]	SD (±)	Highest Read Depth[b]	Lowest Read Depth[b]	Sensitivity (%)[c]	Precision Rate in the Exonic Regions (%)[d]
Pool of 2 (lane 1)	33	14,168	9672	41,065	639	97[e]	100
Pool of 4 (lane 2)	48	19,888	11,998	72,497	397	100	87
Pool of 6 (lane 3)	40	17,702	10,501	56,149	1229	95[e]	91
Pool of 8 (lane 4)	52	27,932	18,537	92,367	2945	92	76
Multiplex of 2×4 (red, lane 5)	42	8209	5340	24,863	315	98[e]	100
Multiplex of 3×4 (red, lane 6)	42	6862	4319	21,452	226	98[e]	93
Multiplex of 4×4 (red, lane 7)	42	2162	1325	7869	129	98[e]	90
Multiplex of 2×4 (green, lane 5)	40	18,943	12,178	60,278	484	97.5[e]	96
Multiplex of 3×4 (green, lane 6)	40	13,986	8956	43,897	535	97.5[e]	92
Multiplex of 4×4 (green, lane 7)	40	9743	6104	31,335	240	97.5[e]	83
Multiplex of 3×4 (yellow, lane 6)	36	6224	3175	15,934	651	97[e]	86
Multiplex of 4×4 (yellow, lane 7)	36	4214	2228	11,256	372	97[e]	96
Multiplex of 4×4 (purple, lane 7)	46	4414	2900	15,575	520	100	100

TP, true positive; TN, true negative; FP, false positive; FN, false negative.
[a]Indicates the number of Sanger-verified control variants (unique and common polymorphisms) available for the corresponding pool.
[b]Read depth is reported as the number of paired-end re-aligned reads per nucleotide site for the control variants (as the average, the highest and the lowest).
[c]Sensitivity was calculated as the number of true positive mutations/number of true positive plus number of false negative mutations (TP/TP+FN).
[d]Precision rate was limited to the exonic regions (in which complete Sanger data were available) and was calculated as number of true positive mutations/number of true positives plus number of false positive mutations (TP/TP+FP). The same mining protocol was used to calculate both the sensitivity and the precision rate.
[e]The two missed variants were two deletions of 38 and 15 bp, which were not detected (one or both, depending on the pool) with the 51-bp long reads used in this proof of principle experiment.

DNA was pooled before PCR amplification (Figure 3B). The two strategies were compared for sensitivity and precision rate by using all of the 48 novel Sanger-verified variants found after analysis of both datasets. Pooling PCR fragments showed that 46 of 48 variants were detected with one false positive, whereas pooling DNA samples showed that 39 of 48 variants were detected with 25 false positives. Hence, pooling of DNA before amplification compared with pooling of PCR amplified fragments led to a substantial loss of sensitivity (96% versus 81%) and a higher number of false positives (1 versus 25).

Identification of Atypical Variants by NGS: Deep Intronic Variants and a *PKD1* Gene Conversion

By amplifying most introns for both genes, NGS allowed a detailed analysis of the pattern of intronic variation in a sizable cohort for the first time, particularly for the duplicated portion of *PKD1*, in which 463 high-confidence intronic variants outside the canonic splice sites were identified (Supplemental Table 3). Although 460 variants were common intronic polymorphisms, *in silico* splicing analysis of three that were unique predicted them as possibly affecting splicing, including patients R1852-*PKD1*/c.7210-10C>A (Table 2), 244111 *PKD2*/c.1094+507G>A, and 100006 *PKD1*/c.216-1198T>G (Table 2). *PKD1*/c.7210-10C>A was the only possibly pathogenic variant detected in patient R1852, and it is described in the PKDB as a variant of indeterminate clinical significance predicted to weaken the polypyrimidine tract (http://pkdb.mayo.edu). Both *PKD2*/c.1094+507G>A and *PKD1*/c.216-1198T>G were identified in patients that were mutation negative in a previous Sanger sequencing analysis (from a group of 28 that were included here for re-analysis by NGS as mutation-negative samples). They are novel variants and predicted to cause pseudoexon activation by creating a new acceptor or new donor site, respectively.

Whereas RNA was not available for R1852, RNA from lymphoblastoid cell lines was analyzed for 244111 and 100006; however, no apparent splicing abnormalities were revealed in these cells.

The deep intronic sequencing obtained by NGS allowed for the first time the fine characterization of a *PKD1* GC event, involving exons 28–32,[33] and 47 variants were identified (12 exonic and 35 intronic) matching one of the *PKD1P1-P6* pseudogenes. Careful comparison with available genomic sequence data showed that this GC likely derives from a conversion event with the *PKD1P6* pseudogene (Figure 7).

DISCUSSION

NGS has critically accelerated the discovery process in human genetics through targeted resequencing,[34] whole-exome analysis,[35–40] and whole-genome analysis.[41,42] To maximize the number of samples that can be run per sequencing lane, targeted resequencing is preferable and more cost-effective when small to moderately sized genomic regions need to be resequenced in large populations. To aid this analysis, methods for genomic partitioning and bar-coding strategies have been developed.[32] Here, we developed an original strategy that uses LR-PCR in association with sample pooling and library bar coding and we applied it to a duplicated genomic region (*PKD1* exons 1–32), for which conventional genomic

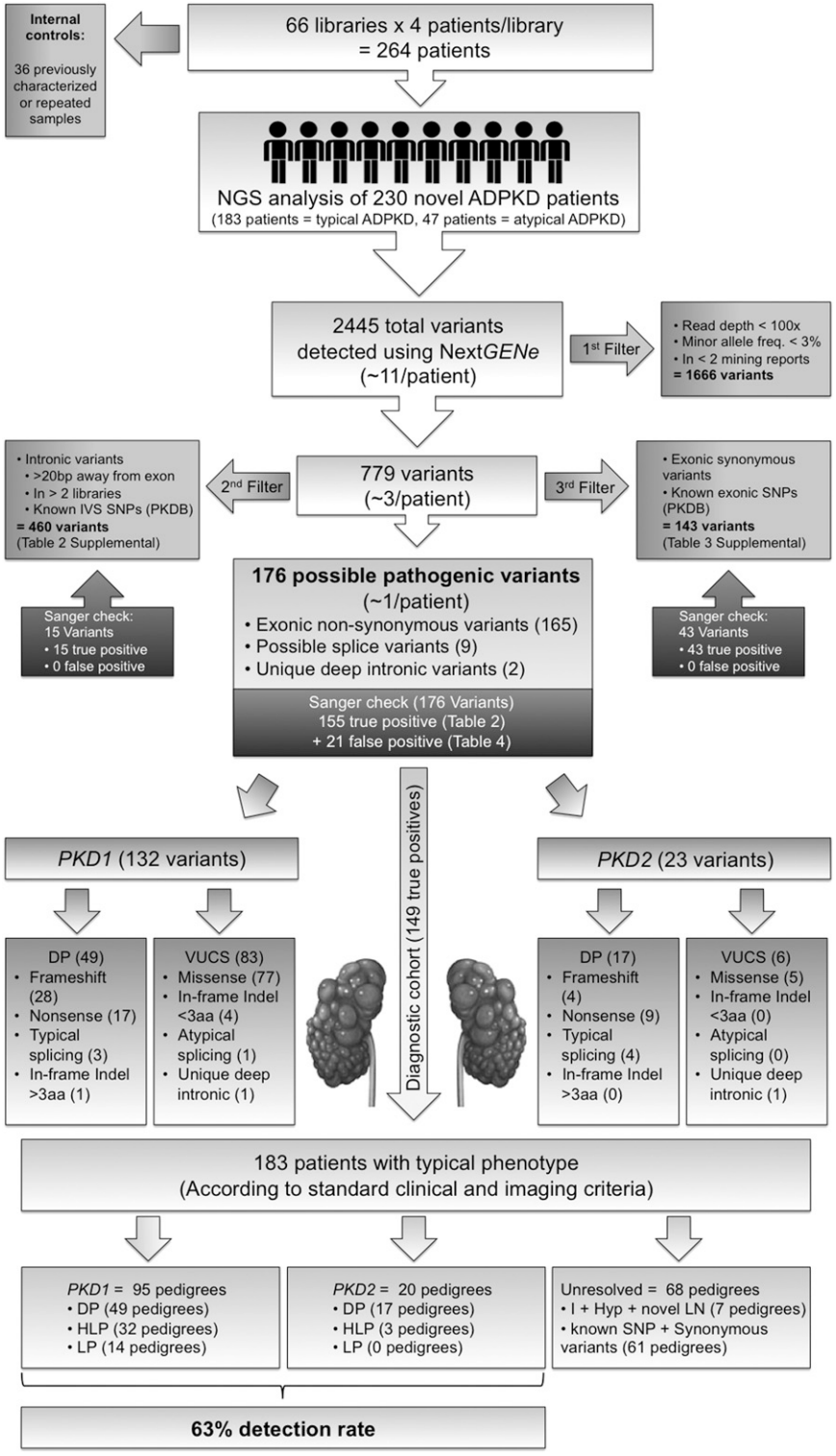

Figure 4. Schematic diagram illustrating the workflow utilized for filtering, parsing, and re-confirming all of the variants derived after the initial data mining in the discovery experiment (Figure 3B). After read re-alignment and variant calling, quality filtering removed 1666 low-confidence variants from the initial pool of 2445 called variants, resulting in 779 high-confidence variants (see Concise Methods for details). This reduced the average number of variants per patient from approximately 10 to 3. Parsing by likelihood of disease association further removed 460 common intronic variants and 143 synonymous or known nonsynonymous exonic polymorphisms, resulting in 176 possible pathogenic variants (approximately 1 per patient). After Sanger re-confirmation, the 155 true positive variants were classified for pathogenicity either as definitely pathogenic (DP) or VUCS, which were further classified as highly likely pathogenic (HLP), likely pathogenic (LP), indeterminate (I), likely hypomorphic (Hyp), and likely neutral (LN) (Table 2 and see Concise Methods). As we focused on the diagnostic cohort of 183 pedigrees (arrow), the genotypes in the pedigrees from this subgroup were classified based on the most pathogenic mutation found as having a DP genotype (49 *PKD1* pedigrees and 17 *PKD2* pedigrees), an HLP genotype (32 *PKD1* pedigrees and 3 *PKD2* pedigrees), an LP genotype (14 *PKD1* pedigrees) (Table 2). Of the 68 pedigrees with unresolved genotype from the diagnostic cohort of 183 patients, 7 carried I, Hyp, or novel LN genotypes (Table 2). The remaining 61 pedigrees from the diagnostic cohort of 183 patients only had synonymous or known polymorphisms (not shown). Hence, within the diagnostic cohort of 183 typical ADPKD according to standard clinical and imaging criteria, the 115 of 183 resolved pedigrees accounted for a final detection rate of 63%. DP, definitely pathogenic; HLP, highly likely pathogenic; LP, likely pathogenic; I, indeterminate; Hyp, likely hypomorphic; LN, likely neutral.

partitioning approaches are not suitable due to the presence of 97.7% identical duplicons.[15]

Pooling and bar coding were utilized as a strategy to achieve cost-efficiency so that NGS could be applied to analyze large ADPKD populations. Although still significant, the cost of NGS is constantly decreasing due to novel strategies for library preparation, decreased hardware cost, and greater sequence output per run.[43] As a more cost-effective workflow is optimized by use of an automated workstation, decreased library preparation costs and introduction of more bar codes to allow up to 96 samples to be analyzed per lane,[43,44] the overall NGS cost will further decrease and make NGS highly competitive compared with conventional Sanger sequencing. The much higher number of reads generated by the recently introduced Illumina HiSeq (approximately 320 million per lane, 8

Table 2. Details of the 155 true positive variants found in the cohort of 230 novel ADPKD pedigrees analyzed by NGS

Pedigree Clinical Features	Ped ID[a]	Pt ID	Gene	Exon	Codon	cDNA Change	Protein Change	PKDB[b]	Classification[c]
Pedigrees with clinical diagnosis compatible with standard clinical and imaging criteria in which DP, HLP, or LP genotypes were found (*n*=115)	DP (*PKD1*=49; *PKD2*=17)								
	M190	R1361	*PKD1*	29	3298	c.9894G>A	p.Trp3298X	Novel	DP
	M201	R1340	*PKD1*	46	4202	c.12604_12631delGGCCGGCT GGGGACAAGGTGTGAGCCTG	p.Gly4202fs146X	Novel	DP
	M254	R1425	*PKD1*	15	1672	c.5014_5015delAG	p.Arg1672fs97X	23×	DP
	M327	R1632	*PKD1*	44	4011	c.12031C>T	p.Gln4011X	4×	DP
	M374	R1557	*PKD1*	5	266	c.796C>T	p.Gln266X	1×	DP
	M375	R1582	*PKD1*	40	3793	c.11379delG	p.Gly3793fs31X	6×	DP
	M388	R1749	*PKD1*	38	3697	c.11090_11091insA	p.His3695fs25X	Novel	DP
	M447	R1918	*PKD1*	15	2142	c.6424C>T	p.Gln2142X	Novel	DP
	M455	R1941	*PKD1*	15	1672	c.5014_5015delAG	p.Arg1672fs97X	23×	DP
	M538	R1804	*PKD1*	36	3603	c.10808G>A	p.Trp3603X	1×	DP
			PKD1	23	2765	c.8293C>T	p.Arg2765Cys	9×	Hyp
	M541	R1450	*PKD1*	3	121	c.360+2T>C	p.Ile120fs	Novel	DP
	M573	R1554	*PKD1*	14	1072	c.3215_3216insA	p.Asn1072fs28X	Novel	DP
	M575	R1572	*PKD1*	IVS17	2403	c.7209+4_+7delAGTG	p.Val2356_Gly2403del	1×	DP
	M576	R1588	*PKD1*	42	3852	c.11554delC	p.Leu3852fs82X	Novel	DP
	M578	R1604	*PKD1*	22	2683	c.8049_8050insGCCGCTCGTGC	p.Cys2683fs4X	Novel	DP
	M580	R1672	*PKD1*	15	1672	c.5014_5015delAG	p.Arg1672fs97X	23×	DP
	M582	R1717	*PKD1*	14	1091	c.3272_3289delTCATGCACACC TACGCTG	p.Val1091_Ala1096delVal-Met-His-Thr-Tyr-Ala	1×	DP
	M586	R1802	*PKD1*	35	3530	c.10588C>T	p.Gln3530X	2×	DP
	M588	R1835	*PKD1*	15	1482	c.4444C>T	p.Gln1482X	Novel	DP
			PKD1	27	3183	c.9548G>A	p.Arg3183Gln	Novel	I
	M591	R1866	*PKD1*	21	2643	c.7927delC	p.Arg2643fs10X	Novel	DP
	M600	R1926	*PKD1*	20	2602	c.7804C>T	p.Gln2602X	Novel	DP
			PKD1	23	2765	c.8293C>T	p.Arg2765Cys	9×	Hyp
	M601	R1939	*PKD1*	11	832	c.2494_2495insC	p.Arg832fs39X	1×	DP
	P19	OX162	*PKD1*	11	874	c.2619_2620insC	p.Cys874fs30X	Novel	DP
			PKD1	11	873	c.2618T>C	p.Val873Ala	Novel	LN
	P21	OX170	*PKD1*	13	1020	c.3058C>T	p.Gln1020X	1×	DP
	P45	OX284	*PKD1*	IVS28	3238	c.9712+1G>T	p.Ser3238fs	Novel	DP
	P94	OX873	*PKD1*	23	2810	c.8428G>T	p.Glu2810X	3×	DP
	P96	OX940	*PKD1*	21	2637	c.7909C>T	p.Gln2637X	Novel	DP
	P104	OX1009	*PKD1*	36	3582	c.10745_10746insC	p.Pro3582fs44X	1×	DP
	P209	OX20	*PKD1*	15	1672	c.5014_5015delAG	p.Arg1672fs97X	23×	DP
	P218	OX1403	*PKD1*	40	3779	c.11337_11338ins AGGAGGCTT CAGCACCAGC	p.Ala3779fs41X	Novel	DP
	P219	OX1394	*PKD1*	11	886	c.2657_2658insACCTTCGTGCC CGGCTGCCC	p.Trp886fs18X	Novel	DP

Table 2. Continued

Pedigree Clinical Features	Ped ID[a]	Pt ID	Gene	Exon	Codon	cDNA Change	Protein Change	PKDB[b]	Classification[c]
	P222	OX1002	*PKD1*	15	1457	c.4369_4370delTC	p.Ser1457fs64X	1×	DP
	P223	OX1051	*PKD1*	15	1960	c.5878C>T	p.Gln1960X	1×	DP
	P226	OX1418	*PKD1*	21	2650	c.7948_7949delCT	p.Leu2650fs9X	1×	DP
	P231	OX969	*PKD1*	26	3082	c.9240_9241delAT	p.Ala3082fs95X	2×	DP
	P238	OX27	*PKD1*	18	2430	c.7288C>T	p.Gln2430X	8×	DP
			PKD1	18	2434	c.7300C>T	p.Arg2434Trp	Novel	LP
			PKD1	10	696	c.2086G>A	p.Ala696Thr	Novel	LN
	P245	OX1948	*PKD1*	10	696	c.2085_2086insC	p.Ala696fs17X	8×	DP
	P246	OX991	*PKD1*	15	1633	c.4897_4898insT	p.Glu1633fs23X	1×	DP
	P272	OX19	*PKD1*	3	117	c.348_352delTTTAA	p.Asn116fs1X	1×	DP
			PKD1	23	2765	c.8293C>T	p.Arg2765Cys	9×	Hyp
	P286	OX1123	*PKD1*	32	3406	c.10216_IVS32+20delAAGAGG TGGGTTCCCTAGAGAAACC	p.Lys3406fs	Novel	DP
	P287	OX1555	*PKD1*	43	3962	c.11885delA	p.Ala3962fs21X	Novel	DP
	P299	OX1650	*PKD1*	15	1117	c.3349C>T	p.Gln1117X	2×	DP
	P369	OX1768	*PKD1*	24	2948	c.8843C>A	p.Ser2948X	Novel	DP
	P437	OX2130	*PKD1*	36	3586	c.10756_10757delGT	p.Val3586fs39X	Novel	DP
	P998	OX1239	*PKD1*	15	1621	c.4861C>T	p.Gln1621X	Novel	DP
	P1001	OX1505	*PKD1*	4	162	c.485delC	p.Ala162fs127X	Novel	DP
	P1008	OX1951	*PKD1*	15	1672	c.5014_5015delAG	p.Arg1672fs97X	23×	DP
	P1009	OX1955	*PKD1*	15	2154	c.6461_6462insCCTGCCGGGA GCCGGAGGTGGACGTGGT	p.Val2154fs29X	Novel	DP
	P1012	OX1246	*PKD1*	18	2444	c.7330_7331insGAT	p.Tyr2444X	Novel	DP
	M365	R1551	*PKD2*	14	845	c.2533C>T	p.Arg845X	2×	DP
	M509	R2020	*PKD2*	IVS5	439	c.1319+1G>A	p.Arg439fs	9×	DP
			PKD1	18	2442	c.7324G>C	p.Glu2442Gln	Novel	I
	M512	R1838	*PKD2*	5	417	c.1249C>T	p.Arg417X	9×	DP
	M567	R1401	*PKD2*	1	68	c.203delC	p.Pro68fs48X	Novel	DP
	M571	R1487	*PKD2*	IVS4	365	c.1094+3_+6delAAGT	p.Ala365fs	7×	DP
	M577	R1601	*PKD2*	14	872	c.2614C>T	p.Arg872X	22×	DP
	M583	R1734	*PKD2*	4	292	c.876C>G	p.Tyr292X	Novel	DP
			PKD2	4	314	c.940C>G	p.Leu314Val	Novel	LN
	M596	R1915	*PKD2*	11	728	c.2182_2183delAG	p.Ser728fs10X	Novel	DP
	P203	OX1358	*PKD2*	1	180	c.538_539insC	p.Leu180fs32X	3×	DP
	P237	OX986	*PKD2*	14	872	c.2614C>T	p.Arg872X	22×	DP
	P288	OX1558	*PKD2*	11	742	c.2224C>T	p.Arg742X	7×	DP
	P300	OX1614	*PKD2*	3	270	c.810delT	p.Phe270fs46X	Novel	DP
	P378	OX1838	*PKD2*	4	306	c.916C>T	p.Arg306X	19×	DP
			PKD1	6	404	c.1211C>G	p.Pro404Arg	1×	I
			PKD1	15	2001	c.6001C>T	p.Arg2001Trp	Novel	I
	P527	OX1432	*PKD2*	IVS4	365	c.1094+3_+6delAAGT	p.Ala365fs	7×	DP

Table 2. Continued

Pedigree Clinical Features	Ped ID[a]	Pt ID	Gene	Exon	Codon	cDNA Change	Protein Change	PKDB[b]	Classification[c]
	P999	OX1244	*PKD2*	IVS5	439	c.1319+1G>T	p.Arg439fs	9×	DP
	P1004	OX1774	*PKD2*	4	320	c.958C>T	p.Arg320X	9×	DP
			PKD1	10	641	c.1922T>C	p.Met641Thr	Novel	LN
	P1006	OX1844	*PKD2*	13	836	c.2508C>G	p.Tyr836X	Novel	DP
	HLP (*PKD1*= 32; *PKD2*= 3)								
	M80	R167	*PKD1*	27	3168	c.9504C>G	p.Phe3168Leu	3×	HLP
	M243	R1364	*PKD1*	15	1503	c.4507G>A	p.Gly1503Arg	2×	HLP
	M351	R1662, R1663	*PKD1*	33	3415	c.10243G>A	p.Glu3415Lys	Novel	HLP
	M378	R1712	*PKD1*	5	381	c.1141G>A	p.Gly381Ser	5×	HLP
	M386	R1738	*PKD1*	46	4150	c.12448C>T	p.Arg4150Cys	4×	HLP
	M412	R1380	*PKD1*	26	3081	c.9241T>C	p.Cys3081Arg	Novel	HLP
	M448	R1909, R1910	*PKD1*	17	2373	c.7118G>A	p.Cys2373Tyr	1×	HLP
	M461	R1962	*PKD1*	4	125	c.373A>G	p.Asn125Asp	Novel	HLP
	M543	R2004	*PKD1*	28	3233	c.9698A>T	p.Glu3233Val	Novel	HLP
			PKD1	28	3232	c.9694A>G	p.Lys3232Glu	Novel	LN
	M581	R1716	*PKD1*	29	3263	c.9787T>C	p.Trp3263Arg	Novel	HLP
	M597	R1917	*PKD1*	5	381	c.1141G>A	p.Gly381Ser	5×	HLP
	M598	R1920	*PKD1*	6	420	c.1259A>C	p.Tyr420Se	Novel	HLP
			PKD1	15	2082	c.2644G>A	p.Ala2082Thr	Novel	LN
	M602	R1944	*PKD1*	26	3130	c.9388_9393delCGGGGC	p.Arg3130_Gly3131del	Novel	HLP
	P20	OX165	*PKD1*	11	845	c.2534T>C	p.Leu845Ser	6×	HLP
	P51	OX341	*PKD1*	15	1699	c.5096C>A	p.Ala1699Asp	Novel	HLP
	P87	OX860	*PKD1*	11	727	c.2180T>C	p.Leu727Pro	7×	HLP
	P93	OX937	*PKD1*	40	3782	c.11345_11346insTTACGA	p.Asp3782_Val3783insTyrAsp	Novel	HLP
	M589	R1841	*PKD1*	5	210	c.629G>A	p.Cys210Tyr	Novel	HLP
	P213	OX1396	*PKD1*	11	727	c.2180T>C	p.Leu727Pro	7×	HLP
	P214	OX1038	*PKD1*	39	3753	c.11258G>T	p.Arg3753Leu	Novel	HLP
	P230	OX824	*PKD1*	39	3751	c.11252A>G	p.Gln3751Arg	1×	HLP
	P281	OX841	*PKD1*	7	508	c.1522T>C	p.Cys508Arg	4×	HLP
	P305	OX1631	*PKD1*	39	3751	c.11252A>G	p.Gln3751Arg	1×	HLP
	P309	OX1947	*PKD1*	15	1999	c.5995G>A	p.Gly1999Ser	2×	HLP
	P312	OX871	*PKD1*	46	4150	c.12448C>T	p.Arg4150Cys	4×	HLP
	P416	OX2062	*PKD1*	11	796	c.2387G>A	p.Tyr796Cys	Novel	HLP
	P435	OX2098	*PKD1*	3	101	c.303_305delCAA	p.N101del	4×	HLP
			PKD1	15	1332	c.3994G>A	p.Asp1332Asn	Novel	LN
	P526	OX1704	*PKD1*	15	1328	c.3982T>C	p.Trp1328Arg	Novel	HLP
	P576	OX2027	*PKD1*	15	1375	c.4124G>A	p.Cys1375Tyr	Novel	HLP
			PKD1	23	2872	c.8615T>G	p.Ile2872Ser	Novel	HLP
			PKD1	15	1292	c.3876C>A	p.Phe1292Leu	Novel	I

Table 2. Continued

Pedigree Clinical Features	Ped ID[a]	Pt ID	Gene	Exon	Codon	cDNA Change	Protein Change	PKDB[b]	Classification[c]
	P601	OX2404	*PKD1*	19	2530	c.7589G>A	p.Hly2530Asp	Novel	HLP
			PKD1	IV15	1254	c.3761C>T	p.Ser1254Leu	Novel	LN
	P1002	OX1509	*PKD1*	23	2771	c.8311G>A	p.Gly2771Lys	18×	HLP
	P1010	OX2057	*PKD1*	11	727	c.2180T>C	p.Leu727Pro	7×	HLP
	M471	R2034	*PKD2*	6	448	c.1343C>A	p.Thr448Lys	Novel	HLP
			PKD1	7	471	c.1412C>T	p.Ser471Leu	Novel	LP
	M604	R1986	*PKD2*	14	886	c.2657A>G	p.Asp886Gly	1×	HLP
			PKD1	23	2779	c.8335G>A	p.Glu2779Lys	2×	I
			PKD1	23	2822	c.8464G>A	p.Val2822Met	1×	I
	M605	R2061	*PKD2*	4	322	c.965G>A	p.Arg322Gln	3×	HLP
			PKD1	43	3906	c.11717G>T	p.Cys3906Phe	Novel	LN
	LP (*PKD1*=14; *PKD2*=0)								
	M127	R1498	*PKD1*	5	271	c.812C>A	p.Ala271Asp	2×	LP
	M152	R1432	*PKD1*	27	3187	c.9561C>A	p.Asp3187Glu	Novel	LP
	M469	R2041	*PKD1*	27	3178	c.9533G>T	p.Ser3178Ile	Novel	LP
	M540	R1558	*PKD1*	18	2423	c.7268C>T	p.Ser2423Phe	2×	LP
	M566	R1040	*PKD1*	23	2816	c.8447T>A	p.Leu2816Gln	Novel	LP
	M572	R1519	*PKD1*	18	2467	c.7400C>T	p.Pro2467Leu	Novel	LP
	M574	R1563	*PKD1*	6	460	c.1379T>C	p.Val460Ala	Novel	LP
	M587	R1832	*PKD1*	18	2423	c.7268C>T	p.Ser2423Phe	2×	LP
	M592	R1881	*PKD1*	15	1914	c.5741G>C	p.Gly1914Ala	Novel	LP
	M595	R1899	*PKD1*	18	2434	c.7301G>A	p.Arg2434Gln	Novel	LP
			PKD1	23	2865	c.8593C>T	p.Arg2865Trp	Novel	I
	M599	R1924	*PKD1*	20	2612	c.7835C>G	p.Ser2612Trp	Novel	LP
	P308	OX1656	*PKD1*	15	1278	c.3834C>A	p.Ser1278Arg	Novel	LP
			PKD1	18	2417	c.7250T>A	p.Leu2417Gln	Novel	LP
	P529	OX832	*PKD1*	15	1278	c.3834C>A	p.Ser1278Arg	Novel	LP
	P1003	OX1701	*PKD1*	15	1610	c.4828_4830delATC	p.Ile1610del	9×	LP
Pedigrees with clinical diagnosis compatible with standard clinical and imaging criteria in which I, Hyp, or novel LN genotypes were found (*n*=7)	M237	R1413	*PKD1*	5	276	c.827C>T	p.Thr276Ile	Novel	I
			PKD1	23	2765	c.8293C>T	p.Arg2765Cys	9×	Hyp
	M267	R181	*PKD1*	23	2742	c.8293C>T	p.Glu2742Lys	2×	Hyp
			PKD1	23	2765	c.8293C>T	p.Arg2765Cys	9×	Hyp
	M281	R1344	*PKD1*	37	3654	c.10960C>G	p.Leu3654Val	2×	I
	M283	R1443	*PKD1*	46	4288	c.12862A>G	p.Ser4288Gly	3×	I
	M590	R1852	*PKD1*	IVS1	2497	c.7210-10C>A	N/A	1×	I
			PKD1	77	532	c.1594C>G	p.Leu532Val	Novel	LN
	P161	OX1211	*PKD1*	10	658	c.1972G>A	p.Ala658Thr	Novel	LN
	M568	R1440	*PKD2*	6	482	c.1445T>G	p.Phe482Cys	3×	Hyp

Table 2. Continued

Pedigree Clinical Features	Ped ID[a]	Pt ID	Gene	Exon	Codon	cDNA Change	Protein Change	PKDB[b]	Classification[c]
Pedigrees with clinical diagnosis noncompatible with standard clinical and imaging criteria in which I or novel LN genotypes were found (n=4)	M414	R1842	*PKD1*	11	927	c.2780C>T	p.Thr927Met	Novel	I
	M593	R1886	*PKD1*	16	2327	c.6969C>T	p.Arg2327Trp	3×	I
	M594	R1890	*PKD1*	29	3240	c.9718G>A	p.Ala3240Thr	Novel	LN
	P57	OX405	*PKD1*	15	1878	c.5633C>T	p.Thr1878Met	Novel	LN
Pedigrees from a group of 28 previous Sanger sequencing mutation negative that were included for re-analysis (and considered as novel) in this experiment, in which unique deep intronic variants predicted to affect splicing were detected	100006	112396	*PKD1*	IVS1	73	c.216-1198T>G	N/A	Novel	LN
	244111	244111	*PKD2*	IVS4	365	c.1094+507G>A	N/A	Novel	LN

DP, definitely pathogenic; HLP, highly likely pathogenic; LP, likely pathogenic; I, indeterminate; Hyp, likely hypomorphic; LN, likely neutral.
[a]Pedigrees are ordered in descending order, for *PKD1* and *PKD2*, respectively; pedigrees with multiple variants are classified based on the most pathogenic, which is shown on top.
[b]Indicates previous description in the PKD Mutation Database (http://pkdb.mayo.edu). The "×" indicates number of pedigrees.
[c]Variants were classified as DP, HLP, LP, I, Hyp, and LN (see Figure 4).

times higher than the Illumina GA2× used in these experiments) and the availability of 96 individual bar codes[44] will soon make it possible to sequence up to 1536 individually barcoded samples for both the *PKD1* and *PKD2* genes concurrently (16 lanes on two flow cells per run), making NGS the method of choice for large population studies. However, the introduction of more scalable, faster turnaround times and simpler workflow instruments like the Illumina MiSeq will likely make NGS an attractive method also for clinical and diagnostic application, particularly when only specific genes or specific sets of genes need to be analyzed, as for ADPKD.

Pooling of DNA before amplification would further simplify the upstream work by greatly reducing the number of amplicons to generate, check, normalize, and pool.[45] However, in our experiments, this did not seem ideal and led to loss of sensitivity and a higher number of false positives. This may be due to variability of DNA quality and errors in DNA concentration measurements due to innate DNA viscosity, as previously suggested.[23] Whole-genome amplification before DNA pooling may be an alternative strategy to overcome this issue. Furthermore, because of the *PKD1* gene duplication, a residual contamination derived from the pseudogenes was detected when mining the data at low stringency (<3% mutant level). Because of the need to filter out the contaminant reads, data mining had to be performed with a 3% mutant level as the minimum threshold, thus posing an additional challenge to pooling samples in the same library.

Detection of small indels has been an issue thus far in NGS due to the short Illumina reads. However, utilizing longer 101-bp reads and the elongation Next*GENe* mining protocol allowed detection of small to medium size indels up to 30% of the read length. This improvement closes the gap with Sanger sequencing, and it is particularly important in ADPKD, in which this mutation type accounts for approximately one-third of all mutations (http://pkdb.mayo.edu). Detection of larger genomic indels by NGS is possible by paired-end mapping[46] and based on variations of the depth of coverage in deleted genomic regions.[47,48] However, this approach was not feasible in these experiments due to the pooling of different samples within the same library and, for *PKD1*, the genomic complexity requiring locus-specific amplification.

In the subset of 183 ADPKD samples with a typical diagnosis according to standard clinical and imaging criteria,[9,11] a maximal detection rate of 63% was achieved, including definitely pathogenic, highly likely pathogenic, and likely pathogenic variants (115 pedigrees) (Figure 4 and Table 2). The remaining unresolved 68 pedigrees (37%) may be due to the phenotypic heterogeneity of the study cohort (purposively mimicking a diagnostic cohort), missed mutations, as well as possible additional genetic heterogeneity. Considering the 19% difference in sensitivity in this experiment compared with Sanger sequencing (Table 5) (corresponding to approximately 13 missed pathogenic mutations in the 68 unresolved

Table 3. Details of cases of discordance between Sanger sequencing and NGS: Cases of NGS false positives after Sanger resequencing (*n*=21)

Gene	Exon/IVS	Codon	Sequence Variant (Amino Acid Change)	Times Detected	Comment
PKD1	1	65	c.194T>C (p.Ile65Thr)	1	Just before homopolymer (**T**CCCC), likely misalignment
PKD1	5	283	c.856_862delTCTGGCC	1	(TCTGGCC) repeated twice, likely misalignment
PKD1	13	1023	c.3068A>G (p.Gln1023Arg)	4	*PKD1P1-P2* contamination
PKD1	15	2236	c.6706T>C (p.Phe2236Leu)	1	Low quality coverage
PKD1	15	2238	c.6713A>C (p.Asp2238Ala)	1	Homopolymer of 4 G nucleotides followed by (AC) repeated twice, likely misalignment
PKD1	23	2872	c.8614delA	2	Palindrome separated by dinucleotide (GCC**A**TCACCG), likely misalignment
PKD1	25	3023	c.9067A>G (p.Met3023Val)	1	Just before a GT motif, likely misalignment
PKD1	33	3454	c.10360delT	1	(TC) dinucleotide repeated twice, likely misalignment
PKD1	33	3456	c.10368_10369insCTC	1	Within a palindrome (GCCAGCC) after a TC motif, likely misalignment
PKD1	40	3780	c.11335_11336insGCGATT	1	Same sequence in wild-type, likely misalignment
PKD1	41	3845	c.11537+2T>GT	1	Low quality coverage
PKD1	41	3864	c.11591A>C (p.His3864Pro)	1	Within palindrome (GC**A**CG), likely misalignment
PKD2	11	720	c.2159delA	4	Homopolymer of 8 A nucleotides, likely misalignment
PKD2	14	843	c.2527delG	1	Homopolymer of 2 G nucleotides, likely misalignment

Bold indicates the site where the corresponding change occurs for single nucleotide or single deletion changes.

Table 4. Details of cases of discordance between Sanger sequencing and NGS

Variant Description	Ped ID	Pt ID	Gene	Exon/IVS	Codon	cDNA change	Protein	Comment
Variants detected by Sanger sequencing but not by NGS (*n*=10)	M368	R1700	*PKD1*	1	37	c.108_109insC	p.Cys37fs76X	Lack of coverage
	M615	R1953	*PKD1*	11	845	c.2534T>C	p.Leu845Ser	Mutant percentage below threshold
	M499	R2001	*PKD1*	12	960	c.2879G>A	p.Gly960Asp	Mutant percentage below threshold
	M152	R1432	*PKD1*	15	1362	c.4084C>T	p.Ser1362Pro	Mutant percentage below threshold
	M118	R95	*PKD1*	15	2212	c.6635G>A	p.Ser2212Asn	Mutant percentage below threshold
					2215	c.6644G>A	p.Arg2215Gln	Mutant percentage below threshold
	M307	R1573	*PKD1*	IVS20	2621	c.7864-2A>G	p.2621fs	Mutant percentage below threshold
	M307	R1581	*PKD1*	IVS20	2621	c.7864-2A>G	p.2621fs	Mutant percentage below threshold
	P387	OX2242	*PKD1*	25	3016	c.9047A>G	p.Gln3016Arg	Mutant percentage below threshold
	P229	OX1056	*PKD1*	IVS31	3390	c.10170+25_+45del CTGGGGGTCCT GGGCTGGG	p.Gln3390fs	Next*GENe* score below threshold
Variants detected by NGS but missed during the original Sanger sequencing analysis (*n*=2)	M453	R1930	*PKD1*	15	2250	c.6749C>T	p.Thr2250Met	Operator-caused error
	M412	R1380	*PKD1*	26	3081	c.9241T>C	p.Cys3081Arg	Allele dropout

pedigrees), a putative detection rate of approximately 70% would be achievable, which is comparable with data recently obtained in a similar diagnostic setting by Sanger sequencing.[49] Sensitivity, specificity, and accuracy will improve by individually bar coding each sample and by further development of the data mining strategies.[50–52]

The application of NGS in these experiments has allowed the discovery and characterization of missed and atypical variants (allele dropout, gene conversion, and deep intronic variants). Allele dropout is a well known cause of missed mutations[53] due to unequal amplification of heterozygote alleles. By utilizing larger amplicons, a previously undetected

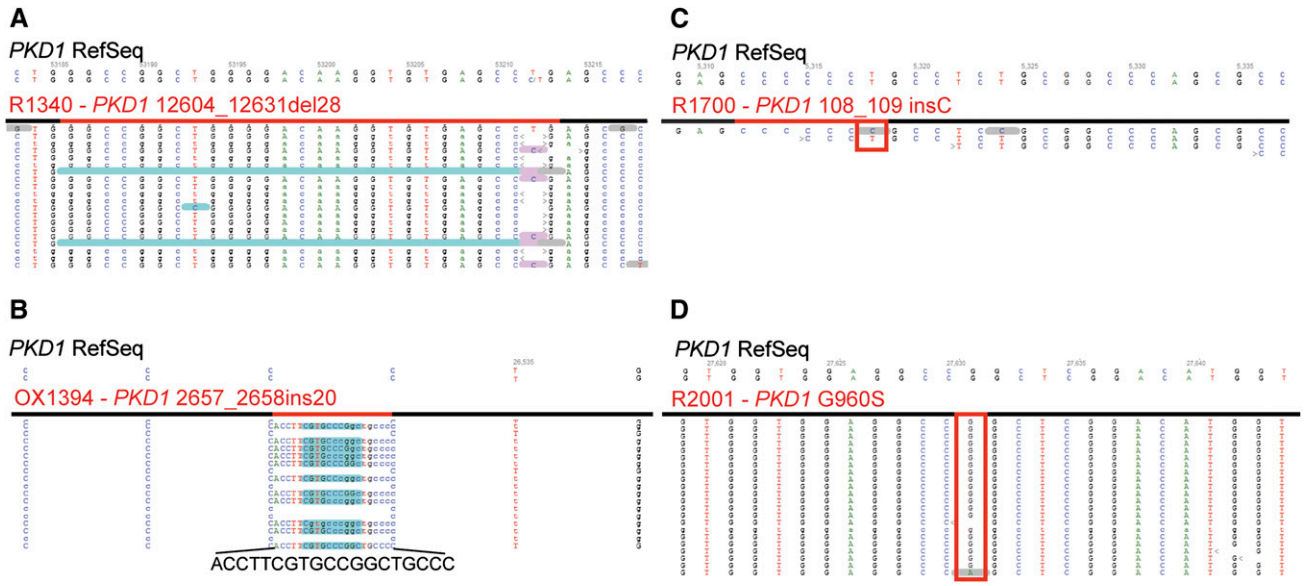

Figure 5. Representative examples of visual inspection of NGS alignments. (A and B) Examples of a medium-sized deletion or insertion. By using the Next*GENe* elongation approach, indels up to one-third of the total read length were detected (c.12604_12631del28 in A and c.2657_2658ins20 in B). (C and D) Examples of missed mutations because of (C) insufficient read depth or (D) low-scoring variant call. (C) Mutation *PKD1* c.108_109insC occurs in a homopolymer of six consecutive C, and it is here covered by a single read that is wild-type for the insertion and shows a below-threshold T>C transition soon after the homopolymer. *PKD1* exon 1 is 85% GC rich and seemed to be often under-represented in these experiments, suggesting that the corresponding amplicon should be added in excess to provide sufficient read depth for confident mutation detection. (D) Mutation *PKD1* p.Gly960Ser is detected by Next*GENe* software (gray underlining, one single mutant read in this screenshot) but is assigned a very low-confidence score because of the low number of mutant reads and is consequently removed as low-confidence variant during data mining.

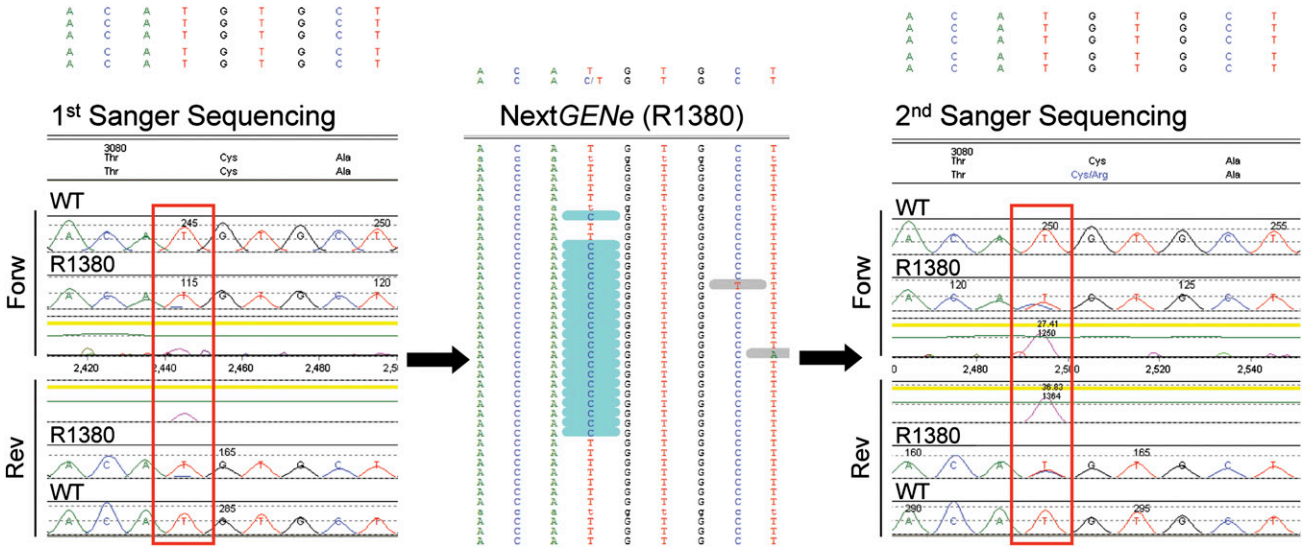

Figure 6. Detection of allele dropout in a previous Sanger mutation-negative sample. NGS manual inspection for the previously Sanger-missed mutation *PKD1* p.Cys3081Arg (middle panel), showing a high-confidence mutation call at a well covered site; following the NGS workflow, this mutation was correctly identified in sample R1380 (right panel, forward and reverse trace). Comparison with the original Sanger screening (left panel) shows that the mutant cytosine is strongly under-represented in both sequencing directions, suggesting unequal amplification and allele dropout rather than a sequencing artifact as the likely cause.

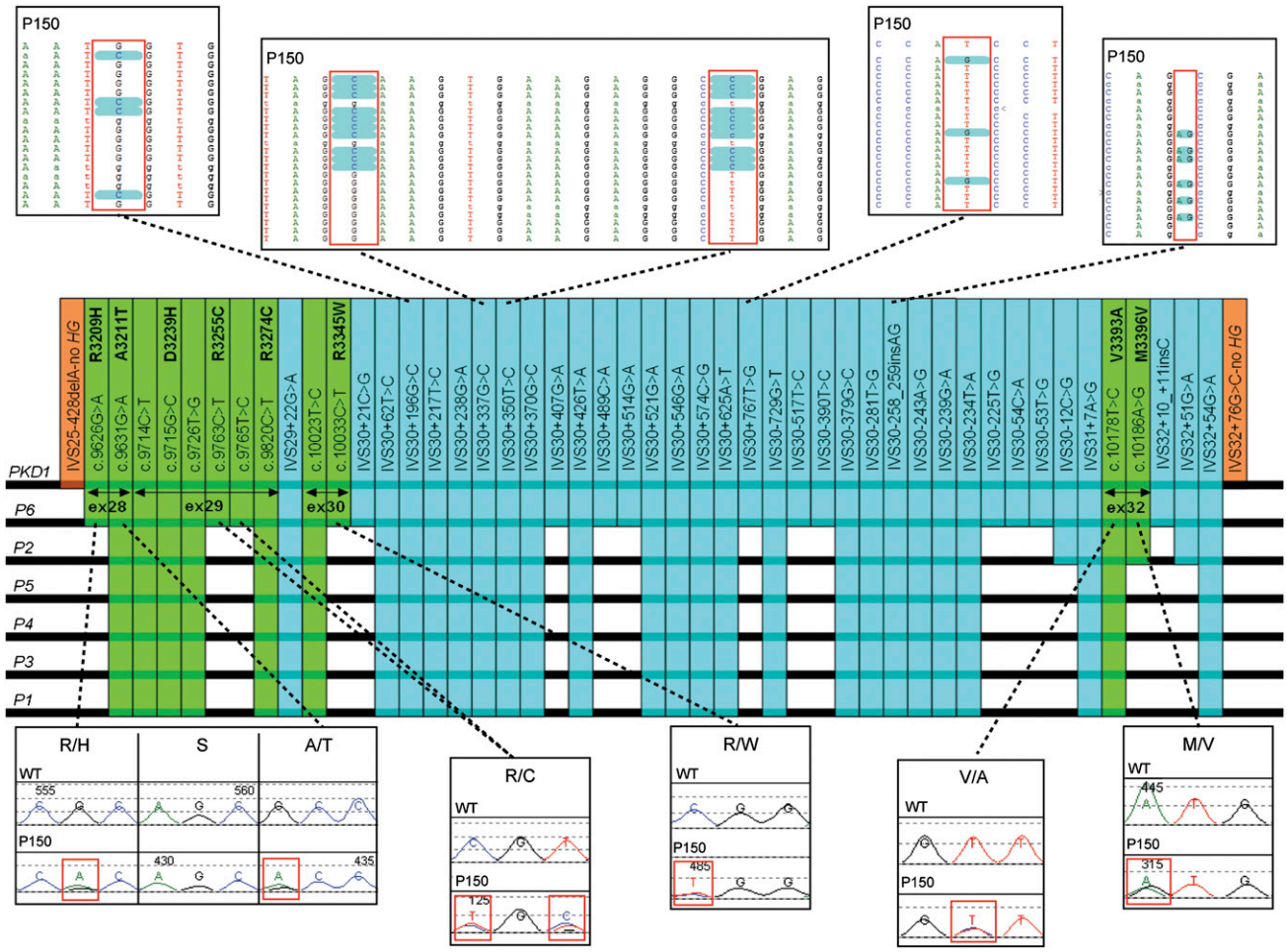

Figure 7. Detailed analysis of a GC event involving *PKD1* exons 28–32. By using long amplicons and achieving deep sequencing of all of the IVS regions, detailed genomic data have been obtained of the entire genomic region putatively involved in this GC event. High-score, high-coverage data mining identified 12 exonic (green) and 35 intronic (light blue) variants that match one of the *PKD1P1-P6*. Careful comparison with available genomic sequence data (*PKD1P1-P6*) shows a complete match with *PKD1P6*, suggesting that the GC event took place between *PKD1* and the duplicon *PKD1P6* over 8.5 kb of genomic sequence. Orange bars are the 5′ and 3′ boundaries of the GC event, before and after which no further *PKD1-P1-6* sequence match is observed. Selected Sanger and NGS chromatograms for some of the variants from two family members are shown in the corresponding panels.

mutation (p.Cys3180Arg) was identified in an otherwise mutation-negative sample. *A posteriori* verification of the original Sanger screening chromatograms revealed an extreme underrepresentation of the mutant allele. The apparent lack of polymorphisms in the binding sites of the Sanger primers[54] suggests stochastic variability in allele amplification efficiency as the likely cause. This may be due to unexpected secondary structures of the DNA template, and may be an underestimated cause of missed mutations in ADPKD due to the GC richness of the *PKD1* gene.

By deep sequencing the *PKD1* introns in the case of a GC involving exons 28–32,[33] an extended haplotype of exonic and intronic variants was generated, which suggested that the GC event likely occurred between *PKD1* and the *PKD1P6* duplicon. The application of this strategy to suspected GC events proved that they are genuine GC events, determined their genomic origin, and precisely defined the extent of the GC event, thereby emphasizing their disease association.

The inclusion of most intronic regions for both genes allowed us to explore the pattern of intronic variation and generate a catalog of 460 high-confidence intronic variants. Notably, this dataset provides an important filter for common intronic variants, which will be useful to pinpoint rare, potential splice-changing deep intronic variants.[55–57] Because of the genomic duplication, these data are not available for the entire duplicated portion of *PKD1* from the 1000 Genomes Project and other sequencing projects.

Lymphoblast-derived cDNA analysis of two private deep intronic variants (*PKD1*/c.216-1198T>G and *PKD2*/c.1094+507G>A) predicted to cause intron exonization in otherwise mutation-negative samples revealed no apparent splicing abnormality, although we cannot definitely exclude such event in

Table 5. Comparison of sensitivity, specificity, and accuracy between Sanger sequencing (as gold standard) and NGS for the 48 patients that were characterized with both methodologies in the cohort of 230 novel ADPKD patients

		Sanger Sequencing (as Gold Standard)	
		+	−
NGS	+	True positive (TP)	Sanger false negative (FN)
		28	1
	−	NGS false negative (FN)	True negative (TN)
		8	11

Sanger sensitivity, 97% (36/37); specificity, 100% (11/11); accuracy 98% (47/48). NGS sensitivity, 78% (29/37); specificity, 100% (11/11); accuracy, 60% (29/48). For both Sanger sequencing and NGS, sensitivity was calculated as TP/TP+FN, specificity as TN/TN+FP, and accuracy as TP+TN/TP+FP+TN+FN.

kidney-derived epithelial cells due to tissue-specific splicing regulation.[58,59]

Allele dropout and unrecognized intronic variants affecting splicing may account for part of the approximately 10% mutation-negative samples in clinically well characterized cohorts,[3,60] and are well suited to be detected and characterized by NGS rather than conventional Sanger sequencing.

In conclusion, we designed a strategy to analyze the *PKD1* gene by NGS, a locus at which conventional genomic partitioning approaches are not possible, and may serve as an example for NGS analysis of other disease-associated segmental duplications in the human genome.[6,61–63] The very high throughput of NGS makes it ideal for large-scale projects and will make genotyping of large ADPKD cohorts feasible (*e.g.*, in association with large ADPKD clinical trials or ADPKD population studies). The NGS deep sequencing of the entire genomic structure of both the *PKD1* and *PKD2* genes holds the potential of discovering atypical mutations like the ones we describe here, and helps to clarify the genetic basis of the 10% of ADPKD pedigrees in which no mutation is detectable through conventional Sanger sequencing.

CONCISE METHODS

Amplification and Pooling of LR-PCR Amplicons

LR-PCR amplicons were amplified as previously described[3,64] using the primers described in Supplemental Table 1. Amplicons were checked and normalized to a control of known concentration by gel densitometry, using a Kodak Gel Logic 200 Instrument. The amplicon standards were previously amplified ones with no smearing or extra-bands and the concentration measured using a fluorescence-based method (Quant-It, Invitrogen Inc). For the purpose of amplicon equimolar assembly, variations in concentration up to three-folds were tolerated. Thirty picomoles of each amplicon were assembled for each sample, and equimolar amounts of assembled samples (2–8) were pooled to generate 2–3 μg of total material for library preparation (proof of principle and libraries 1–66 of the second experiment).

For libraries 67–74, genomic DNA of each individual sample were diluted to 10 ng/μl, gently agitated overnight, pooled in equal amounts, mixed again overnight, and then PCR amplified.

Preparation of Illumina Bar-Coded, Paired-End Libraries, Clusters Formation, and Reads Generation on the Illumina GA2X Sequencer

Input pooled DNA was fragmented by sonication using the Covaris E210 ultrasonicator (Covaris Inc), and Illumina bar-coded, paired-end libraries prepared using the NEBNext kit (New England Biolabs). Bar-coded libraries were subjected to cluster formation using the Illumina cluster station (Illumina Inc), and clusters loaded at an average density of 250,000 clusters per lane.

Low-level bar coding (<8 indexes per lane) was performed following the bar-code combinations suggested by Illumina. Clusters were sequenced with 51- or 101- bp long paired-end reads.

Data Mining, Sequence Variant Call, and Confirmation of Mutations

Reads were exported as FASTAQ files and deconvoluted by bar code, and each bar code was imported separately into the Next*GENe* software version 1.99 (SoftGenetics Inc) for data mining. Mining was performed using the Next*GENe* elongation algorithm. This polishes the reads (reducing the machine error rate well below 1%); elongates them by approximately 30% of the original length by read-to-read comparison, facilitating variants and indels calls; and preserves the paired-end information, reducing the false positives rate. Next*GENe* software assigns a confidence score to each variant call on a scale from 0 to 30, which was used for further variant filtering.[50]

Each bar code was mined at 4% mutant allele percentage–100× read depth (high stringency), 4% mutant allele percentage–100× read depth on trimmed, 76-bp long reads (medium stringency), and 3% mutant allele percentage–100× read depth (low stringency). These three mining protocols were typically linked within Next*GENe* for each bar code, and several bar codes linked in series, for automatic execution. Reports were exported as spreadsheets and merged to generate a consensus report. Variants were called if present in two of three of the reports (high-confidence variants). Low-confidence variants were defined as variants present in only one report, which was always the low stringency one as defined above.

All novel and likely pathogenic exonic nonsynonymous/typical splicing variants and a select group of other variants were Sanger verified by re-amplifying the four original samples contained in each library.

Exonic synonymous and intronic variants in Supplemental Table 2 that were not tested by Sanger sequencing were validated by filtering at 1000× minimum read depth and progressively higher stringency (using Next*GENe* scores of 27, 28, 29, and 30). To ensure high confidence in these datasets, variants were included only if present in at least two different libraries with a minimum Next*GENe* score of 27, unless it was previously described (PKDB or dBSNP) or Sanger verified (high-confidence variants). Similarly, high read depth and high Next*GENe* score (approximately 8000× and 27 minimum score) were used to validate all the intronic variants involved in the *PKD1-P6* gene conversion.

Scoring of VUCS

VUCS were classified based on the scores of the ADPKD Mutation Database (http://www.pkdb.mayo.edu) for previously described variants, and as previously described for novel variants.[17] Hypomorphic alleles were classified as previously described.[65]

Analyses of Splicing Variants

Private intronic variants were analyzed for splicing potential using the following tools: BDGP (http://www.fruitfly.org/seq_tools/splice.html),[66] Spliceport (http://spliceport.cs.umd.edu/),[67] GeneSplicer (http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml),[68] NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/),[69] and HSF (http://www.umd.be/HSF/).[70] cDNA analysis was performed as previously described.[17]

ACKNOWLEDGMENTS

We thank all of the patients and families who agreed to participate in this study; Mr. Bernard and Mrs. Edith Waterman for their generous support; Dr. Franklyn G. Prendergast (Mayo Center for Individualized Medicine), Dr. Eric D. Wieben (Mayo Advanced Genomics Technology Center), and Dr. Fernando Fervenza (Mayo Clinic) for their support; and Mrs. Linda Pelleymounter (Mayo Clinic) for her collaboration. We thank Dr. Jared Grantham and Dr. Arlene Chapman of the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease for kindly supplying samples 100006 and 244111, in which the two deep intronic variants described in this manuscript were found.

This work was supported by Grant R21DK083669, the Robert W. Fulk Career Development Award, a generous gift of the Bernard and Edith Waterman Charitable Foundation through the Mayo Center for Individualized Medicine, Grant R01DK58816, and by the Mayo Translational PKD Center (P30DK090728).

This work was partially presented as an abstract at the Annual Meeting of the American Society of Nephrology, November 16–21, 2010, Denver, Colorado, as well as at the Polycystic Kidney Disease: From Bench to Bedside meeting of the Federation of American Societies for Experimental Biology, June 26 to July 1, 2011, Saxtons River, Vermont.

DISCLOSURES

None.

REFERENCES

1. Dalgaard OZ: Bilateral polycystic disease of the kidneys; a follow-up of two hundred and eighty-four patients and their families. *Acta Med Scand Suppl* 328: 1–255, 1957
2. Iglesias CG, Torres VE, Offord KP, Holley KE, Beard CM, Kurland LT: Epidemiology of adult polycystic kidney disease, Olmsted County, Minnesota: 1935-1980. *Am J Kidney Dis* 2: 630–639, 1983
3. Rossetti S, Consugar MB, Chapman AB, Torres VE, Guay-Woodford LM, Grantham JJ, Bennett WM, Meyers CM, Walker DL, Bae K, Zhang QJ, Thompson PA, Miller JP, Harris PC CRISP Consortium: Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. *J Am Soc Nephrol* 18: 2143–2160, 2007
4. Ravine D, Walker RG, Gibson RN, Forrest SM, Richards RI, Friend K, Sheffield LJ, Kincaid-Smith P, Danks DM: Phenotype and genotype heterogeneity in autosomal dominant polycystic kidney disease. *Lancet* 340: 1330–1333, 1992
5. Harris PC, Torres VE: Polycystic kidney disease. *Annu Rev Med* 60: 321–337, 2009
6. Bischof JM, Chiang AP, Scheetz TE, Stone EM, Casavant TL, Sheffield VC, Braun TA: Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum Mutat* 27: 545–552, 2006
7. Watnick TJ, Piontek KB, Cordal TM, Weber H, Gandolph MA, Qian F, Lens XM, Neumann HP, Germino GG: An unusual pattern of mutation in the duplicated portion of *PKD1* is revealed by use of a novel strategy for mutation detection. *Hum Mol Genet* 6: 1473–1481, 1997
8. Watnick TJ, Gandolph MA, Weber H, Neumann HPH, Germino GG: Gene conversion is a likely cause of mutation in *PKD1*. *Hum Mol Genet* 7: 1239–1243, 1998
9. Ravine D, Gibson RN, Walker RG, Sheffield LJ, Kincaid-Smith P, Danks DM: Evaluation of ultrasonographic diagnostic criteria for autosomal dominant polycystic kidney disease 1. *Lancet* 343: 824–827, 1994
10. Nascimento AB, Mitchell DG, Zhang XM, Kamishima T, Parker L, Holland GA: Rapid MR imaging detection of renal cysts: Age-based standards. *Radiology* 221: 628–632, 2001
11. Pei Y, Obaji J, Dupuis A, Paterson AD, Magistroni R, Dicks E, Parfrey P, Cramer B, Coto E, Torra R, San Millan JL, Gibson R, Breuning M, Peters D, Ravine D: Unified criteria for ultrasonographic diagnosis of ADPKD. *J Am Soc Nephrol* 20: 205–212, 2009
12. Harris PC, Rossetti S: Molecular diagnostics for autosomal dominant polycystic kidney disease. *Nat Rev Nephrol* 6: 197–206, 2010
13. Hogan MC, Masyuk TV, Page LJ, Kubly VJ, Bergstralh EJ, Li X, Kim B, King BF, Glockner J, Holmes DR 3rd, Rossetti S, Harris PC, LaRusso NF, Torres VE: Randomized clinical trial of long-acting somatostatin for autosomal dominant polycystic kidney and liver disease. *J Am Soc Nephrol* 21: 1052–1061, 2010
14. Loftus BJ, Kim UJ, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, Deslattes Mays A, Cao Y, Xu RX, Kang HL, Mitchell S, Eichler EE, Harris PC, Venter JC, Adams MD: Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics* 60: 295–308, 1999
15. Symmons O, Váradi A, Arányi T: How segmental duplications shape our genome: Recent evolution of ABCC6 and PKD1 Mendelian disease genes. *Mol Biol Evol* 25: 2601–2613, 2008
16. Bogdanova N, Markoff A, Gerke V, McCluskey M, Horst J, Dworniczak B: Homologues to the first gene for autosomal dominant polycystic kidney disease are pseudogenes. *Genomics* 74: 333–341, 2001
17. Rossetti S, Strmecki L, Gamble V, Burton S, Sneddon V, Peral B, Roy S, Bakkaloglu A, Komel R, Winearls CG, Harris PC: Mutation analysis of the entire *PKD1* gene: Genetic and diagnostic implications. *Am J Hum Genet* 68: 46–63, 2001
18. Tucker T, Marra M, Friedman JM: Massively parallel sequencing: The next big thing in genetic medicine. *Am J Hum Genet* 85: 142–154, 2009
19. Mardis ER: Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9: 387–402, 2008
20. Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31–46, 2010
21. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ: Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* 5: 887–893, 2008
22. Margraf RL, Durtschi JD, Dames S, Pattison DC, Stephens JE, Mao R, Voelkerding KV: Multi-sample pooling and illumina genome analyzer sequencing methods to determine gene sequence variation for database development. *J Biomol Tech* 21: 126–140, 2010
23. Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PE, Tops CM, Breuning MH, van Ommen GJ, den Dunnen JT, Devilee P, Hes FJ: Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat* 30: 1703–1712, 2009
24. Melum E, May S, Schilhabel MB, Thomsen I, Karlsen TH, Rosenstiel P, Schreiber S, Franke A: SNP discovery performance of two second-generation sequencing platforms in the NOD2 gene region. *Hum Mutat* 31: 875–885, 2010

25. Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C: Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* 38: e142, 2010
26. Nijman IJ, Mokry M, van Boxtel R, Toonen P, de Bruijn E, Cuppen E: Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat Methods* 7: 913–915, 2010
27. Kenny EM, Cormican P, Gilks WP, Gates AS, O'Dushlaine CT, Pinto C, Corvin AP, Gill M, Morris DW: Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res* 18: 31–38, 2011
28. Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, Wilson CJ, Altshuler D, Gabriel SB, Daly MJ, Thorburn DR, Mootha VK: High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet* 42: 851–858, 2010
29. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, Orr N, Matthews C, Qi L, Crenshaw A, Markovic Z, Fredrikson KM, Jacobs KB, Amundadottir L, Jarvie TP, Hunter DJ, Hoover R, Thomas G, Harkins TT, Chanock SJ: Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum Genet* 124: 161–170, 2008
30. Goossens D, Moens LN, Nelis E, Lenaerts AS, Glassee W, Kalbe A, Frey B, Kopal G, De Jonghe P, De Rijk P, Del-Favero J: Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. *Hum Mutat* 30: 472–476, 2009
31. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7: 111–118, 2010
32. Turner EH, Ng SB, Nickerson DA, Shendure J: Methods for genomic partitioning. *Annu Rev Genomics Hum Genet* 10: 263–284, 2009
33. Rossetti S, Chauveau D, Kubly V, Slezak JM, Saggar-Malik AK, Pei Y, Ong AC, Stewart F, Watson ML, Bergstralh EJ, Winearls CG, Torres VE, Harris PC: Association of mutation position in polycystic kidney disease 1 (PKD1) gene and development of a vascular phenotype. *Lancet* 361: 2196–2201, 2003
34. Otto EA, Hurd TW, Airik R, Chaki M, Zhou W, Stoetzel C, Patil SB, Levy S, Ghosh AK, Murga-Zamalloa CA, van Reeuwijk J, Letteboer SJ, Sang L, Giles RH, Liu Q, Coene KL, Estrada-Cuzcano A, Collin RW, McLaughlin HM, Held S, Kasanuki JM, Ramaswami G, Conte J, Lopez I, Washburn J, Macdonald J, Hu J, Yamashita Y, Maher ER, Guay-Woodford LM, Neumann HP, Obermüller N, Koenekoop RK, Bergmann C, Bei X, Lewis RA, Katsanis N, Lopes V, Williams DS, Lyons RH, Dang CV, Brito DA, Dias MB, Zhang X, Cavalcoli JD, Nürnberg G, Nürnberg P, Pierce EA, Jackson PK, Antignac C, Saunier S, Roepman R, Dollfus H, Khanna H, Hildebrandt F: Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet* 42: 840–850, 2010
35. Biesecker LG: Exome sequencing makes medical genomics a reality. *Nat Genet* 42: 13–14, 2010
36. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42: 30–35, 2010
37. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 106: 19096–19101, 2009
38. Haack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, Boehm D, Uziel G, Lamantea E, Invernizzi F, Poulton J, Rolinski B, Iuso A, Biskup S, Schmidt T, Mewes HW, Wittig I, Meitinger T, Zeviani M, Prokisch H: Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat Genet* 42: 1131–1134, 2010
39. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42: 790–793, 2010
40. Teer JK, Mullikin JC: Exome sequencing: The sweet spot before whole genomes. *Hum Mol Genet* 19[R2]: R145–R151, 2010
41. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 362: 1181–1191, 2010
42. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328: 636–639, 2010
43. Garber K: Fixing the front end. *Nat Biotechnol* 26: 1101–1104, 2008
44. Kozarewa I, Turner DJ: 96-plex molecular barcoding for the Illumina Genome Analyzer. *Methods Mol Biol* 733: 279–298, 2011
45. Otto EA, Ramaswami G, Janssen S, Chaki M, Allen SJ, Zhou W, Airik R, Hurd TW, Ghosh AK, Wolf MT, Hoppe B, Neuhaus TJ, Bockenhauer D, Milford DV, Soliman NA, Antignac C, Saunier S, Johnson CA, Hildebrandt F GPN Study Group: Mutation analysis of 18 nephronophthisis associated ciliopathy disease genes using a DNA pooling and next generation sequencing strategy. *J Med Genet* 48: 105–116, 2011
46. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318: 420–426, 2007
47. Medvedev P, Stanciu M, Brudno M: Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 6[Suppl]: S13–S20, 2009
48. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 19: 1586–1592, 2009
49. Hoefele J, Mayer K, Scholz M, Klein HG: Novel PKD1 and PKD2 mutations in autosomal dominant polycystic kidney disease (ADPKD). *Nephrol Dial Transplant* 26: 2181–2188, 2011
50. Pelleymounter LL, Moon I, Johnson JA, Laederach A, Halvorsen M, Eckloff B, Abo R, Rossetti S: A novel application of pattern recognition for accurate SNP and indel discovery from high-throughput data: Targeted resequencing of the glucocorticoid receptor co-chaperone FKBP5 in a Caucasian population. *Mol Genet Metab* 104: 457–469, 2011
51. Bansal V: A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 26: i318–i324, 2010
52. Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA: Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res* 20: 537–545, 2010
53. Laios E, Glynou K: Allelic drop-out in the LDLR gene affects mutation detection in familial hypercholesterolemia. *Clin Biochem* 41: 38–40, 2008
54. Ward KJ, Ellard S, Yajnik CS, Frayling TM, Hattersley AT, Venigalla PN, Chandak GR: Allelic drop-out may occur with a primer binding site polymorphism for the commonly used RFLP assay for the -1131T>C polymorphism of the Apolipoprotein AV gene. *Lipids Health Dis* 5: 11, 2006
55. King K, Flinter FA, Nihalani V, Green PM: Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. *Hum Genet* 111: 548–554, 2002
56. Clendenning M, Buchanan DD, Walsh MD, Nagler B, Rosty C, Thompson B, Spurdle AB, Hopper JL, Jenkins MA, Young JP: Mutation

deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* 10: 297–301, 2011
57. Lo YF, Nozu K, Iijima K, Morishita T, Huang CC, Yang SS, Sytwu HK, Fang YW, Tseng MH, Lin SH: Recurrent deep intronic mutations in the SLC12A3 gene responsible for Gitelman's syndrome. *Clin J Am Soc Nephrol* 6: 630–639, 2011
58. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476, 2008
59. Venables JP: Downstream intronic splicing enhancers. *FEBS Lett* 581: 4127–4131, 2007
60. Rossetti S, Harris PC: Genotype-phenotype correlations in autosomal dominant and autosomal recessive polycystic kidney disease. *J Am Soc Nephrol* 18: 1374–1380, 2007
61. Pulkkinen L, Nakano A, Ringpfeil F, Uitto J: Identification of ABCC6 pseudogenes on human chromosome 16p: Implications for mutation detection in pseudoxanthoma elasticum. *Hum Genet* 109: 356–365, 2001
62. Lee HH, Niu DM, Lin RW, Chan P, Lin CY: Structural analysis of the chimeric CYP21P/CYP21 gene in steroid 21-hydroxylase deficiency. *J Hum Genet* 47: 517–522, 2002
63. Khurana E, Lam HY, Cheng C, Carriero N, Cayting P, Gerstein MB: Segmental duplications in the human genome reveal details of pseudogene formation. *Nucleic Acids Res* 38: 6997–7007, 2010
64. Rossetti S, Chauveau D, Walker D, Saggar-Malik A, Winearls CG, Torres VE, Harris PC: A complete mutation screen of the ADPKD genes by DHPLC. *Kidney Int* 61: 1588–1599, 2002
65. Rossetti S, Kubly VJ, Consugar MB, Hopp K, Roy S, Horsley SW, Chauveau D, Rees L, Barratt TM, van't Hoff WG, Niaudet P, Torres VE, Harris PC: Incompletely penetrant PKD1 alleles suggest a role for gene dosage in cyst initiation in polycystic kidney disease. *Kidney Int* 75: 848–855, 2009
66. Reese MGEF, Eeckman FH, Kulp D, Haussler D: Improved splice site detection in Genie. *J Comput Biol* 4: 311–323, 1997
67. Dogan RI, Getoor L, Wilbur WJ, Mount SM: SplicePort–an interactive splice-site analysis tool. *Nucleic Acids Res* 35: W285–W291, 2007
68. Pertea M, Lin X, Salzberg SL: GeneSplicer: A new computational method for splice site prediction. *Nucleic Acids Res* 29: 1185–1190, 2001
69. Brunak S, Engelbrecht J, Knudsen S: Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* 220: 49–65, 1991
70. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C: Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37: e67, 2009

This article contains supplemental material online at http://jasn.asnjournals.org/lookup/suppl/doi:10.1681/ASN.2011101032/-/DCSupplemental.