Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing

Borràs, Daniel M; Vossen, Rolf; Liem, Michael; Buermans, Henk P J; Dauwerse, Hans; van Heusden, Dave; Gansevoort, Ron T.; den Dunnen, Johan T; Janssen, Bart; Peters, Dorien J M

Published in:
Human Mutation

DOI:
10.1002/humu.23223

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Methods

Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing

Daniel M. Borràs1,2,3 | Rolf H. A. M. Vossen4 | Michael Liem4
Henk P. J. Buermans4 | Hans Dauwerse5 | Dave van Heusden5 | Ron T. Gansevoort6∗
Johan T. den Dunnen4,5,7 | Bart Janssen1 | Dorien J. M. Peters5∗ | Monique Losekoot7∗ | Seyed Yahya Anvar4,5

1GenomeScan B.V, Leiden, The Netherlands
2Institut National de la Santé et de la Recherche Médicale (INSERM), Institut of Cardiovascular and Metabolic Disease, Toulouse, France
3Université Toulouse III Paul-Sabatier, Toulouse, France
4Leiden Genome Technology Center (LGTC), Department of Human Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands
5Department of Human Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands
6Department of Nephrology, University Hospital Groningen, University Medical Center Groningen, Groningen, The Netherlands
7Department of Clinical Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands

Correspondence
Seyed Yahya Anvar, Department of Human Genetics, Leiden University Medical Center, Postzone: S-04-P, Postbus 9600, 2300 RC Leiden, the Netherlands.
Email: s.y.anvar@lumc.nl

∗On behalf of the DIPAK consortium.
Communicated by Graham R. Taylor

Abstract
A genetic diagnosis of autosomal-dominant polycystic kidney disease (ADPKD) is challenging due to allelic heterogeneity, high GC content, and homology of the PKD1 gene with six pseudogenes. Short-read next-generation sequencing approaches, such as whole-genome sequencing and whole-exome sequencing, often fail at reliably characterizing complex regions such as PKD1. However, long-read single-molecule sequencing has been shown to be an alternative strategy that could overcome PKD1 complexities and discriminate between homologous regions of PKD1 and its pseudogenes. In this study, we present the increased power of resolution for complex regions using long-read sequencing to characterize a cohort of 19 patients with ADPKD. Our approach provided high sensitivity in identifying PKD1 pathogenic variants, diagnosing 94.7% of the patients. We show that reliable screening of ADPKD patients in a single test without interference of PKD1 homologous sequences, commonly introduced by residual amplification of PKD1 pseudogenes, by direct long-read sequencing is now possible. This strategy can be implemented in diagnostics and is highly suitable to sequence and resolve complex genomic regions that are of clinical relevance.

Keywords
ADPKD, complex genomic regions, DNA diagnostics, long-read sequencing, PacBio, PKD1, single-molecule real-time sequencing, variant detection

1 | INTRODUCTION

DNA sequencing technologies have widely been applied in biomedical and biological research as well as diagnostics. Relatively low-cost and high-throughput are major advantages of next-generation sequencing (NGS) over standard diagnostic assays (Mardis, 2013; Oliver, Hart, & Klee, 2015; Su et al., 2011). However, despite widespread use of NGS-based diagnostics strategies (Chang & Li, 2013; Codina-Solà et al., 2015; Dewey et al., 2014; LaDuca et al., 2014; Ligt et al., 2012; Ozsolak & Milos, 2011; Sun et al., 2015; von Kanel & Huber, 2013; Willig et al., 2015; Yang et al., 2013), short-read sequencing approaches such as whole-genome sequencing (WGS) and whole-exome sequencing (WES), often fail at reliably characterizing complex regions of the human genome (Chaisson et al.,...
2 | MATERIALS AND METHODS

2.1 | Selection of subjects and DNA isolation

Nineteen genotyped patient samples were selected for this study from the diagnostic laboratory in which at least one pathogenic mutation was detected by Sanger sequencing or MLPA. The selection aimed to include different types of variants (e.g., SNVs, as well as small and larger insertions and deletions [indels]) that are located in exons or in immediately flanking intronic sequences, for both the duplicated regions as well as the unique part of PKD1. Although PKD2 is not a complex gene and is not the focus of this study, the sequencing of LR-PCR fragments for PKD2 was performed as a proof of principle of long-read sequencing and detection of variants also for PKD2. Genomic DNA isolation was performed from peripheral blood samples using PUREGENE nucleic acid purification chemistry on the AUTOPURE LS 98 Instrument (Qiagen).

2.2 | Long-read sequencing and variant identification for ADPKD genes

2.2.1 | LR-PCR amplification

To cover the entire PKD1 and PKD2 coding regions (including exon boundaries), a total of five and nine LR-PCR fragments were designed, respectively. Primers were optimized to produce amplicons of similar sizes (>4Kb) that could be pooled to improve sequencing efficiency and loading capacity for SMRT sequencing (Supp. Table S1; Supp. Fig. S1). The major part of PKD1 intron 1 was excluded from the design due to its large size and the lack of previously reported pathogenic variants in this region. Fragments were amplified from 50 ng of genomic DNA using 1x Extensor Hi-Fidelity Long Range PCR Master Mix (Thermo Scientific, Massachusetts, USA) on a 25 µl of PCR reaction volume with 200 nM of M13-tagged primers. Initial denaturation was performed for 10 min at 98°C, followed by 35 cycles of 15 sec at 98°C and 10 min at 68°C, respectively. Final extension was 10 min at 68°C. Products were size selected using the BluePippin DNA size selection system to classify them in three different groups of sizes 4.3–6.1, 7.1–7.5, and 7.6–8.1 Kb (Supp. Table S1; Supp. Fig. S1). Fragments of equal size were pooled equimolar, and were visually inspected by band intensity on agarose gel. Finally, all pools were purified with a 0.6x v/v ratio of AMPure XP Beads (Beckman-Coulter, Woerden, Netherlands).

2.2.2 | SMRT sequencing library preparation

Sample indexes for patient tracking were added to the LR-PCR fragments using an additional five-cycle PCR with the previous LR-PCR conditions. Barcoded pools were then purified with AMPure XP Beads, and pooled equimolar according to their size. Molar concentration was verified on a Bioanalyzer 12000 chip (Agilent, California, USA). For each barcoded pool, a SMRT-bell library was prepared according to the PacBio’s 5- or 10-Kb Template Preparation procedures. Pooled amplicons were sequenced on five SMRT cells on the PacBio RSII system with the P6 sequencing chemistry. Data collected from 360-min movie time was preprocessed using the standard primary analysis tools (Fig. 1).

In this study, we aim to show the value of single-molecule long-read sequencing as a tool to characterize genetic variants associated with autosomal-dominant polycystic kidney disease (ADPKD). ADPKD is a common inherited disease that accounts for 5%–10% of end-stage renal disease (Harris & Rossetti, 2010; Spithoven et al., 2014). Most ADPKD pathogenic variants occur in PKD1 (MIM# 601313) and PKD2 (MIM# 173910) genes with a reported prevalence of 85% and 15%, respectively (Barua et al., 2009; Harris & Rossetti, 2010). The mutation spectrums in PKD1 and PKD2 are highly heterogeneous, with no mutation hotspots present, indicating that pathogenic variants in either PKD1 or PKD2 are usually private (Gout, Martin, Brown, & Ravine, 2007; Harris & Rossetti, 2010). The screening of PKD1 is challenging due to difficulties in amplification and low resolution of its complex locus (Qi et al., 2013; Rossetti et al., 2007; Tan et al., 2009). This is partly due to its high homology for most of PKD1 sequence with six pseudo-genes as well as high GC content (Qi et al., 2013; Rossetti et al., 2007; Tan et al., 2009). In this study, we used PKD1 as an excellent example of a challenging and complex locus.

Several attempts have been made to improve the screening of PKD1 gene by using short-read NGS approaches to replace the standard diagnostics based on Sanger sequencing and multiplex ligation-dependent probe amplification (MLPA) assays (Eisenberger et al., 2015; Mallawaarachchi et al., 2016; Qi et al., 2013; Rossetti et al., 2012; Tan et al., 2014; Trujillano et al., 2014). These strategies provided a clear diagnosis with high sensitivity and specificity (97%–100%) for ADPKD diagnostics. We show that direct sequencing of long-range PCR (LR-PCR) products eliminates the interference of residual amplification of PKD1 pseudogenes, as well as alignment ambiguities. This also enabled a reliable identification of pathogenic variants, from single-nucleotide variants (SNVs) to large deletions.

2015; Lee & Schatz, 2012). These regions are often associated with extreme GC content, segmental duplications (SDs), low-complexity sequences, and gaps in the human reference sequence (Chaisson et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014). Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Laver et al., 2016; Loomis et al., 2013; Qiao et al., 2015; Lee & Schatz, 2012; Steinberg et al., 2014).
2.2.3 Sequence alignment and variant calling

We used the RS_Resequencing protocol from SMRT Analysis Suite v.2.3 to align long reads against the human reference genome, downloaded from the Genome Reference Consortium version 37 patch 13 (GRCh37.p13). Samples were demultiplexed into individual files (H5 and BAM formats) using known barcode sequences and a minimum barcode identity score of 22 (Fig. 1). Alignments were filtered to contain mapped reads with a mapping quality threshold of 30 Phred score using Samtools v.1.2 (Supp. Table S2). Read coverage and targeted PCR statistics for LR-PCR amplicons were computed with BedTools v.2.25.0, and PicardTools v.1.1.40 (Supp. Tables S2–S4). Variant calling was performed using Quiver (allowing for diploid calling and maximum coverage of 10,000). Variants with a Quiver confidence score lower than 40 were filtered out from downstream analysis (Fig. 1).

2.2.4 Reconstruction of allelic sequences

PKD1 and PKD2 allelic sequences were reconstructed using the Long Amplicon Analysis available in SMRT Analysis Suite 2.3. Only reads longer than 3,000 base pairs (bp) and average signal to noise ratio of three were used for the reconstruction (Fig. 1). Based on this reference-free subread (full-length and nonfull-length reads) clustering, chimeric sequences were identified and comprehended ≤0.85% (6,288/738,822) of subreads that were subsequently removed from the analysis. Allelic sequences of PKD1 and PKD2 were aligned to the human reference genome GRCh37.p13 using BLASR (Chaisson & Tesler, 2012), and reporting a single best-scoring alignment. Variants were extracted by comparison between the reconstructed alleles and the human reference sequence with the Variant Description Extractor from the Mutalyzer Suite 2.0.21 (Vis, Vermaat, Taschner, Kok, & Laros, 2015).

2.2.5 Loss of heterozygosity analysis

Loss of heterozygosity (LoH) for each amplified fragment was assessed to identify patients with potential large deletions for PKD1. We first identified heterozygous substitutions within the amplified fragments with a variant frequency between 25% and 75%. Amplified fragments with zero heterozygous substitutions were identified as LoH. Large deletions produce multiple LR-PCR fragments dropouts, and were identified by the detection of consecutive LoH fragments. The detection of consecutive LoH fragments was not a direct identification of large deletions per se, rather than an indication of the presence of large deletions in the amplified LR-PCR fragment regions. Identified LoH regions were then compared with large deletions detected by MLPA.

3 ADPKD VARIANT NOMENCLATURE AND GENOTYPING

Variant descriptions were standardized to concord with HGVS guidelines (den Dunnen et al., 2016), using the Mutalyzer Name Checker tool (Wildeman, van Ophuizen, Dunnen, & Taschner, 2008). Genomic HGVS descriptions were converted to coding notations using the Position Converter from Mutalyzer (Wildeman et al., 2008). Only changes in RefSeq-annotated canonical transcripts for PKD1 (NM_001009944.2) and PKD2 (NM_000297.3) were further analyzed. HGVS descriptions of deletion–insertions (delins) were manually inspected to avoid variant redundancies and undesired clustering of neighboring independent events (Fig. 1; Supp. Table S5). Then, standardized variants were annotated using the Variant Effect Predictor (VEP), from Ensembl tools v83 (McLaren et al., 2010), with additional parameters “-everything,” and “-refseq” (Fig. 1). All variant annotations reported by VEP are fully disclosed in the raw VCF files (EGAS00001002106). Variant frequency
and coverage were used to filter low-confidence variants by applying thresholds for: (1) sequencing depth of $\geq 50x$ subreads and $\geq 15x$ reads that ensures a sufficient control over the SMRT sequencing random error rate (1% mismatches and 13% indels) and (2) minimum variant frequency of 10% for substitutions and 15% for delins (Fig. 1). For interpreting insertion and deletion frequencies, neighboring bases were also examined. The selection of strong pathogenic variant candidates or variants of unknown clinical significance (VUCS) was based on the following criteria: (1) high predicted effect on the coding sequence or splice-site region (e.g., missense, in-frame indels, frameshifts, and splice-site acceptor or donor variants); (2) population frequency in the 1000 genomes project <1%; (3) unique occurrence (1/19) (~5%) in the patient cohort since, in ADPKD, no single disease-causing variant accounts for more than 2% of affected families (Harris & Rossetti, 2010), or more than 1.7% of ADPKD reported cases in the ADPKD database (PKDB) (http://pkdb.mayo.edu/; accessed version 3.1) (Gout et al., 2007) (Fig. 1).

3.1 | Clinical diagnostics pipeline for ADPKD genotyping

3.1.1 | Sanger sequencing

The current diagnostics pipeline for ADPKD genotyping, including Sanger sequencing and MLPA, uses a different set of LR-PCR primers to target the duplicated part of PKD1 (exons 1–32) (Supp. Table S6). The nonduplicated region of PKD1 (exons 33–46), and PKD2 regions (exons 1–5) were amplified using targeted standard PCR reactions (Supp. Fig. S1), with 100 ng of input genomic DNA with M13 tail primers. The nested and standard PCR amplicons were designed to cover the complete coding regions and splice sites with at least 20 bp of flanking intronic sequences (Supp. Tables S7 and S8). The duplicated part of PKD1, which includes exons 1–32, was amplified using four different LR-PCR fragments that covered exons 1, 2–13, 14–21, and 22–32, respectively (Supp. Table S6). LR-PCR amplification was performed using Thermo Scientific (Massachusetts, USA) 2× Exensor Long Range PCR Master Mix on 50 ng of DNA. Then, a nested PCR was carried out on 4 µl 100–250x of diluted product to obtain the final Sanger sequencing fragments. The nested PCR primers with an M13 tail were used to amplify the coding region including 5–20 bp of intronic sequences (Supp. Fig. S1). Large exons such as exon 5, 10, 11, 15, and 23 were amplified using overlapping nested PCR products, although 10 bases of exon 15 (c.6503–6514) were not covered. Nested PCR and standard PCR of the nonduplicated part of PKD1, and PKD2, was carried out in a final volume of 15 µl in GoTaq Colorless Taq Reaction buffer with 0.6 U of Taq DNA polymerase (Promega, Leiden, Netherlands) at a final concentration 5 mM for each primer, 200 µM of each dNTP. After a hot start at 95°C, a denaturation was performed for 5 min at 95°C, followed by 35 cycles of 45 sec at 94°C, 45 sec at 60°C, and 30 sec at 72°C. The final extension was of 5 min at 72°C in a T-Professional Thermocycler (Biometra, Göttingen, Germany; Westburg, Leusden, Netherlands). All liquid handling steps were automated using the SciClone (ALH-HV96 pipetting station; Perkin Elmer, Massachusetts, USA) or Biomek FX workstation (Beckman-Coulter, Woerden, Netherlands). PCR products (20–50 ng) were purified using an Ampure XP PCR purification kit and sequenced using BigDye Terminator v3.1 sequencing reactions (Applied Biosystems, California, USA) with PAGE purified ~21M13 or M13REV sequencing primer. The excess of dye terminators was removed by gel filtration using the Edge Biosystem Dye Terminator Removal (DTR) with a 96-well plate. After electrophoresis on an ABI Prism 3730 (XL) DNA analyzer (Life technologies, California, USA; Applied Biosystems, California, USA), data processing was automated using SeqPatient software (Sequence Pilot, JSI Medical Systems GmbH, Ettenheim, Germany).

3.1.2 | MLPA

To detect large deletions and duplications, two commercially available MLPA kits (P351-B2 and P352-C1; MRC-Holland, Amstermd, The Netherlands) were used following manufacturer’s protocols and manuals.

3.2 | Comparative analysis of SMRT sequencing and current ADPKD diagnostic assay

The overlap between identified variants based on PacBio and Sanger sequencing data was achieved by assessing identical standardized HGVS descriptions. Only variants with predicted effects on coding DNA or splice-site regions were considered (Supp. Table S5). PacBio and Sanger variants were manually inspected to detect overlapping variants with discordant descriptions between the two datasets. To facilitate interpretation, each unique variant was further annotated with its PKDB clinical significance, single-nucleotide polymorphism database version 144 (dbSNP) identifier, and the number of patients where it was detected in the cohort. Surrounding bases were evaluated to identify and remove potential sequencing artifacts occurring in homopolymer stretches. Finally, variants were considered as high-confidence variants if previously reported in PKDB or dbSNP, showed strong PacBio sequencing evidence of being present, or detected in any patient by both Sanger and PacBio sequencing.

3.3 | Short-read loss of power for known PKD1 pathogenic variants in WGS and WES

Previously known pathogenic variants for PKD1 gene were obtained from PKDB. Only variants that were classified as “definitely pathogenic” were selected for further analysis. Large deletions (few hundred bp to several Kbp long) were excluded from the analysis as they are not usually detected with common variant calling algorithms. For the genomic position of each pathogenic variant, sequencing depth was extracted from nine publicly available WGS and WES datasets (Sun et al., 2015). In addition, we included the sequencing depth of nine randomly selected libraries from the study of Rossetti et al. (2012), in which the authors used a similar strategy based on LR-PCR and followed by short-read sequencing. Each library represents an equimolar pool of DNA from four different patient samples that were not possible to further demultiplex because individuals were not barcoded. Variant positions with low sequencing depth (<8 reads, or <32 for the short-read LR-PCR approach) were marked as inaccessible positions of clinical significance using BedTools v2.25.0. Finally, variant positions were classified into three categories based
FIGURE 2  SMRT sequencing and variant calling of LR-PCR amplicons. A: Sequencing depth (DP; in number of reads) of the alignments to chromosome 16 and chromosome 4. Number of uniquely aligned reads (y axis, blue line) sequenced with PacBio that mapped to PKD1 and PKD2. Off-target amplification is discriminated from the main PKD1 gene sequences showing alignments to pseudogene homologous sequences at proximal loci (e.g., PKD1P1, PKD1P5, PKD1P6) (blue boxes). B: Mapping quality (MQ; in Phred quality scores; values >90 were scaled down for visualization purposes), and sequencing depth (DP; in number of reads) of uniquely aligned molecules to PKD1 (NM_001009944.2) for the five LR-PCR fragments amplified. Mapping quality of alignments with even coverage distribution along the amplified fragments (fragments), including regions with SDs, repetitive elements (repeats), and high GC content (GC%). Despite fragments A and E showing lower coverage, compared with the average sequencing depth of $\geq 421 \times$ (minimum $\geq 19 \times$; maximum $1,528 \times$), they had sufficient coverage for variant calling within the exon regions, including the first exons of PKD1, with average coverage of $\geq 55 \times$ (minimum $\geq 24 \times$; maximum $91 \times$) (Supp. Table S4). C: We detected 1,506 intron variants (blue) and 177 coding or splice-site variants (yellow). The predicted transcript effects of coding and splice-site variants were quantified (bar chart) as log10 count (x axis) on the number of individuals with poor coverage at each position: (1) variants with sufficient coverage in all nine individuals; (2) variants reported inaccessible in two to four individuals; and (3) variants reported inaccessible in five or more individuals.

3.4 | Data availability

Sequencing data and alignments in BAM format can be accessed through the European Genome-phenome Archive (EGA), as well as raw variants in VCF file format, under the EGA study identifier EGAS00001002106. Coding or splice-site variants were also uploaded to the Leiden Open Variation Database (LOVD). Description and examples of custom scripts used in this manuscript are accessible upon request from a local GitLab repository.

4 | RESULTS

4.1 | Targeted sequencing of ADPKD genes

Direct sequencing of LR-PCR fragments (designed to specifically and uniquely amplify PKD1 and PKD2 gene regions) (Supp. Fig. S1; Supp. Table S1) was performed to evaluate the utility of long-read sequencing in resolving ADPKD for molecular diagnostics. All PKD1 and PKD2 exons (including the duplicated part of PKD1, as well as 20 bp of flanking intron regions) from 19 ADPKD patients could be completely covered using long reads, sequenced on the PacBio RS II platform (Fig. 2; Supp. Fig. S2). Most of the long reads (94.4%) were uniquely mapped to PKD1 and PKD2 (Supp. Table S2). Reads originating from residual off-target amplification (5.6%; Supp. Table S2) introduced during the
LR-PCR steps were identified, and discriminated, by their unique alignment to the PKD1 pseudogenes (Fig. 2; Supp. Table S2). All PKD1 and PKD2 protein coding and flanking intron sequences (±20 bp) were covered at average sequencing depth ≥421× (minimum ≥19×; maximum 1,528×), with ≥97.36% of bases over ≥30×, which was well above the applied threshold of ≥15× reads (Supp. Tables S2 and S4). Amplicons that cover the first and last exons of PKD1 were underrepresented when compared with other LR-PCR fragments, with a total of ≥593 average reads (minimum ≥300; maximum 1,580) and ≥87 (minimum ≥35; maximum 153) for PKD1 fragments A and E, respectively (Fig. 2B; Supp. Table S3). The usually difficult to sequence first exons of both PKD1 and PKD2 genes were covered, on average ≥55× (minimum ≥24×; maximum 91×) and ≥71× (minimum ≥43×; maximum 111×), respectively (Supp. Table S4). Most of the sequenced reads (>99.9%) were uniquely mapped to PKD1 and PKD2 (Fig. 2B; Supp. Fig. S2).

Large deletions in PKD1

Detection of allele dropouts and large deletions in PKD1 was assessed by performing a LoH analysis for each of the amplified regions (Fig. 3). We identified 17 LR-PCR fragments with LoH among all 19 patients sequenced. Most of LoH regions (10) were identified in fragment E (Fig. 3). Only two patients showed consecutive LoH fragments indicating the presence of large deletions spanning between two or more LR-PCR fragments. These consecutive LoH fragments are not a direct identification of the deletions per se but an indication of the presence of large deletions in the amplified region. The two patients that showed two or more consecutive fragments with LoH (Fig. 3) were in concordance with large deletions identified by MLPA as pathogenic variants in the same ADPKD patients. A deletion of ≥1,543 bp (c.(2097+1_2098-1)_3640del; exons 11–15) and a deletion of ≥9,108 bp (c.(287+1_288-1)_(9397+1_9398-1)del; exons 3–26) were detected by MLPA for patient sample 7 and 13, respectively. With the current experimental design, however, the exact location of the breakpoints for each large deletion could not be determined with either method.

Comparative analysis between SMRT-Seq and the ADPKD diagnostic assay

The evaluation of 167 coding or splice-site variants identified by standard ADPKD diagnostic assay showed that 159 out of 167 were correctly detected by PacBio (Supp. Fig. S3C). The overall observed sensitivity and specificity in detecting coding variants was of 95.2% (159/167) and 88.8% (159/179), respectively. Eight variants were...
Loss of BORRÀS ET AL. (Illumina) NGS genetic regions such as Accurate diagnosis is a difficult task when performed in complex regions may be from the observed high variability in coverage ranging from >30,000x (Supp. Fig. S6). Moreover, several exonic regions may be expected to be missed in many samples, irrespective of the short-read sequencing strategy used (Supp. Fig. S6).

4.5 Loss of PKD1 diagnostic power in short-read (Illumina) NGS

The potential loss of diagnostic power when resolving PKD1 by short-read NGS was evaluated based on 797 pathogenic variants that were previously reported and validated, and are publicly available in PKDB. The repetitive nature of PKD1 gene hampers proper alignment of short Illumina NGS reads (Supp. Fig. S5). Over 12% of the reported pathogenic variants would have been missed in WGS and WES data purely due to poor sequencing depth (Supp. Fig. S6). In comparison, other short-read approaches based on LR-PCR enrichment show lower percentage (1.3%) of reported pathogenic variants that would have been missed because of low sequencing depth. However, this approach required very high sequencing depth, which can be appreciated from the observed high variability in coverage ranging from <8x to >30,000x (Supp. Fig. S6).

5 DISCUSSION

Accurate diagnosis is a difficult task when performed in complex genetic regions such as PKD1 (Qi et al., 2013; Rossetti et al., 2007; Tan et al., 2009). To facilitate the diagnosis, we have developed and applied a new methodology using direct long-read sequencing of amplified LR-PCR fragments on PacBio. Because of the repetitive nature of PKD1, current diagnostics is performed by Sanger sequencing using LR-PCR fragments generated for approximately two thirds of the PKD1 gene that serve as a template for the exon-specific nested PCR amplification. In contrast, in this study, we directly sequenced all LR-PCR fragments amplified from the duplicated and unique parts of PKD1 gene as well as PKD2. On top of reducing the PCR amplification steps required and limiting the implicit PCR artifacts, single-molecule sequencing improves sequence alignments and aids in discriminating between homologous or repeated sequences, such as PKD1 pseudogenes. This provides a cleaner dataset for variant calling, free of chimeric (0.85%) and pseudogene (5.6%; Supp. Table S2) reads that are introduced by the LR-PCR amplification. Finally, by using this approach, we identified 20 out of 21 (95.2%) PKD1 disease-causing variants diagnosed by Sanger sequencing or MLPA, providing a correct diagnosis for 18 out of 19 ADPKD patients (94.7%) with at least one pathogenic variant in PKD1.

In contrast to current ADPKD diagnostic assays, based on Sanger sequencing and MLPA, we show that direct long-read sequencing can aid in resolving PKD1 for ADPKD diagnostics. Longer sequencing reads discriminate between PKD1 and pseudogenes (Fig. 2A), and improve the mapping quality of PKD1 (Fig. 2B). The improved mappability reduced the interference of homologous sequences, high GC content, or repetitive elements for ADPKD diagnosis (Qi et al., 2013). This allowed us to develop a long-read-based sequencing assay for detecting a broad spectrum of variants, from SNVs to large deletions (Table 1). In contrast, Sanger sequencing is very labor-intensive and requires many phases of overlapping PCR amplification steps prior to sequencing, including LR-PCR and nested PCR. Despite the amplification of PKD1 being based on unique PCR primers, these are of limited number for PKD1 and have been shown to produce residual amplification of homologous regions that would still interfere with the aggregated signal of Sanger sequencing (Rossetti et al., 2012; Tan et al., 2014). Based on our approach, we confirmed the presence of residual amplification of PKD1 pseudogenes, introduced by the LR-PCR (5.6%) (Fig. 2A; Supp. Table S2). This, most likely, led to the identification of 24 false-positive or false-negative variants detected by Sanger sequencing (Supp. Table S9; Supp. Fig. S4). One of the major drawbacks of our method, however, is the noise associated with PacBio sequencing, and the sophisticated algorithms required to overcome it. This noise is likely to be the cause of most of the 324 homopolymer deletion artifacts that were solely identified by PacBio (Supp. Table S10). In addition, this noise was the most likely cause of the single pathogenic insertion that was missed despite ample sequencing depth. However, based on a recent release of the new circular consensus calling algorithm for PacBio sequencing data (www.pacb.com: “An improved circular consensus algorithm with an application to detect HIV-1 drug-resistance associated with mutations (DRAMS)”), we expect that calling of true homopolymer-associated variants will be significantly improved.

In recent years, several attempts have been made to replace the standard ADPKD diagnostics by NGS approaches that would improve the screening of PKD1 gene (Eisenberger et al., 2015; Mallawaarachchi et al., 2016; Qi et al., 2013; Rossetti et al., 2012; Tan et al., 2014; Trujillano et al., 2014). These screenings were based on
TABLE 1  Uniquely identified pathogenic variants or variants of unknown clinical significance identified by PacBio sequencing

<table>
<thead>
<tr>
<th>Patient</th>
<th>Genomic position</th>
<th>Exon</th>
<th>c. notation</th>
<th>p. notation</th>
<th>SNP ID</th>
<th>Freq (%)</th>
<th>Depth</th>
<th>PolyPhen</th>
<th>VEP impact</th>
<th>Comparison with Sanger sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chr16</td>
<td>25</td>
<td>c.9034_9039del</td>
<td>p.(Thr3012_Ser3013del)</td>
<td></td>
<td>40.9</td>
<td>314</td>
<td>Moderate</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>2</td>
<td>chr16</td>
<td>18</td>
<td>c.7214G&gt;T</td>
<td>p.(Trp2405Leu)</td>
<td></td>
<td>29.4</td>
<td>666</td>
<td>Possibly damaging (0.983)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>3</td>
<td>chr16</td>
<td>15</td>
<td>c.3643C&gt;G</td>
<td>p.(Leu1215Val)</td>
<td>rs144338515</td>
<td>49.2</td>
<td>576</td>
<td>Moderate</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>3</td>
<td>chr16</td>
<td>15</td>
<td>c.4475G&gt;C</td>
<td>p.(Arg1492Pro)</td>
<td></td>
<td>32.8</td>
<td>563</td>
<td>Possibly damaging (0.665)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>3</td>
<td>chr16</td>
<td>16</td>
<td>c.6986G&gt;A</td>
<td>p.(Arg2329Gln)</td>
<td>rs575211353</td>
<td>43.3</td>
<td>538</td>
<td>Benign (0.37)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>3</td>
<td>chr16</td>
<td>26</td>
<td>c.9324del</td>
<td>p.(Ile3109SerfsTer207)</td>
<td>rs780284643</td>
<td>26</td>
<td>302</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>4</td>
<td>chr16</td>
<td>11</td>
<td>c.2681_2690del</td>
<td>p.(Phe894Ter)</td>
<td></td>
<td>53.7</td>
<td>143</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>5</td>
<td>chr16</td>
<td>1</td>
<td>c.182C&gt;T</td>
<td>p.(Pro61Leu)</td>
<td></td>
<td>28.8</td>
<td>43</td>
<td>Benign (0.119)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>5</td>
<td>chr16</td>
<td>26</td>
<td>c.9397 + 1G&gt;A</td>
<td></td>
<td></td>
<td>23.3</td>
<td>330</td>
<td>Moderate</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>6</td>
<td>chr16</td>
<td>6</td>
<td>c.1261C&gt;T</td>
<td>p.(Arg421Cys)</td>
<td></td>
<td>38.7</td>
<td>273</td>
<td>Possibly damaging (0.836)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>6</td>
<td>chr16</td>
<td>21</td>
<td>c.7940C&gt;T</td>
<td>p.(Thr2647Met)</td>
<td>rs748496650</td>
<td>44</td>
<td>357</td>
<td>Probably damaging (1)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>7</td>
<td>chr16</td>
<td>21</td>
<td>c.(2097 + 1_2098-1)_3640del</td>
<td></td>
<td></td>
<td>44.1</td>
<td>583</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>8</td>
<td>chr16</td>
<td>24</td>
<td>c.8859dup</td>
<td>p.(Glu2954Ter)</td>
<td></td>
<td>41.5</td>
<td>466</td>
<td>Moderate</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>9</td>
<td>chr16</td>
<td>15</td>
<td>c.6657_6671del</td>
<td>p.(Arg2220_Pro2224del)</td>
<td></td>
<td>27.7</td>
<td>20</td>
<td>Low</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>10</td>
<td>chr16</td>
<td>40</td>
<td>c.11412-3C&gt;A</td>
<td></td>
<td></td>
<td>32.7</td>
<td>313</td>
<td>Possibly damaging (0.999)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>11</td>
<td>chr16</td>
<td>26</td>
<td>c.8293C&gt;T</td>
<td>p.(Arg2765Cys)</td>
<td>CM092156_rs144979397</td>
<td>41.1</td>
<td>572</td>
<td>Probably damaging (0.988)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>11</td>
<td>chr16</td>
<td>25</td>
<td>c.9187C&gt;T</td>
<td>p.(Arg3063Cys)</td>
<td>rs145906459</td>
<td>36.1</td>
<td>557</td>
<td>Benign (0.39)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>12</td>
<td>chr16</td>
<td>21</td>
<td>c.8017-2_8017-1del</td>
<td></td>
<td></td>
<td>48.5</td>
<td>527</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>13</td>
<td>chr16</td>
<td>25</td>
<td>c.(287 + 1_288-1)_9397 + 1_9398-1del</td>
<td></td>
<td></td>
<td>41.3</td>
<td>46</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>14</td>
<td>chr16</td>
<td>42</td>
<td>c.11554del</td>
<td>p.(Leu3852TrpfsTer93)</td>
<td>rs724159823</td>
<td>41.3</td>
<td>46</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>14</td>
<td>chr16</td>
<td>13</td>
<td>c.3100A&gt;G</td>
<td>p.(Asn1034Asp)</td>
<td>rs369180760</td>
<td>36.5</td>
<td>321</td>
<td>Benign (0.098)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>15</td>
<td>chr16</td>
<td>15</td>
<td>c.6223_6224insTT</td>
<td>p.(Arg2075LeufsTer42)</td>
<td></td>
<td>43.3</td>
<td>1,203</td>
<td>High</td>
<td>Sanger</td>
<td>PacBio</td>
</tr>
<tr>
<td>15</td>
<td>chr16</td>
<td>11</td>
<td>c.2269del</td>
<td>p.(Gln757SerfsTer28)</td>
<td>rs775710328</td>
<td>28.4</td>
<td>519</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>17</td>
<td>chr16</td>
<td>15</td>
<td>c.4968_4969delinsC</td>
<td>p.(Arg1657GlyfsTer65)</td>
<td></td>
<td>41.2</td>
<td>1,120</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>18</td>
<td>chr16</td>
<td>16</td>
<td>c.6994_7000dup</td>
<td>p.(Ala2332GlyfsTer90)</td>
<td></td>
<td>23.4</td>
<td>913</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
<tr>
<td>19</td>
<td>chr16</td>
<td>46</td>
<td>c.12890A&gt;G</td>
<td>p.(Lys4297Arg)</td>
<td>rs758833703</td>
<td>14.1</td>
<td>46</td>
<td>Benign (0.07)</td>
<td>Moderate</td>
<td>Overlap</td>
</tr>
<tr>
<td>19</td>
<td>chr16</td>
<td>15</td>
<td>c.4248dup</td>
<td>p.(Gly1417TrpfsTer14)</td>
<td></td>
<td>24.8</td>
<td>979</td>
<td>High</td>
<td>Overlap</td>
<td>PacBio</td>
</tr>
</tbody>
</table>

Notes: Sanger-detected pathogenic variants are shown in bold. PacBio variants were filtered by coding sequence predicted effects (frameshifts, missense, in-frame deletions, and splicing variants), as well as DP >15 and >50 subreads, and variant frequency (>10% for substitutions, and >15% for insertions and deletions) (RefSeq NM_001009944.2).

Additional information of each variant including SIFT classification, and 1000G frequencies among other annotations can be obtained from the VCF files uploaded to EGA with accession number EGAS00001002106.
This study was performed within the scope of the iMODE-CKD Initial Training Network (ITN) (Clinical and system-omics for the identification of the Molecular Determinants of established Chronic Kidney Disease).

Data obtained from Rossetti et al. (2012) was kindly provided by Peter C. Harris and Christina M. Heyer (Mayo Clinic College of Medicine).

Patient samples were kindly provided by the DIPAK Consortium, an interuniversity collaboration in The Netherlands, established to study ADPKD and to develop rational treatment strategies for this disease (www.nierstichting.nl/dipak). Principal investigators are (in alphabetical order): J. P. H. Drenth (Department of Gastroenterology and Hepatology, Radboud University Medical Center Nijmegen), J. W. de Fijter (Department of Nephrology, Leiden University Medical Center), R. T. Gansevoort (Department of Nephrology, University Medical Center Groningen), D. J. M. Peters (Department of Human Genetics, Leiden University Medical Center), J. Wetzel’s (Department of Nephrology, Radboud University Medical Center Nijmegen), and R. Zietse (Department of Internal Medicine, Erasmus Medical Center Rotterdam).

We acknowledge the continued support of Joost P. Schanstra (Inserm; Université Toulouse III Paul Sabatier, Institut de Métérologie Moléculaire de Rangueil).

**DISCLOSURE STATEMENT**

The authors declare no conflict of interest.

**REFERENCES**


**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.