Functional alterations due to amino acid changes and evolutionary comparative analysis of ARPKD and ADPKD genes

Burhan M. Edrees, Mohammad Athar, Zainularifeen Abduljaleel, Faisal A Al-Allaf, Mohiuddin M. Taher, Wajahatullah Khan, Abdellatif Bouazzaoui, Naffaa Al-Harbi, Ramzia Safar, Howaida Al-Edressi, Khawala Alansary, Abulkareem Anazi, Naji Altyayeb, Muawia A. Ahmed

A targeted customized sequencing of genes implicated in autosomal recessive polycystic kidney disease (ARPKD) phenotype was performed to identify candidate variants using the Ion torrent PGM next-generation sequencing. The results identified four potential pathogenic variants in PKHD1 gene [c.4870C>T, p.(Arg1624Trp), c.5725C>T, p.(Arg1909Trp)], c.1736C>T, p.(Thr579Met) and c.10628T>G, p.(Leu3543Trp)] among 12 out of 18 samples. However, one variant c.4870C>T, p.(Arg1624Trp) was common among eight patients. Some patient samples also showed few variants in autosomal dominant polycystic kidney disease (ADPKD) disease causing genes PKD1 and PKD2 such as c.12433G>A, p.(Val4145Ile) and c.1445T>G, p.(Phe482Cys), respectively. All causative variants were validated by capillary sequencing and confirmed the presence of a novel homozygous variant c.10628T>G, p.(Leu3543Trp) in a male proband. We have recently published the results of these studies (Edrees et al., 2016). Here we report for the first time the effect of the common mutation p.(Arg1624Trp) found in eight samples on the protein structure and function due to the specific amino acid changes of PKHD1 protein using molecular dynamics simulations. The computational approaches provide tool predict the phenotypic effect of variant on the structure and function of the altered protein. The structural analysis with the common mutation p.(Arg1624Trp) in the native and mutant modeled protein were also studied for solvent accessibility, secondary structure and stabilizing residues to find out the stability of the protein between wild type and mutant forms. Furthermore, comparative genomics and evolutionary analyses of variants observed in PKHD1, PKD1, and PKD2 genes were also performed in some mammalian species including human to understand the complexity of genomes among closely related mammalian species. Taken together, the results revealed that the evolutionary comparative analyses and characterization of PKHD1, PKD1, and PKD2 genes among various related and unrelated mammalian species will provide important insights into their evolutionary process and understanding for further disease characterization and management.

© 2016 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Polycystic kidney disease (PKD) is a clinically and genetically heterogeneous disorder with different modes of inheritance. There are two fundamental types of hereditary polycystic kidney disorders, autosomal dominant (ADPKD) and autosomal recessive (ARPKD). The ADPKD is the most frequent and life-threatening genetic disease with a prevalence of one in 500–1000, affecting >12 million individuals worldwide. The clinical symptoms usually not appear until adulthood; however, about 2%–5% of patients with ADPKD show early clinical manifestations,
which are often indistinguishable from the ARPKD. The ADPKD has three types viz., PKD1 gene on chromosome 16, PKD2 gene on chromosome 4 and ADPKD 3 gene on an unknown chromosomal site, whereas, the ARPKD gene is present on chromosome 6.

The ARPKD is considerably rare compared to its dominant counterpart ADPKD with incidence rate of about one in 20,000 live births [1]. It could be diagnosed in utero or prenatally by sonography displaying bilateral large echogenic kidneys and oligohydranmios in severe cases. It has been reported that mutations in the polycystic kidney and hepatic disease 1 (PKHD1) gene are responsible for ARPKD and its severity depends on the type of mutations [2].

The ADPKD is usually diagnosed through renal imaging using age-related cyst quantity criteria; yet it is not effective to exclude disease in vulnerable individuals under 40 years of age [3]. The mutation evaluation of the ADPKD is hampered due to the large size multi-exons of PKD1 and PKD2, occurrence of genomic duplication in PKD1, marked allelic heterogeneity, and presence of common missense variations with hypomorphic alleles [4]. About two-thirds of the PKD1 gene (exons 1–33) is duplicated six times on chromosome 16 (pseudogenes PKD1P1–P6) [5]. The PKD1 and PKD2 variation studies may provide prognostic and diagnostic insights including pre-implantation heredity diagnostics for early-onset of ADPKD (Fig. 1A–B) [5].

The PKHD1 gene is located on the short arm of chromosome 6 (chr6p21) and encodes for fibrocystin protein, which is present on the primary cilium of the renal epithelial cells. There is a high risk of fetal manifestation and neonatal death if the fetus has two truncating mutations [6]. The PKHD1 gene is approximately 470 kb long encoding various transcripts with sizes ranging from 9 to16 kb. The largest uninterrupted open reading frame translates a denominated fibrocystin protein [7] or polyductin [7] of 447 kD having 4074 amino residues. The fibrocystin/polyductin protein has a single transmembrane (TM)-spanning domain near carboxyl end having six to seven TIG/IPT (immunoglobin-like fold shared by plexins and transcription factors) domains. Such structural motif with unidentified function has also been recognized in a number of transmembrane receptors. Various transcripts translate into truncated products lacking the TM and may be secreted if translated [8]. The fibrocystin is believed to be a transmembrane receptor/ligand that has a role in collecting duct and biliary differentiation. Remarkably, in cultured renal epithelial cells the primary location of expression was found to be the primary cilium specifically the basal body of the cilium [9–10].

Characteristically, neonates have fusiform dilations of the collecting ducts and have portal and interlobular fibrosis of the liver along with the biliary duct hyperplasia. Furthermore, respiratory failure, hypertension and urinary tract infections are common occurrence. The disease typically clinically manifest either neonatal or in early childhood (Fig. 1C), and about 30% of affected children die within the first year of life [11]. However, appearance of ARPKD at later ages and survival into adulthood has been noticed in a considerable number of cases [12]. In these cases, ARPKD can be clinically indistinguishable from autosomal dominant polycystic kidney disease (ADPKD) [13]. The ARPKD gene, denominated PKHD1 for polycystic kidney and hepatic disease, is situated on the short arm of chromosome 6 [14]. The disease is clinically heterogeneous nonetheless so far all families have been linked to this locus, including patients with severe or with relatively mild progression of the disease [15–16]. Nevertheless, congenital polycystic kidney disease can be part of a number of syndromes, such as Meckel–Gruber Syndrome, and Zellweger Syndrome [17].

Recently, we have validated the disease-causing variants due to amino acid substitution [18] among ARPKD (PKHD1) and ADPKD (PKD1 and PKD2) genes based on the exons and flanking regions study in Saudi patients using Ion-PGM sequencing. This study show that based on the computational predictions an amino acid substitution may eventually affect the protein function due to the structural alteration causing stable changes in the involved domain. Results of this study show that the ARPKD may be also caused as a result of partial or complete loss of polyductin/fibrocystin function. These methodological strategies are useful tools before gene based and correlative analyses to expand the understanding of ARPKD variant spectrum among Saudi populations. The study also expounded on the challenges related to interpretation of the pathogenicity of harmful variants in this large and complex gene. Hence, a detailed study was conducted to investigate the nucleotide sequences and their molecular organization of the three genes among various mammalian species to understand the evolution process of these genes.

2. Materials and methods

2.1. Ion torrent PGM next-generation sequencing (NGS)

Library construction, target enrichment, template preparation and sequencing were performed as described in Edrees et al. [18]. In brief, the DNA libraries and enrichment of targeted sequences of the genes (PKHD1, PKD1 and PKD2) were achieved using Ion Plus Fragment Library kit, Ion Xpresso bar-code adapters 1–16 kit and TargetSeq customized panel (Life Technologies). Template preparation and sequencing were done using the Ion PGM Template Kit v2.0 and Ion PGM sequencing 200-kit v2 (Life Technologies) respectively. NGS data analysis was performed using CLC Genomics Workbench v9, USA (http://www.clcbio.com). The BAM binary format sequence data raw reads went through adapter trimming, hence removal of reads shorter than 20 bp and removal of exact duplicates, as well as quality trimming were performed. The pre-processed reads were aligned with the reference genome (hg19) sequences corresponding to the customized genes. This was followed by SNV and indel detection and all variants detected within the exons of the customized genes were considered for subsequent analyses using probabilistic variant detection method.

2.2. Multiple sequences alignment in phylogenetic analysis

DNA sequences for PKD1 and PKD2 and most of the PKHD1 proteins analyzed were obtained by blastp or tblastn searches from the public database NCBI (National Center for Biotechnology Information http://blast.ncbi.nlm.nih.gov/Blast.cgi). The Protein sequence alignments were performed using ClustalW (http://www.clustal.org/clustal2) and
2.3. Solvent accessibility of amino acid residues

Amino acid solvent accessibility (ASA) was used to suggest solvent accessibility of amino acid residues in proteins structures. The protein structure was retrieved from the Define Secondary Structure of Proteins (DSSP) database of secondary structure assignments [19–20]. The DSSP software characterizes secondary structure, geometrical appearance and solvent structures of proteins in Protein Data Bank. The software DSSP estimates entries from PDB (protein data bank) through the entry implementation of ASA view for all the protein or individual chain. The standard used was to add chain or PDB query in the input file, exhibit ASA view plot for co-ordinate or PDB file. The amino acid residues verified three forms of solvent accessibility i.e., buried, partially buried and exposed indicating i.e., low, moderate and high accessibility respectively [21]. The Secondary structure was used for studying the association between amino acid and protein structure.

2.4. Protein stability changes

For many mutants (single amino acid changes or nsSNPs) in humans and their impact on protein function remains unknown. Using computational approach provides binary classifications (impact/neutral) along with a more rational score. Additionally, we get information about the protein’s stability through Schrodinger (BioLuminate, USA). In the mutated structure, the mutant residue was targeted in the native protein structure. Many known disease-related nsSNPs in proteins with identified 3D protein structure will have an effect on the structurally significant residues and sites significant for the protein function. The disease causing mutations typically occur in the protein residues that are buried and at the hydrogen bonding residues [22]. In protein kinases, comparisons show a cluster inside the functionally fundamental catalytic core after residue scanning to repair the polar and neutral residues in the protein stability analysis and solvent accessibility. In addition, it was also compared with the predictions of the functional effect that were determined by the Screening for Non-Applicable Polymorphisms (SNAP) [23]. The SNAP scores range from −100 (strongly predicted as neutral) to 100 (strongly predicted to change function); the gap is instantly related to the binary determination boundary (0), which measures the reliability of the effect [24]. To demonstrate, disease-related mutations might possibly have an impact on protein interactions [25]. The protein function is generally associated with the evolutionarily conserved residues [26]. A damaging signal corresponds to a mutation that is predicted to be stabilizing. The changes within the folding free energy upon mutation (ΔΔG) support the idea that a mutation may cause an alteration in the protein structure and function eventually leading to disease. The SNAP score is related to additional functional effects [27].

2.5. Molecular dynamics simulation

The structure of PKHD1 protein modeling was generated by iTasser program [28]. The molecular dynamics (MD) simulation programs used were CHARMM + + [29] and Gromacs [30]. The simulation to obtain the solvated process was neither minimized nor equilibrated, nevertheless 0.15 M ions were added within the simulation box by specifying ions (KCl) and concentration (C). The ions had been spontaneously established by means of the ion-accessible volume (V), whole charge of the conduct (Qsys), and by means of the confident ion (z+) valency to neutralize the whole system charge, (z + N + −N − = − Qsys). The ion-available volume (V) used to be anticipated with promote of subtracting molecular volume from the entire system and determining the ion-placing method of Monte Carlo (MC). The solvation free energy was once expressed as nonpolar and electrostatic contributions, but the nonpolar contribution was again partitioned into repulsive and dispersive contributions utilizing the weeks. The preface configuration of ions confirmed the short MC simulations with a primitive model equivalent to Van der Waals (VdW) interactions. The free energy simulations had been carried out with few specific solvent water molecules in close proximity to the solute, even as the influence of the intermersion of the solvent mass used to be proven implicitly as an effective solvent boundary potential (SSBP). The KCl used to be included in the box to neutralize the overall poor charge of the PKHD1 protein model. MD simulations had been conduct out with a 2 fs time step at a consistent temperature of 300 K and a constant pressure of 1 atm below periodic solvent boundary stipulations. The Particle Mesh Ewald (PME) approach was utilized for electrostatics, and a 12 Å cutoff was once utilized for VdW interactions. The TIP3P water model was used to model the solvent [31].

3. Results

3.1. High-throughput sequencing, mapping and coverage

To discover DNA variants of disease causing genes involved in genetic disorders, we designed a unique high throughput density capture array 538,751 targeted coverage exons sequences of 20 genes including three genes (PKHD1, PKD1 and PKD2) causing polycystic kidney disease. A total of 18 patient’s samples were analyzed. The functional variant detection was performed using targeted customized genes, where a total of 543,366 (P16)–2,444,050 (P10) high-quality map reads were obtained, encompassing 2,009,135 (P6)–2,444,050 (P10) high-quality bases per patient. Following mapping with the reference human genome, about 90% yielded clean reads uniquely coordinated to the target regions and about 98% of the targeted region covered with at least 95% folds mean depth coverage for each sample. The average depth coverage for exons among the 12 patient samples was 98% with highest depth coverage of 99%, adequate to reliably detect DNA variants within the majority of the targeted regions.

3.2. Functional regions are conserved in different species

The sequence similarity based on conserved regions in a region provides valuable information, therefore the human PKHD1, PKD1, PKD2 genes conserved regions with significant orthologous among other species using NCBI-BLAST were examined. We found that in 71 out of 76 (93%) showed conserved regions and 167 out of 270 (61%) showed human query sequences that were experimentally verified to be functional. It is most likely that all conserved regions matching the domain in a particular region, therefore NCBI-BLAST may exclude many other domains by considering conserved segments individually. The identified conserved domains in a region indicated new or same function of the conserved regions in three genes targeted. The SMART BLAST was used to identify some homologous regions for these human functional regions. Discontinuous SMART BLAST has been able to identify more divergent sequence similarity than NCBI-BLAST. Using a recommended parameter combination (-A 50 -t 21 -W 11 -N 1), we identified hits in 80 out of 156 clustered conserved domain regions. Among them, only 76 of clustered conserved region domains have NCBI-BLAST hits with an E-value less than 1E-10. In all these regions, our method identified all NCBI-BLAST hits as conserved segments. Among 71 out of the 76
(93%) clustered regions and in 80 out of the 156 (30%) conserved regions, the results identified more conserved segments than Smart Blast.

3.3. Phylogenetic orthologs of genomic conserved regions

The single nucleotide variation (SNV) position affected targets in three genes i.e., *PKHD1*, *PKD1*, and *PKD2* were studied and used for phylogenetic analysis. The results obtained by using the two different methods ([CLCbio and PhyML (http://atgc.lirmm.fr/phyml)]) were almost identical, in addition to their corresponding internal bootstrap support, demonstrating the reliability of the phylogenetic analysis. The phylogenetic tree showed *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), and Gorilla within one group and showed a clustering topology in general well supported with high bootstrap values. The percentage of coding nucleotides in the *PKHD1* gene varied across taxa with the highest detected in *Chlorocebus sabaeus* (green monkey), *Macaca mulatta* (rhesus macaque), *Macaca nemestrina* (pig tailed macaque), and *Mandrillus leucophaeus* (drill) (100%) in one taxon and lowest in *Pan paniscus* (bonobo) (79%). This validated the fact that the coding region of *PKHD1* gene in variants (c.4870C→T, p.(Arg1624Trp), c.5725C→T, p.(Arg1909Trp), c.1736C→T, p.(Thr579Met) and c.10628T→G, p.(Leu3543Trp) position were absolutely conserved (Fig. 2A and B). To view and understand the phylogenetic inter-relationship among all the 14 mammalian species at genetic level, an un-rooted Neighbor-joining (NJ) tree was constructed. As expected, all mammal species fell in one group and rodents in another. *Homo sapiens* (human) and *Pan troglodytes* (chimpanzee) were closer (100%) to each other. Moreover, *Camelus dromedarius* (arabian camel), and *Bos taurus* (cattle), belonged to one clade, and the positions of donor-related conserved amino acids in the human *PKHD1* are illustrated in the red. For each positions if conserved at N99%, are shown with a different color residue, and the highly conserved related amino acid positions are shown within each domain as clear highlight letters on a solid background. Unrooted phylogenetic tree was included all known *PKHD1* protein from various mammals. B). The tree is based on a ClustalW using Neighbor joining algorithm alignment of the amino-acid sequences. The distance scale and bootstrapping values are shown at each branch point (100 replicates) among the orthologs of *PKHD1*.

![Fig. 2. Phylogenetic analysis of PKHD1 ClustalW](image)
long-tailed chinchilla) slightly diverging from these two taxa. Unexpectedly, Canis lupus familiaris (dog) and Felis catus (cat) belonging to the super order placental mammals, to which Bos taurus (cattle) and Camelus dromedarius (arabian camel) also belong fell separately to clade superorder and placental mammals super order organisms. Being the out-groups, Chinchilla laniger (long-tailed chinchilla) and Rhinopithecus roxellana (golden snub nosed monkey) were in separate groups. Interestingly, Homo sapiens (human) and Pan troglodytes (chimpanzee) were in the same group and matched for all the studied genes however, PKD1 gene did not match with any regions and aligned with the Homo sapiens (human). Even though the original Homo sapiens (human) gene nucleotide as well as protein sequences matched with all the mammalian species that were included for the phylogeny tree construction, it was apparent that all altered variant forms of three genes PKHD1, PKD1 and PKD2 in Saimiri boliviensis (black capped squirrel monkey) and Callithrix jacchus (common marmoset) formed a separate clade, but in reality all three genes and protein in Homo sapiens (human) showed similarity with Chlorocebus sabaues (green monkey), Macaca mulatta (rhesus macaque), Macaca nemestrina (pig-tailed macaque), and Mandrillus leucophaeus (drill). This result also signifies that Homo sapiens and Pan troglodytes showed similarity to each other for PKD1 and PKD2 genes. It was already known that functionally both these mammalian species are closer to each other than to any other Hominoidea, hence sequence similarities in both these taxa at the gene and protein level is no surprise. It has already been reported that the PKD1 and PKHD1 of Placental mammalian organisms exhibits a propensity for recombinational exchange with the closely linked genes. Hence, apart from characterization at the gene level, we also have taken into consideration the protein sequences in all the 12 mammalian species including Equus caballus (alpaca) and Equus caballus (horse) to ascertain the functional level changes across the species. Results of the PKD1 protein alignment p.(Phe482Cys) in all 12 species were found to be remarkably similar with Homo sapiens (human) and Pan troglodytes (chimpanzee) (Fig. 3A and B). The Papio anubis (olive baboon) sequences were found to be similar to only one substitution at codon position 482, signifying that plenty of conservation exists at the PKD1 protein level in closely related mammalian species, unlike at the genetic level, we did not find “Phe” or the “Cys” amino acid position in Homo sapiens (human) PKD1 protein sequence. Instead, Phe was observed at the different codon position. Similar to human PKD1 protein, Phe was also detected at the same position for Pan troglodytes (chimpanzee, Papio anubis (olive baboon), and Macaca fascicularis (crab eating macaque). Absolute protein homology was observed with the rodent species. However, interestingly chimpanzee showed missing codons in PKD2 (Fig. 4A and B). To further assess the evolutionary relationship among all the 14 mammalian species, an Unrooted phylogenetic tree was also constructed based on the protein sequences and inferences were drawn.

3.4. Protein stability changes

To realize the thermodynamic protein stability alterations as a consequence of the p.(Arg1624Trp) common mutation in protein, PopMuSic-2.0 [32] and Schrodinger-BioLuminate (http://www.Schrodinger.Com/BioLuminate/) was utilized based on the statistical potentials of narrow sequence coefficients to check the solvent accessibility model after the mutation. This algorithm confirmed that the amino acid change from “Arg” to “Trp” at position 1624 would affect the protein function due to the excessive free energy ($\Delta\Delta G = 0.64$).
3.5. Molecular dynamics and simulation of common variant site may alter the proper function of a protein when it is un-stabilize.

The structural weak the Acc for mutant p.(Arg1624) was 39.01 (31%) compared to 30.82 (24%) accessibility (Acc) modeling alter the protein structure, indicating that larly in the region of mutation and caused stability changes. The solvent

**Based on large scale benchmarking analysis, we have observed that a**

**binding site** and predicted binding site in our targeted protein structure. 

**measure of local similarity (sequence & structure) between template**

**protein PKHD1 protein structure and mutant type protein structure**

**based on the protein model RMSD. The complete energy of native structure**

**of PKHD1 and mutant model structure residue of common mutation** (Arg1624Trp) have been calculated and showed 0.25 to 0.50 RMSD variation between the wild and mutant protein structures (Supplementary Fig.1). The entire energy for the native PKHD1 protein structure after energy minimization was −244,080.5 kJ/mol with a 0.52 score, compared to before energy minimization structure which was 207,844,375.1 kJ/mol with −3.14 score. The CHARMM graphical interface (GUI) for MD simulations revealed the consequence of the p.(Arg1624Trp) mutation under specific solvent conditions, by performing molecular dynamics and stability investigations for the predicted PKHD1 protein and compared it with the native structure. The solvate was used to create a realistic aqueous solvent environment around the protein model PKHD1 with water. The solvate determines the dimension of process with octahedral shapes of water box fitting to fully solvate the molecule with side distance 10.0.

**3.5. Molecular dynamics and simulation of common variant**

The program of iTasser PKHD1 protein structure modeling and predicted binding site amino acid residues were 144, 174, 175, ref; common mutation p.(Arg1624) was highly binding with peptide reference based on this protein PDB ID: 1TDQ. The CscoreLB is the con-

**Polycystic kidney disease (PKD) is a common genetic disease charac-

terized by the accumulation of multiple fluid-filled cysts in each kidney and other organs. The renal cysts from the renal tubular epithelial cells lined by way of a single layer of cells that have better rates of cellular proliferation and apoptosis, and are much less differentiated than the normal tubular cells. Progression of cysts in the kidneys ultimately**
causes end-stage renal disease (ESRD), which makes PKD one of the leading causes of ESRD in children and adults [33]. Pooling and bar coding process to reap price-effectiveness so that NGS would be utilized to analyze large ADPKD and ARPKD genes. The findings of this study indicate that the ARPKD may be also caused as a result of partial or complete loss of polyductin/fibrocystin function.

Although still very high, the cost of NGS is gradually decreasing due to novel strategies for library preparation, decreased hardware cost, and greater sequence output per run [34]. Ion-torrent probably makes NGS an attractive approach additionally for clinical and diagnostic application, chiefly when simplest particular genes or specific sets of genes through to be analyzed, such as we investigated for ADPKD and ARPKD genes [18].

In an effort to maximize the accuracy of our pathogenicity estimates for the missense variants, we combined various computational methods to predict its effects on the protein stability due to disease-causing missense mutations as it has been reported that the function of a protein can be affected in a variety of ways [35]. Among them, the most common effect is changing protein stability, i.e. destabilizing or stabilizing the wild type protein fold [36–37]. The validation of the identified variants in our previous study [18] due to amino acid substitution and its impact of the protein structure and function due to the stability changes were studied using computational simulation techniques. The six identified variants displaying clinical significance were studied further using computational methods and showed that the predicted protein with mutant variant structure will have influence on the protein function. However, the predictions about the changes of the folding energy not only indicate whether they favor the stability or not, but the predicted absolute magnitude should be accurate as well to allow to distinguish between disease-causing and harmless mutations. Because of this significant effort were devoted to develop methods and approaches to evaluate the stability changes upon amino acid substitutions, but despite of the efforts, accurate calculations of folding free energy are still a challenge [38].

The study also investigated the integrative and cross-species comparative genomics analysis among the APRKD and ADPKD genes. The results showed that the APRKD and ADPKD genes were evolutionarily conserved across different mammalian species apart from the genomic alterations. It has been shown that the human genes PKHD1, PKD1 and PKD2 are highly conserved in the mouse and suggested that the intricate splicing pattern is likely to be functionally important [39]. Furthermore, the larger the evolutionary distance between the aligned different species, the lower the risk of over predicting pathogenicity. The study was based on the alignment ranging from tolerance to variation for the amino acid position in question by aligning the homologues of the protein from various species, and compares the tolerance to the magnitude of the chemical alteration caused as a result of particular variant. The three genes PKHD1, PKD1 and PKD2 alignment of protein and amino acid sequences when compared with 14 mammalian species including *Pan troglodytes* (chimpanzee). The observation that the human and mouse genes shared complex patterns of splicing prompted us to examine whether specific features were conserved. We compared the specific patterns of splicing, the relative position and sequence of exons that occasionally used alternative splice sites, and the relative position and sequence of unique exons that were not part of the 67 exon longest ORF transcript. These three genes are thought to give rise, through alternative splicing, to thousands of isoforms. Interestingly, the pattern of alternative exon and splice-site usage is very similar to what we have observed for human and mouse Pkd1. In the case of the three exons, alternative splicing has been shown to result in products with different ligand binding properties [40–41].

The MD simulation under solvated conditions were also integrated to the predicted protein domain structure based on the amino acid sequences of *PKHD1* gene. The predicted protein domain macromolecular structure was constructed using Charm++ and Gromacs computational programs and was studied for the stability and structural consequences compared to the mutant structure. The computational analysis is not aimed at predicting the absolute value of the expected energy changes, but rather to predict whether it is stabilizing or destabilizing and more importantly to reveal the details of the suggested changes. The foremost advantage of the structure-based approaches is that they could illustrate the details of the changes causing the malfunction in a corresponding protein and in principle these findings could be used to develop therapeutics to neutralize the ill effects as result of such alterations. In the future, it may be feasible to analyze naturally occurring variations in the genome sequence, predict the biophysical effect of the mutation, and estimate the likelihood of a given variation to be tolerating or damaging. Such studies could also suggest further research strategies into protein function, and possibly, mutation-specific therapies. Although, many of these proteins are experimentally intractable, the strategy presented here could make not only mutation studies in several human proteins possible for further analyzing them but also for designing efficient disease management strategies.

5. Conclusion

The study demonstrated on the challenges related to interpretation of the pathogenicity of harmful variants in large and complex genes such as causing ARPKD. Hence, a detailed study was conducted to investigate the nucleotide sequences and their molecular organization of the three genes among various mammalian species to understand the evolution process of these genes. The study validated the disease-causing common variant p.(Arg1624Trp), which was highly associated with autosomal recessive polycystic kidney disease (ARPKD) using Ion-PGM sequencing. Overall, the results were discussed with an inference on the role of evolutionary forces in maintaining such close comparisons and variations across closely related taxa. Further studies are needed to utilize more comparative approaches to understand the disease causing genes evolution in closely related mammalian species. These methodological strategies are useful tools before gene based and correlative analyses to expand the understanding of ARPKD variant spectrum among Saudi populations.

Supplementary data to this article can be found online at [http://dx.doi.org/10.1016/j.gdata.2016.10.009](http://dx.doi.org/10.1016/j.gdata.2016.10.009).

**Author contributions**

Conceived and designed the experiments: B.E, M.A, F.A, Performed the experiments: M.A, Z.A, M.T, Analyzed the data: W.K, A.B, N.H; contributed reagents, materials and analysis tools: R.S, H.E, KA, AA, A.M, writing of the manuscript: M.A, Z.A., W.K.

**Financial disclosure**

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Conflict of interest statement**

All other authors have no conflicts of interest to declare.

**Acknowledgment**

We are indebted to the subjects under this study and family members for their cooperation. The authors would like to thank the Science and Technology Unit at Umm Al-Qura University for the continuous support of the research grant. This work was supported by a research grant (No.10-BIO1250-10) from National Science, Technology and Innovation Plan (NSTIP) of Saudi Arabia.
References

[1] C. Bergmann, Autosomal recessive polycystic kidney disease. in: T. Kenny, P. Beales (Eds.), Ciliopathies: A Reference for Clinicians, Oxford University Press, Oxford, 2014, pp. 194–217.

[2] E. Denamur, A.L. Delezoide, C. Alberti, A. Bourillon, M.C. Gubler, R. Bouvier, O. Pascaud, J. Eliot, B. Grandchamp, L. Michel-Caemard, et al., Genotype-phenotype correlations in fetuses and neonates with autosomal recessive polycystic kidney disease. Kidney Int. 77 (4) (2010) 350–358.

[3] Y. Pei, J. O'haji, A. Dupuis, A.D. Paterson, R. Magistrini, E. Dicks, P. Parfrey, B. Cramer, E. Coto, R. Torra, et al., Unified criteria for ultrasonographic diagnosis of ADPKD. J. Am. Soc. Nephrol. 20 (1) (2009) 205–212.

[4] P.C. Harris, S. Rossetti, Molecular diagnostics for autosomal dominant polycystic kidney disease. Nat. Rev. Nephrol. 6 (4) (2010) 197–206.

[5] O. Symmons, A. Váradi, T. Arányi, How segmental duplications shape our genome: recent evolution of ARCG6 and PKD1 Mendelian disease genes. Mol. Biol. Evol. 25 (12) (2008) 2601–2613.

[6] C. Bergmann, J. Sendeder, E. Windelen, F. Küpper, I. Middendorf, F. Schneider, C. Dornia, S. Rudnik-Schöneborn, M. Konrad, C.P. Schmitt, et al., Clinical consequences of PKHD1 mutations in 164 patients with autosomal-recessive polycystic kidney disease (ARPKD). Kidney Int. 67 (3) (2005) 829–848.

[7] C.J. Ward, M.C. Hogan, S. Rossetti, D. Walker, T. Sneddon, X. Wang, V. Kubly, J.M. Cunningham, R. Bacallao, M. Ishihashi, D.S. Milliner, V.E. Torres, P.C. Harris, The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. Nat. Genet. 30 (2002) 259–269.

[8] L.F. Onuchic, L. Furu, Y. Nagaawa, X. Hou, T. Eggermann, Z. Ren, C. Bergmann, J. Sendeder, E. Esquivel, R. Zeltner, et al., PHHD1, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple immunoglobulin-like plexin-transcription-factor domains and parallel beta-helix 1 repeats. Am. J. Hum. Genet. 70 (5) (2002) 1305–1317.

[9] S. Wang, Y. Luo, P.D. Wilson, G.B. Witman, J. Zhou, The autosomal recessive polycystic kidney disease protein is localized to primary cilia, with concentration in the basal body area. J. Am. Soc. Nephrol. 15 (3) (2004) 592–602.

[10] M.Z. Zhang, W. Mai, C. Li, S.Y. Cho, C. Hao, G. Moeckel, R. Zhao, I. Kim, J. Wang, H. L.M. Guay-Woodford, R.A. Desmond, Autosomal recessive polycystic kidney disease. Kidney Int. 77 (4) (2010) 350–358.

[11] S. Wang, Y. Luo, P.D. Wilson, G.B. Witman, J. Zhou, The autosomal recessive polycystic kidney disease protein is localized to primary cilia, with concentration in the basal body area. J. Am. Soc. Nephrol. 15 (3) (2004) 592–602.

[12] B.M. Edrees et al., / Genomics Data 10 (2016) 127–134.