

## DATABASES

# PKDB: Polycystic Kidney Disease Mutation Database—A Gene Variant Database for Autosomal Dominant Polycystic Kidney Disease

Alexander M. Gout,<sup>1,2</sup> Neilson C. Martin,<sup>3</sup> Alastair F. Brown,<sup>4</sup> and David Ravine<sup>2\*</sup>

<sup>1</sup>The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia; <sup>2</sup>Western Australian Institute for Medical Research, School of Medicine and Pharmacology, The University of Western Australia, Crawley, Western Australia, Australia; <sup>3</sup>School of Psychology, Division of Health Sciences, Curtin University of Technology, Bentley, Western Australia, Australia; <sup>4</sup>Medical Research Council Human Genetics Unit, Edinburgh, Scotland, United Kingdom

Communicated by A. Jamie Cuticchia

Autosomal dominant polycystic kidney disease (ADPKD) arises from mutations in the *PKD1* and *PKD2* genes. The Polycystic Kidney Disease Mutation Database (PKDB) is an internet-accessible relational database containing comprehensive information about germline and somatic disease-causing variants within these two genes, as well as polymorphisms and variants of indeterminate pathogenicity. The PKDB database structure incorporates an interface between these gene variant data and any associated patient clinical data. An initiative of the Polycystic Kidney Disease Foundation, PKDB is a publicly accessible database that aims to streamline the evaluation of *PKD1* and *PKD2* gene variants detected in samples from those with ADPKD, as well as to assist ongoing clinical and molecular research in the field. As the accurate reporting of nucleotide variants is essential for ensuring the quality of data within PKDB, a mutation checker has been mounted on the PKDB server allowing contributors to assess the accuracy of their *PKD1* and *PKD2* variant reports. Researchers and clinicians may submit their *PKD1/PKD2* gene variants and any associated deidentified clinical data via standardized downloadable data entry forms accessible through the PKDB site. PKDB has been launched with the full details of *PKD1* and *PKD2* gene variant reports published in 73 peer-reviewed articles. Through a series of user-friendly advanced search facilities, users are able to query the database as required. The PKDB server is accessible at <http://pkdb.mayo.edu>. Hum Mutat 0, 1–6, 2007. © 2007 Wiley-Liss, Inc.

KEY WORDS: *PKD1*; polycystin-1; *PKD2*; polycystin-2; germline gene variants; somatic gene variants; locus-specific database; clinical data

## INTRODUCTION

### Polycystic Kidney Disease

Autosomal dominant polycystic kidney disease (ADPKD) is the most common of the inherited nephropathies with an estimated occurrence of 1 in 1,000. It is characterized by multiple fluid-filled cysts in the kidneys and other organs. ADPKD-causing mutations occur in two genes, with approximately 85% of cases arising from mutations within the *PKD1* (MIM# 601313) gene [Peters and Sandkuijl, 1992] and the remainder due to *PKD2* (MIM# 173910) gene mutations. A small number of reports of families apparently not linked to the chromosomal locations of either gene raise the possibility of ADPKD-causing mutations in other genes. However, to date, the *PKD3* gene remains unmapped. Should another ADPKD-related gene be found, it is expected that it would account for only a small minority of cases.

*PKD1* is a 46-exon gene located at 16p13.3 adjacent to the tuberous sclerosis *TSC2* gene. It encodes a large RNA transcript (> 14 kb) from which a 4,302–amino acid protein, polycystin-1 is translated [European Polycystic Kidney Disease Consortium, 1994]. *PKD1* genomic analysis is complicated by the presence of highly homologous genomic duplications, which contain approximately six *PKD1*-like pseudogenes that have high homology with

the first two thirds of the *PKD1* gene (exons 1 to 33) [Rossetti et al., 2002]. Care thus needs to be taken to avoid simultaneous amplification of these homologous duplications. By contrast, *PKD2* located at 4q22.1, identified 2 years after *PKD1* [Mochizuki et al., 1996], is a smaller 15-exon gene encoded by a single-copy genomic sequence on chromosome 4 that generates a 5-kb transcript from which a 968–amino acid protein, polycystin-2 is translated.

A large number of ADPKD-causing variants, polymorphisms and variants of uncertain clinical significance have now been reported in both genes. Disease-causing variants disrupt protein

The Supplementary Material referred to in this article can be accessed at <http://www.interscience.wiley.com/jpages/1059-7794/suppmat>.

Received 13 July 2006; accepted revised manuscript 1 December 2006.

Grant sponsor: Polycystic Kidney Disease Foundation.

\*Correspondence to: David Ravine, Western Australian Institute for Medical Research, School of Medicine and Pharmacology, MBDP: M570, The University of Western Australia, Crawley, Western Australia 6009, Australia. E-mail: david.ravine@uwa.edu.au

DOI 10.1002/humu.20474

Published online in Wiley InterScience (www.interscience.wiley.com).

function predominately by premature truncation arising either from nonsense mutations or frameshifting deletions, duplications, and other more complex genomic rearrangements [Rossetti et al., 2002; Deltas, 2001]. Importantly, a considerable number of reported variants incur missense mutations at the amino acid level, which are difficult to assess for disease causality. Furthermore, locus and allelic genetic heterogeneity of ADPKD together with the ongoing reporting of novel nucleotide variants creates significant challenges for those seeking to identify *PKD1* and *PKD2* variants, and also for those evaluating the functional and therefore the clinical significance of identified gene variants. Substantial ethnogeographic variation in the population frequency of polymorphic variants [Phakdeekitcharoen et al., 2000; Rossetti et al., 2002] is also a particular challenge for those evaluating the clinical significance of gene variants, particularly as the existing approach used to assess likely pathogenicity includes a weighting based on the rarity of the nucleotide variant in question [Cotton and Scriver, 1998].

### Purpose of PKDB

PKDB is a centralized, curated, online repository of published and unpublished information about *PKD1* and *PKD2* genomic variants. Like the PKHD1 database for autosomal recessive polycystic kidney disease ([www.humgen.rwth-aachen.de](http://www.humgen.rwth-aachen.de)) and many other locus-specific databases that act as publicly accessible storage sites of information about variants in other genes, PKDB aims to assist both researchers and clinicians seeking information about the likely clinical significance of variants found within these genes. It should be noted that the Human Gene Mutation Database (HGMD) [Krawczak and Cooper, 1997; Krawczak et al., 2000; Stenson et al., 2003] includes details of published *PKD1* and *PKD2* gene variants in its collection. However, HGMD involves no record of recurrent variants, nor details of *PKD1* and *PKD2* gene variants considered to be clinically benign. To this end, we have ensured that all *PKD1* and *PKD2* gene variants within the publications listed in Supplementary Table S1 (available online at <http://www.interscience.wiley.com/jpages/1059-7794/suppmat>) have been entered into PKDB, irrespective of whether they have been categorized as likely pathogenic or clinically benign. In keeping with this theme, future contributing researchers and clinicians are encouraged to submit all *PKD1* and *PKD2* variants identified during their detection activities, irrespective of their likely clinical significance status. Prospective submission of all identified gene variants to the database will progressively generate a comprehensive picture of the range and frequency of *PKD1* and *PKD2* gene variants detected on both ADPKD- and wild-type-linked alleles. As well as becoming a useful tool for epidemiological research, it is anticipated that PKDB will mature into a valuable resource for the clinical community, especially as an aid for those evaluating the clinical significance of rarer *PKD1* and *PKD2* gene variants.

### SOFTWARE DESIGN AND IMPLEMENTATION PKDB Server and Database

The database server and custom-designed web interface have been established on a Debian Linux 2.4/AMD XP 32-bit environment. The server has several interacting components, as shown in Figure 1. The dynamic pages created as search results are served by a secure server, which uses a secure sockets layer to encrypt and verify data transfers between the server and users. The database is implemented as a MySQL relational database. Each search request through the web interface search page generates a custom relational command using Structured Query Language

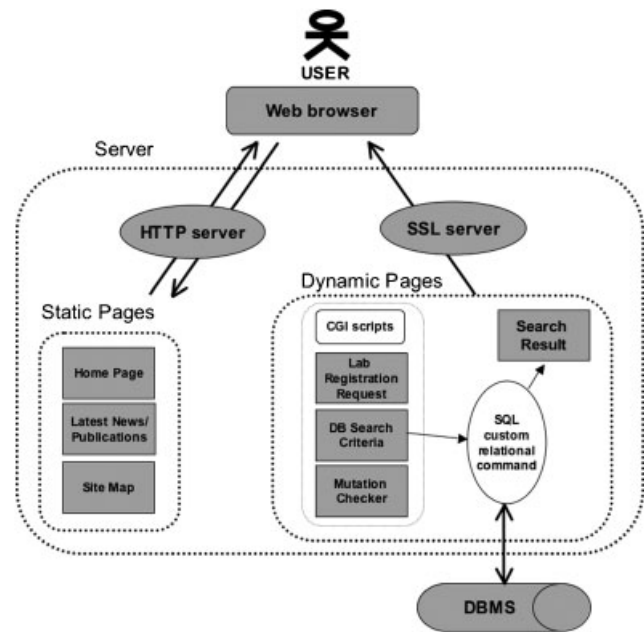


FIGURE 1. Component organization of the PKDB server. Abbreviations used are: HTTP (Hypertext Transfer Protocol), SSL (Secure Sockets Layer), CGI (Common Gateway Interface), SQL (Structured Query Language), DB (Database), and DBMS (Database Management System).

(SQL) to interrogate the database. Commands to the database management system are implemented by common gateway interface (CGI) scripts.

A modular configuration has been adopted for the server (Fig. 1) so that future changes can easily be made to isolated components within the system. Future improvements and updates to CGI scripts can thus be made without changing the database or the static pages. The server configuration also allows changes to be made to its structure, which are easily propagated throughout the server via simple alterations to the CGI scripts.

### Database Structure

The database structure is composed of eight interconnected tables, which are presented as a Unified Modeling Language<sup>TM</sup> class diagram in Figure 2. This structure was arrived at via normalization of the published *PKD1* and *PKD2* gene variant dataset to the third normal form. The normalization served to ensure a structure was arrived at that eliminates storage redundancy (and hence data corruption), ensures efficient data management (optimized performance of search requests) and enables the future addition of gene variant data to the preexisting dataset. Data are organized around the entered set of known *PKD1* and *PKD2* gene variants, which are located in the *gene variant* table. Each reported instance of a variant detected in a patient sample is recorded within the *sample* table. Corresponding deidentified patient details, including any available clinical information, can be recorded within the *patient* table. Variants reported in allele frequency studies, rather than reported from an individual patient sample, are recorded within the *population* table. Additional details about each published variant, including source reference, corresponding author, associated laboratory, and cDNA nucleotide reference sequence are housed within the *reference*, *corresponding author*, *lab*, and *cDNA* tables, respectively.

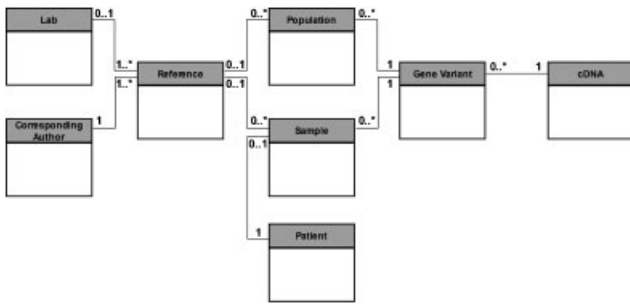


FIGURE 2. The relational database structure of PKDB represented as a Unified Modeling Language™ class diagram. Each class contains a differing number of attributes: Lab (7), Corresponding Author (4), Reference (11), Population (15), Sample (15), Patient (66), Gene Variant (29), and cDNA(12). The relationships between each of the classes are described via an interconnecting line. Numbers at the junctions of the interconnecting lines specify the possible number of class instances that may participate in a relationship with a single instance of the adjoining class.

**Class attributes.** Attributes for each gene variant have been selected to ensure efficient query handling and also to facilitate future nomenclature changes that may be required for the description of a variant at the genomic and protein level. The attributes within the *gene variant* class permit both recording and selective retrieval of details of all genomic variants and the predicted change at the amino acid level. Ancillary attributes permit the inclusion of data that may aid future clinical and research applications (e.g., the likely clinical significance of the variant, whether the variant was detected in germline or somatic tissues, and variant population frequencies) have also been included. Attributes within the patient table serve to capture a limited set of clinical features that have proven sufficient for the purpose of describing the natural history of PKD1 and PKD2-type polycystic kidney diseases [Hateboer et al., 1999].

### Retrospective Entry of Published Variants

To launch PKDB, details of all *PKD1* and *PKD2* variants reported in 73 cited peer reviewed publications (Supplementary Table S1) have been entered within the database. Prior to their inclusion, all variant nucleotide and protein level descriptions were curated and altered, where necessary, to ensure conformity to the current HGVS nomenclature standards ([www.hgvs.org/mut-nomen](http://www.hgvs.org/mut-nomen)) [den Dunnen and Antonarakis, 2000]. Standardization of the variant report numbering was achieved via representing each with respect to the translational start sites of the *PKD1* and *PKD2* NCBI RefSeq sequences (*PKD1*: NM\_000296.2; *PKD2*: NM\_000297.2). Consequently, variants previously reported with reference to genomic sequences or those only described at protein level are now described by their cDNA position (i.e., c.1232\_1234delATC in lieu of g.44003delATC or p.LEU440del). In addition, variants positioned within introns are now also reported in terms of their cDNA representation in accordance with current nomenclature standards (e.g., c.1234+12A>T in lieu of IVS4+12A>T). The accuracy of each reported variant was assessed with the use of an in-house mutation checker tool in conjunction with the Artemis software package [Rutherford et al., 2000]. Any inconsistency noted in the description of variants was flagged and subjected to closer investigation until resolved. As a final verification step, corresponding authors were invited to check the recorded details of variants reported within their

publications. Any additional amendments made corresponding authors have been incorporated by the curator into the database. At the time of launching the PKDB, it contained the details of 1,028 published variants, of which 796 are *PKD1* variants (512 of which are unique) and 232 are *PKD2* variants (131 of which are unique).

### Future Data Submission

**Gene variants.** Researchers and clinicians who have unreported instances of *PKD1* and *PKD2* gene variants (novel or otherwise) and who are continuing to characterize variants within the two genes are invited to submit these details using the submission forms available for download from the PKDB web server. Gene variants must be described using the above-mentioned current HGVS nomenclature standards. Numbering within gene variant descriptions should be with respect to the translational start site of the current NCBI RefSeq mRNA sequence for the gene involved, the accession number of which should be provided. To assist with the accurate reporting of variants, the in-house mutation checker software is mounted in the PKDB server (Fig. 3). To use the mutation checker, users simply need to select the RefSeq for the ADPKD gene involved, enter the variant cDNA start position, choose the type of mutation and provide the variant nucleotide(s) involved. Clicking the “check” button prompts a results page that describes the entered mutation and any resultant amino acid changes. If the entered nucleotide change results in a frameshifting mutation, the alternatively translated nucleotide sequence is displayed. The accuracy of variant data submitted via the downloadable submission forms will be further assessed by the database curator prior to inclusion within the database to ensure the high standard of data quality is maintained.

**Clinical details.** PKDB offers an option for standardized entry of deidentified clinical data, which can interface with genomic variant data stored within the database. This is primarily intended to assist clinical research, particularly forthcoming therapeutic intervention studies that will require standardized longitudinal documentation of clinical severity indicators in large numbers of ADPKD-affected individuals. Researchers and clinicians are encouraged to consider recording clinical details in an internationally standardized way, aided by downloadable clinical data entry forms present on the PKDB web server.

### Search Interface

As well as a number of useful links to ADPKD and general genomic resources, the PKDB web interface incorporates a search facility that permits searches for specific genomic variants within PKDB as required. The search interface has been designed for ease of use by those who have no prior knowledge of relational query operations. To this end, “pull down menus” have been included and, where advantageous, “radio buttons” have been incorporated into the search interfaces. Search results are displayed on a results page, which consists of a table of matches to the specified search parameters.

A wide variety of searches are made possible through the web interface, comprising: *reference searches*, performed by providing a particular publication criterion (for a given author and/or publication year and/or journal); *patient details searches*, through the provision of a particular patients details; *clinical details searches*, by selecting a particular clinical data type; or *advanced molecular searches*, by providing desired gene variant characteristics either at the nucleotide or protein level. With the exception of the patient details search (which displays a given patient’s clinical

# Mutation Checker

Site Map Latest news/publications Database

Use the form below to check mutations in the PKD genes.

REFERENCE SEQUENCE	Select PKD Gene: PKD1
MUTATION DESCRIPTION	Start position: 12258 Type: Point Variant nucleotide(s): A
ACTION	Check Reset

Help (in new window)  
Mutation entered was T => A at nucleotide position 12258 in PKD1.  
New sequence is: CTGTG **A** GTGGG  
Predicted change is Cysteine (C) to Terminator (\*) at amino acid 4086.

FIGURE 3. Illustrating the mutation checker interface accessible through the PKDB site. Upon providing details describing the reference sequence and gene variant, clicking the “Check” button generates a description of the predicted effect at the protein level. [Color figure can be viewed in the online issue, which is available at [www.interscience.wiley.com](http://www.interscience.wiley.com).]

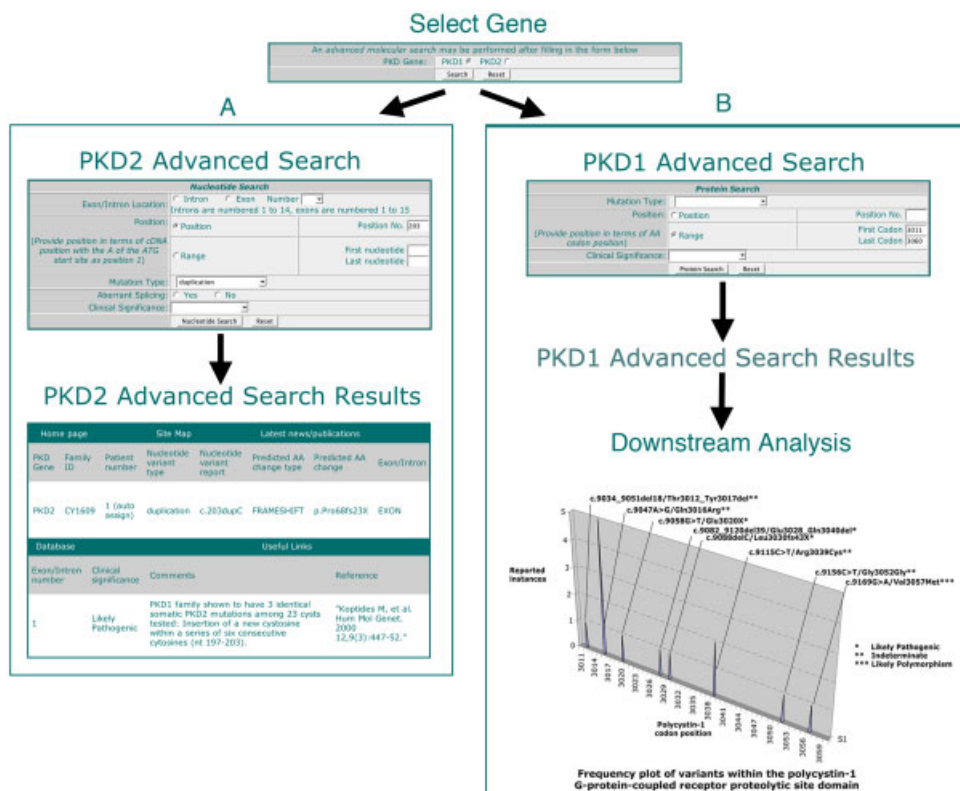


FIGURE 4. Illustrates the possible steps involved in an advanced molecular search. **A:** To determine the novelty of PKD2 variant c.203dupC, a user may enter the appropriate search criteria (Nucleotide Search box), which returns a previous reporting of the variant by “Koptides et al.” in 1999. **B:** Analysis of the frequency of reported variants within the G-protein-coupled receptor proteolytic site domain of polycystin-1 (codon range 3011–3060). Following the submission of the search criteria (Protein Search box), the results indicate the presence of eight distinct variants within PKDB reported with varying frequencies. [Color figure can be viewed in the online issue, which is available at [www.interscience.wiley.com](http://www.interscience.wiley.com).]

details), each search returns a list of reported gene variants, displaying in each case the gene involved, family ID, patient number, nucleotide variant type, nucleotide variant report, predicted amino acid change type, predicted amino acid change, exon/intron location, clinical significance, curator comments, and publication details.

### Example PKDB Queries

To illustrate the breadth of investigation made possible with the above mentioned search interface, a number of example searches are presented as follows. When an unknown PKD2 variant is detected in a sample from a person diagnosed with ADPKD, an

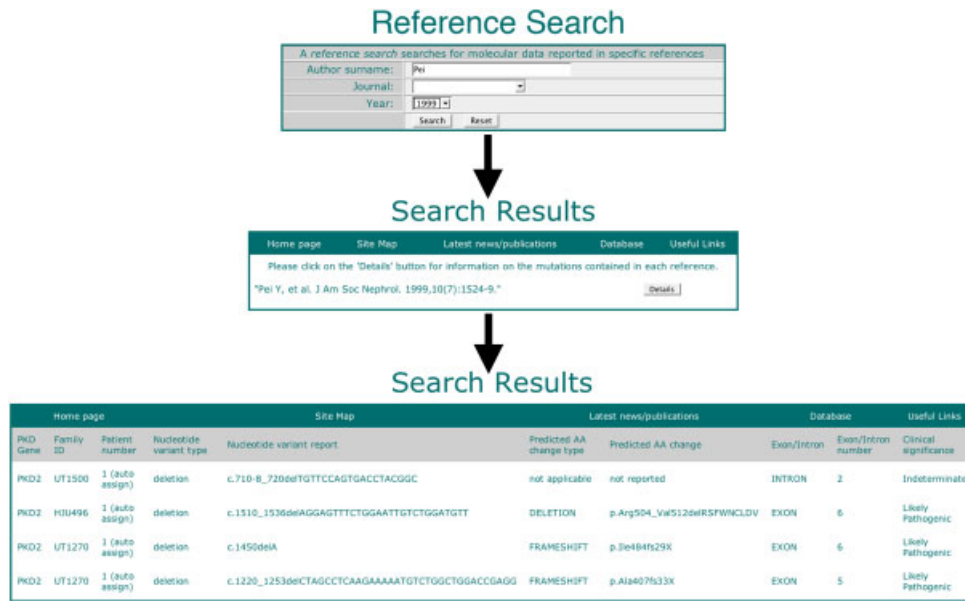


FIGURE 5. Displaying the retrieval of PKD1 variants reported by the author “Pei” in the year 1999. Once the user has entered the relevant search criteria, a page displaying the publications matching the search criteria within PKDB is generated. Clicking on the “Details” button on this new page yields a table of gene variants reported in the parent publication (not all shown here). [Color figure can be viewed in the online issue, which is available at [www.interscience.wiley.com](http://www.interscience.wiley.com).]

*advanced molecular search* may be performed on all reported PKD2 variants to check whether the variant has been previously reported. For example, if an investigator identifies a c.203dupC variant in the PKD2 gene of a patient suffering from ADPKD, an *advanced molecular search* providing nucleotide position 203 within the “Nucleotide Search” box reveals that this variant has been reported previously (Fig. 4A). For an overview of the likely pathogenic complement of PKD1 variants reported thus far, an investigator can perform an *advanced molecular search* similar to that presented in Figure 4A, through selecting “Likely Pathogenic” from the pathogenicity drop-down list in the nucleotide search box. Such a search provides the full list of reported likely pathogenic PKD1 variants. Alternatively, an investigator wishing to view the PKD2 variants reported by Pei in 1999 can perform a *reference search* as presented in Figure 5, revealing a list of 34 gene variants published by the author that year.

The search interface may also be used to facilitate more complex investigation. For instance, a researcher interested in viewing the frequency of variants across the G-protein-coupled receptor proteolytic site domain (codon range 3011–3060) [Qian et al., 2002] of *polycystin-1* may perform an *advanced molecular search* using appropriate search criterion within the “Protein Search” box (Fig. 4B). Analysis of the results shows eight distinct published gene variants (four likely pathogenic, three indeterminate, and one likely polymorphic) in this region occurring with varying frequency (one to five individuals). Using similar means, a researcher could examine the different protein variant type complements within each of the different protein domains in each gene, the distribution of likely pathogenic vs. likely polymorphic variants across the PKD1 and PKD2 nucleotide sequences and so forth.

## SUMMARY AND FUTURE DIRECTIONS

PKDB has been established as a publicly accessible database, which aims to streamline the evaluation of PKD1 and PKD2 gene variants detected in samples from those with ADPKD, as well as to

assist ongoing clinical and molecular research in the field. The unique features of this locus-specific database include downloadable data forms to facilitate the recording of clinical and molecular details in a standardized format, which will be essential for any forthcoming clinical trials; a mutation checker to aid the provision of accurate descriptions of gene variants; and a database structure providing an interface between gene variant data and associated patient clinical data. Finally, PKDB has been launched with a large number of retrospectively curated published variants, which permits immediate use by those who are interested. Critical feedback is encouraged, either directly to the corresponding author or through the website.

## ACCESSIBILITY AND USAGE

The database may be accessed through the PKDB URL (<http://pkd.waimr.uwa.edu.au>) and is now publicly available for searching. Those wishing to enter data must make contact with the database curator who can be contacted by email through the above URL.

## CITATION

Publications arising from the use of PKDB should cite this article and the accompanying URL (<http://pkdb.mayo.edu>).

## ACKNOWLEDGMENTS

This work would not have been possible without funding from the PKD Foundation and the feedback from PKD1 and PKD2 publication corresponding authors, which is greatly appreciated. We are also grateful for the support of Dr. Peter Harris and Dr. Sandro Rosetti, to whom the curatorship of PKDB has been transferred. Our thanks go to Heikki Lehväslaiho (EMBL Outstation, European Bioinformatics Institute, Cambridge, UK) for help and advice on the creation of the mutation checker on the PKDB web server.

## REFERENCES

- Cotton RG, Scriver CR. 1998. Proof of “disease causing” mutation. *Hum Mutat* 12:1–3.
- Deltas CC. 2001. Mutations of the human polycystic kidney disease 2 (PKD2) gene. *Hum Mutat* 18:13–24.
- den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 15:7–12.
- European Polycystic Kidney Disease Consortium. 1994. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. *Cell* 77:881–894.
- Hateboer N, v Dijk MA, Bogdanova N, Coto E, Sagggar-Malik AK, San Millan JL, Torra R, Breuning M, Ravine D. 1999. Comparison of phenotypes of polycystic kidney disease types 1 and 2. European PKD1-PKD2 Study Group. *Lancet* 353:103–107.
- Krawczak M, Cooper DN. 1997. The Human Gene Mutation Database. *Trends Genet* 13:121–122.
- Krawczak M, Ball EV, Fenton I, Stenson PD, Abeyasinghe S, Thomas N, Cooper DN. 2000. The Human Gene Mutation Database: a biomedical information and research resource. *Hum Mutat* 15:45–51.
- Mochizuki T, Wu G, Hayashi T, Xenophontos SL, Veldhuisen B, Saris JJ, Reynolds DM, Cai Y, Gabow PA, Pierides A, Kimberling WJ, Breuning MH, Deltas CC, Peters DJ, Somlo S. 1996. PKD2, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* 272:1339–1342.
- Peters DJ, Sandkuijl LA. 1992. Genetic heterogeneity of polycystic kidney disease in Europe. *Contrib Nephrol* 97:128–139.
- Phakdeekitcharoen B, Watnick TJ, Ahn C, Whang DY, Burkhart B, Germino GG. 2000. Thirteen novel mutations of the replicated region of PKD1 in an Asian population. *Kidney Int* 58:1400–1412.
- Qian F, Boletta A, Bhunia AK, Xu H, Liu L, Ahrabi AK, Watnick TJ, Zhou F, Germino GG. 2002. Cleavage of polycystin-1 requires the receptor for egg jelly domain and is disrupted by human autosomal-dominant polycystic kidney disease 1-associated mutations. *Proc Natl Acad Sci USA* 99:16981–16986.
- Rossetti S, Chauveau D, Walker D, Sagggar-Malik A, Winearls CG, Torres VE, Harris PC. 2002. A complete mutation screen of the ADPKD genes by DHPLC. *Kidney Int* 61:1588–1599.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* 16:944–945.
- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeyasinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 Update. *Hum Mutat* 21: 577–581.