Remarkable Sequence Conservation of the Last Intron in the PKD1 Gene

Marianna Rodova,* M. Rafiq Islam,† Kenneth R. Peterson,* and James P. Calvet*

*Department of Biochemistry and Molecular Biology, and the Kidney Institute, University of Kansas Medical Center, Kansas City, Kansas; and †Department of Chemistry/Physics, Northwest Missouri State University, Maryville, Missouri

The last intron of the PKD1 gene (intron 45) was found to have exceptionally high sequence conservation across four mammalian species: human, mouse, rat, and dog. This conservation did not extend to the comparable intron in pufferfish. Pairwise comparisons for intron 45 showed 91% identity (human vs. dog) to 100% identity (mouse vs. rat) for an average for all four species of 94% identity. In contrast, introns 43 and 44 of the PKD1 gene had average pairwise identities of 57% and 54%, and exons 43, 44, and 45 and the coding region of exon 46 had average pairwise identities of 80%, 84%, 82%, and 80%. Intron 45 is 90 to 95 bp in length, with the major region of sequence divergence being in a central 4-bp to 9-bp variable region. RNA secondary structure analysis of intron 45 predicts a branching stem-loop structure in which the central variable region lies in one loop and the putative branch point sequence lies in another loop, suggesting that the intron adopts a specific stem-loop structure that may be important for its removal. Although intron 45 appears to conform to the class of small, G-triplet-containing introns that are spliced by a mechanism utilizing intron definition, its high sequence conservation may be a reflection of constraints imposed by a unique mechanism that coordinates splicing of this last PKD1 intron with polyadenylation.

Introduction

Genes encoding eukaryotic mRNAs are usually interrupted by introns, whose origins and functions still remain mysteries. Theories for the origins of introns are that they coevolved with the earliest genes (intron-early) (Fedorov et al. 2001; Roy et al. 2001) and/or that they evolved after the earliest genes by insertion or duplication mechanisms (intron-late) (Logsdon, Stoltzfus, and Doolittle 1998; Fedorov et al. 2001). Most genes contain numerous introns, which can range from very large (>10,000 bp) to very small (<100 bp). Typically, introns are much larger than exons and are composed of sequences that are much less conserved than those of exons (Hawkins 1988; Kolkman and Stemmer 2001). In some cases, introns have been attributed functions, including transcriptional regulation (van Haasteren et al. 2000), the coding of intron-derived small RNAs (Tycowski, Shu, and Steitz 1988), and alternative splicing (Maniatis and Tasic 2002). In most cases, however, introns have no known function.

Mechanisms for intron removal by pre-mRNA splicing require the precise identification of intron-exon borders (Reed 2000). Sequence elements within introns that are involved in the splicing mechanism are conserved donor (5'') and acceptor (3'') splice sites and an internal branch point sequence containing a conserved adenosine residue. Splice-site recognition and splicing efficiency are also regulated by intron and exon splicing enhancer sequences (McCullough and Berget 2000). In general, spliceosome formation is thought to first involve “exon definition,” whereby upstream acceptor and downstream donor splice sites flanking an exon are identified by the splicing machinery before branch-site recognition and subsequent splicing of the upstream intron (Berget 1995). This mechanism is thought to apply to the majority of splicing events, which typically involve small exons flanked by much larger introns. In contrast, small introns may be recognized by a process of “intron definition” (Talerico and Berget 1994), and their splicing may be facilitated by the presence of multiple G-triplet sequences within these introns (McCullough and Berget 1997). The removal of the last intron of some pre-mRNAs may involve a mechanism that defines the last exon by coupling splicing with polyadenylation (Berget 1995; Cooke and Alwine 2002).

Autosomal dominant polycystic kidney disease (ADPKD) is one of the most common, potentially lethal inherited diseases worldwide, with a frequency of 1 in 200 to 1,000 in the population (Calvet and Grantham 2001; Peters and Breuning 2001; Qian, Harris, and Torres 2001; Peters and Breuning 2001; Qian, Harris, and Torres 2001; Igarashi and Somlo 2002). The disease is characterized by the growth of innumerable, large, fluid-filled cysts arising from kidney tubules, which can ultimately lead to renal failure in midlife. In some patients, the disease is also characterized by extrarenal manifestations, such as liver and pancreatic cysts, and cerebral and aortic aneurysms. Mutations in the PKD1 gene account for 85% to 90% of ADPKD cases. The human PKD1 gene, which encodes a very large multimeric-spanning receptor termed polycystin-1, is greater than 50 kb in length and has 46 exons (The European PKD Consortium 1994; Burn et al. 1995; Harris et al. 1995; Hughes et al. 1995; The International PKD Consortium 1995). PKD1 sequences have been characterized from human, mouse, rat, dog, and pufferfish (Lohning, Nowicka, and Frischau 1997; Sandford et al. 1997; Xu et al. 2001; Dackowski et al. 2002). PKD1 orthologs have also been identified in sea urchin and nematode (Moy et al. 1996; Barr and Sternberg 1999). We previously reported a very high sequence conservation in the amino acid sequence flanking the last intron in the PKD1 gene, intron 45 (Parnell et al. 1998). We subsequently examined the genomic sequence and found striking nucleotide sequence conservation of this intron across the four known mammalian PKD1 genes.

Materials and Methods

DNA clones. A mouse PKD1 genomic clone was isolated by screening a 129/SVJ mouse genomic library in lambda FIXII (Stratagene) using a labeled 790-bp PCR
product generated from a cDNA template using primers for sequences in exons 45 and 46. A portion of this genomic clone containing the 3′ region of the PKD1 gene was subcloned into pBlueScript and sequenced on both strands (GenBank accession number AF333927). Rat genomic DNA was prepared using the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer’s instructions. PCR primer sequences were introns 43 and 44 sense 5′-CCC-GTT-TGC-ACG-ACC-ATC-CAC-3′; introns 43 and 44 antisense 5′-GGG-CGG-TAA-AGC-TCC-CCA-CGC-AAG-GC-3′; and intron 45 antisense 5′-GGA-AGG-GTG-TGA-AGC-TTC-TGA-ACC-3′; PCR products were purified and cloned into the T-Easy vector (Promega) and sequenced on both strands.

Sequence Analysis

PKD1 and α-globin-2 sequences were obtained from GenBank as follows: rat PKD1 cDNA accession number AF277452; dog PKD1 genomic accession number AF483210; human PKD1 genomic accession number L39891; mouse α-globin-2 genomic accession number AF333927. Multiple sequence alignments were carried out using ClustalW version 1.8. For the α-globin-2/intron-2 alignment, the human and mouse introns were anchored with flanking exon sequences. Pairwise identities for the four PKD1 exons (exon 46 coding) and for intron 45 were determined using Blast version 2 (NCBI), and all six pairwise combinations were averaged. Pairwise identities for introns 43 and 44 were calculated directly from the ClustalW alignments (fig. 1). Translations were carried out using 6 Frame Translation (Baylor College of Medicine Search Launcher; http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html). RNA secondary structure predictions were carried out using RNAfold (Vienna RNA Package; http://www.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

Results

Intron 45 of the mammalian PKD1 gene (the last intron) lies within the region of the gene that encodes the C-terminal, cytosolic tail of polycystin-1. Our previous studies identified an approximately 50–amino acid region within the C-tail of polycystin-1 that is among the most highly conserved sequence regions in the approximately 4,300–amino acid protein (Parnell et al. 1998). In analyzing the genomic sequence of the mouse PKD1 gene, we found that intron 45 (which splits this region) exhibits striking sequence similarity with the human counterpart (fig. 1). For further comparison, we sequenced intron 45 from rat and we obtained the dog intron 45 sequence from GenBank (accession number AF483210). As shown in figure 1 (top), these sequences display
Figure 1 shows a comparison of intron 45 from different species, highlighting the conservation of G-triplets and the splicing site sequence. The intron's secondary structure is also depicted, indicating a conserved structure across species.

The analysis of intron 45 conservation across species reveals unusually high conservation. In fact, the mouse and rat intron 45 sequences are 100% identical over their 94-bp length. Conservation was recently noted between human and dog intron 45 sequences (Dackowski et al. 2002); however, the identity was reported to be only 83% instead of 91%, which we found in our analysis.

To further analyze the sequence conservation over the 3' terminal region of the PKD1 gene, we compared the last four exons (exons 43 to 46) and the last three introns (introns 43 to 45) of the PKD1 gene by carrying out a ClustalW alignment for these sequences from the four mammalian species (see Supplementary Material online). Figure 2 summarizes the degree of sequence identity for the four exons and three introns. As shown, all three introns are small (<100 bp). The percent identities are the averages of all six pairwise sequence comparisons. As expected, the coding regions are at least 80% identical, whereas introns 43 and 44 are only 57% and 54% identical, respectively. In contrast, the most conserved region in this greater than 1,400-bp sequence is intron 45, with an average identity of 94% across all four species.

It was recently shown that small vertebrate introns can contain a high proportion of G-triplet sequences, which may function as intronic splicing enhancers (McCullough and Berget 1997). As shown in figure 1 (top), PKD1 intron 45 has five G-triplets (overlined). For comparison, figure 1 (bottom) shows a pairwise alignment for the human and mouse a-globin-2 intron 2 sequences, the intron first used to investigate G-triplet function. Each contains numerous G-triplets: nine G-triplets for human (overlined) and eight G-triplets for mouse (underlined). In contrast to PKD1 intron 45, the globin intron sequences are only 57% identical.

Analysis of the intron 45 pre-mRNA sequences from the four mammalian species using RNAfold shows that the intron is capable of forming a very stable secondary structure (fig. 3). The length of the intron is within the approximately 110-nucleotide, naked RNA “window” that would allow intramolecular base pairing to occur (Eperon et al. 1988), making it very likely that such a structure would form in the nascent RNA transcript immediately after being synthesized, before the binding of hnRNP and splicingosomal proteins. It is notable that the one variable region near the middle of the intron (see fig. 1, top) occurs in a loop and thus does not alter the base-paired structure. In addition, the two single base-pair transitions in the human intron, which lie upstream (A→G) and downstream (C→T) of the variable region, are within base-paired stems, preserving the secondary structure. Furthermore, the likely branch-point sequence (Zhuang, Goldstein, and Weiner 1989) lies in another loop, which would make it accessible for interaction with the splicing factor SF1 (Liu et al. 2001) and with U2 snRNA (Calvet, Meyer, and Pederson 1982; Zhuang, Goldstein, and Weiner 1989; Pascolo and Seraphin 1997). Thus, intron 45 may adopt a specific secondary structure conformation that is required to facilitate its splicing.

Discussion

Intron 45 was found to be the most conserved region in the 3' terminal sequence of the PKD1 gene, with significantly higher conservation than the exon 43 to 46 coding regions. It is also striking that the 94-bp mouse and rat intron 45 sequences are 100% identical. By contrast, the overall sequence conservation between the mouse and rat exon 43 to 46 coding regions is 92%. The conservation of intron 45 does not extend to fish, as the comparable intron in the pufferfish PKD1 gene (intron 53) shows no sequence similarity whatsoever with the mammalian sequences (data not shown), despite being the last intron and lying in the identical position in the coding sequence (Sandford et al. 1997). Whereas the mammalian intron 45

Figure 2.—Average pairwise identities for the exon 43 to 46 region of the PKD1 gene. Boxes are exons, lines are introns, and shading represents coding sequence. The relative lengths of the coding regions and introns are represented by the lengths of the respective boxes and lines. Pairwise identities for the four PKD1 exons (exon 46 coding) and for intron 45 were determined using Blast 2, and all six pairwise combinations were averaged to provide the numbers shown. Pairwise identities for introns 43 and 44 were calculated directly from ClustalW alignments (see Supplementary Material online).

Figure 3.—Secondary structure prediction for PKD1 intron 45. The four mammalian PKD1 intron 45 sequences were analyzed by RNAfold; the complete structure for the human sequence is shown. Differences in the base-paired stems between human and the other sequences are shown circled in lowercase italicized letters. Differences in the variable loop region in the mouse, rat, and dog sequences are shown in lowercase italicized letters. The putative branch-point sequence and branch-point adenosine (arrow) are indicated. This adenosine lies 23 nt upstream of the acceptor splice site and therefore is within the optimal distance of 19 to 23 nt (Chua and Reed 2001). The minimum free energies of the predicted intron 45 structures are −50.4 kcal/mol (human), −52.6 kcal/mol (mouse and rat), and −54.6 kcal/mol (dog).
sequences are less than 100 bp and 75% GC, the pufferfish intron 53 sequence is 207 bp and only 44% GC.

The very high sequence conservation of intron 45 suggests that it has been selected for, and therefore may have a unique function. There are many possibilities, and several are worthy of comment. One function may be that the intron (and perhaps the flanking conserved sequence) has a role in regulating transcription. Although no known promoter lies close to intron 45, it is well established that enhancers can function over quite long distances. We tested intron 45 and its flanking sequences for enhancer activity by transient transfection in HEK293T cells using the PKD1 promoter cloned upstream of a luciferase reporter (Rodova et al. 2002) but were unable to demonstrate an effect on transcriptional activity (data not shown). It is also possible that the unique tail-to-tail organization of the PKD1 and TSC2 genes, which are directly abutted at their 3’ ends (The European PKD Consortium 1994; Olsson et al. 1996), requires that transcription from the TSC2 gene be efficiently terminated so as to not extensively overlap with and therefore interfere with PKD1 transcription. If so, the intron 45 sequence, which lies only approximately 1,500 bp downstream of the 3’ end of the TSC2 gene, may function as a termination signal. Arguing against this possibility is that the pufferfish PKD1 and TSC2 genes are also in a tail-to-tail organization (Sandford et al. 1996) and that RNA polymerase II terminators usually have a run of T residues (Kérrpolka and Kane 1991), which is not found in or around intron 45. Alternatively, the PKD1 and TSC2 genes may share a locus control region, or there may be a transcriptional insulator between the promoters of these two genes, which may include intron 45.

Intron 45 may contain within it, or may be part of, a transcription unit that produces a small RNA. Inspection of the sequence shows that there is an RNA polymerase III box A motif (Geiduschek and Tocchini-Valentini 1988) near the 5’ end of the intron (fig. 1, top). It is also possible that a small RNA is processed from the intron after its excision by splicing, as has been shown for a number of modification guide snoRNAs, which are encoded within introns (Seraphin 1993; Tycowski, Shu, and Steitz 1993). In general, however, snoRNAs are encoded in genes having some functional connection with the process of translation, making this possibility less likely. We have also considered that the intron has a protein-coding function, such that if the intron is retained by alternative splicing, it would encode a short insertion in the polycystin-1 C-tail. We feel that this can be ruled out because, whereas the human intron 45 sequence has the required length and open reading frame to encode an in-frame 30–amino acid insertion, those of the other species do not.

Intron 45 appears to conform to the class of small vertebrate introns that are spliced by a mechanism utilizing intron definition (McCullough and Berget 1997, 2000). These introns often contain multiple G-triplets, which serve as intron enhancer sequences. It has been shown that the presence of G-triplets can overcome “weak” acceptor splice sites. Intron 45 has a very C-rich acceptor site, with no contiguous T residues, and therefore would be considered to have a “weak” acceptor splice site by these criteria. The G-triplet density (number of G residues in G-triplets divided by intron length) of intron 2 of α-globin-2 (fig. 1, bottom) is 18% to 20% and the G-triplet density of intron 45 is 20% to 21%. In contrast, the comparable pufferfish intron has a G-triplet density of only 7%. Furthermore, the pufferfish intron has a T-rich polypyrimidine track with multiple runs of contiguous T residues, suggesting that it is spliced by a different mechanism. The presence of G-triplets in intron 45 does not explain the very high sequence conservation; however, as the “prototypic” G-triplet–containing intron, α-globin–2 intron 2, shows very little sequence conservation (fig. 1, bottom).

Another possibility is that the conserved intron 45 sequence is required for the regulation of its own splicing. It has been shown that hnRNA contains regions of intramolecular base pairing in vivo (Calvet and Pederson 1977, 1978, 1979a, 1979b) and that secondary structure can play a role in the regulation of splicing (Balvay, Libri, and Fiszman 1993; Charpentier and Rosbash 1996). The predicted secondary structure of the intron 45 transcript (fig. 3) places the variable region of the intron within a loop. Furthermore, the two nonloop nucleotide substitutions preserve the base-paired stems, suggesting that the intron adopts a highly specific stem-loop conformation in which the stems are functionally important. The predicted structure also places the putative branch-point sequence within another loop, where it should be accessible for interaction with splicing factors such as SF1 and U2 snRNA, which both require single-stranded RNA for binding. The conservation of intron 45 suggests that the sequence has a function in addition to (or other than) intron definition, since other small G-triplet introns do not show sequence conservation. Possibly, the tertiary structure of intron 45 may be important for defining it as the last intron, by perhaps creating a unique platform for factors that couple its splicing with polyadenylation (Cooke and Alwine 2002).

Other examples of conserved introns or intron regions exist (Bourbon and Amalric 1990; Aruscavage and Bass 2000; Sun et al. 2000; Yatsuki et al. 2000); however, in no case is the reason for this conservation understood. Further insight into whether PKD1 intron 45 has a special function could be obtained by testing the effect of germline mutation of the intron in transgenic mice, and by searching for otherwise unexplained intron 45 mutations in ADPKD patients.

Acknowledgments

This research was supported by National Institutes of Health grants P50 DK57301 and P01 DK53763 (J.P.C.), P50 DK57301 (K.R.P.), and a grant from the KUMC Research Institute (J.P.C.).

Literature Cited


Sandford, R., B. Sgotto, T. Burn, and S. Brenner. 1996. The tuberin (TSC2), autosomal dominant polycystic kidney disease (PKD1), and somatostatin type V receptor (SSTR5) genes form a synteny group in the Fugu genome. Genomics 38:84–86.


Claudia Kappen, Associate Editor

Accepted May 29, 2003