Mammalian Genome 11, 213–219 (2000).

Mammalian Genome
Incorporating *Mouse Genome*

Comparative analysis of the PCOLCE region in *Fugu rubripes* using a new automated annotation tool

Giorgia Romina Riboldi Tunnicliffe,[1]* Gernot Gloeckner,[1] Greg S. Elgar,[2] Sydney Brenner,[3] André Rosenthal[1]

[1]Institute for Molecular Biotechnology, Department of Genome Analysis, Beutenbergstrasse 11, 07745 Jena, Germany
[2]UK HGMP Resource Centre, Hinxton, Cambridge, CB10 1SB, UK
[3]The Molecular Sciences Institute, 2168 Shattuk Ave., Berkeley, California 94074, USA

Received: 13 October 1998 / Accepted: 25 July 1999

Abstract. The Japanese pufferfish *Fugu rubripes* with a genome of about 400 Mb is becoming increasingly recognized as a vertebrate model organism for comparative gene analysis (see Elgar 1996 for review). We have isolated and sequenced two *Fugu* cosmids spanning a genomic region of 66 kb containing the *Fugu* homolog to the human PCOLCE-I (Glöckner et al. 1998). We then examined if RUMMAGE-DP, a newly developed analysis tool for gene discovery which was designed for human and mouse genomic DNA, can be used for automatic annotation of *Fugu* genomic sequence. The exon prediction programs contained in RUMMAGE-DP performed better overall for the human sequence than for the *Fugu* contig. The GENSCAN program was the only exon prediction programme that performed equally well for both organisms. We show that RUMMAGE-DP is very useful in automatic analysis of *Fugu* sequences. Comparative analysis of the genomic structure of the PCOLCE-I genes in *Fugu* and human reveals that the exon/intron structure throughout the protein coding region is almost identical. We defined an additional domain based on the high degree of similarity of 26 aa between mammals and *Fugu*. The PCOLCE-I protein in both organisms contains two highly conserved CUB domains. Exons 6 and 7 are the only coding exons that differ in length between the two species. We assume that these exons do not code for any catalytic domain of the protein. Analysis of the remaining five *Fugu* genes within the 66 kb interval revealed no conserved synteny with the corresponding human 7q22 region.

Introduction

Recently *Fugu rubripes*, a Japanese pufferfish, was proposed as a vertebrate model organism for comparative analysis (Brenner et al. 1993). With a size of 400 Mb (Hinegardner 1968), the *Fugu* genome is very compact and comparative analysis of several *Fugu* and human genes reveals a high degree of gene structure conservation between the two species as manifested by the number of exons, exon lengths, and intron phases. Several examples have been studied including the gene for Huntington's disease (Baxendale et al. 1995), L1CAM (Coutelle et al. 1998, mutations in this gene are associated with X-linked hydrocephalus, spastic paraplegia (SPG1), MASA syndrome), G6PD (Riboldi-Tunnicliffe et al., in preparation), murine *whn* (associated with congenial athymia and hairless Schuddekopf et al. 1996), as well as the Hox gene clusters (for review, Holland 1997) and others. Generally, *Fugu* genes are much smaller than their human or murine counterparts due to shorter introns. Up to now only one exception has been reported: L1CAM (Coutelle et al. 1998). Comparison of syntenic regions demonstrates that intragenic distances are smaller in *Fugu* than in humans (Trower et al. 1996). Comparative analysis between *Fugu* and humans can be also used to highlight conserved protein domains (Coutelle et al. 1998) which remain undiscovered if only mouse/human comparisons are performed. Recently, a survey sequencing project of about a thousand *Fugu* cosmids with 50–100 reads for each investigated cosmid has been completed (http://fugu.hgmp.mrc.ac.uk). Due to its importance for gene discovery and protein structure prediction it can be anticipated that the whole *Fugu* genome may be sequenced in the near future. First pass analysis of *Fugu* genomic sequence will be important to use this genomic sequence information efficiently.

We have recently developed an automated first pass annotation tool named RUMMAGE-DP to analyze large human and mouse genomic sequences (http://genome.imb-jena.de/, Glöckner et al. 1998). We now report on the use of RUMMAGE-DP for the analysis of *Fugu* genomic DNA sequence. In particular, we were interested to examine this tool for gene prediction. As a target we have sequenced and analyzed a 66 kb genomic region in *Fugu* containing the PCOLCE-I homologue and five further *Fugu* genes and compared it with the human PCOLCE locus on 7q22 (Accession No.: AF053356, Glöckner et al. 1998).

Collagen provides the structural framework for tissues and organs. Collagens I, II, and III are synthesized as precursor molecules called procollagen (Kivirikko and Millyla 1984); the cleavage of procollagen to collagen is catalyzed by the Procollagen Proteinase (Kadler and Watson 1995) and enhanced by the Procollagen Proteinase Enhancer Proteins: one for the C-terminus and one for the N-terminus (Kessler and Adar 1989). The C-terminal enhancer protein (PCOLCE-I), in human, localized on chromosome 7 band q21-22 (Takahara et al. 1996), has been shown to enhance the cleavage of the C-terminus of procollagen type I, II, and III. PCOLCE-I is highly conserved among the species so far analyzed (human, rat, mouse: Accession Nos.: AB008549, AB008534, AB008548). The evolutionary distance between *Fugu* and humans (400 My) may allow us to more precisely define the conserved domains of the PCOLCE gene than by comparing only mammalian sequences. Furthermore, we wanted to know whether the PCOLCE-I is contained in a syntenic region shared between humans and *Fugu*.

Materials and methods

DNA sequencing and assembling of shotgun reads. The two overlapping *Fugu* cosmids sequenced in this analysis were isolated as previously described (Elgar et al. 1995). They were sequenced using shotgun cloning in M13mp18 as described elsewhere (Platzer et al. 1997). In brief, after sonication of cosmid DNA and size selection, 0.6–1.6 kb fragments were subcloned in the SmaI site of M13mp18. M13 DNA was obtained by

Correspondence to: A. Rosenthal

* *Present address:* Division of Molecular Genetics, IBLS, University of Glasgow, Glasgow, G11 6NU, UK

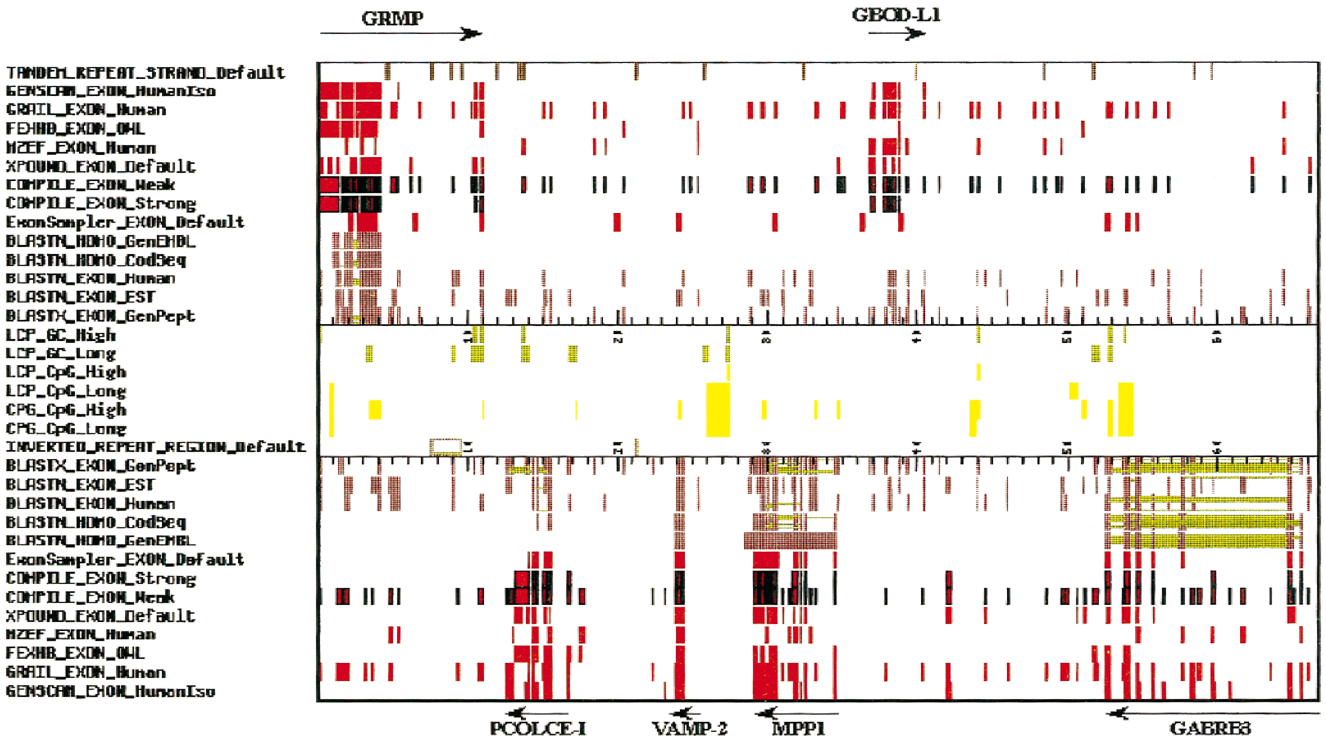


Fig. 1. Graphical view of RUMMAGE-DP results. In this figure the results of several features of RUMMAGE-DP are presented: repeats (brown), exon prediction programmes (red), BLAST searches (green) and GC content (yellow). The arrows show the direction of transcription for the six genes present in this locus. All genes are presented with their short name above the arrows. More detailed description of the features and output formats of RUMMAGE-DP can be found at http://genome.imb-jena.de

the Triton method (Mardis 1994) and then sequenced using dye terminator chemistry and the ABI 377 sequencer.

Computational analysis. RUMMAGE-DP integrates more than 20 different programmes (http://genome.imb-jena.de/, Glöckner et al. 1998). The COMPILE_EXON features summarize the results of the five programs for exon prediction: XGRAIL version 1.3c (Uberbacher and Mural 1991; Uberbacher et al. 1996), XPOUND (Thomas and Skolnick 1994), MZEF (Zhang 1997), GENSCAN (Burge and Karlin 1997), and FEXHB (Solovyev and Salamov 1997). The ExonSampler programme (Weber J. et al., unpublished) searches for homologies of the genomic sequence with entries in the CodSeq (Weber J., unpublished) and EST databases and tries to accommodate positive results on the genomic sequence obeying orf constraints (est2genome: Mott 1997). The results obtained using the RUMMAGE-DP can be formatted into the ACEDB format. The PROCRUSTES programme (http://www-hto.usc.edu/software/procrustes/qpn.html; Gelfand et al. 1996) was used to compare protein sequences from other species with the homologous genes in the *Fugu* cosmid sequence in order to confirm the predicted splicing sites of the exons. Several programs were used to predict the secondary structures of the putative *Fugu* proteins (nnPredict: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html, Kneller et al. 1990; GOR: http://molbiol.soton.ac.uk/compute/GOR.html, Garnier et al. 1996; PredictProtein server: http://www.embl-heidelberg.de/predictprotein/predictprotein.html). These proteins were then compared with the human, mouse, and rat counterparts to hunt for conserved domains. The ProDom database (http://protein.toulouse.inra.fr/prodom.html) was used to check the presence of specific domains such as CUB in all analyzed proteins.

Exon confirmation by RT-PCR. Total RNA was isolated from several *Fugu* tissues: liver, brain, and muscle using the method of Chomczynski and Sacchi (Chomczynski and Sacchi 1987). First strand synthesis using 1.5 µg of total RNA was performed using the ‘SuperScript plasmid System for cDNA synthesis’ (Gibco BRL). RT-PCR was carried out using primers inside different exons of each gene. Both PCRs were performed as follows: a denaturation step of 1′30″ at 95°C was followed by thirty cycles of denaturation 30″ at 95°C, annealing 30″ at 50–55°C (depending on the primer couple) and elongation 45″ at 72°C. All PCR products were directly sequenced to confirm the predicted splice sites.

Results

Sequence analysis of 66 kb of contiguous Fugu *sequence around the PCOLCE gene homolog.* Two overlapping *Fugu* cosmids (155N11, 192F24) containing the homolog of the human PCOLCE-I gene were sequenced using shotgun cloning in M13 and ABI fluorescent dye chemistry. Approximately 1,800 sequences were assembled using GAP4 (Bonfield et al. 1995). Subsequent editing and gap closure resulted in a 66,721 bp long consensus sequence (Accession No. AF16494). The accuracy of the final sequence was above 99.99% as determined by the 20 kb overlap of the two cosmids.

The whole *Fugu* genomic region containing PCOLCE-I was analyzed using the first pass sequence annotation tool RUMMAGE-DP (Fig. 1). Microsatellites (di- tri- tetra- and pentanucleotide repeat sequences) were found using RepeatMasker, they account for 1.48% of the genomic sequence and are uniformly distributed along the sequence. In addition, there are only three inverted repeats clustered in the first 22 kb. The overall GC content of the *Fugu* region is 45%. Using several exon prediction engines within RUMMAGE-DP six distinct exon clusters comprising 54 exons were predicted along the *Fugu* contig. Each exon cluster showed significant homology to human and murine genes suggesting the existence of six *Fugu* homologous genes. A total of 23 exons, derived from four of the six genes, were independently confirmed by RT-PCR using *Fugu* cDNA pools (results not shown). The *Fugu* specific cDNA sequences as well as the human

Table 1. The *Fugu* and human PCOLCE-I gene. Comparison of exon and intron sizes in the genes encoding for the PCOLCE-I protein in *Fugu* and humans.

	Exon size (bp)			Intron size (bp)	
No.	*Fugu*	Human	Intron phases	*Fugu*	Human
1	86	95	2	1217	878
2	109[a]		0	94	419
3	259[a]		1	370	873
4	125[a]		0	97	460
5	137[a]		2	208	603
6	266[a]	215	1	761	821
7	69	72	1	87	113
8	171		1	88	129
9	131	167			

[a] Exons confirmed by sequencing the amplified cDNA from *Fugu.*

and murine cDNAs were used to determine the exon/intron structure of all genes (Tables 1 and 2).

GENSCAN, XGRAIL, and COMPILE_EXON_Strong detected 53 of the 54 exons which were confirmed by cDNA sequence analysis. Thirty-seven of the true exons could also be confirmed by the existence of EST data from other species. False positive EST data were excluded by using only spliced ESTs for the analysis. The number of false positive exons presented by the COMPILE_EXON_Strong algorithm in both *Fugu* and human PCOLCE-I contigs is remarkably low (see Table 3).

The other exon prediction programmes used were less effective: FEXHB and MZEF detected fewer true exons whereas XPOUND, MZEF and XGRAIL found a large number of false positives (Table 3).

Comparison of the Fugu *and human PCOLCE-I gene.* For the *Fugu* PCOLCE gene nine exons were predicted from the cosmid sequence (Table 1). Five of these exons were confirmed by the partial cDNA sequences obtained by nested PCR on cDNA libraries. Using PROCRUSTES we analyzed the PCOLCE-I genomic region in *Fugu* and humans by alignment with different mammalian PCOLCE protein sequences (human Accession No.: AB008549, rat Accession No.: AB008534, and mouse Accession No.: AB008548). These alignments revealed that the *Fugu* and human PCOLCE genes contain nine coding exons distributed along a genomic region of 4.3 kb in *Fugu* and 5.8 kb in human. The intron phases are identical (see Table 1) in both genes. The first intron in *Fugu* is 1.5 times larger than in humans. In *Fugu,* a polyadenylation signal was predicted 189 bp after the stop codon. The predicted *Fugu* PCOLCE coding region is 3 bp (1 amino acid) longer than that of the human gene. Exons 6 and 7 have different lengths in *Fugu* and humans as well as among several mammalian species. The last two thirds of exon 6 and exon 7 show low homology at the amino acid sequence level (Table 1, Fig. 2). This part of the protein appears to be either non-essential for protein function or it has evolved a different function. All PCOLCE-I proteins contain two highly conserved CUB (Complement-Uefg-Bmp1) domains. Exon 2 and exon 3 (from aa 34 to aa 143 in the *Fugu* protein) encode for the first CUB domain. The second CUB domain (from aa 156 to aa 267 in the *Fugu* protein sequence) is encoded by the exon 4 up to the first half of exon 6. The presence of these CUB domains was verified by a search in the ProDom database. The CUB motifs consist of approximately 110 residues (110 and 112, respectively), mainly found in developmentally regulated proteins (Bork 1991; Bork and Beckmann 1993). Cysteine residues that are involved in the formation of disulfide bridges are conserved (Fig. 2). Short stretches of beta sheets predicted by the nnpredict programme (data not shown) were present in the rest of the *Fugu,* human, mouse, and rat protein. There was no homology in the ProDom database for the third homologous domain present in exon 9 (from aa 405 to aa 430 in the *Fugu* protein, Fig. 2) which we defined from the high similarity in a 26 aa region, between the species analyzed.

Other Fugu *genes.* Besides the PCOLCE homolog, five further genes were detected in the 66 kb *Fugu* genomic sequence. Orientation and exon/intron structure of the coding part of all six genes are described in Fig. 1, Table 1, and Table 2. 5′ UTRs and 3′ UTRs of these genes could not be verified since complete *Fugu* mRNA sequences were not available and could not be obtained by RT-PCR. The genes were detected from left to right in the following order: GRMP (Glucose Repressor Mediator Protein), PCOLCE-I VAMP2 (Vesicle Associated Membrane Protein type 2), MPP1 (erythroid p55), GDOB-L1 (gamma-butyrobetaine, 2-oxoglutarate dehydrogenase-like1), and GABRB3 (GABA beta 3 receptor). PCOLCE-I, VAMP2, MPP1 and GABRB3 are transcribed on the same strand (reverse); GRMP and GDOB-L1 are transcribed in the forward strand of the contig.

The first predicted exon cluster (GRMP) is localized downstream the *Fugu* PCOLCE gene on the opposite strand (forward), it spans a region of over 11 kb and consists of 15 exons. It shows homology with a human mRNA for the KIAA0346 gene (Accession No.: AB002344, Nagase et al. 1997). This human protein is homologous to the yeast Glucose Repressor Mediator Protein (Accession No. SP: P14922). Seven of the fifteen predicted exons (from exon 4 to exon 10) were confirmed by a partial *Fugu* cDNA sequence (Table 2). The first 12 predicted exons are homologous to different human EST sequences (Accession Nos.: AF000992-5) which contain a TPR (tetratricopeptide) domain (Ohira et al. 1996), a degenerate consensus sequence of 34 amino acids found in multiple copies in different proteins. At the beginning of exon 5 an in frame deletion of 6 nucleotides (AACGAG), leading the loss of two amino acids (NE) in the protein, was detected at cDNA level. The cosmid sequence contains these six bases. The predicted cDNA of the 15 exons of the *Fugu* GRMP gene is 2,270 bp long, with an average exon size of 151 bp (exons range from 69 to 288). The intron lengths range from 70 to 4,963 bp (mean of 535 bp).

The third exon cluster (VAMP2) was found upstream to *Fugu* PCOLCE-I on the same strand and shows homology to the Vescicle Associated Membrane Protein type 2, also known as synaptobrevin 2. The coding sequence of the *Fugu* VAMP2 gene spans less than 3 kb with four of its five exons. It was not possible to find exon 1 as the coding part of this exon is only 2 bp long (AT) and the known non-coding part of the analysed genes show no similarity between each other. Alignment of the human VAMP2 (Accession No.: AJ225044), mouse VAMP2 (Accession No.: AF007168), and the VAMP1 homolog from *Torpedo californica* (Accession No.: J03777, Trimble et al. 1988) proteins with the *Fugu* VAMP2 gene using PROCRUSTES revealed the position of the last exon. Three of the five exons, which contain the synaptobrevin motif (exon 2 to 4), were confirmed at the sequence level by RT-PCR. There was no detectable homology between the different 5′ UTR of VAMP2 proteins in the database. The VAMP2 coding sequence is 333 bp long with the coding part of the exons ranging between 2 and 159 bp and introns between 91 and 128 bp.

Upstream of the VAMP2 gene on the same strand, another cds (MPP1) that shows homology to the human MPP1 (P55) protein (Kim et al. 1996) was detected. The twelve exons of the *Fugu* MPP1 gene (Elgar et al. 1995) span 5.5 kb. The complete cds of *Fugu* MPP1 is 1,404 bp long. Upstream of MPP1, on the forward strand, the fifth exon cluster (GBOD-L1) is located. This gene shows only weak homology to other proteins and probably represents a new gene. Based on its homology with two *C. elegans* proteins (Accession Nos.: Z66523_7; Z36948_5; Wilson et al. 1994) and a *Pseudomonas sp.* protein (Accession No. SP; P80193) we named it GBOD-L1 (gamma-butyrobetaine, 2-oxoglutarate de-

Table 2. Exon/intron lengths and intron phases of the *Fugu* genes present in the PCOLCE-I locus.

	GRMP			VAMP-2			MMP1[b]			GBOD-L1			GABA beta3		
No.	Exon	Intron[c]	Intron-	Exon	Intron[c]	Intron-	Exon	Intron[c]	Intron-	Exon	Intron[c]	Intron-	Exon	Intron[c]	Intron-
1	181	111	2	2	un	2	120	1871	0	148	562	2	68	686	2
2	127	176	0	103[a]	128	0	132	106	0	250	85	0	221[a]	7066	1
3	106	148	1	159[a]	122	0	79	243	1	120	184	0	83[a]	1781	0
4	206[a]	180	0	52[a]	91	1	86	72	0	237	135	0	138[a]	940	0
5	137[a]	88	2	17			69	1558	0	139	1568	1	153[a]	485	0
6	69[a]	89	2				197	119	1	162			257[a]	638	2
7	149[a]	96	1				104	80	0				90[a]	234	2
8	115[a]	70	2				81	84	0				354[a]		
9	188[a]	70	1				81	79	0						
10	142[a]	121	1				203	140	2						
11	127	70	1				75	82	2						
12	171	1167	1				177								
13	73	4963	2												
14	191	249	1												
15	288														

[a] Exons confirmed by sequencing the amplified cDNA from *Fugu*.
[b] Gene already reported in literature (Accession No.: X81359).
[c] Exon and intron lengths are expressed in base pairs (bp).
un, Unknown length.

Table 3. Comparison of the exon prediction efficiency of different programs between the PCOLCE-I loci in *Fugu* and human.

	Fugu[a]				Human[b]			
	# exons predicted		Analysis on nucleotide level		# exons predicted		Analysis on nucleotide level	
Programs	TP	FP	Sn[d]	Sp[d]	TP	FP	Sn[d]	Sp[d]
MZEF	22	16	0.340	0.507	108	16	0.549	0.891
XPOUND	32	23	0.625	0.684	103	29	0.577	0.817
XGRAIL	47	58	0.887	0.497	127	60	0.721	0.802
FEXHB	26	6	0.565	0.630	76	63	0.527	0.537
GENSCAN	51	6	0.963	0.790	149	20	0.834	0.910
COMPILE_E_Strong	44	2	0.718	0.677	126	1	0.769	0.833
ExonSampler	37	—[c]	0.650	0.779	131	—[c]	0.768	0.682

[a] *Fugu* number of true exons: 54.
[b] Human number of true exons: 161.
[c] Only spliced ESTs were considered.
[d] Sn and Sp as defined by Burset and Guigo (1996).

hydrogenase-like1). Several ESTs from mouse and human were also found (three human ESTs Accession Nos.: AA011424, AA130866, and AA677541; one mouse EST and Accession No.: AA270095) to be homologous to this exon cluster. The six predicted exons of the *Fugu* GBOLD-L1 gene span over 3.5 kb of sequence. Since amplification of cDNA was not successful all predicted exons remain unconfirmed. The hypothetical length of the CDS is 1087 bp. A polyadenylation signal was found 536 bp after the stop codon.

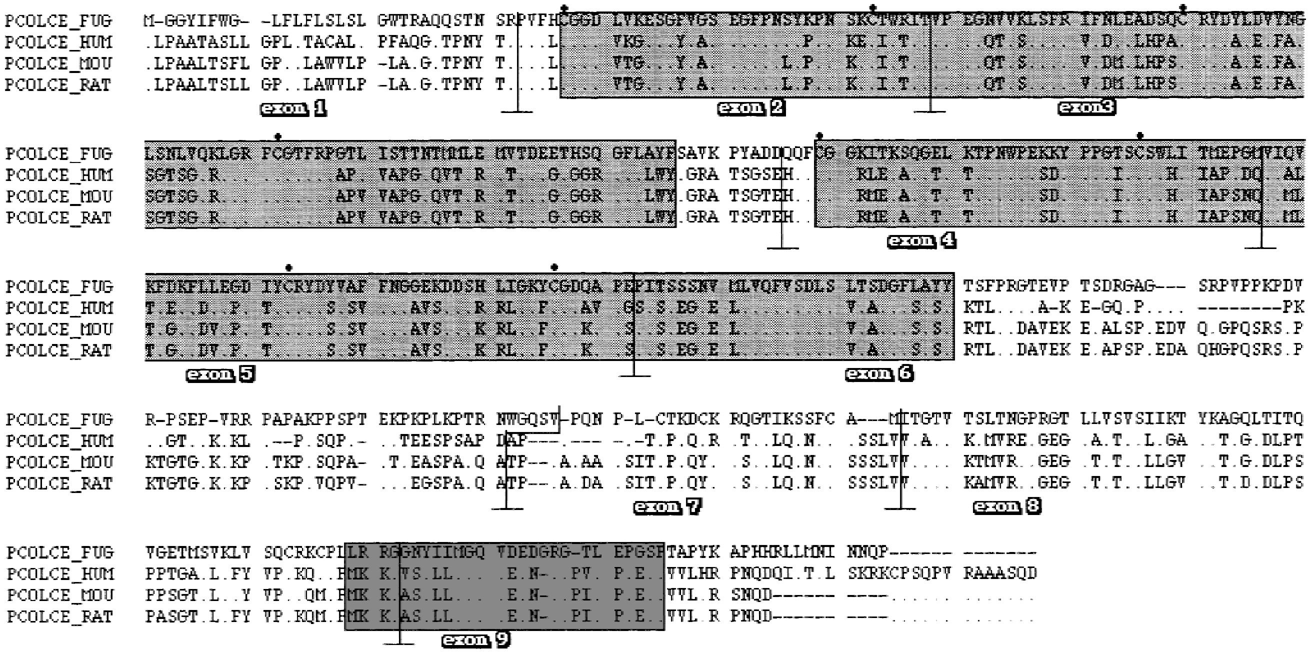
Downstream of the GBOD-L1 gene on the reverse strand, an exon cluster which shows significant homology to human GABRB3 (Accession No.: M82919, Wagstaff et al. 1991) was detected. Only the last 8 exons of this gene are present in the sequence. Seven exons (exon 1 to 6 and exon 8) were predicted using GABRB3 from human, mouse (Accession No.: U14420, Kamatchi et al. 1995), rat (Accession No.: X15468, Ymer et al. 1989), and chicken (Accession No.: X54243, Bateson et al. 1990) as target proteins in the PROCRUSTES programme. None of the exon prediction programs detected this exon (exon 7), but the sequencing of the RT-PCR products confirmed its presence in *Fugu*. The sequenced cds of the *Fugu* GABRB3 (eight exons) is 1,364 bp long.

Discussion

The last two years have seen a dramatic increase in the number of *Fugu rubripes* sequences in the databases. Several *Fugu* genes corresponding to human disease genes have been sequenced and characterized. In addition, there are over 32,000 genomic survey sequences (GSS) originating from the project at the Human Genome Mapping Project Resource Centre (HGMP-MRC, http://fugu.hgmp.mrc.ac.uk/).

Here we investigated whether RUMMAGE-DP—a new first pass annotation tool developed in our laboratory for mouse and human genomic DNA—can also be applied to *Fugu* genomic DNA. For this purpose we sequenced a 66 kb genomic segment of *Fugu* DNA containing the *Fugu* homolog of the human PCOLCE-I gene and analyzed the obtained data with RUMMAGE-DP. The results were compared with a 226 kb human genomic sequence localised on chromosome 7q22, which contained the PCOLCE gene (Glöckner et al. 1998). As a single exon prediction tool GENSCAN performed equally well with both organisms. Ninety-five percent of all *Fugu* exons and 93% of all human exons were detected, the percentage of false positives was in the same range for both organisms (11% versus 12%). The other prediction programs when used alone were less successful in the analysis of *Fugu* sequences. COMPILE_EXON_Strong summarizes the results of all five exon prediction engines. It enabled the detection of 44 confirmed exons (82%) in the *Fugu* contig and 78% of the human exons. In both cases it produced a remarkably low rate of false positives (3.7% in *Fugu* and 0.6% in human sequences) compared with GENSCAN. Thus, it gives a satisfactory overview on the coding capacity of human sequences as well as *Fugu* sequences.

Since there are no ESTs available yet for *Fugu* in any database ExonSampler, which searches for EST matches on the genomic



```
PCOLCE_FUG   M-GGYIFWG- -LFLFLSLSL GWTRAQQSTN SRPVFHCGGD LVKESGFVGS EGFPNSYKPN SKCTWRITVP EGNVVKLSFR IFNLEADSQC RYDYLDVYNG
PCOLCE_HUM   .LPAATASLL GPL.TACAL. PFAQG.TPNY T....L.... .VKG..Y.A. .......P.. KE.I.T.... ...QT.S... .V.D..LHPA ...A.E.FA.
PCOLCE_MOU   .LPAALTSFL GP..LAWVLP -LA.G.TPNY T....L.... .VTG..Y.A. ......L.P. .K..I.T... ...QT.S... .V.DM.LHPS ...A.E.FA.
PCOLCE_RAT   .LPAALTSLL GP..LAWVLP -LA.G.TPNY T....L.... .VTG..Y.A. ......L.P. .K..I.T... ...QT.S... .V.DM.LHPS ...A.E.FA.

PCOLCE_FUG   LSNLVQKLGR FCGTFRPGTL ISTTNTMMLE MVTDEETHSQ GFLAYFSAVK PYADDQQFCG GKITKSQGEL KTPNWPEKKY PPGTSCSWLI THEPGMVIQV
PCOLCE_HUM   SGTSG.R... .......AP. VAPG.QVT.R .T...G.GGR ...LWY.GRA TSGSEH.... ...RLE.A.T .T......SD .....I...H IAP.DQ..AL
PCOLCE_MOU   SGTSG.R... .......APV VAPG.QVT.R .T...G.GGR ...LWY.GRA TSGTEH.... ...RME.A.T .T......SD .....I...H IAPSNQ..ML
PCOLCE_RAT   SGTSG.R... .......APV VAPG.QVT.R .T...G.GGR ...LWY.GRA TSGTEH.... ...RME.A.T .T......SD .....I...H IAPSNQ..ML

PCOLCE_FUG   KFDKFLLEGD IYCRYDYVAF FNGGEKDDSH LIGKYCGDQA PEPITSSSNV MLVQFVSDLS LTSDGFLAYY TSFPRGTEVP TSDRGAG--- SRPVPPKPDV
PCOLCE_HUM   T.E..D..P. T.....S.SV ...AVS...R RL..F...AV .GS.S.EG.E L......... V.A...S.S KTL....A-K E-GQ.P.... --------PK
PCOLCE_MOU   T.G..DV.P. T.....S.SV ...AVS...K RL..F...K. .S..S.EG.E L......... V.A...S.S RTL..DAVEK E.ALSP.EDV Q.GPQSRS.P
PCOLCE_RAT   T.G..DV.P. T.....S.SV ...AVS...K RL..F...K. .S..S.EG.E L......... V.A...S.S RTL..DAVEK E.APSP.EDA QHGPQSRS.P

PCOLCE_FUG   R-PSEP-VRR PAPAKPPSPT EKPKPLKPTR NWGQSV-PQN P-L-CTKDCK RQGTIKSSFC A---MITGTV TSLTNGPRGT LLVSVSIIKT YKAGQLTITQ
PCOLCE_HUM   ..GT..K.KL .--P.SQP.- ..TEESPSAP DAP---.--- -.-T.P.Q.R .T..LQ.N.. .SSLVV.A.. K.MVRE.GEG .A.T..L.GA ..T.G.DLPT
PCOLCE_MOU   KTGTG.K.KP .TKP.SQPA- .T.EASPA.Q ATP--.A.AA .SIT.P.QY. .S..LQ.N.. SSSLVV.... KTMVR..GEG .T.T..LLGV ..T.G.DLPS
PCOLCE_RAT   KTGTG.K.KP .SKP.VQPV- ...EGSPA.Q ATP--.A.DA .SIT.P.QY. .S..LQ.N.. SSSLVV.... KAMVR..GEG .T.T..LLGV ..T.D.DLPS

PCOLCE_FUG   VGETMSVKLV SQCRKCPLLR RGGNYIIMGQ VDEDGRG-TL EPGSFTAPYK APHHRLLMNI NNQP------ -------
PCOLCE_HUM   PPTGA.L.FY VP.KQ..FMK K.VS.LL... .E.N-..PV. P.E..VVLHR PNQDQI.T.L SKRKCPSQPV RAAASQD
PCOLCE_MOU   PPSGT.L..Y VP..QM.FMK K.AS.LL... .E.N-..PI. P.E..VVL.R SNQD------ ----...... .......
PCOLCE_RAT   PASGT.L.FY VP.KQM.FMK K.AS.LL... .E.N-..PI. P.E..VVL.R PNQD------ ----...... .......
```

Fig. 2. Alignment of the PCOLCE-I protein sequences. The alignment was made using the CLUSTALW programme from the GCG package. The *Fugu* sequence is used as a consensus: dots (.) show the conserved amino acids, gaps are represented by hyphens (-). The exon/intron boundaries are drawn as vertical lines (|) and the exon numbers are written under the sequences. Dark grey boxes show the localisation of the two CUB domains, the light grey box is a newly detected domain localised in exon 9.

sequence, has to be used with a lower blast threshold (60% versus 75%) than for human sequences. With this modification ExonSampler is useful also for *Fugu* sequences for the confirmation of exons with experimental data from other species available in the databases.

To confirm gene structures further analyses are needed including cDNA and EST sequencing, expression analysis, and comparative studies. *Fugu* and human exons of homologous genes usually share the same intron/exon boundaries and often have the same exon length. Only the intron lengths are not conserved between the two species (Brenner et al. 1993). A detailed comparative analysis was undertaken for the PCOLCE-I gene of *Fugu* and human. The genomic sequence of the human PCOLCE-I gene (Accession No.: AF053356, Glöckner et al. 1998) has been previously reported. Aligning the human coding sequences with the *Fugu* genomic sequence using PROCRUSTES revealed the exon/intron structure of the PCOLCE-I locus in both organisms. COMPILE_EXON_ Strong detected only the first six exons. Only GENESCAN and MZEF predicted exon 7 in the *Fugu* cosmid sequence, reflecting perhaps the low similarity that this exon shows between species. Analysis of the cognate genes in rat and mouse show that this region of the PCOLCE-I protein (exon 6 and 7) is highly variable also between closely related mammals (see Fig. 2). It is likely that the two splice sites have shifted subsequently during evolution: i.e., the changes of the exon/intron boundaries were independent events. No further homologous regions were found outside the coding part of the gene.

A more detailed analysis showed two highly conserved domains present in both proteins (CUB domains). They have been preserved over 400 million years—the evolutionary distance between *Fugu* and humans—and obviously represent the main functional feature of this class of enhancer proteins. Previous studies showed that the CUB domains are necessary for the protein functions (Hulmes et al. 1997). In the C-terminal of the protein 26 aa are highly conserved between all organisms analysed (Fig. 2). The similarity in this region of 26 aa is 100%, between mammals; in *Fugu* the similarity drops to 80–85%. The surrounding regions show only 44% similarity. This is clearly a critical region for protein function. This highly conserved domain has never been detected before, due to the high overall homology in the C-terminal region between the mammalian PCOLCE-I proteins.

The *Fugu* GRMP gene is partially present in the analyzed sequence. Using the partial cds of a human protein KIAA0346 (Accession No.: AB002344) we were able to confirm some of the predicted exons. The human and *Fugu* GRMP proteins do not share homology at the 3′ end. It is known that the KIAA sequences (Nagase et al. 1997) are artificially deduced by overlapping ESTs. An error during the assembling of the KIAA0346 mRNA may explain the lack of homology at 3′ end with our *Fugu* protein and with the other human EST (containing the TRP motif). VAMP2 (Vescicle Associated Membrane Protein type 2) is a membrane spanning protein present in the synaptic vescicle. In mammals, four different forms of synaptobrevin (VAMP) have been described and all of them are highly conserved during evolution (DiAntonio et al. 1993). Interestingly, the VAMP-like proteins in yeast are more complex than in vertebrates (Linial and Parnas 1996). The *Fugu* MPP1 gene, also known as P55, was previously described (Elgar et al. 1995). The *Fugu* GDOB-L1 gene shows some degree of similarity with two *C. elegans* proteins, which are present on two discrete cosmids; these proteins do not share identical amino acid residues. Thus, we hypothesise that *Fugu* GBOD-L1 is a member of a new protein family of at least two members (the two different *C. elegans* proteins). The last gene is the *Fugu* homolog to the GABRB3 of which only 8 exons were located on the 66 kb sequence. Exon 7 was not detected by any of the prediction programs and it seems to be absent in the human and mouse homologous genes.

We also checked if the PCOLCE-I *Fugu* locus could be de-

Table 4. Genomic localization of genes homologous to the *Fugu* genes present at the PCOLCE-I locus.

	Localization[a]	
Gene name	*Homo sapiens*	*Mus musculus*
GRMP	—	—
PCOLCE-I	7q21	5 (80.0 cM)
VAMP2	17pter-p12	11 (S)
MPP1	Xq28	X
GBOD-L1	—	—
GABRB3	15q11-q13	7 (28.6 cM)

[a] All localization data were found at http://www.ncbi.nlm.nih.gov/Homology/.

fined as a syntenic region between the two organisms. While in human, the MPP1 gene and the factor VIII gene are located in close proximity, the two genes are not linked in *Fugu* since we could not identify a *Fugu* factor VIII homolog in proximity of the *Fugu* MPP1 (see Table 4). The five other genes detected in *Fugu* show no similarities with human genes present in the PCOLCE-I region on chromosome 7q22 (Glöckner et al. 1998). In mouse, the genes around PCOLCE-I are present in a syntenic region to the human chromosome 7q22 localized on the mouse chromosome 5.

In summary, our analysis shows that the pufferfish *Fugu rubripes* is a good model for studying gene structures. Comparative analysis between *Fugu* and humans is very useful for identification of functionally conserved protein motifs and domains.

References

Bateson AN, Harvey RJ, Bloks CC, Darlison MG (1990) Sequence of the chicken GABAA receptor beta 3-subunit cDNA. Nucleic Acids Res 18, 5557

Baxendale S, Abdulla S, Elgar G, Buck D, Berks M, et al. (1995) Comparative sequence analysis of the human and pufferfish Huntington's disease genes. Nat Genet 10, 67–76

Bonfield JK, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23, 4992–4999

Bork P (1991) Shuffled domains in extracellular proteins. FEBS Lett 286, 47–54

Bork P, Beckmann G (1993) The CUB domain. A widespread module in developmentally regulated proteins. J Mol Biol 231, 539–545

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S (1993) Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 366, 265–268

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268, 78–94

Burset M, Guigo R (1996) Evaluation of gene structure prediction programs. Genomics 34, 353–367

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162, 156–159

Coutelle O, Nyakatura G, Taudien S, Elgar G, Brenner S, et al. (1998) The neural cell adhesion molecule L1: genomic organisation and differential splicing is conserved between man and the pufferfish *Fugu.* Gene 208, 7–15

DiAntonio A, Burgess RW, Chin AC, Deitcher CL, Scheller RH, Schwarz TL (1993) Identification and characterization of Drosophila genes for synaptic vesicle proteins. J Neurosci 13, 4924–4935

Elgar G (1996) Quality not quantity: the pufferfish genome. Hum Mol Genet 5, 1437–1442

Elgar G, Rattray F, Greystrong J, Brenner S (1995) Genomic structure and nucleotide sequence of the p55 gene of the puffer fish *Fugu rubripes.* Genomics 27, 442–446

Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266, 540–553

Gelfand MS, Mironov AA, Pevzner PA (1996) Gene recognition via spliced sequence alignment. Proc Natl Acad Sci USA 93, 9061–9066

Glöckner G, Scherer S, Schattevoy R, Boright A, Weber J, Tsui LC, Rosenthal A (1998) Large scale analysis of two regions in human chromosome 7q22: annotation of 650 Kb of genomic sequence around the PCOLCE and CUTL1 loci reveals 17 genes. Genome Research 8, 1060–1083

Hinegardner R (1968) Evolution of cellular DNA content in teleost fishes. Am Nat 102, 517–523

Holland PW (1997) Vertebrate evolution: something fishy about Hox genes. Curr Biol 7, R570–R572

Hulmes DJ, Mould AP, Kessler E (1997) The CUB domains of procollagen C-proteinase enhancer control collagen assembly solely by their effect on procollagen C-proteinase/bone morphogenetic protein-1. Matrix Biol 16, 41–45

Kadler KE, Watson RB (1995) Procollagen C-peptidase: procollagen C-proteinase. Methods Enzymol 248, 771–781

Kamatchi GL, Kofuji P, Wang JB, Fernando JC, Liu Z, et al. (1995) GABAA receptor beta 1, beta 2, and beta 3 subunits: comparisons in DBA/2J and C57BL/6J mice. Biochim Biophys Acta 1261, 134–142

Kessler E, Adar R (1989) Type I procollagen C-proteinase from mouse fibroblasts. Purification and demonstration of a 55-kDa enhancer glycoprotein. Eur J Biochem 186, 115–121

Kim AC, Metzenberg AB, Sahr KE, Marfatia SM, Chishti AH (1996) Complete genomic organization of the human erythroid p55 gene (MPP1), a membrane-associated guanylate kinase homologue. Genomics 31, 223–229

Kivirikko KI, Millyla R (1984) *Extracellular Matrix Biochemistry.* (New York, Elsevier Science Publishing Co., Inc. New York)

Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214, 171–182

Linial M, Parnas D (1996) Deciphering neuronal secretion: tools of the trade. Biochim Biophys Acta 1286, 117–152

Mardis ER (1994) High-throughput detergent extraction of M13 subclones for fluorescent DNA sequencing. Nucleic Acids Res 22, 2173–2175

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477–478

Nagase T, Ishikawa K, Nakajima D, Ohira M, Seki N, et al. (1997) Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. DNA Res 4, 141–150

Ohira M, Ootsuyama A, Suzuki E, Ichikawa H, Seki N, et al. (1996) Identification of a novel human gene containing the tetratricopeptide repeat domain from the Down syndrome region of chromosome 21. DNA Res 3, 9–16

Platzer M, Rotman G, Bauer D, Uziel T, Savitsky K, et al. (1997) Ataxia-telangiectasia locus: sequence analysis of 184 kb of human genomic DNA containing the entire ATM gene. Genome Res 7, 592–605

Schuddekopf K, Schorpp M, Boehm T (1996) The whn transcription factor encoded by the nude locus contains an evolutionarily conserved and functionally indispensable activation domain. Proc Natl Acad Sci USA 93, 9661–9664

Solovyev V, Salamov A (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Ismb 5, 294–302

Takahara K, Osborne L, Elliott RW, Tsui LC, Scherer SW, Greenspan DS (1996) Fine mapping of the human and mouse genes for the type I procollagen COOH-terminal proteinase enhancer protein. Genomics 31, 253–256

Thomas A, Skolnick MH (1994) A probabilistic model for detecting coding regions in DNA sequences. IMA J Math Appl Med Biol 11, 149–160

Trimble WS, Cowan DM, Scheller RH (1988) VAMP-1: a synaptic vesicle-associated integral membrane protein. Proc Natl Acad Sci USA 85, 4538–4542

Trower MK, Orton SM, Purvis IJ, Sanseau P, Riley J, et al. (1996) Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). Proc Natl Acad Sci USA 93, 1366–1369

Uberbacher EC, Mural RJ (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA 88, 112261–11265

Uberbacher EC, Xu Y, Mural RJ (1996) Discovering and understanding genes in human DNA sequence using GRAIL. Methods Enzymol 266, 259–281

Wagstaff J, Chaillet JR, Lalande M (1991) The GABAA receptor beta 3 subunit gene: characterization of a human cDNA from chromosome 15q11q13 and mapping to a region of conserved synteny on mouse chromosome 7. Genomics 11, 1071–1078

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, et al. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature 368, 32–38

Ymer S, Schofield PR, Draguhn A, Werner P, Kohler M, Seeburg PH (1989) GABAA receptor beta subunit heterogeneity: functional expression of cloned cDNAs. Embo J 8, 1665–1670

Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA 94, 565–568