RESEARCH

Complete Genomic Sequence and Analysis of the Prion Protein Gene Region from Three Mammalian Species

Inyoul Y. Lee,[1] David Westaway,[4] Arian F.A. Smit,[1,6] Kai Wang,[1,7] Jason Seto,[1,8] Lei Chen,[1,7] Chetana Acharya,[1] Mike Ankener,[1] Dale Baskin,[1,9] Carol Cooper,[2] Hong Yao,[4] Stanley B. Prusiner,[2,3] and Leroy E. Hood[1,5]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA; Departments of [2]Neurology and [3]Biochemistry and Biophysics, University of California, San Francisco, California 94143-0518 USA; [4]Centre for Research in Neurodegenerative Diseases, University of Toronto, Toronto, Ontario M5S 3H2, Canada

The prion protein (PrP), first identified in scrapie-infected rodents, is encoded by a single exon of a single-copy chromosomal gene. In addition to the protein-coding exon, *PrP* genes in mammals contain one or two 5′-noncoding exons. To learn more about the genomic organization of regions surrounding the *PrP* exons, we sequenced 10^5 bp of DNA from clones containing human, sheep, and mouse *PrP* genes isolated in cosmids or λ phage. Our findings are as follows: (1) Although the human PrP transcript does not include the untranslated exon 2 found in its mouse and sheep counterparts, the large intron of the human *PrP* gene contains an exon 2-like sequence flanked by consensus splice acceptor and donor sites. (2) The mouse $Prnp^a$ but not the $Prnp^b$ allele found in 44 inbred lines contains a 6593 nucleotide retroviral genome inserted into the anticoding strand of intron 2. This intracisternal A-particle element is flanked by duplications of an AAGGCT nucleotide motif. (3) We found that the *PrP* gene regions contain from 40% to 57% genome-wide repetitive elements that independently increased the size of the locus in all three species by numerous mutations. The unusually long sheep *PrP* 3′-untranslated region contains a "fossil" 1.2-kb mariner transposable element. (4) We identified sequences in noncoding DNA that are conserved between the three species and may represent biologically functional sites.

[The nucleotide sequence data reported in this paper have been submitted to the GenBank sequence database and have been assigned the accession numbers U29185 (human), U29186 (mouse), and U67922 (sheep).]

The prion diseases include scrapie of sheep and goats, bovine spongiform encephalopathy (BSE), and Creutzfeldt-Jakob disease (CJD) of humans (Prusiner 1997). The wide variety of presentations manifested by the human prion diseases as genetic, sporadic, and infectious illnesses is unprecedented. As early as 1930, familial clusters of CJD were recognized (Meggendorfer 1930; Stender 1930), but the subsequent transmission of CJD to apes shifted thinking away from a genetic etiology to that of an atypical virus (Gibbs et al. 1968; Masters et al. 1981). Subsequently, the identification of mutations in the prion protein (*PrP*) gene (PRNP) of patients with familial prion disease demonstrated that these diseases are inherited while also transmissible (Hsiao et al. 1989; Prusiner 1989).

Over the past 15 years, many lines of evidence have converged to argue the transmissible prion particle is composed of an abnormal isoform ($PrPS^{Sc}$) of a normal cellular protein (PrP^C). PrP^{Sc} is formed from PrP^C by a post-translational process that occurs in caveolae-like domains (CLDs) near the surface of the cell (Gorodinsky and Harris 1995; Taraboulos et al. 1995; Vey et al. 1996). Through an as yet undefined process, PrP^{Sc} acts as a template for the refolding of PrP^C into a nascent molecule of PrP^{Sc}. The interaction of PrP^C with PrP^{Sc} is optimal

Present addresses: [6]Axys Pharmaceuticals, La Jolla, California 92087 USA; [7]Chiroscience, Bothell, Washington 98021 USA; [8]Institute of Biosciences, Bioinformatics, and Biotechnology, Manassass, Virginia 20110 USA; [9]PE Applied Biosystems, Foster City, California 94404 USA
[5]Corresponding author.
E-MAIL leehood@u.washington.edu; FAX (206) 616-5197.

 8:1022–1037 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

when the sequences of the two isoforms are the same as with homologous prion transmission within a single species (Prusiner et al. 1990). Heterologous transmission between species requires crossing the prion "species barrier" where the sequences of the two isoforms are different (Scott et al. 1997). The high degree of sequence similarity among ungulates seems to have facilitated transmission of ovine prions to cattle (Wilesmith et al. 1991).

The presence of the human and mouse *PrP* genes within conserved syntenic groups (Sparkes et al. 1986) and the presence of a *PrP* gene in chicken (Gabriel et al. 1992) argue that the *PrP* gene existed before the speciation of mammals. In mammals, DNA sequences of the ORFs encoding PrP generally exhibit ~90% similarity. As expected, the degree of similarity at the amino acid level increases to >95% when PrPs of different primates are compared (Schätzl et al. 1995) but is much lower when human PrP is compared with that of a marsupial (~70%) (Windl et al. 1995). An even lower degree of homology is found when human PrP is compared with that of the chicken (~30%) (Harris et al. 1989; Gabriel et al. 1992). Attempts to find *PrP*-related genes in lower eukaryotes have, to date, been unsuccessful (Oesch et al. 1991). In all species studied, the PrP ORFs encode proteins of ~250 amino acids. All PrP molecules seem to be post-translationally modified by removal of an amino-terminal signal peptide, cleavage of a carboxy-terminal signal sequence upon addition of a GPI anchor, and addition of two or three Asn-linked sugar chains (Stahl et al. 1987; Gabriel et al. 1992).

Because functionally related mammalian genes are often organized into gene complexes, one may use large-scale DNA sequencing to identify adjacent, perhaps functionally related, genes. With this in mind, we sequenced phage and cosmid clones encompassing the human, sheep, and mouse *PrP* genes. Primary objectives included the search for genes adjacent to *PrP* that might encode PrP-like proteins, "private" chaperones such as protein X that might interact with PrP^C (Telling et al. 1995; Kaneko et al. 1997), or a putative scrapie incubation time gene. We were also interested in analyzing the comparative chromosomal organizations of the *PrP* genes in humans, sheep, and mice. These sequence data define an exon 2-like sequence within the human *PrP* gene and two large transposon insertions within the transcribed regions of the mouse and sheep *PrP* genes. The organization of the repetitive elements in each species is complex and permits interesting evolutionary deductions.

RESULTS

Principal Features of the Nucleotide Sequence

Human PrP

To determine the complete genomic structure of the human *PrP* region, cosmid clone pGPrP1 (Puckett et al. 1991) was sequenced by shotgun sequencing methods using fluorescence-based automated DNA sequencers. The nucleotide sequence determined from cosmid clone pGPrP1 was 35,522 bp in length. This cosmid clone had two known *PrP* exons (134 and 2355 nucleotides in length) separated by an intron of 12,696 bases, and had a dG + dC content of 44.1%. Genomic sequencing allowed the identification of an additional exon-like sequence analogous to exon 2 of sheep and mouse (described below in Comparison of the Mammalian *PrP* Genes). Plotting the observed CG frequency/expected CG frequency ratio shows that a 1-kb region around exon 1 forms a "CpG island" where CG dinucleotides have been protected from their usual fast decay (Fig. 1A).

The human *PrP* region was evaluated for interspersed repetitive elements using a library of human repeat sequences and the program cross_match as described in Methods. The analysis revealed that at least 40% of the human locus seems to be derived from mobile elements (see Table 1). The region has an equal density (11%) of *Alu* and LINE1 elements, which is close to the genome-wide average for both. None of the short interspersed repetitive elements (SINEs) or long interspersed repetitive elements (LINEs) are expected to be dimorphic in the human population; in fact, 14 of the 15 young SINEs (*Alu* sequences) belong to subfamilies (*Alu*Sx, *Alu*J, and monomers) that are thought to be >35 million years old (Kapitonov and Jurka 1996). The youngest L1 element belongs to the PA13 subfamily that probably spread before this time as well (Smit et al. 1995). In addition to the interspersed repetitive elements, potentially polymorphic simple sequence repeats (SSRs) were identified using the sputnik program (Abajian 1994). SSRs or microsatellites are used as genetic markers because of their high degree of polymorphism in length (Tautz 1989). There were eight SSRs at least 10 nucleotides in length detected (shown in Fig. 1A). Of these, one tetranucleotide repeat, AAAC, was found in intron 1 and one dinucleotide repeat, AG, was found in large intron 2. Two of the SSRs appear to be associated with *Alu* elements.

Sheep PrP

The genomic sequence of 31,412 bp corresponding

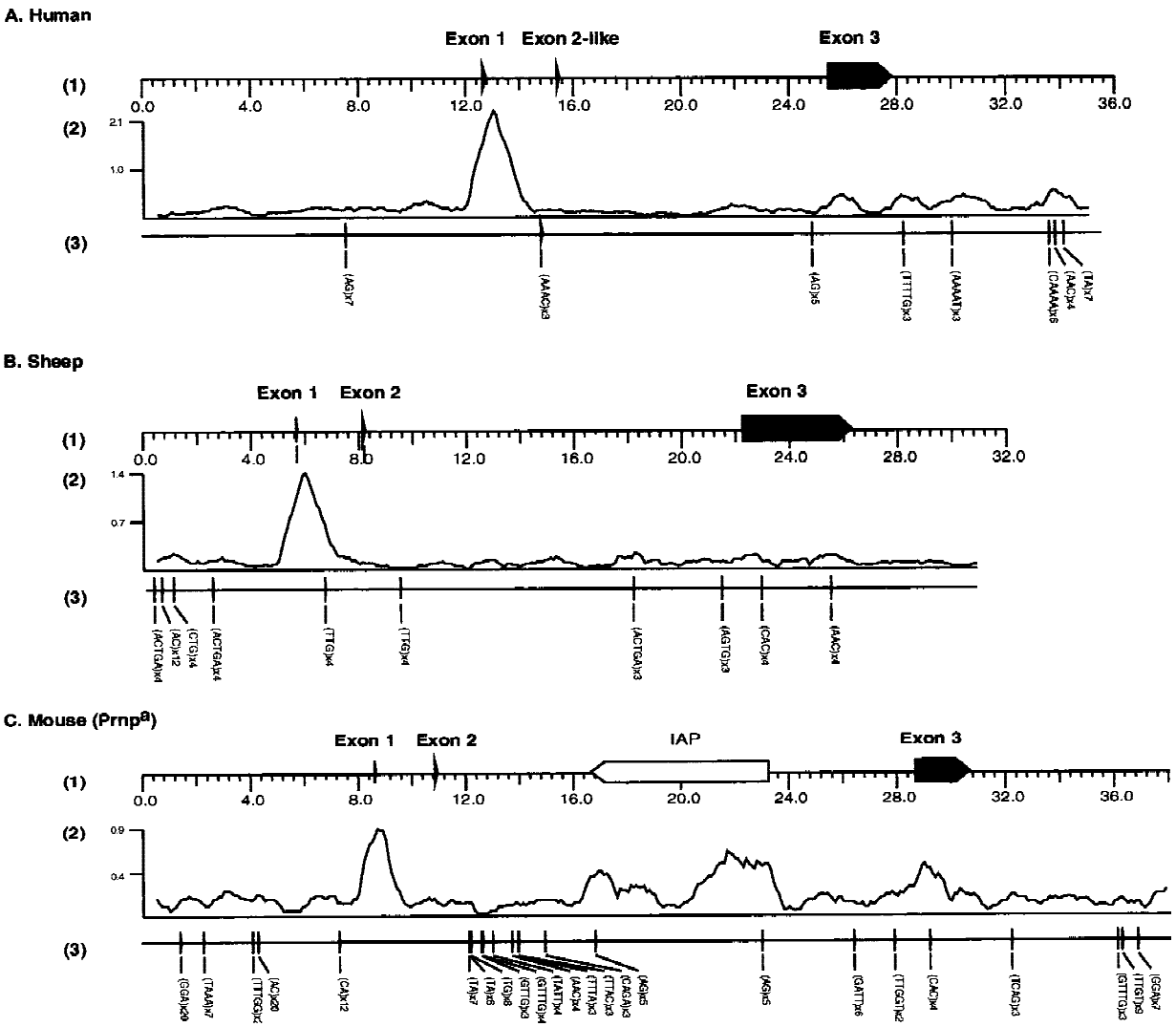

Figure 1 A summary of features found in *PrP* loci. (*A*) The human *PrP* locus. (*B*) The sheep *PrP* locus. (*C*) The mouse locus (*Prnp*[a]). (*A–C*) (*1*) The positions of the exons in each species. The scale is in kb. (*2*) The frequency of obseved CpG dinucleotides divided by frequency of expected CpG . CpG frequencies were calculated in a window length of 1000 base positions and an offset length of 100 bp. Note that an unusually high CpG frequency is found in the promoter region of each *PrP* gene. The mouse *PrP* gene has a second high peak in intron 2 owing to the IAP insertion. (*3*) Simple repeats at least 10 nucleotides in length.

to the sheep PrP cosmid from the scrapie-susceptible Suffolk breed confirms and extends earlier studies (Westaway et al. 1994c) (Fig. 1B). The sheep *PrP* gene has three exons (52, 98, and 4028 nucleotides in length) separated by two introns (2421 and 14,031 nucleotides in length). The 3′-untranslated region (UTR) of the sheep *PrP* mRNA is much longer than that of other analyzed species (3246 bp). Surprisingly, these sequences differed at 82 positions from those reported by Goldmann and coworkers (1990) (GenBank accession no. M31313). Although errors in our sequence cannot be excluded, at least one highly "variable" region between nucleotides ~25,800 and 26,200 was covered by six to eight sequences of independent M13 clones. Furthermore, our sequence is fully concordant with the partial 3′ UTR sequence of Shinagawa et al. (unpubl.; GenBank accession no. D38179) between nucleotides 23,049 and 23,725, an area that includes four disparities with the sequence of Goldmann et al. Alternatively, errors may exist in the earlier sequence for this region, or the 3′ UTR, presumably under minimal selection pressure, may evolve rapidly.

The extended length of the sheep *PrP* mRNA appears because of the integration of three transposable elements in the 3′ UTR during artiodactyl evo-

Table 1. Repeat Insertions in the *PrP* Loci

	SINEs		LINE1		other LINEs		DNA *trans*		LTR elements		Other		Total	
Total														
Human	18	11.5%	6	11.7%	4	6.5%	7	3.4%	4	7.1%	0	0.0%	39	40.2%
Mouse	32	13.3%	5	9.5%	2	2.7%	4	1.7%	10	9.8%	1	3.6%	53	40.4%
Sheep	12	5.9%	4	13.1%	9	10.8%	6	6.1%	3	20.3%	1	0.7%	34	57.1%
Order specific (young)														
	Bov-B												**Total young**	
Human	15	10.8%	3	5.7%	0	0.0%	1	1.3%	1	2.5%	0	0.0%	20	20.2%
Mouse	29	12.5%	2	3.2%	0	0.0%	0	0.0%	8	7.5%	1	3.6%	39	26.7%
Sheep	10	5.2%	1	7.1%	5	5.9%	1	3.9%	2	18.5%	1	0.7%	19	41.3%
Shared (old)														
	SINEs (MIR)		**LINE1**		**LINE2**								**Total old**	
Human	3	0.7%	3	6.0%	4	6.5%	6	2.2%	3	4.7%	0	0.0%	19	20.0%
Mouse	3	0.8%	3	6.2%	2	2.7%	4	1.7%	2	2.3%	0	0.0%	14	13.7%
Sheep	2	0.7%	3	6.0%	4	4.9%	5	2.3%	1	1.8%	0	0.0%	15	15.7%

Character of the insertions in the *PrP* loci. Almost all are insertions of SINEs, LINEs, retrovirus-like elements, and DNA transposons (see Smit 1996). The two exceptions are a laminin receptor mRNA that retroposed in the mouse locus, and the unclassified interspersed repeat MER21 in the sheep locus. Listed are the number of copies of these elements in the three *PrP* loci and the fraction of these loci they comprise (mouse DNA limited to the region for which the human orthologous region is known). The data have been subdivided in order-specific elements, which are only found in one species, and elements that are shared between species, representing the offspring of transposable elements that were active after and before the mammalian radiation, respectively. The mouse data do not include the IAP insert in the $Prnp^a$ allele. With the IAP, it brings the total transposon-derived fraction to 52%.

lution (Fig. 2): (1) a LINE-like element Bov-B (Szemraj et al. 1995; Smit 1996), (2) an artiodactyl SINE, Bov-tA (= art-2) (Duncan 1987), and (3) a mariner DNA transposon relic (Smit and Riggs 1996) (see Fig. 3). The deduced amino acid sequence of the transposase from this relic places the element in the *Mellifera* (honeybee) subfamily of mariner elements (data available on request). The mariner fossils recently described in the human genome belong to two other subfamilies (Auge-Gouillou et al. 1995; Morgan 1995; Oosumi et al. 1995; Smit and Riggs 1996). The mariner transposase pseudogene contains seven or eight frameshifts and five stop codons (Fig. 3), suggesting that the insertion is relatively old and probably is shared by all ruminants. A recently submitted bovine *PrP* cDNA sequence (GenBank accession no. AB001468) contains the mariner element as well, although the original bovine cDNA entries in GenBank (accession nos. D10612 and D90545) (Yoshimoto et al. 1992) were partial and did not extend to the mariner integration site.

Mouse PrP: An IAP Insertion

Mouse genomic sequence of the $Prnp^a$ region was obtained through shotgun sequence analysis of three overlapping clones: two λ clones, λ 4 and λ 7, and one plasmid clone derived from a λ clone 6 (all derived from a genomic library constructed from inbred strain 129 Sv DNA) (Westaway et al. 1994a). Three overlapping contigs were assembled independently and merged to a 38,418-bp contig (Fig. 1C). The $Prnp^a$ gene has three exons (47, 98, and 2008 nucleotides in length) separated by two introns (2190 and 17,733 nucleotides in length). The GC content of the $Prnp^a$ region was 45%. Repetitive sequence analysis of the mouse *PrP* gene region revealed that a larger number (29) of more variable SINEs have integrated in this locus than in the human and sheep loci (Table 1). Whereas all 15 primate-specific SINEs are members of the *Alu* family, and the 10 SINEs in sheep belong to three subfamilies of ArtSINE1, at least five rodent-specific types of SINEs are present in the mouse locus, two of which are newly described here (see below; Fig. 4). Microsatellite repeats at least 10 nucleotides in length were also identified. There are 24 microsatellite repeats in the $Prnp^a$ region (Fig. 1C).

Several restriction site polymorphisms have

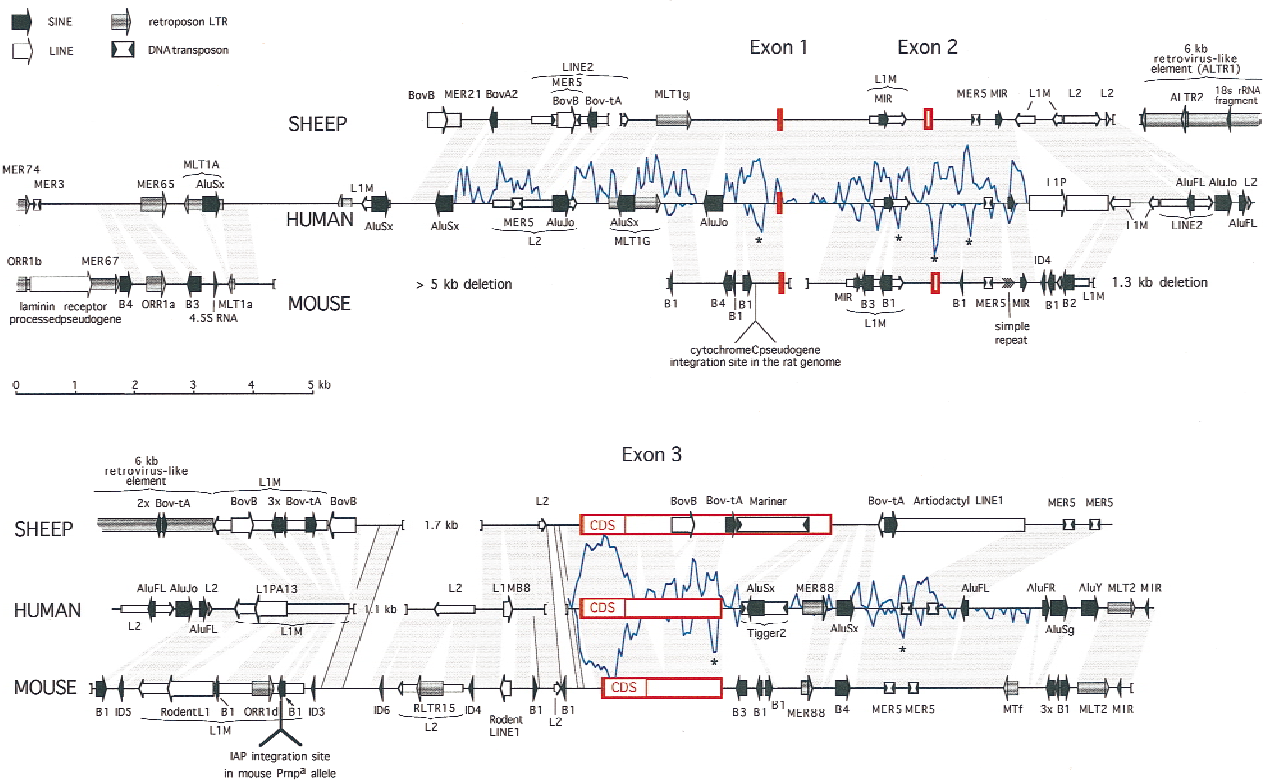

Figure 2 Comparison of the human, sheep, and mouse *PrP* locus sequences. The complete sheep and human sequences are shown, whereas the *Prnp*^{a} mouse cosmid sequence extends 2.5 kb upstream of the aligned regions. Sequences orthologous between human and either mouse or sheep are indicated by shaded areas. In two regions deleted in the human sequence, orthologous sequences between sheep and mouse are indicated with outlined shading. Deletions in the genome >250 bp are indicated by brackets. Repeats were identified as described in Methods. Identified interspersed repeats are represented by four differently shaded arrows for the four major classes of repeats: SINEs, LINEs, LTR elements, and DNA transposons. Wide arrows are used for elements that are absent at the orthologous sites in the other two species and thus probably integrated after the mammalian radiation. Older elements are indicated with thinner arrows. The indicated shared interspersed repeats are usually not detectable in the mouse genome (and sometimes not in the sheep genome) by direct comparison with the repeat consensus sequence but could be inferred from the alignment with human DNA. The mariner and large retroviral insert in sheep and the multiple SINEs in mouse are discussed in the text and Figs. 3 and 4. Briefly, other nonfamiliar, but common, inserts are as follows: *Alu*FL and *Alu*FR in human are the older, free left and right arm monomeric precursors of the *Alu* dimer, Bov-B in sheep is a LINE-like element specific to true ruminants (Szemraj et al. 1995; Smit 1996), and the mammalian-wide LINE2 (L2) elements were responsible for the distribution of the MIR SINEs (A. Smit, unpubl.). All newly derived repeat consensus sequences and descriptions have been incorporated in RepBase Update (Jurka and Smit 1997). The location of the cytochrome *c* pseudogene in rat is based on sequence data of Saeki et al. (1996). A measure of similarity between the human and either sheep or mouse sequences, depicted by a graph above and below the human sequence, respectively, is derived from raw cross match (Smith–Waterman) scores of 100-bp fragments (overlapping by 50 bp) of human sequence with their orthologous sites in sheep and mouse using a matrix. The lowest score above which virtually all matches had been found to be significant during construction of the RepeatMasker program (A. Smit and P. Green, unpubl.), was used as a cutoff score and subtracted from the total score to form the baseline. Regions outside the coding region that are conserved between all species are indicated with an asterisk (*) and were studied in detail. Alignments of the most interesting regions are shown in Fig. 6.

been mapped 5′ to the mouse *PrP* coding region within intron 2 (Westaway et al. 1987; Carlson et al. 1988). Subsequent studies of cloned DNA revealed that *a* and *b* alleles of *Prnp* differed in size by 6.6 kb in the polymorphic region. The *a* allele, found in 44 inbred mouse strains, corresponds to the larger size variant (Carlson et al. 1988). The smaller version of intron 2 present in *Prnp*^{b} mice (e.g., the I/LnJ strain) was interpreted to represent a deletion event (Westaway et al. 1994a). Whereas the complete genomic

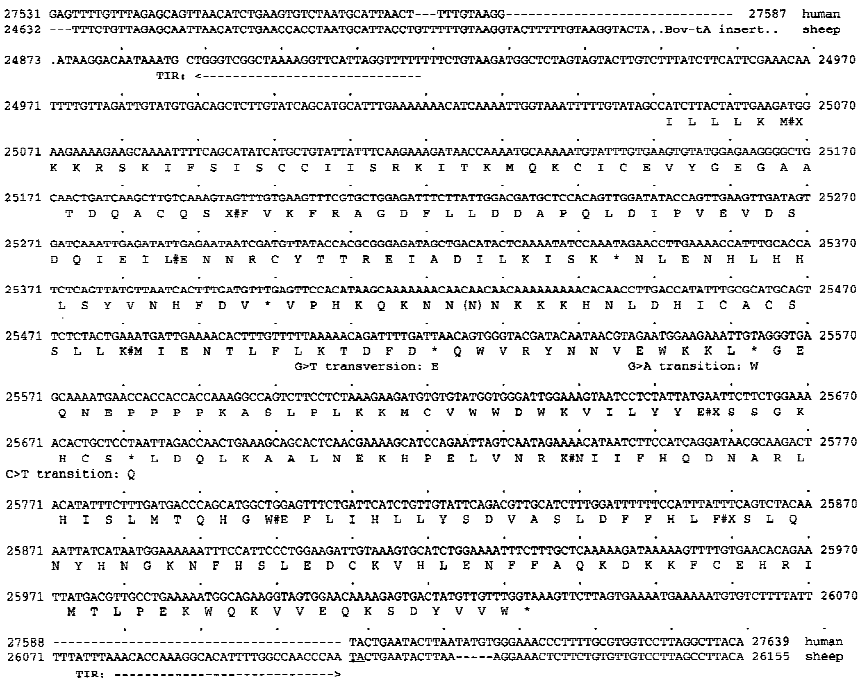

Figure 3 The mariner relic in the 3′ UTR of the sheep *PrP* transcript. Translation of this sequence was guided by closely related mariner transposases of the *Mellifera* subfamily. Introduced frameshifts are indicated with a pound sign (#). The likely origin of stop codons is indicated. The terminal inverted repeats are indicated by arrows. The TG...TA flanking dinucleotides may be derived from the typical mariner TA target site. The orthologous human sequence is aligned with the flanking sites.

sequence of a $Prnp^b$ cosmid clone will be described elsewhere, genomic sequencing of $Prnp^a$ clones defines this intron 2 size polymorphism (17,733 bp long in $Prnp^a$ mice vs. ~11,000 bp in $Prnp^b$ mice) and excludes a deletion event, instead revealing an intracisternal A-particle (IAP) element insertion in the wild-type *a* allele 5404 bp upstream of the translation start codon of exon 3 (Fig. 1C).

IAPs are defective murine retroviruses, unable to spread by horizontal infection, and encoded by a gene family of ~1000 members present in the genome of many mouse species (for review, see Kuff and Lueders 1988). The IAP element inserted in the large intron 2 of mouse $Prnp^a$ ($Prnp^a$–IAP) was 6593 bp long, in very close agreement with restriction mapping studies noted above. The IAP element is inserted in the opposite transcriptional orientation of the *PrP* gene (bases 16,687–23,279 of GenBank accession no. U29186) and is flanked by a 6-bp duplication of cellular sequences, AAGGCT.

$Prnp^a$–IAP is 98.6% similar over its full length to an IAP element integrated in the mouse T cell receptor α locus (bases 31220–38277 of GenBank accession no. AC003993), including a 558-bp deletion

We also compared the $Prnp^a$–IAP sequence with a cDNA of an IAP-related mRNA that is upregulated in scrapie-infected neuroblastoma cells (Doh-ura et al. 1995); these analyses revealed minimal homology centered around the pol genes (not presented). The cDNA described by Doh-ura et al. was 97% similar with a different class of IAP-related retroviral elements (the IAP-E element, GenBank accession no. M73818), indicating that the cDNA does not originate from within Prnp.

Other Retrovirus-like Elements in the *PrP* Regions

Database searches with the repeat-masked sequences revealed the presence of several not yet described genome-wide elements resembling retroviral elements or having features characteristic of retroviral LTRs, that is, TG...CA terminal dinucleotides flanked by 4- to 6-bp insertion site duplications, a consensus polyadenylation signal, and easily distinguishable subfamilies that often show mosaic relationships (Smit et al. 1995).

Intron 2 of the sheep *PrP* gene appears to contain a 6-kb insertion of a retrovirus-like element, only 500 bp from the IAP insertion site in the $Prnp^a$ mouse intron 2 (Fig. 2). The insert has 80% conserved terminal direct repeats [here named artiodactyl LTR 1, (ALTR1)] that are found interspersed at

A

```
                              Box A                                  Box B
                            TGGCNAGGTGG                            GGTTCGANNCC

 B2            GGGCTGGAGAGATGGCTCAGTGGTTAAGAGCACCTGACTGCTCTTCCAGCGGTCCTGAGTTCAATTCCCAGCAACCACA 79
 B3           GGGGCTGGAGAGATAGCTCAGCGGTTAAGAGCAC-TGGCTGCTCTTCCAGAGGACCCGGGTTCGGTTCCCAGCACCCACA 79

 B4           GGGGCTGGGGAGATGGCTCAGTCGATAA-AGTGCTTGCCGTACAAGCATGAGGACCTGAGTTCGATCCCCAGAACCCACG 81
 ID_mouse     GGGGTTGGGGATTTAGCTCAGTGGT--AGAGCGCTTGCCTAGCAAGCGCAAGGCCCTGGGTTCGGTCCTCAGCTCTGGAA 78
 ID5          GGGGCTGGGGATGTAGCTCAGTGGT--AGAGCGCTTGCCTAGCATGTATGAGGCCCTGGGTTCGATCCCCAGCATCTCCA 78
 4.5S RNA  38 GGGCTGGAGAGATGGCTCAGCCGTTAA-AG-GCTAGGCTCACAACCAAAAATATAAGAGTTCGGTTCCCAGCACCCACG 114

tRNA-Gln          GGTTCCATGGTGTAATGGTTA-GCACTCTGGACTCTGAATCCAGCGA-TCCGAGTTCAAATCTCGGTGGAACCT(CCA)
tRNA-Lys          GCCCGGCTAGCTCAGTCGGTA-GTGCATGGGACTCTTAATCCCAGGGTTGTGGGTTCGAGCCCCACGTTGGGCGCCA
MIR                RCAGYATAGYATAGTGGTTAAGAGCAC-GGRCTCTGGAGCCAGACTGCCTGGGTTCGAATCCCGGCTCTGCCA
```

B

```
B4      1 GGGGCTGGGGAGATGGCTCAGTCGATAAAGTGCTTGCCGTACAAGCATGA 50
                    vi i       - i  i  i      vvi  v i
ID5     1 GGGGCTGGGGATGTAGCTCAGT-GGTAGAGCGCTTGCCTAGCATGTATGA 49

B4     51 GGACCTGAGTTCGATCCCCAGAACCCACGTAAAAAA 86
            v    i              v i iv iv
ID5    50 GGCCCTGGGTTCGATCCCCAGCATCTCCAAAAAA 83 (end)

B4     87 GCCGGGCGTGGTGGTGTACGCTTGTAATCCCGGCGCTGGAGAGGCGGAGA 16
                 i      i i    i v      i  i  v i     i   i
B1      1 GCCGGGCATGGTGGCGCACGCCTTTAATCCCAGCACTTGGGAGGCAGAGG 50

B4    137 CAAGCGGATCCCTGGGGCTCGCTGGCCAGCCAGCCTAGCCGAMTCGGCGA 186
            i      ii   i -i   v-----i         i i v ?v?i i
B1     51 CAGGCGGATTTCTGAG-TTCGA-----GGCCAGCCTGGTCTACANAGTGA 94

B4    187 GTTCCAGGTCAGT------------GAGAGACCCTGTCTCAAAAAACAAR 224
                  v   i------------    i           i    v i
B1     95 GTTCCAGGACAGCCAGGGCTACACAGAGAAACCCTGTCTCGAAAAAAAAA 144 (end)

B4    225 GTGGAGAGTGCCTGAGGAAGACACCCGGGGTTGACCTCTGACCTCCACA(CA)n 275 end
```

Figure 4 Two novel tRNA-derived SINEs in the rodent genome denoted B3 and B4. (*A*) Comparison of the tRNA-like region in rodent SINEs. It is noteworthy that the tRNA-like regions of five rodent SINEs are more closely related to each other and the mammalian-wide MIR SINE (Smit and Riggs 1995) than to any known tRNA. Similarity of the B3 to B2 elements stretches beyond the tRNA-like region. The first 135 bp of the B3 consensus sequence is 84% similar to bp 1–124 of the B2 element, but the terminal 93 bp is unrelated. (4.5s RNA) The rat gene for a small nuclear RNA in rodents that has produced many pseudogenes (Saba et al. 1985; Takeuchi and Harada 1986). Indicated above the alignments are the consensus polymerase III promoter A and B boxes. (*B*) The B4 element appears to be a fusion product of an ID element, a B1 element, and an unrelated flanking region. ID5 is a consensus sequence of one of six ID subfamilies identified in the mouse genome, where higher numbers indicate higher divergence and older age (ID2–ID6 are all present in both rat and mouse) (Smit et al. 1995). Mismatches in the alignment are indicated with i (transition) and v (transversion).

some other loci in the cow and pig genome. The high divergence of these LTRs suggests that the insertion was a relatively early event in the artiodactyl evolution of the *PrP* locus, a notion corroborated by the interruption of the internal sequence by a second LTR-like interspersed repeat (ALTR2) in the internal sequence that appears elsewhere at orthologous sites in the sheep, cow, and pig CYP11A (P-450-scc) mRNA. Both the LTRs and the internal sequence of the retrovirus-like insert in the sheep *PrP* gene show similarity to members of the "MER4 group" of LTR elements (Smit 1996). These elements are characterized by 4-bp target site duplications and up to 7.5-kb-long internal sequences apparently lacking any coding capacity (V. Kapitonov and A.F.A. Smit, in prep.). The MER65 LTR in the human *PrP* gene and the MER67 LTR in the mouse gene (Fig. 2) are members of this group as well. Although both represent families of nonautonomous LTR elements, the MER4 group is not related to the mammalian LTR-retroposon (MaLR) family (Smit 1993).

Two additional new LTR-like repeats identified in the human locus (MER74 and MER88; Fig. 2) are related to two consensus sequences in the human repeat database (MER73 and MER54). We found that all these elements are LTR-like and have 5-bp insertion site duplications, but we could not yet identify an internal sequence.

Search for Protein-Coding Regions Adjacent to *PrP*

Three approaches were used to locate possible coding regions adjacent to the *PrP* gene. First, BLAST searches with the repeat-masked PrP region against the nonredundant databases identified *PrP* coding regions in each DNA sequence and pseudogenes: an 18S rRNA pseudogene in intron 2 of the sheep *PrP* gene and a laminin receptor pseudogene upstream of the mouse *PrP* gene.

The second type of analysis identified large ORFs >300 bp in length. Including *PrP* and IAP protein-coding regions, there are 39 ORFs in the human *PrP* gene (1 ORF/902 bp), 42 ORFs in the mouse *Prnp*[a] (1 ORF/915 bp), and 33 ORFs in the sheep gene (1 ORF/952 bp). With one exception, that is, *PrP*, none of these ORFs are conserved between species, suggesting that they do not have functional significance.

Third, we searched for potential exons using the gene-finding program GRAIL 2 (Uberbacher and Mural 1991) and a program accessed through the GENSCAN Web server (http://gnomic.stanford.edu/GENSCANW.html). Both programs identified the *PrP* coding regions (but not the anticoding strand ORF, present in three *PrP* alleles analyzed here) but did not reveal any other candidate coding regions

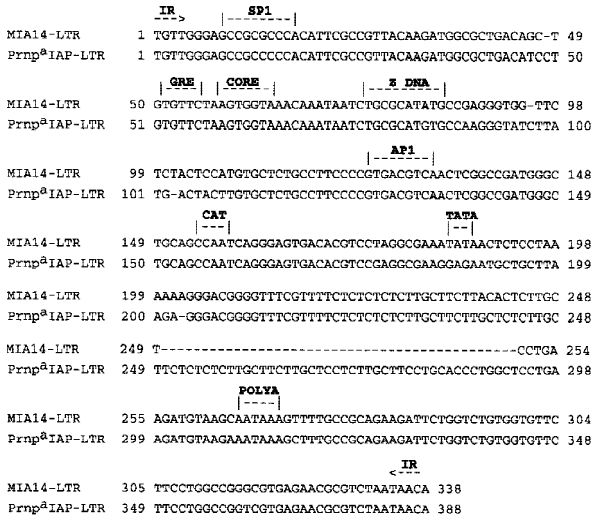

Figure 5 Sequence alignment of IAP LTRs. ''IR'' indicates 4-bp inverted repeat. Note that *Prnp*^{a}–IAP LTRs contained several putative binding sites found in other IAP LTRs: (1) a consensus binding site (GCCGCCCCCA) for the transcription factor SP1; (2) a binding site (TGTTCT) for the glucocorticoid responsive element (GRE) of mouse mammary tumor virus (MMTV) (Scheidereit et al. 1983); (3) (CORE), Transcription enhancer sequence (AGTGGTAAA) that shares homology with the SV40 enhancer motif [TGTGGTA(T/A)] (Weiher et al. 1983); and (4) a consensus binding site (GTGACGTCA) for the transcription factor AP1 (Lee et al. 1987). (CAT and TATA) Promoter boxes that together may determine the transcriptional initiation site. (POLY A) indicates polyadenylation site.

common to the three species. Experiments to extend this analysis to more distal areas of the 5′- and 3′-flanking sequences are in progress.

Comparison of the Mammalian *PrP* Genes

Multiple Alignment of the PrP *Regions of Human, Mouse, and Sheep*

Figure 2 shows a comparison of the PrP loci in human, sheep, and mouse. Full alignment of both the sheep and the mouse locus with the human sequence was possible using the Smith–Waterman-based program cross_match (P. Green, unpubl.) and the alignment parameters optimized in the development of RepeatMasker. The alignments were facilitated by deletion of all order-specific repeats prior to alignment. We deleted the order-specific repeats in each sequence and aligned 100-bp fragments of the human sequence with the orthologous sequences in mouse and sheep using the program cross_match (for details, see legend to Fig. 2). The Smith–Waterman score for these alignments is a better (relative) measure of conservation than simply the percent similarity between the sequences, because this score reflects, besides the percent matches, the number and length of insertions and deletions and the relative amount of transitions and transversions. The score for each alignment is plotted in Figure 2. From the comparison of three species, one can usually infer whether the absence of sequences in one species is owing to deletions in this species or to insertions in the others. All insertions in the three species are the result of integration of known classes of transposable elements or processed pseudogenes (see Fig. 2; Table 1). Insertion of these elements has increased the size of the human *PrP* locus by 20% and both the mouse and sheep loci by 41% since the mammalian radiation (Table 1), although some large deletions have partially reversed this size increase. In almost all cases, repetitive sequences belonging to subfamilies predicted to predate the mammalian radiation are found at orthologous sites in all species, whereas predicted younger elements are found in only one species. An apparent exception is a short L1MB8 element at the end of intron 2 that is predicted to be fixed in all mammals (Smit et al. 1995) but is precisely absent in both the sheep and mouse sequences. Another exception is the MER88 LTR 3′ of the human *PrP* gene, which is precisely absent in sheep but partially present in the mouse locus. The first observation may indicate more recent activity of a branch of LINE1 elements resembling L1MB8, and the second could signify a precise deletion in the artiodactyl lineage or integration in an ancestor common to rodents and primates only.

Intron 2 has a particularly high density (50%–71%) of interspersed repeats and has undergone multiple insertions and deletions in all species (Table 2), suggesting that changes in length and base composition of this intron contribute little to the function of the gene. Thus, the insertion of an IAP element in this intron in the *Prnp*^{a} allele of mouse may have little effect on the *PrP* gene activity. Another informative deletion is the large (at least 5-kb) deletion 1.5 kb 5′ to the first exon in mouse, which makes it less likely that the corresponding region in human and sheep contains essential regulatory sequences. Recently, the rat *PrP* promoter region sequence has been published, revealing the presence of a processed 1.1-kb cytochrome *c* pseudogene just upstream of the conserved promoter region (Saeki et al. 1996). These data also suggest that most of the upstream regulatory elements of the *PrP* gene promoters may lie within the 1.5-kb region immediately 5′ to the first exon.

Table 2. Insertions and Deletions in *PrP* Gene Intron 2

Locus	Length (bp)	Percent repeats	Events since mammalian radiation (kb)			original length (kb)
			inserts	deletions	net gain	
Human	9,970	60	2.5	1.7	0.9	9.1
Mouse	18,012	69	10.2	1.3	8.9	9.1
Sheep	14,031	71	7.4	2.5	4.9	9.1

Insertions and deletions in intron 2 of the *PrP* genes. For each species the total number of bases in intron 2 occupied by species-specific interspersed repeats was calculated. The size of the deletions was calculated from the size of the orthologous regions in the other species, disregarding species-specific repeats in those regions. In all three species one comes to an estimate for the length of this intron before the mammalian radiation of 9.1 kb.

Table 3 lists the divergence between the different *PrP* gene regions in the three species. In all regions, the human and sheep sequences are more similar to each other than to the mouse sequence. Also noteworthy is that the divergence between mouse and human is less than between mouse and sheep in all regions but exon 2 (see below).

In addition to multiple sequence alignments, dot-matrix analyses were performed to compare the DNA sequences of the entire *PrP* regions from different species (not shown). The diagonal line of similarity across the entire *PrP* region of human and sheep indicates that noncoding as well as coding sequences are conserved in these two species. This observation stands in contrast to analogous pairwise analyses of human/mouse and sheep/mouse *PrP* genes, which do not show such high levels of similarity in noncoding sequences.

Identification of a New Human PrP *Exon and Putative Functional Elements of the* PrP *Gene Region*

Regions that are conserved between all three species may indicate functionally conserved sites. Regions

Table 3. Divergence and Substitution Levels Between the *PrP* Gene Regions in the Three Species

	Human–sheep			Human–mouse			Sheep–mouse		
	length	div.	*K*	length	div.	*K*	length	div.	*K*
5′ flanking	4,317	0.253	0.317	1,727	0.283	0.369	648	0.290	0.373
Exon 1	120	0.242	0.292	82	0.366	0.533	0	>0.370	>0.530
Intron 1	1,920	0.249	0.308	1,330	0.307	0.403	1,048	0.327	0.450
Exon 2	97	0.206	0.251	98	0.265	0.340	95	0.221	0.275
Intron 2	4,723	0.247	0.307	4,586	0.302	0.402	4,227	0.347	0.494
Coding region	733	0.119	0.131	738	0.131	0.146	747	0.165	0.190
Synonymous Substitutions[a]			0.489			0.594			0.709
Nonsynonymous Substitutions[a]			0.050			0.063			0.080
Exon 3 UTR	1,325	0.197	0.236	1,150	0.265	0.338	1,125	0.283	0.372
3′ flanking	2,101	0.281	0.362	3,127	0.314	0.425	1,416	0.342	0.488
Overall	15,336	0.243	0.302	12,838	0.290	0.383	9,306	0.316	0.438

Divergence and substitution levels between the *PrP* gene regions in the three species. Insertions and deletions are ignored in these numbers. The indicated length is the total number of bases matched or mismatched over the given region between the given species (i.e., bases opposite gaps are excluded). The divergence (div.) is the level of mismatches among these bases. *K* is the actual substitution level calculated using the two-parameter method of Kimura (1980).

[a]For the synonymous and nonsynonymous calculations, we used divergence in the GCG package (v 8).

showing unusually high Smith–Waterman scores for both the human–sheep and the human–mouse alignments, indicated with asterisks in Figure 2, were further investigated. Clearly, the coding and promoter regions are highly conserved. However, the first exon and the beginning of intron 1 are poorly conserved. The mouse exon 1 could hardly be aligned with its human or sheep counterparts. Among the apparently conserved regions is a 99-bp region in human DNA, which starts 2303 bp after exon 1 and corresponds to exon 2 in sheep and mouse (Fig. 6B). This region is not present in human *PrP* mRNAs analyzed to date. Like the authentic mouse and sheep sequences, the human exon 2-like sequence is flanked by consensus splice donor and acceptor sites. It does not contain a translation start codon (ATG) in the proper translational reading frame with the human *PrP* exon 3 and forms part of the 5′ UTR of the mRNA. The sequence of the new human exon 2 is 82% and 81% identical to exon 2 in the sheep and mouse, respectively. To determine whether mRNA containing this new exon 2 was expressed, Southern blot analysis was performed on RT–PCR products generated by a number of permutations of exon 1, 2, and 3 primers. To date, exon 2-containing mRNAs have not been identified in cDNA from whole brain and cerebrum cortex (Strategene). These data indicate that either this is a functional exon expressed in mRNA not yet analyzed or the transcript concentrations are too low to detect under conditions we used. Alternatively, the environment of the human exon 2-like sequence precludes splicing into mature mRNA.

There are three other conserved regions: one 500 bp upstream of "exon 2" (Fig. 6A), one at the 3′ end of the transcripts (Fig. 6C), and one 3–4 kb downstream of the polyadenylation site (Fig. 6D). We also looked for the longest exact matches outside the coding region, almost all of which coincide with regions conserved in all three species (Fig. 2).

Sequence data presented in this paper did not reveal any elements >10 nucleotides in length conserved in the 5′-flanking region of all three species beyond the four motifs in the *PrP* gene promoter described previously (Westaway et al. 1994a). At this time, the transcription factor LyF-1 can be considered a candidate for binding to the *PrP* promoter region, because the consensus binding site TCAGGGAG is identical to motif 4 (TCAGGGAG) (Hahm et al. 1994). Introns contain three regions that are conserved in all species, two of which are shown in Figure 6A. The third conserved block was found in the beginning of intron 2 (between bp 15,781 and 16,135 in human, 8541 and 8873 in sheep, and 11,398 and 11,668 in mouse). These regions could be involved in transcriptional regulation, especially since Baybutt and Manson (1997) have recently shown that the mouse intron 1 contains both sections that promote (between bp 10,114 and 10,307) and suppress (bp 9744 and 9932) transcription.

DISCUSSION

A Cryptic Exon 2-like Sequence in the Human *PrP* Gene?

Three-way alignment of *PrP* gene sequences revealed several blocks of conserved nucleotides clustered around exon 2. Previous PCR analysis of *PrP* cDNAs has demonstrated that this exon 2 is present in the majority of mature *PrP* mRNAs in brain RNA of sheep and mice but not in the human *PrP* cDNAs sequenced to date. Though three human *PrP* cDNAs (accession nos. M13899, M13667, and X82545) isolated to date demonstrate exon 1 is spliced directly to exon 3 (using the exon numbering system of the mouse and sheep *PrP* genes), it is notable that this 99-bp exon is conserved in all species, most markedly in the 5′ half, and indeed is far more conserved than the exon 1 sequences adjacent to the *PrP* promoter. It is also notable that the human homolog of exon 2 is flanked by consensus splice sites (Fig. 6B).

Despite this conservation, our experimental data have not provided evidence that exon 2 is part of an alternative human mRNA. It is possible that the substitutions in the human donor splice site (as indicated in Fig. 6B) are responsible for the absence of exon 2. Although the acceptor splice site is highly conserved and close to the consensus splice site, the lack of a good donor site is expected to result in skipping of exon 2 during splicing according to the exon-definition model of splice site recognition (Berget 1995). The conservation of the first half of exon 2 in the human sequence may simply reflect the decreased mutation level in human evolution. This is supported somewhat by the observation that exon 2, unlike any other region of the *PrP* gene, is better conserved between sheep and mouse than between human and mouse (Table 3). The conservation could also be caused by the presence of transcription factor binding sites. Baybutt and Manson (1997) have shown recently that the last 550 bp of intron 1 participates in the promoter activity of the mouse *PrP* gene, and it is possible that this functional region extends into exon 2. The conserved sites blocked at the end of intron 1 could thus represent transcription factor binding sites.

Alternatively, the human exon 2 could be

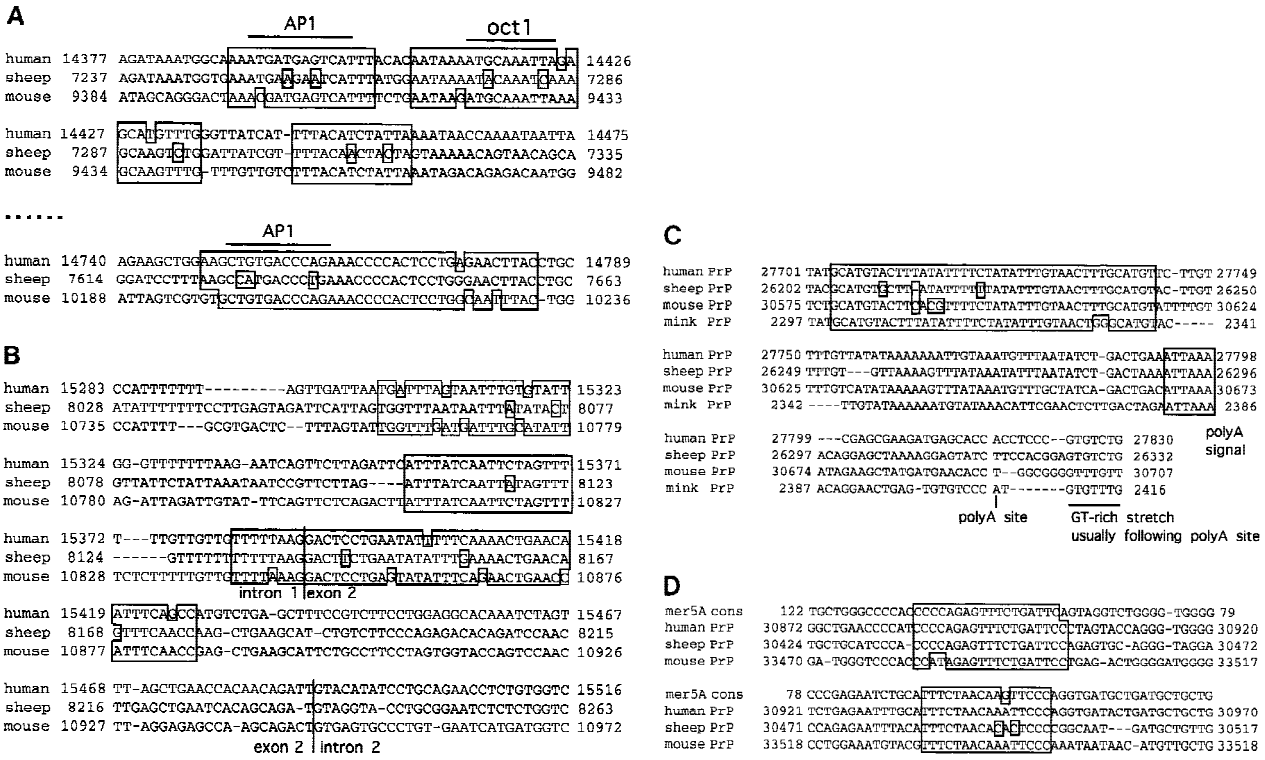

Figure 6 Alignments of sequences outside the *PrP* coding regions conserved between the human, mouse, and sheep. The most conserved regions are blocked. (*A*) Two conserved sections of intron 1 that may be involved in transcriptional regulation. Both sites are within an old LINE1 element (Fig. 2). This L1 element is probably ancient, because a notoriously old element (MIR) has been inserted in it. Potential AP-1 and Oct-1 binding sites were identified with the program TFSEARCH (http://pdap1.trc.rwcp.or.jp/research/db/TFSEARCH.html). However, in the sheep gene they all diverge from the consensus binding sites. Using a cat assay in neuroblastoma cells, Baybutt and Manson (1997) identified transcriptional suppressor activity between bases 9744 and 9932 and promoter activity between bases 10,114 and 10,307 of the mouse sequence, suggesting that the second, highly conserved block (bases 10,200–10,233 in mouse) may be involved in transcriptional activation. (*B*) Alignment of the human, sheep, and mouse *PrP* gene exon 2 regions. Two sites that are conserved between mouse and sheep but have diverged in the human genome are highlighted in gray. Comparison of the scores given by GeneFinder (C. Wilson and P. Green, unpubl.) for the exon 2 splice sites shows that the donor site is quite distinct in sheep and mouse but has become indistinct in the human genome because of the presence of a C and A at positions 4 and 5. The donor site consensus and that of the other species (including cow, rat, and hamster; data not shown) contain a purine and G. (*C*) A 40-bp block ~100 bp before the polyadenylation site in the *PrP* gene is highly conserved between four mammalian orders. (*D*) In the 3′ flanking region of the *PrP* gene are two small blocks of conserved sequences lying within a MER5a DNA transposon fossil. Selective conservation of these blocks is further suggested by their almost complete similarity to the MER5a consensus sequence; MER5 elements are ancient components of the mammalian genome, and sequences are, on average, 25% diverged from the consensus (Smit and Riggs 1996).

spliced and included within a subset of mRNAs from other tissues that have not yet been analyzed. These possibilities are compatible with the apparent ubiquity of exon 2 splicing in other mammals; exon 2-containing cDNAs have now been described in cattle (Yoshimoto et al. 1992) and recent reanalysis of the Syrian hamster (SHa) *PrP* gene, for which exon 1 originally appeared directly spliced to exon 3 (Basler et al. 1986). The reanalysis showed that whereas 90% of SHaPrP mRNAs in the brain exhibit exon 1–3 splicing, 10% include exon 2 sequences closely related to those of sheep and mouse (Li and Bolton 1997). An increased abundance of exon 2-containing mRNAs was observed in brain mRNA of scrapie-infected hamsters, which may reflect preferential expression in astrocytes (which become activated during the course of prion infections).

The function of the alternatively spliced mRNA form that includes exon 2 remains an intriguing unknown, particularly because the protein-coding sequence remains unchanged and humans may be exceptional in lacking it.

Mutations in Mammalian *PrP* Genes

Sequence analysis of the coding sequence of the human *PrP* gene, PRNP, from patients with the familial prion diseases CJD, Gerstmann–Sträussler–Scheinker (GSS) disease, and fatal familial insomnia (FFI) identified 20 nonconservative missense substitutions, as well as expansions of the tandemly repeated octapeptide motifs (Hsiao et al. 1989; for review, see Prusiner 1997). Genetic linkage has been established for 5 of the 20 known mutations that segregate with the inherited prion diseases (Hsiao et al. 1989; Dlouhy et al. 1992; Petersen et al. 1992; Poulter et al. 1992; Gabizon et al. 1993). Recent studies suggest that mutations destabilize PrP^{C}, which leads to its conversion into PrP^{Sc}. Once PrP^{Sc} is formed, then it acts as a ligand that binds to PrP^{C}, resulting in its conversion into a second molecule of PrP^{Sc} (Huang et al. 1994; Riek et al. 1996; James et al. 1997; Prusiner 1997). In addition to the documented role of point mutations, data presented here reveal that transposition events have contributed to the evolution of PrP genes and should also be considered among the "forward" mutational events in mammalian PrP loci. This type of mutation is of particular interest as it may result in misregulated gene expression, which has been discussed as a potential pathogenic mechanism underlying sporadic CJD (Westaway et al. 1994b).

The IAP insertion in the $Prnp^{a}$ allele defines the structural basis for an intron polymorphism, as it distinguishes (in addition to missense substitutions at codons 108 and 189) the *a* and *b* alleles of the mouse *PrP* gene. DNA sequences reveal that the IAP insertion site lies 565 bp 5′ to a *Bst*EII site (GGTNACC) polymorphism that also differs between *a* and *b* alleles ($Prnp^{a}$ = GGTGACC, $Prnp^{b}$ = GGTGGCC). Extrapolating from published restriction mapping data, RIIIS/J ($Prnp^{c}$ haplotype) and Molf/Ei mice ($Prnp^{f}$ haplotype) may also contain an IAP genome within intron 2. Cast/Ei ($Prnp^{e}$ haplotype) mice bear a constellation of restriction sites inconsistent with an IAP insertion but instead suggestive of a crossover event between a *Taq*I site and the aforementioned *Bst*EII site. Finally, MaMy/J mice ($Prnp^{d}$ haplotype) may contain an extensively deleted or rearranged IAP genome because they exhibit a diagnostic *Sac*I site (perhaps corresponding to nucleotides 17,469 and 21,248 of the $Prnp^{a}$ sequence), while exhibiting an IAP–intron 2 *Bam*HI fragment smaller than that of $Prnp^{a}$ mice. In summary, because four out of six *Prnp* haplotypes carry full-length or partial IAP elements, disrupted alleles are representative of the wild-type *Prnp* gene in laboratory mice. Thus, although the absence of the IAP in the $Prnp^{b}$ allele must represent the ancestral state, most laboratory mouse strains appear to contain (part of) the IAP insert in intron 2.

The polymorphic status and the identical LTRs argue for a recent integration of the IAP element in $Prnp^{a}$. These considerations beg the question of transcriptional status, both for the IAP element and $Prnp^{a}$ itself. Regarding the former, because the TATA box in the LTRs is heavily mutated (see Fig. 5), transcriptional activity appears unlikely. However, some enhancer elements in the LTRs still appear to be intact (Fig. 5) and may influence transcription of the PrP gene. Regarding the $Prnp^{a}$ transcripts, steady-state levels of the $Prnp^{a}$ mRNA are known to approximate those of $Prnp^{b}$ mice (Westaway et al. 1987). This result suggests that the IAP element has little effect on transcription on the other DNA strand. However, the IAP element may have tissue-specific effects. Again, this conclusion must be approached with a degree of caution as the $Prnp^{a}$ and $Prnp^{b}$ alleles differ, aside from the IAP, by three substitutions and one 1-bp deletion in the putative promoter (Westaway et al. 1994a) and two substitutions and one 1-bp deletion in the 3′ UTR. Therefore, although precedents exist for effects of IAP insertions on target genes (for review, see Amariglio and Rechavi 1993), extrapolation from these examples to the specific case of the $Prnp^{a}$ gene is inappropriate. Functional studies will have to be undertaken to address interactions between the IAP element and Prnp transcription.

Evolution of *PrP* Genes

This study is one of the first to provide large-scale genomic comparisons in three species. A three-way comparison has several advantages over a simple two-species comparison.

First, one can determine whether an unaligned sequence is an insert or a deletion. This distinction allowed us to estimate that the second intron of the *PrP* gene was ~9.1 kb in length before the mammalian radiation (~85 million years ago) and has since independently increased considerably in size in all three species by numerous mutations (Table 2). It is also noteworthy that the type (and therefore mechanism) of all insertions, which now represent 40%–57% of the loci, could be deduced (Table 1). A preliminary related observation is that independent insertions of related transposable elements may occur at similar locations in different species (Fig. 2). If this is so, it suggests possible integration site preferences and conservation of local chromosomal struc-

ture. There are at least four examples: (1) The IAP and ALTR1 elements were inserted at almost the same location in the intron 2 L1M element in mouse and sheep; (2) in primates (L1PA13) and rodents (Rodent L1), a LINE1 element integrated at almost the same position and in the same orientation in this intron 2 L1M element; (3) a mariner and a Tigger 2 DNA transposon (both relatively rare elements) were inserted within 1 kb of each other in the sheep and human loci; and (4) five *Alu*s and five "rodent *Alu*s" (B1s) were inserted at corresponding regions in the 3′-flanking region of the *PrP* gene in the human and the mouse loci, respectively.

Second, general differences in the rates of mutation among mammalian orders become apparent in a three-way comparison. As presented in Table 3, the sheep and mouse *PrP* loci are more similar to human DNA than to each other, and the human *PrP* locus is much more similar to sheep than to mouse. These observations cannot be explained by any phylogenetic tree with constant mutation rates (per time unit) on all branches, but they fit the now broadly accepted model that the neutral mutation rate has decreased in the lineage leading to humans (Goodman et al. 1971) and has been higher in rodents than in artiodactyls and primates (Laird et al. 1969). The coding regions of the three species and the 3′ UTR are, as expected, more highly conserved than the introns and flanking regions. It is surprising to note that the synonymous substitution levels are higher than those of the introns. This may reflect the presence of conserved regions in the introns (see Fig. 2).

Third and most importantly, a sequence comparison across three species offers a powerful approach to identifying highly conserved and presumably functional regions. A region that is highly conserved in three rather than two species is much less likely to arise by coincidence and is more likely to reflect a functionally constrained sequence. The conserved regions indicated in Figures 2 and 6 raise fascinating questions as to their possible functions. The conserved regions lying outside the processed transcripts could bind regulatory factors. The conservation of exon 2 in human suggests an alternatively spliced mRNA product. The high conservation of the 3′ end of the transcript could reflect an important role in mRNA stability. The exon 2 conservation may be deceptive in the case of human DNA and may reflect the presence of transcriptionally regulatory elements and low mutation rate, rather than a functional exon; however, comparison with the other species allowed us to suggest a reason for the apparent absence of human transcripts that include exon 2.

METHODS

Cosmid Purification, M13 Subcloning, and DNA Instability

To obtain nucleotide sequence for the *PrP* regions from human, mouse, and sheep, we sequenced two cosmid clones for the human *PrP* and sheep *PrP* gene, and two λ clones and one plasmid clone for the mouse *Prnp*[a] gene (Puckett et al. 1991; Westaway et al. 1994a,c). Cosmid and plasmid DNA were prepared by the alkaline lysis method (Birnboim and Doly 1979). To use the random or shotgun sequencing strategy, each DNA was randomly sheared by sonication (20–40 sec at 100% continuous power), size fractionated (0.7–1.5 kb), and ligated into blunt-end (*Hin*cII or *Sma*I) cleaved, dephosphorylated M13mp9 vector. Phage λ DNA was prepared using standard protocols, and M13 clones containing λ vector-derived inserts were identified by hybridization of gridded phage arrays on membranes to nonradioactive digoxigenin-labeled λ DNA and discarded. Difficulties were experienced retrieving M13 subclones from intron 2 of the sheep PrP cosmid, and the resulting gap in the sequence contig was closed by PCR products obtained directly from the cosmid with flanking primers.

DNA Sequencing

Phage M13 clones bearing *PrP* inserts were sequenced on ABI 373 sequencers by dideoxy chain termination method using *Taq* DNA polymerase (ABI) or Sequenase (U.S. Biochemical) along with fluorescent dye primers. Individual sequences were initially edited to remove M13 vector sequences at the 5′ end and ambiguities at the 3′ end and assembled into contigs using the "SeqMan" program (DnaStar, Madison, WI). Remaining gaps were closed by directed sequencing procedures using fluorescent dideoxynucleotide terminators or using PCR amplification, cloning, and sequencing. The consensus sequence was generated by the final editing process that includes manual visual confirmation with the original chromatograms using SeqMan. Assembly accuracy was verified by comparison of restriction enzymes digests of the DNA to computer generated restriction maps and by alignments with known genomic DNA sequences. The average redundancy for the presented sequences is 6.5 to 8.5. GenBank accession numbers are U29185 (human), U29186 (mouse), and U67922 (sheep).

Genomic DNA Analysis Programs

Genome-wide repeats and low complexity regions were identified and masked using the program RepeatMasker (A.F.A. Smit and P. Green, unpubl.), which uses the program cross_match (P. Green, unpubl.) and expanded versions of the publicly available databases of repetitive sequences (Jurka et al. 1996). This utility can be accessed at the Web server (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Subsequent database searches (GenBank, release 91) were performed with BLAST (Altschul et al. 1990) programs running under Inherit Analysis (Foster City, CA), DnaStar, and

cross_match. Consensus sequences for newly identified interspersed repeats were derived as described previously (Smit 1993), now using the program cross_match. These consensus sequences have been included in the human, rodent, and mammalian repeat reference databases (Jurka and Smit 1997).

We searched for potential coding regions beside the *PrP* gene with the GRAIL gene finding program (Uberbacher and Mural 1991) and by comparing the sequences to the protein databases with BLASTX (Gish and States 1993). The complete overlapping regions of the clones of all three species were pairwise-aligned using cross_match. These alignments were significantly aided by splicing out the order-specific interspersed repeats. Promoter analysis was performed with the help of the programs Web Signal Scan (Prestridge 1991) and Web Promoter Scan (Prestridge 1995). CpG islands and microsatellite repeats were analyzed using the DiNucleotides (T. Smith, unpubl.) and Sputnik program (Abajian 1994), respectively.

ACKNOWLEDGMENTS

We thank Dr. Lee Rowen for helpful discussion, Dr. Todd Smith for computer programs, and Drs. Bruce Chesebro and Sylvia Perryman for communicating the sequence of an IAP-related cDNA up-regulated in scrapie-infected neuroblastoma cells. This work was supported by grants from the National Institutes of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abajian, C. 1994. sputnik. http://www.geospiza.com/people/chris/software/sputnik.html.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Amariglio, N. and G. Rechavi. 1993. Insertional mutagenesis by transposable elements in the mammalian genome. *Environ. Mol. Mutagen.* **21:** 212–218.

Auge-Gouillou, C., Y. Bigot, N. Pollet, M.H. Hanmelin, M. Meunier-Rotival, and G. Periquet. 1995. Human and other mammalian genomes contain transposons of the mariner family. *FEBS Lett.* **368:** 541–546.

Basler, K., B. Oesch, M. Scott, D. Westaway, M. Wälchli, D.F. Groth, M.P. McKinley, S.B. Prusiner, and C. Weissmann. 1986. Scrapie and cellular PrP isoforms are encoded by the same chromosomal gene. *Cell* **46:** 417–428.

Baybutt, H. and J. Manson. 1997. Characterisation of two promoters for prion protein (PrP) gene expression in neuronal cells. *Gene* **184:** 125–131.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411–2414.

Birnboim, H.C. and J. Doly. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7:** 1513–1523.

Carlson, G.A., P.A. Goodman, M. Lovett, B.A. Taylor, S.T. Marshall, M. Peterson-Torchia, D. Westaway, and S.B. Prusiner. 1988. Genetics and polymorphism of the mouse prion gene complex: The control of scrapie incubation time. *Mol. Cell. Biol.* **8:** 5528–5540.

Dlouhy, S.R., K. Hsiao, M.R. Farlow, T. Foroud, P.M. Conneally, P. Johnson, S.B. Prusiner, M.E. Hodes, and B. Ghetti. 1992. Linkage of the Indiana kindred of Gerstmann-Sträussler-Scheinker disease to the prion protein gene. *Nat. Genet.* **1:** 64–67.

Doh-ura, K., S. Perryman, R. Race, and B. Chesebro. 1995. Identification of differentially expressed genes in scrapie-infected mouse neuroblastoma cells. *Microb. Pathog.* **18:** 1–9.

Duncan, C.H. 1987. Novel Alu-type repeat in artiodactyls. *Nucleic Acids Res.* **15:** 1340.

Gabizon, R., H. Rosenmann, Z. Meiner, I. Kahana, E. Kahana, Y. Shugart, J. Ott, and S.B. Prusiner. 1993. Mutation and polymorphism of the prion protein gene in Libyan Jews with Creutzfeldt-Jakob disease. *Am. J. Hum. Genet.* **53:** 828–835.

Gabriel, J.-M., B. Oesch, H. Kretzschmar, M. Scott, and S.B. Prusiner. 1992. Molecular cloning of a candidate chicken prion protein. *Proc. Natl. Acad. Sci.* **89:** 9097–9101.

Gibbs, C.J., Jr., D.C. Gajdusek, D.M. Asher, M.P. Alpers, E. Beck, P.M. Daniel, and W.B. Matthews. 1968. Creutzfeldt-Jakob disease (spongiform encephalopathy): Transmission to the chimpanzee. *Science* **161:** 388–389.

Gish, W. and D.J. States. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Goldmann, W., N. Hunter, J.D. Foster, J.M. Salbaum, K. Beyreuther, and J. Hope. 1990. Two alleles of a neural protein gene linked to scrapie in sheep. *Proc. Natl. Acad. Sci.* **87:** 2476–2480.

Goodman, M., J. Barnabaas, G. Matsuda, and G.W. Moore. 1971. Molecular evolution in the descent of man. *Nature* **233:** 604–613.

Gorodinsky, A. and D.A. Harris. 1995. Glycolipid-anchored proteins in neuroblastoma cells form detergent-resistant complexes without caveolin. *J. Cell Biol.* **129:** 619–627.

Hahm, K., P. Ernst, K. Lo, G.S. Kim, C. Turck, and S.T. Smale. 1994. The lymphoid transcription factor LyF-1 is encoded by specific, alternatively spliced mRNAs derived from the Ikaros gene. *Mol. Cell. Biol.* **14:** 7111–7123.

Harris, D.A., D.L. Falls, W. Walsh, and G.D. Fischbach. 1989. Molecular cloning of an acetylcholine receptor-inducing protein. *Soc. Neurosci.* **15:** 70.7.

Hsiao, K., H.F. Baker, T.J. Crow, M. Poulter, F. Owen, J.D. Terwilliger, D. Westaway, J. Ott, and S.B. Prusiner. 1989. Linkage of a prion protein missense variant to Gerstmann-Sträussler syndrome. *Nature* **338:** 342–345.

Huang, Z., J.-M. Gabriel, M.A. Baldwin, R.J. Fletterick, S.B. Prusiner, and F.E. Cohen. 1994. Proposed three-dimensional structure for the cellular prion protein. *Proc. Natl. Acad. Sci.* **91:** 7139–7143.

James, T.L., H. Liu, N.B. Ulyanov, S. Farr-Jones, H. Zhang, D.G. Donne, K. Kaneko, D. Groth, I. Mehlhorn, S.B. Prusiner, and F.E. Cohen. 1997. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci.* **94:** 10086–10091.

Jurka, J. and A.F.A. Smit. 1997. Collection of human repetitive elements. http://www.girinst.org/server/repbase.html.

Jurka, J., V.V. Kapitonov, P. Klonowski, J. Walichiewicz, and A.F. Smit. 1996. Identification of new medium reiteration frequency repeats in the genomes of Primates, Rodentia and Lagomorpha. *Genetica* **98:** 235–247.

Kaneko, K., L. Zulianello, M. Scott, C.M. Cooper, A.C. Wallace, T.L. James, F.E. Cohen, and S.B. Prusiner. 1997. Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation. *Proc. Natl. Acad. Sci.* **94:** 10069–10074.

Kapitonov, V. and J. Jurka. 1996. The age of Alu subfamilies. *J. Mol. Evol.* **42:** 59–65.

Kuff, E.L. and K.K. Lueders. 1988. The intracisternal A-particle gene family: Structure and functional aspects. *Adv. Cancer Res.* **51:** 183–276.

Laird, C.D., B.L. McConaughy, and B.J. McCarthy. 1969. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224:** 149–154.

Lee, W., P. Mitchell, and R. Tjian. 1987. Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell* **49:** 741–752.

Li, G. and D.C. Bolton. 1997. A novel hamster prion protein mRNA contains an extra exon: Increased expression in scrapie. *Brain Res.* **751:** 265–274.

Masters, C.L., D.C. Gajdusek, and C.J. Gibbs Jr. 1981. Creutzfeldt-Jakob disease virus isolations from the Gerstmann-Sträussler syndrome. *Brain* **104:** 559–588.

Meggendorfer, F. 1930. Klinische und genealogische Beobachtungen bei einem Fall von spastischer Pseudosklerose Jakobs. *Z. Gesamte Neurol. Psychiatr.* **128:** 337–341.

Morgan, G.T. 1995. Identification in the human genome of mobile elements spread by DNA-mediated transposition. *J. Mol. Biol.* **254:** 1–5.

Oesch, B., D. Westaway, and S.B. Prusiner. 1991. Prion protein genes: Evolutionary and functional aspects. *Curr. Top. Microbiol. Immunol.* **172:** 109–124.

Oosumi, T., W.R. Belknap, and B. Garlick. 1995. Mariner transposons in humans. *Nature* **378:** 672.

Petersen, R.B., M. Tabaton, L. Berg, B. Schrank, R.M. Torack, S. Leal, J. Julien, C. Vital, B. Deleplanque, W.W. Pendlebury, D. Drachman, T.W. Smith, J.J. Martin, M. Oda, P. Montagna, J. Ott, L. Autilio-Gambetti, E. Lugaresi, and P. Gambetti. 1992. Analysis of the prion protein gene in thalamic dementia. *Neurology* **42:** 1859–1863.

Poulter, M., H.F. Baker, C.D. Frith, M. Leach, R. Lofthouse, R.M. Ridley, T. Shah, F. Owen, J. Collinge, G. Brown, J. Hardy, M.J. Mullan, A.E. Harding, C. Bennett, R. Doshi, and T.J. Crow. 1992. Inherited prion disease with 144 base pair gene insertion. 1. Genealogical and molecular studies. *Brain* **115:** 675–685.

Prestridge, D.S. 1991. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Appl. Biosci.* **7:** 203–206.

———. 1995. Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249:** 923–932.

Prusiner, S.B. 1989. Scrapie prions. *Annu. Rev. Microbiol.* **43:** 345–374.

———. 1997. Prion diseases and the BSE crisis. *Science* **278:** 245–251.

Prusiner, S.B., M. Scott, D. Foster, K.-M. Pan, D. Groth, C. Mirenda, M. Torchia, S.-L. Yang, D. Serban, G.A. Carlson, P.C. Hoppe, D. Westaway, and S.J. DeArmond. 1990. Transgenetic studies implicate interactions between homologous PrP isoforms in scrapie prion replication. *Cell* **63:** 673–686.

Puckett, C., P. Concannon, C. Casey, and L. Hood. 1991. Genomic structure of the human prion protein gene. *Am. J. Hum. Genet.* **49:** 320–329.

Riek, R., S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wüthrich. 1996. NMR structure of the mouse prion protein domain PrP(121–231). *Nature* **382:** 180–182.

Saba, J.A., H. Busch, and R. Reddy. 1985. A new moderately repetitive rat DNA sequence detected by a cloned 4.5 S1 DNA. *J. Biol. Chem.* **260:** 1354–1357.

Saeki, K., Y. Matsumoto, Y. Matsumoto, and T. Onodera. 1996. Identification of a promoter region in the rat prion protein gene. *Biochem. Biophys. Res. Commun.* **219:** 47–52.

Schätzl, H.M., M. Da Costa, L. Taylor, F.E. Cohen, and S.B. Prusiner. 1995. Prion protein gene variation among primates. *J. Mol. Biol.* **245:** 362–374.

Scheidereit, C., S. Geisse, H.M. Westphal, and M. Beato. 1983. The glucocorticoid receptor binds to defined

nucleotide sequences near the promoter of mouse mammary tumour virus. *Nature* **304:** 749–752.

Scott, M.R., J. Safar, G. Telling, O. Nguyen, D. Groth, M. Torchia, R. Koehler, P. Tremblay, D. Walther, F.E. Cohen, S.J. DeArmond, and S.B. Prusiner. 1997. Identification of a prion protein epitope modulating transmission of bovine spongiform encephalopathy prions to transgenic mice. *Proc. Natl. Acad. Sci.* **94:** 14279–14284.

Smit, A.F. 1993. Identification of a new, abundant superfamily of mammalian LTR transposons. *Nucleic Acids Res.* **21:** 1863–1872.

Smit, A.F.A. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.

Smit, A.F. and A.D. Riggs. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23:** 98–102.

———. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci.* **93:** 1443–1448.

Smit, A.F., G. Tóth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE1 repetitive sequences. *J. Mol. Biol.* **246:** 401–417.

Sparkes, R.S., M. Simon, V.H. Cohn, R.E.K. Fournier, J. Lem, I. Klisak, C. Heinzmann, C. Blatt, M. Lucero, T. Mohandas, S.J. DeArmond, D. Westaway, S.B. Prusiner, and L.P. Weiner. 1986. Assignment of the human and mouse prion protein genes to homologous chromosomes. *Proc. Natl. Acad. Sci.* **83:** 7358–7362.

Stahl, N., D.R. Borchelt, K. Hsiao, and S.B. Prusiner. 1987. Scrapie prion protein contains a phosphatidylinositol glycolipid. *Cell* **51:** 229–240.

Stender, A. 1930. Weitere Beiträge zum Kapitel "Spastische Pseudosklerose Jakobs". *Z. Gesamte Neurol. Psychiatr.* **128:** 528–543.

Szemraj, J., G. Pluciennicczak, J. Jaworski, and A. Pluciennicczak. 1995. Bovine Alu-like sequences mediate transposition of a new site-specific retroelement. *Gene* **152:** 261–264.

Takeuchi, Y. and F. Harada. 1986. Cloning and characterization of rat 4.5S RNAI genes. *Nucleic Acids Res.* **14:** 1643–1656.

Taraboulos, A., M. Scott, A. Semenov, D. Avrahami, L. Laszlo, and S.B. Prusiner. 1995. Cholesterol depletion and modification of COOH-terminal targeting sequence of the prion protein inhibits formation of the scrapie isoform. *J. Cell Biol.* **129:** 121–132.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17:** 6463–6471.

Telling, G.C., M. Scott, J. Mastrianni, R. Gabizon, M. Torchia, F.E. Cohen, S.J. DeArmond, and S.B. Prusiner. 1995. Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell* **83:** 79–90.

Uberbacher, E.C. and R.J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* **88:** 11261–11265.

Vey, M., S. Pilkuhn, H. Wille, R. Nixon, S.J. DeArmond, E.J. Smart, R.G. Anderson, A. Taraboulos, and S.B. Prusiner. 1996. Subcellular colocalization of the cellular and scrapie prion proteins in caveolae-like membranous domains. *Proc. Natl. Acad. Sci.* **93:** 14945–14949.

Weiher, H., M. König, and P. Gruss. 1983. Multiple point mutations affecting the simian virus 40 enhancer. *Science* **219:** 626–631.

Westaway, D., P.A. Goodman, C.A. Mirenda, M.P. McKinley, G.A. Carlson, and S.B. Prusiner. 1987. Distinct prion proteins in short and long scrapie incubation period mice. *Cell* **51:** 651–662.

Westaway, D., C. Cooper, S. Turner, M. Da Costa, G.A. Carlson, and S.B. Prusiner. 1994a. Structure and polymorphism of the mouse prion protein gene. *Proc. Natl. Acad. Sci.* **91:** 6418–6422.

Westaway, D., S.J. DeArmond, J. Cayetano-Canlas, D. Groth, D. Foster, S.-L. Yang, M. Torchia, G.A. Carlson, and S.B. Prusiner. 1994b. Degeneration of skeletal muscle, peripheral nerves, and the central nervous system in transgenic mice overexpressing wild-type prion proteins. *Cell* **76:** 117–129.

Westaway, D., V. Zuliani, C.M. Cooper, M. Da Costa, S. Neuman, A.L. Jenny, L. Detwiler, and S.B. Prusiner. 1994c. Homozygosity for prion protein alleles encoding glutamine-171 renders sheep susceptible to natural scrapie. *Genes & Dev.* **8:** 959–969.

Wilesmith, J.W., J.B.M. Ryan, and M.J. Atkinson. 1991. Bovine spongiform encephalopathy—Epidemiologic studies on the origin. *Vet. Rec.* **128:** 199–203.

Windl, O., M. Dempster, P. Estibeiro, and R. Lathe. 1995. A candidate marsupial *PrP* gene reveals 2 domains conserved in mammalian PrP proteins. *Gene* **159:** 181–186.

Yoshimoto, J., T. Iinuma, N. Ishiguro, M. Horiuchi, M. Imamura, and M. Shinagawa. 1992. Comparative sequence analysis and expression of bovine PrP gene in mouse L-929 cells. *Virus Genes* **6:** 343–356

Received June 17, 1998; accepted in revised form August 28, 1998.

Complete Genomic Sequence and Analysis of the Prion Protein Gene Region from Three Mammalian Species

Inyoul Y. Lee, David Westaway, Arian F.A. Smit, et al.

Genome Res. 1998 8: 1022-1037
Access the most recent version at doi:10.1101/gr.8.10.1022