Mammalian Genome 12, 700–712 (2001).
DOI: 10.1007/s00335-001-2044-8

Mammalian Genome
Incorporating Mouse Genome

A genomic sequence analysis of the mouse and human microtubule-associated protein tau

Parvoneh Poorkaj,[1,2] Arnie Kas,[3] Ian D'Souza,[1,2] Yang Zhou,[3] Quynh Pham,[3] Mariam Stone,[3] Maynard V. Olson,[3,4] Gerard D. Schellenberg[1,2,5]

[1]Geriatric Research Education Clinical Center 182-B, Veterans Affairs Puget Sound Health Care System, Seattle Division, 1660 S. Columbian Way, Seattle, Washington 98108, USA
[2]Division of Gerontology and Geriatric Medicine, University of Washington, Seattle, Washington 98195, USA
[3]Department of Medicine, University of Washington, Seattle, Washington 98195, USA
[4]Department of Genetics, University of Washington, Seattle, Washington 98195, USA
[5]Departments of Neurology and Pharmacology, University of Washington, Seattle, Washington 98195, USA

Received: 1 April 2001 / Accepted: 20 April 2001

Abstract. Microtubule associated protein tau (MAPT) encodes the microtubule associated protein tau, the primary component of neurofibrillary tangles found in Alzheimer's disease and other neurodegenerative disorders. Mutations in the coding and intronic sequences of MAPT cause autosomal dominant frontotemporal dementia (FTDP-17). MAPT is also a candidate gene for progressive supranuclear palsy and hereditary dysphagic dementia. A human PAC (201 kb) and a mouse BAC (161 kb) containing the entire MAPT and *Mtapt* genes, respectively, were identified and sequenced. Comparative DNA sequence analysis revealed over 100 conserved non-repeat potential *cis*-acting regulatory sequences in or close to MAPT. Those islands with greater than 67% nucleotide identity range in size from 20 to greater than 1700 nucleotides. Over 90 single nucleotide polymorphisms were identified in MAPT that are candidate susceptibility alleles for neurodegenerative disease. The 5′ and 3′ flanking genes for MAPT are the corticotrophin-releasing factor receptor (CRFR) gene and KIAA1267, a gene of unknown function expressed in brain.

Introduction

Tau is a member of the microtubule-associated protein (MAP) family found primarily in neurons, and at lower levels in oligodendrocytes, astrocytes, and in some non-nervous system tissues (LoPresti et al. 1995; Gu et al. 1996; Vanier et al. 1998). In vitro, tau binds to microtubules and stimulates microtubule assembly. In vivo, tau promotes microtubule assembly and stability and may participate in axonal extension and maintenance (Caceres and Kosik 1990; Caceres et al. 1991).

Expression of the gene encoding tau (MAPT in human, *Mtapt* in mouse) is highly regulated, particularly at the RNA splicing stage, and this regulation differs between rodents and humans. MAPT has 15 exons, where 6 of 14 coding exons undergo alternative splicing (Fig. 1) (Himmler et al., 1989; Himmler, 1989; Andreadis et al., 1992). In the fetal central nervous system (CNS), a single tau isoform is produced lacking all alternatively spliced exons. In the adult human CNS, six splice variants are produced by inclusion of alternative exons 2, 3, and 10 (Goedert et al. 1989) (Fig. 1). In contrast, in the adult rodent brain, only three isoforms are present, with all forms containing exon 10 (E10; Kosik et al. 1989) and E2 and E3 being alternatively spliced (Collet et al. 1997). Tau has microtubule-binding domains that are imperfect 18 amino acid repeats separated by 13–14 amino acid inter-repeat regions that are dissimilar; E10 encodes one binding repeat and one inter-repeat. Depending on whether E10 is excluded or included, tau has either three (3R tau) or four (4R tau) microtubule-binding repeats, respectively. The functional consequence of adding E10 is that 4R tau binds microtubules with a higher affinity compared with 3R tau(Butner and Kirschner 1991; Gustke et al. 1994). In adult human brain, the 3R/4R ratio is approximately 1 (Hong et al., 1998), while in rodent brain only 4R tau is made (Kosik et al. 1989). The use of other alternatively spliced exons (4a, 6, and 8) appears to be confined to the peripheral nervous system in humans, though low levels of E4a- and E6-containing transcripts are found in human and rodent brain (Georgieff et al. 1991, 1993; Mavilia et al. 1993, 1994; Boyne et al. 1995; Wei and Andreadis 1998). In addition, in some mouse transcripts, the intron between E13 and E14 is removed by RNA splicing (Lee et al. 1988), while in other mouse transcripts and in all rat and human transcripts described to date, the equivalent sequences are retained. Poly-adenylation site usage is also regulated, and MAPT transcripts have either a short 200- to 250-nt 3′ untranslated region (3′UTR) or a much longer ~4kb 3′UTR (Goedert et al. 1988; Sadot et al. 1994).

Mutations in MAPT cause frontotemporal dementia, Chromosome (Chr) 17 type (FTDP-17) (Clark et al. 1998; Hutton et al. 1998; Spillantini et al. 1998; Poorkaj et al. 1998), an autosomal dominant neurodegenerative disease. Different MAPT mutations cause FTDP-17 by different mechanisms. Some mutations alter the biochemical properties of tau, resulting in decreased microtubule-binding capacity or decreased rates of tau-stimulated microtubule polymerization (e.g., P301L, V337M; Hong et al. 1998). Other mutations disrupt the normal regulation of E10 splicing, either increasing or decreasing the inclusion of E10 (Hutton et al. 1998; D'Souza et al. 1999; D'Souza and Schellenberg 2000). For some FTDP-17 families, genetic linkage analysis has clearly localized the disease-causing defect to the MAPT region of Chr 17 and yet no mutations have been identified in the MAPT open reading frame or in the intronic sequences immediately flanking exons [e.g., the HDDD2 kindred (Lendon et al. 1998)]. Mutations in these families are presumably in regulatory sequences within introns or in regulatory sequences flanking the gene.

MAPT is a candidate gene for a number of other neurodegenerative disorders including corticobasal degeneration (CBD), progressive supranuclear palsy (PSP), Picks disease, and amyotrophy lateral sclerosis parkinsonism dementia complex of Guam. A ge-

[1] To whom correspondence should be addressed at GRECC 182-B, Veterans Affairs Puget Sound Health Care System, 1660 S. Columbian Way, Seattle, WA 98108. Telephone: (206) 764-2701. FAX: (206) 764-2569.

Correspondence to: G.D. Schellenberg; email: Zachdad@U.Washington.edu.

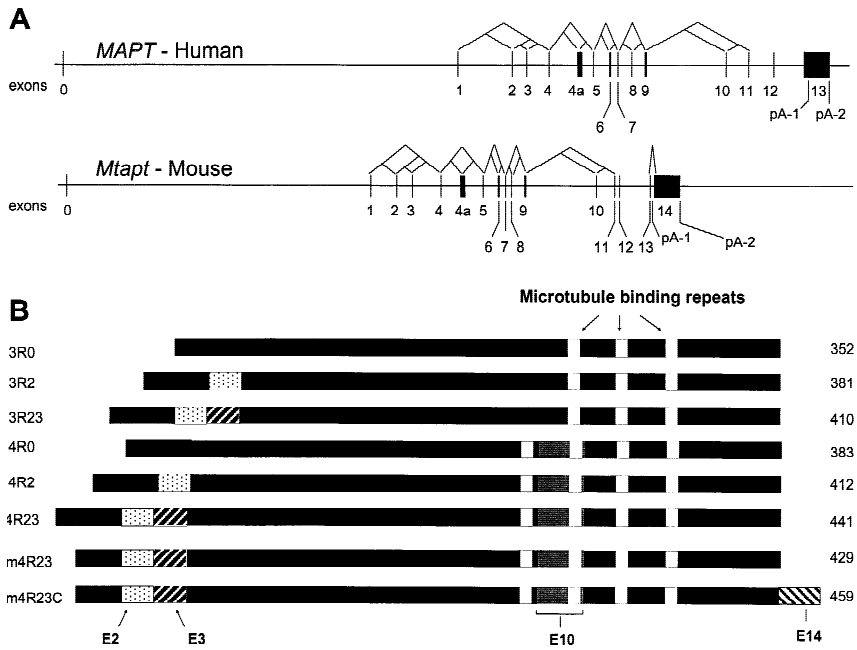

Fig. 1. A. Genomic organization and alternative splicing of MAPT and *Mtapt*. The human and mouse tau gene exon structures are shown with exons represented by a vertical bar. Exon numbers are given below each exon. E0 encodes the 5′UTR, and E1 encodes that ATG codon that begins the coding region of the gene. Splicing patterns of alternatively utilized exons (E2, E3, E4a, E6, E8, E10, and E14) are shown with connecting lines above the exons. Note that E3 is observed only in constructs containing E2. Distances between exons are proportional to the sizes of the introns as determined from the genomic sequence. Abbreviations are; pA-1, first polyadenylation site; pA-2, second polyadenlation site. **B.** Structure of tau isoforms found in the CNS. The amino acid length of each isoform is shown on the right. On the left is nomenclature for each isoform with R referring to the microtubule binding repeat and "m" indicating mouse-specific isoforms. The stippled, gray, and cross-hatched bars represent alternatively spliced exons. In adult human CNS, six isoforms are present (3R0, 3R2, 3R23, 4R0, 4R2, and 4R23) while in adult mouse only three isoforms are found (4R0, 4R2, and 4R23). Presently it is not known what the distribution is of mouse proteins containing the 23 amino acids encoded by mouse E14, shown as m4R23C.

netic association between MAPT alleles and PSP was reported (Conrad et al. 1997) and consistently confirmed (Oliva et al. 1998; Higgins et al. 1998; Baker et al. 1999), suggesting that there is a MAPT low-penetrance PSP-susceptibility allele. However, although mutation analysis of MAPT coding and directly flanking intronic regions has been performed for PSP subjects, the identity of the PSP-susceptibility site is unknown. Presumably, the PSP-susceptibility site is in a regulatory sequence located in an intron or in sequences flanking the gene.

To identify potential *cis*-acting regulatory regions, we compared genomic sequences for MAPT and *Mtapt*. Over 100 conserved intronic sequences were identified that are not repeats and are not exons. To determine whether MAPT is part of a gene cluster, the flanking genes were identified. Assuming that MAPT regulatory elements are not within adjacent genes, we have identified the genetic boundaries of where *cis*-acting MAPT elements may be located.

Materials and methods

Isolation of genomic clones. Mouse BAC (bacterial artificial chromosome) and human PAC (P1 derived artificial chromosome) clones containing MAPT and *Mtapt* were identified by hybridization screening of high-density arrays of a mouse embryonic stem cell release I BAC library and a human male white blood cell PAC library (Genome Systems, St. Louis, Mo.). Filters were hybridized with probes generated from genomic DNA by PCR (polymerase chain reaction) amplification using the primer pairs 1BF/1BR, 4F/4R, 10F/10R, and 14F/14R (Poorkaj et al. 1998) as described previously (Levy-Lahad et al. 1996). Five human PAC clones (16C10, 61D6, 223A9, 231I12, and 246L12) and three mouse BAC clones (191P19, 35D9, and 35D14) containing MAPT and *Mtapt* exons, respectively, were identified. PAC and BAC DNA was isolated from individual colonies as described previously (Poorkaj et al. 1998). The exon content of the individual clones was determined by PCR amplification and direct sequencing using primers 1BF/1BR, 10F/10R, and 14F/14R (Poorkaj et al. 1998). Mouse BAC clone 35D9 (m35D9) and human PAC clone 61D6 contained the entire coding sequence for tau and were selected for further analysis. Flanking and overlapping human genomic clones were identified by BLASTN searches of the high-throughput sequencing database (NCBI) using 61D6 end sequences as queries.

DNA sequence analysis. PAC and BAC DNA's were isolated by a modified alkaline lysis procedure. DNA shearing was minimized by eliminating vortexing and mixing steps after the bacterial cell lysis. The DNA was treated with an RNase A/T1 RNase mix and isopropanol precipitated. Re-suspended DNA was re-precipitated with two volumes of 100% ethanol, 1/10 volume 3 M Na-acetate, and resuspended in 10 mM Tris, 0.5 mM EDTA, pH 8.0. The DNA was sonicated in 3 μg aliquots using four different sonication times (4, 6, 8, and 10 s), and DNA (120 ng) from each time point was analyzed by electrophoresis with 1% agarose, 1X TAE gels. BAC/PAC inserts ranging from 1.6 to 3 kb were end-filled with T4 DNA polymerase and inserts isolated from 1% Nusieve/1X TAE electrophoresis gels and purified (Wizard, Promega). Purified inserts were ligated into the Novagen M13mp18 vector. Libraries where >85% of the clones contained inserts and <11% of the clones contained *E. coli* DNA were used for shotgun sequencing by using an ABI 377 sequencer and PHRED base calling software (http://www.genome.washington.edu). Sequence was assembled from a minimum of 3000 lanes using PHRAP (A phragment assembly program) and Consed (sequence assembly editor companion to PHRAP; http://www.genome.washington.edu). Direct sequencing of human PAC and mouse BAC DNA was used to close minimal gaps in the contigs and to clarify sequence ambiguities.

Sequence comparison methods. Completed human and mouse DNA sequences were compared by using the Crossmatch program, a general-purpose utility based on an efficient implementation of the Smith-Waterman algorithm (http://www.genome.washington.edu/phrap.docs/phrap.html). Repeats were identified and masked using Repeatmasker (Smit and Green 1996–1997). The minimum score for the initial matrix was set at 50 to identify large conserved regions. Additional searches were performed with lower matrix cutoff levels to identify minimally conserved regulatory elements. Individual intron/exon files were generated, and additional searches were performed by using lower matrix cutoff levels.

Single nucleotide polymorphism (SNP) analysis. Single nucleotide polymorphisms were detected by direct sequencing (Poorkaj et al. 1998). Regions sequenced were the 150–250 nt flanking E1, E2, and E3, one conserved region of 136 nt in I11 (61D6 nt #129,351–129551), one conserved region of 160 nt in I12 (61D6 nt #136,751–136,961), all of I9, I10, and all of the 3′ UTR. Both strands of all regions were sequenced in 12 normal Caucasian controls. In addition, 6 Guamanian chamorro ALS and 5 PDC subjects and 12 normal Guamanian chamorro controls were sequenced for all exons and portions of I9 and I10. Intronic and 3′ UTR sequences were amplified as 500–600 nt fragments (with a 50–100 nt overlap) yielding contiguous sequence data.

Exon splicing assays. Tau minigene constructs were generated to assess the potential splicing regulatory elements present in intronic sequences that directly flank tau exons 9, 10, and 11. Splicing was assayed using vector

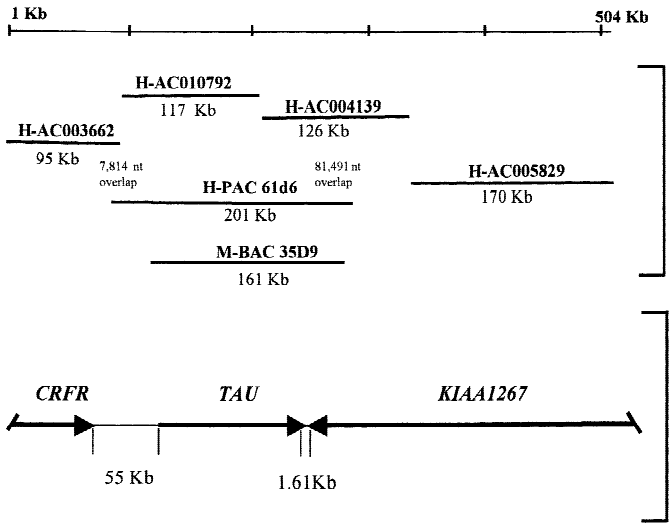

Fig. 2. Contig of MAPT region. Sequenced BACs and PACs are identified by accession number except for 61D6 and mouse BAC m35D9. Accession numbers for these are AC091628 and AC091629, respectively. All BACs and PACs are human except for m35D9. AC003662 contains 1.33 kb of *E. coli* sequence 1760 nt from the right end of the assembled BAC sequence. When this sequence is removed, AC003662 and 61D6 overlap by 7.8 kb. For each gene, the arrow points towards the 3′ end.

RSV/9-10-11, which contains human E9, E10, and E11 and portions of I9 and I10. Three genomic fragments containing E9, E10 and E11, each with adjacent intronic sequences, were amplified separately from PAC 61D6. The amplified products were digested at unique restriction sites engineered into the primers (underlined nt in the primer sequences below) and inserted individually into the multiple cloning site of expression vector pRcRSV (Invitrogen). The E9 segment, including 209 nt of downstream I9, was amplified with primer pair E9AF (5′- CCCAAGCTTGCTAGCCCGCGGGTGAACCTCCAAAATCAG -3′) and I9CR (5′- GGACTAGTAACTGAACTTCCTTGAAGAGGGTCC -3′) and digested with *Hin*dIII/*Spe*I before insertion into pRcRSV. The E10 segment with 567 bp and 190 bp of flanking I9 and I10, respectively, was amplified with the primer pair I9AF (5′- GGACTAGTGAGACTGAAGCCAGACTCCTAGATT -3′) and I10AR (5′- ATGCATCCTCACACTGGGAACAGTGGACCATG -3′). The E10 product was digested with *Spe*I/*Bst*XI and ligated just downstream of the E9 segment in pRcRSV. The E11 segment containing 640 bp of adjacent I10 was amplified with the primer pair I10BF (5′- CAGGAATGTCCATCACACTGGGTGTGCAGGTGCCTG -3′) and E11AR (5′- TGCTCTAGACTGGTTTATGATGGATGTTGCCTA -3′) digested with *Bst*XI/*Xba*I before insertion downstream of the E10 segment in pRcRSV.

Minigene AAH3 is a variant derived from RSV/9-10-11 by deleting 98 bp of the I10 sequence between the *Bst*XI and *Hin*dIII sites at the 5′ end of the E11 segment. The deleted sequence removes the first 95 bp of the 132-bp human-mouse conserved sequence in I10. AAH3 was generated by digesting pRSV/9-10-11 with *Spe*I/*Hin*dIII to remove the entire E10 segment, including the first 98 bp of adjoining E11 segment. Finally, the *Bst*XI site at the 3′ end of the E10 segment was replaced with a *Hin*dIII site by PCR mutagenesis to facilitate reinsertion of *Spe*I/*Hin*dIII-digested E10 fragment between the E9 and E11 segments. Minigene constructs AH3, S1AH3, and S1H3 are modifications of AAH3 where I9 and/or I10 sequences immediately flanking E10 are shortened.

The resulting constructs were used to transiently transfect COS-7 cells. Cell culture and transfection reagents were from Gibco-BRL. COS-7 cells were maintained in DMEM supplemented with 10% fetal calf serum. Splicing was assayed essentially as previously described (D'Souza et al., 1999). Transient transfections were performed in triplicate with 1 μg plasmid DNA with 6 μl Lipofectamine in a total of 700 μl OptiMEM per 35 mm well. Cells were exposed to the lipid/DNA complex for 5 h at 37°C in a 5% CO_2 incubator and allowed to recover with 700 μl of DMEM containing 20% fetal calf serum. Total cellular RNA was isolated 48 h later with TRIzol (BRL). RNA samples were DNase I-treated (Pharmacia) prior to reverse transcription. This RNA (2–2.5 μg) was reverse transcribed with random hexamers by using the GeneAmp RNA PCR kit (Perkin Elmer). E10+ (502 bp) and E10− (386 bp) spliced products were amplified by PCR by using vector-specific forward and reverse primers pREP (5′-GCTCGATACAATAAACGCCA-3) and BGHPA (5′-TAGAAGGCACAGTCGAGGC-3′), respectively. PCR reactions contained 1 ng of ^{32}P-labeled BGHPA and were performed for 18 cycles to obtain linear amplification before resolving by electrophoresis with 4% acrylamide gels. Quantitation was performed with a PhosporImager. For each mutant construct, values presented are the average of at least three different transfection experiments. Statistical comparisons were made using a two-tailed Student's *t*-test.

Results

MAPT region genomic sequence. The genomic sequence of MAPT and flanking genes was assembled from PAC 61D6, which was sequenced for this project, and from available database sequences (Fig. 2). The total contig described is 504 kb and extends 95.8 kb 5′ to the MAPT promoter and 274.5 kb 3′ to the MAPT terminal polyadenylation signal (Fig. 1, pA-2). A 161-kb mouse BAC (m35D9) containing *Mtapt* was also sequenced for comparison with the human gene. The mouse genomic sequence extends 28 kb 5′ to the *Mtapt* promoter and 23 kb 3′ to the *Mtapt* terminal polyadenylation signal (Fig. 1, pA-2). The human and mouse genes have similar GC-contents of 46.9% and 47.2%, respectively, and similar repetitive element compositions of 33.8% and 31.4%, respectively. There are three CpG islands associated with MAPT (Fig. 3, orange bars). The first island encompasses E0 spanning greater than 3 kb, while the second and third overlap E4A and E9, respectively. The E9 CpG island is potentially associated with an incomplete ALUSg segment.

Gene identification. To identify genes flanking or possibly nested within MAPT, the entire 504 kb of human sequence was analyzed by GENSCAN, BLASTN searches of the EST and EPD (eukaryotic promoter database) databases, and Powerblast searches of the NCBI non-redundant nucleotide and dbEST databases. All known MAPT/*Mtapt* exons were identified (Fig. 3, Table 1) by comparison with known cDNA sequences. MAPT, from E0, which encodes the 5′ UTR, to the end of the 3′ UTR, spans 133.9 kb. The mouse gene is smaller at 102.8 kb. GENSCAN identified 12 of the 15 MAPT exons, missing E2, E5, and E8. No additional exons were predicted within either the human or mouse tau genes.

The corticotropin releasing factor receptor gene (CRFR) is the next gene 5′ to MAPT. BLASTN searches identified 13 exons matching the CRFR cDNA sequence (Chen et al. 1993); Table 2). A potential polyadenylation signal for CRFR is located 55 kb upstream of the MAPT promoter. The initial CRFR exon(s) is not present in the available human genomic sequence (Table 2). In the 55 kb of sequence between the 3′ end of CRFR and the 5′ end of MAPT, no additional exons or genes were predicted by GENESCAN or identified by BLAST searches.

The gene located 3′ to MAPT and encoded by the opposite DNA strand was predicted by GENSCAN (Fig. 2, Table 3). The

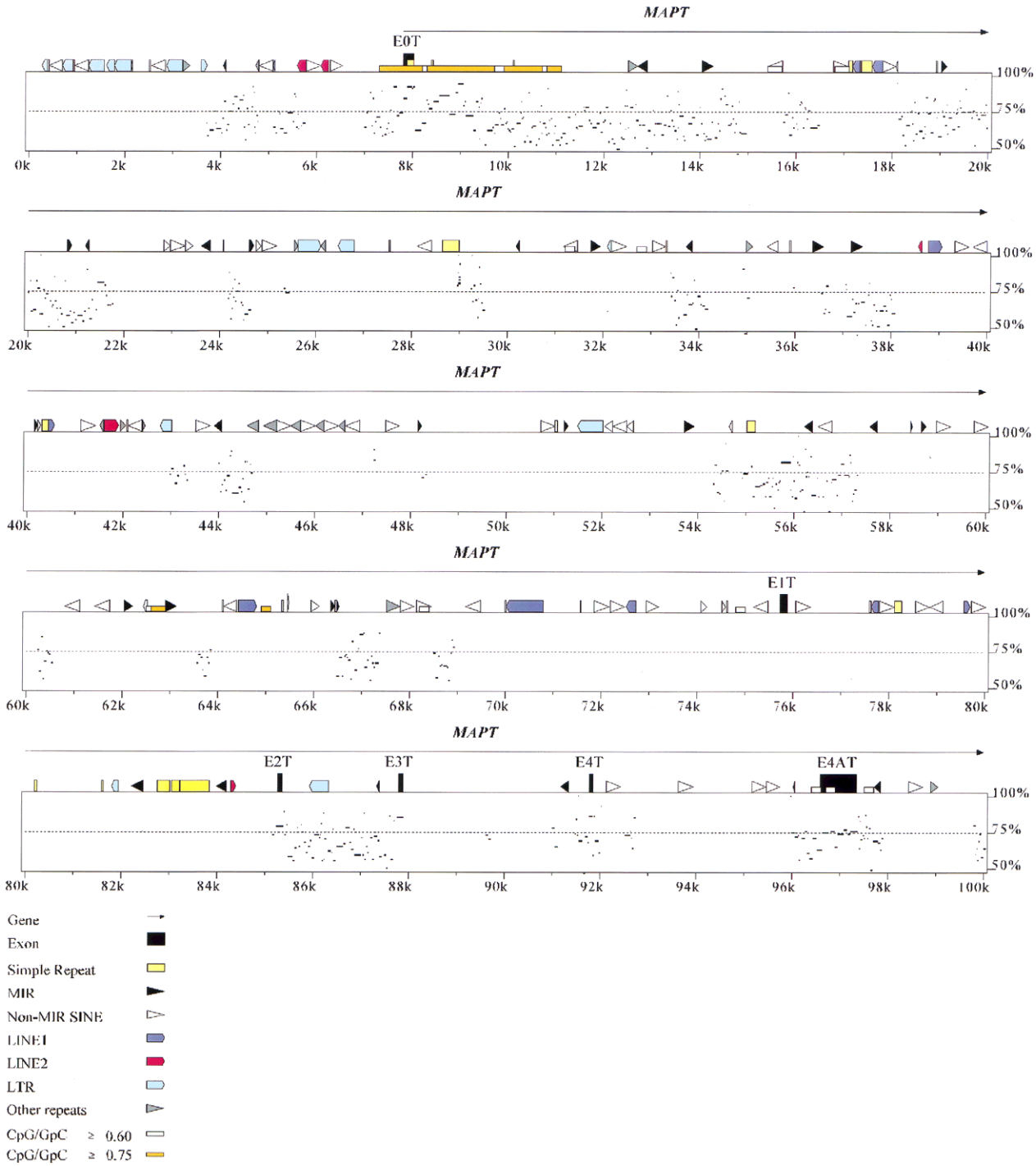

Fig. 3. Genomic organization and repeat content of MAPT and comparison with *Mtapt*. The location and orientation of MAPT and KIAA1267 are shown with an arrow above the sequence pointing to the 3′ end of each gene. Exons are black boxes, and repeat sequences are colored symbols with orientation indicated by the direction of the symbol. Simple repeats (including microsatellites, di- tri and tetra-nucleotide repeats) are represented by yellow boxes. Other repeats, including low complexity repeats that are primarily poly-purine or poly-pyrimidine stretches, are represented by gray boxes. Percentage identity plots (PIP) showing identity between MAPT and *Mtapt* were generated with Pipmaker (http://bio.cse.psu.edu; (Schwartz et al. 2000). The genomic region and nucleotide numbering shown are based on the MAPT sense strand sequence from PAC 61D6. PIP plots extend from 1 to approximately 180 kb. No mouse sequence is available for comparison for the final 20 kb region from 180 k to 200 k.

predicted coding sequence is identical to a 4730 bp cDNA sequence KIAA1267 from adult human brain (Nagase et al. 1999) and also matches a number of other ESTs. The terminal polyadenylation signal of KIAA1267, as predicted from the cDNA sequence, is 1.6 kb from the terminal polyadenylation signal of MAPT. Northern blot analysis of KIAA1267 shows two transcript sizes of 7.5 and 2.5 kb in brain (data not shown). KIAA1267 consists of 14 exons and spans a genomic distance of 143,973 nt (Table 3). The 5′ end of the KIAA1267 cDNA sequence is not fully characterized, and there may be additional 5′ exons not contained in the present contig. No sequence homology between human MAPT, KIAA1267, or CRFR was identified, and thus MAPT is not part of a cluster of genes encoding functionally related proteins.

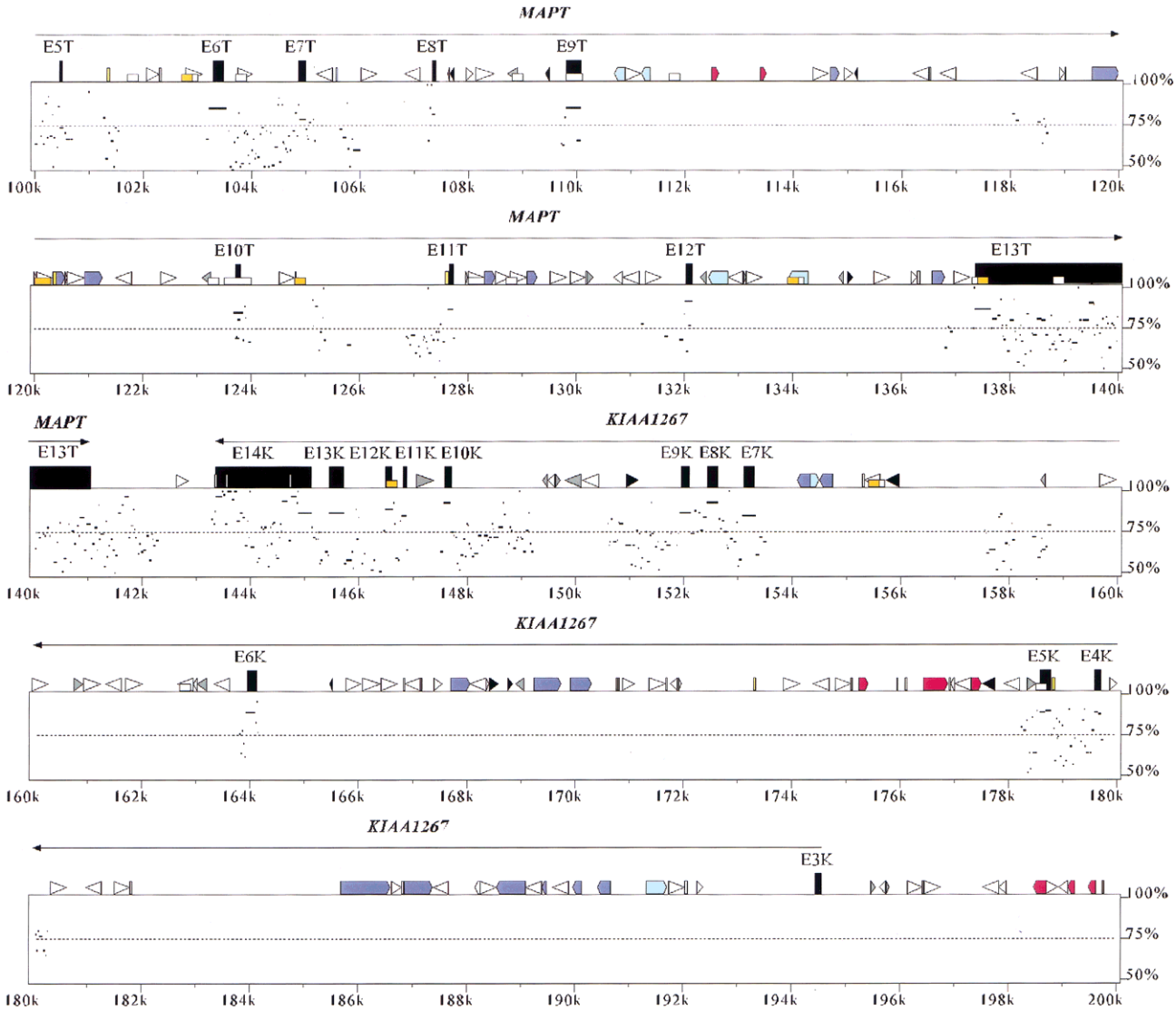

Fig. 3. *Continued.*

Human and mouse tau coding regions. The overlapping sequences from 61D6 and 35D9 were compared to identify similarities and differences between MAPT and *Mtapt.* Mouse and human tau protein sequences are 88% identical for the longest tau CNS isoform (4R23, Fig. 1). The greatest divergence is in the amino terminal end, where human tau has 11 additional amino acids in exon 1 (human amino acids 17–27) compared with mouse tau. Other differences are that the human protein has one additional amino acid in E2 (amino acid 50), two additional amino acids in E5 (amino acids 129–130), and two fewer amino acids in E7 (between amino acids 149 and 150). The highest amino acid conservation is within the microtubule binding domains and the carboxyl terminus with 98% identity in the last 252 amino acids encoded by E9–E13. Human and mouse coding regions for 4R23 are 83% identical at the nucleotide level, with 266 identical codons. Third position wobble results in amino acid conservation at 119 non-identical codons (26.9%). At four other non-identical codons, wobble at the first (2 codons) and second position (2 codons) results in amino acid conservation. The GC3 contents (GC content in the third codon position) of the longest human and mouse tau isoforms are 62% and 60%, respectively, indicating a potentially gene-rich iso-

Table 1. Comparative genomic structures of MAPT and *Mtapt.* Intron and exon sizes of MAPT and *Mtapt.*

Exon	*Mtapt* Exon Length (nt)	MAPT Exon Length (nt)	Intron	*Mtapt* Intron Length (nt)	MAPT Intron Length (nt)
E0	218	224	0	52,168	67,698
E1	100	150	1	4,539	9,388
E2	84	87	2	2,696	2,439
E3	87	87	3	4,865	3,904
E4	66	66	4	3,602	4,738
E4A	753	753	4A	3,664	3,109
E5	50	56	5	2,764	2,782
E6	198	198	6	1,217	1,384
E7	133	127	7	1,121	2,337
E8	54	54	8	2,322	2,421
E9	266	266	9	12,116	13,640
E10	93	93	10	3,100	3,840
E11	82	82	11	994	4,303
E12	113	113	12	5,247	5223
E13	208	4,363[a]	13	922	NA[b]
E14	2,969[a]	3,229[a]	NA	NA	NA

[a] Includes both coding and 3′UTR sequence. Length is up to the last nt in the AATAAA polyadenylation site.
[b] NA, not applicable.

Table 2. CRFR gene structure. Intron/exon structure of the CRFR gene.

Exon	cDNA nt[a]	Intron Size (nt)	Splice Acceptor[b]	Splice Donor
1[c]	1–33	>8,435	—	—
2	34–122	9,366	cctgcagGCC..	..CAGgtgagtc
3	123–241	4,772	cccccagGAC..	..CAAgtaagga
4	242–327	7,774	tttccagACA..	..GAGgtgagg
5	328–434	239	gctccagAAA..	..CAGgtgaga
6	435–521	446	tacccagGCC..	..AAGgtacctg
7	522–642	202	gcaccagGAG..	..GTGgtacgtc
8	643–796	309	gtggcagGGC..	..GGGgtgagc
9	797–857	2,197	cccccagGTG..	..GAAgtaagtc
10	858–930	241	cttctagGTG..	..CAGgtaaccg
11	931–1016	176	aattgcagATC..	..CTGgtaagaa
12	1017–1152	157	accccagGAA..	..CAGgtacagc
13	1153–1185	575	tccccagGGC..	..GAGgtgagg
14	1186–[d]	NA[e]	cccacagGTC..	NA

[a] Nt numbering based on the human CRFR mRNA (L23332).
[b] Small letters represent intronic sequences; capital letters represent exonic sequences.
[c] Exon 1 and the beginning of the subsequent intron are not present on BAC AC003662. Therefore, nt 1–33 may be contained in more than one exon. The intron between E1 and E2 is predicted to be greater than the remaining BAC sequence 5′ to E2. The CRFR gene is greater than 40 kb in size.
[d] The cDNA sequence L23332 does not contain the 3′UTR.
[e] NA, not applicable.

chore (Zoubak et al. 1996), although the overall GC content of the genomic sequence for MAPT and *Mtapt* is not high at 47% for both genes.

3′UTR. Tau-encoding transcripts of approximately 2 and 6 kb are observed in RNA from rodent and human CNS and neuroblastoma cell lines (Goedert et al. 1988; Wang et al. 1993; Sadot et al. 1994). The difference in length is due to the use of alternative polyadenylation sites (Fig. 1). In both human and rodents, the shorter transcript is the result of using polyadenylation sites 228 and 233 nt (human, mouse/rat, respectively), downstream of the stop codon shown in Fig. 4A. The resulting 3′UTR is highly conserved between rodents and human up to the polyadenylation signal, after which the sequences diverge. In mouse, transcripts have been described with I13 retained and with I13 spliced out, joining E13 and E14 (m4R23C, Fig. 1, Lee et al. 1988). When I13 is removed, the I13 polyadenyation site is gone and 23 amino acids are added to the C-terminal end of the protein (Figs. 1, 4). Thus, in mouse, alternative removal of I13 could regulate polyadenylation site selection and the type of 3′UTR utilized. However, for human and rat, no transcripts lacking I13 have been identified as a cDNA clone or an EST. Sequence divergence between species at splice sites could potentially result in loss of the ability to remove I13. While human, rat, and mouse have identical 5′-splice-site sequences at the end of E13 (Fig. 4A), the 3′-splice site at the beginning of a potential E14 in humans is not conserved compared with mouse and rat (Fig. 4B). The human sequence aca-GAA is missing a single nt compared with the mouse acagGAA (Fig. 4B). The result is that the human 3′splice site does not match the typical 3′ splice site consensus sequence (cagG), and thus may be a suboptimal splice site compared with the mouse/rat sequence, though the human sequence could still potentially function in splicing. While rat and mouse are identical at the I13/E14 junction, the sequences of potential RNA splicing branch-points upstream of E14 differ and may prevent I13 removal in the rat and human (Fig. 4B).

The long 3′UTR sequence found in 6-kb transcripts was reported for rat (Sadot et al. 1994), but not mouse or human. The long rat 3′UTR sequence, determined from a poly-A tailed cDNA clone includes a sequence equivalent to mouse I13 and E14 (Fig. 1A) and extends 3752 nt beyond the termination codon shown in Fig. 4A. Comparison of the rat 3′UTR with mouse and human genomic sequences identified polyadenylation signals (AATAAA) in approximately the same position as in rat, 4148 nt and 4363 nt past the first nt of E13 for mouse and human, respectively, in a region highly conserved in all three species (Fig. 5B). No other polyadenylation signals are present in either the human or mouse sequence between the first site in I13 and the second site predicted from the rat cDNA sequence. Thus the polyadenylation signal appears conserved for all three species. Consistent with this interpretation, numerous human ESTs align without gaps across the human 3′UTR, but none extend significantly past the polyadenylation site shown in Fig. 5B.

Tau protein in neurons is found in the axonal compartment and the cell body, but is not in dendritic compartments. Functional analysis of rat mRNA trafficking suggests that the 6-kb form contains a *cis*-element that targets tau mRNA to the axonal hillock by a microtubule-dependent process (Litman et al. 1993, 1996; Litman 1994; Behar et al. 1995; Aranda-Abreu et al. 1999). Previous work with rat tau indicates this targeting element is in the first 2744 nt of the long 3′UTR (Behar et al. 1995). Within this region, a 91-nt segment was identified containing a T-rich element that binds HuD, a human emybryonic lethal abnormal vision (ELAV)-like, RNA-binding protein (Aranda-Abreu et al. 1999). ELAV proteins stabilize mRNA by binding to U-rich sequences and are possibly part of a microtubule-associated RNA-protein particle involved in mRNA trafficking. However, in the human MAPT 3′UTR, the equivalent region is not well conserved and is missing the poly-T track present in rat (Fig. 5A). The equivalent T-rich region in mouse is also not well conserved compared with the rat sequence. Thus, human and possibly mouse tau mRNA may not bind HuD, at least not at the same site as in rat.

Comparison of MAPT and Mtapt introns. The distribution and orientation of repetitive DNA elements within MAPT introns are shown in Fig. 3. The highest density of SINE elements in the human and mouse genes are in I0 and I9. The density of 0.5 *Alu* repeats/kb for 61D6 is only slightly above the predicted average density for the human genome (0.25Alu/kb) (Moyzis et al. 1989). The MAPT introns with the lowest densities of repeats are in I4 and I10, each with a single SINE element. Human/mouse intron-size ratios are approximately 1:1 with the exception of I1, I7, and I11 where the ratios are 1.8, 2.0, and 4.3, respectively (Table 1). The inter-species differences in intron lengths are due in part to insertion of repeat elements within the human introns. For example, human I9 contains 5006 nt of repeat sequence (23 repeat elements), while mouse I9 contains only 1970 nt of repeat sequence (20 repeat elements, Fig. 6). The non-repeat element, non-conserved unique sequence content also differs between human and mouse. For I9, the non-repeat non-conserved sequence content is 7298 nt for the human gene and 8955 nt for the mouse gene (Fig. 6).

Within the introns, there are 105 islands of conserved sequence between mouse and human (Fig. 3) defined as regions exceeding 67% identity. These islands range in size from 20 nt to greater than 1700 nt. The 1-kb region 5′ to E0 is extensively conserved including the 335 nt promoter immediately 5′ to E0 (Andreadis et al. 1996; Sadot et al. 1996; Oliva et al. 1998). The promoter/E0 region is within an extensive CpG island. This region contains adjacent islands of sequence conservation that extend 7 kb into I0. I0, which separates constitutively spliced E0 and E1, is extensively conserved and contains 11 islands of conserved sequences downstream of the promoter/CpG island region. Upstream of the promoter region is another area of conserved sequence (at 4–6 kb; Fig. 3) that may potentially regulate gene expression. Another extensively conserved intron is I2, which is 67% identical between mouse and human. Since both E2 and E3 are alternatively spliced, this intron may be critical for regulation of usage of these two exons. Likewise, I6 is highly conserved with E6 being an alternatively spliced exon. Introns with the most sequence conservation are those located 3′ to alternatively spliced exons (I2, 63%; I3, 38%; I4A, 42%; I6, 31.4%; and I10, 35.6%; where the percentage

Table 3. KIAA1267 comparative gene structure[a]. Intron/exon structure of the KIAA1267 gene.

Exon	Human Exon Size (nt)	Human Intron Size (nt)	Human Splice Acceptor Site[a]	Human Splice Donor Site[b]	Mouse Exon Size (nt)	Mouse Intron Size (nt)	Exon Sequence Conservation (% Identical nt's)
1	1,377	78,110		..CCTgtga	NS	NS	NS
2	142	12,012	tttagGAG..	..AAGgtaa	NS	NS	NS
3	102	14,726	ggagGTA..	..AGGgtgt	NS	NS	NS
4	119	816	atcagATT..	..CACgtaa	116	803	86
5	196	14,453	tatagTCT..	..AAGgtcg	193	6,090	88
6	172	10,651	acagGTT..	..ATGgtaa	166	6,771	87
7	183	1,015	gcagATG..	..CCAgtaa	183	444	90
8	189	340	ttcagAGC..	..AAGgtag	189	347	92
9	149	4,252	gcagTGT..	..TCGgtaa	149	4,166	87
10	125	700	acagCAG..	..CAGgtag	122	1,296	94
11	58	210	gcagCTG..	..GAGgtaa	58	215	93
12	113	780	gcagATT..	..CAGgtga	113	521	92
13	253	383	gcagGTC..	..CAGgtaa	264	229	89
14	1,766	NA[c]	acagTCT..	NA[c]	405	NA	NA

[a] The human KIAA1267 gene is greater than 143 kb.
[b] NS = no sequence available.
[c] NA = not applicable.

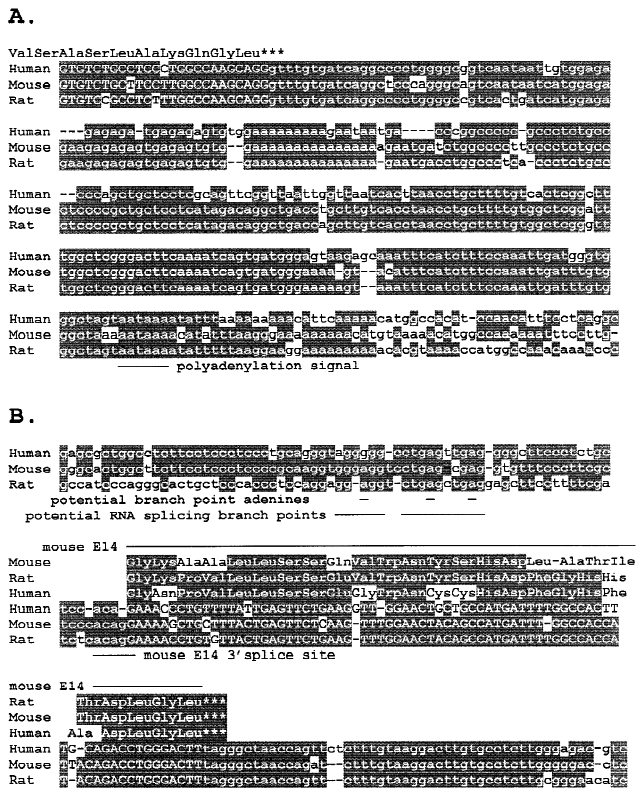

Fig. 4. Sequence comparison for human, mouse, and rat tau genes for E13 and the first polyadenylation site. The rat sequence accession number is X79321 (Sadot et al. 1994). **A.** Alignment for the end of E13 and the beginning of I13. The amino acid sequence for the c-terminal end of tau is identical in the three species. The position of the polyadenylation site is based on the rat cDNA sequence and location of the rat terminal poly-A sequence. **B.** Alignment of the end of I13 (mouse) and beginning of E14. The removal of I13 has been observed in mouse cDNA clones but not in rat or human cDNA clones. Thus, the predicted human and rat amino acid sequence of the region corresponding to mouse E14 is presented only for purposes of comparison to the mouse sequence.

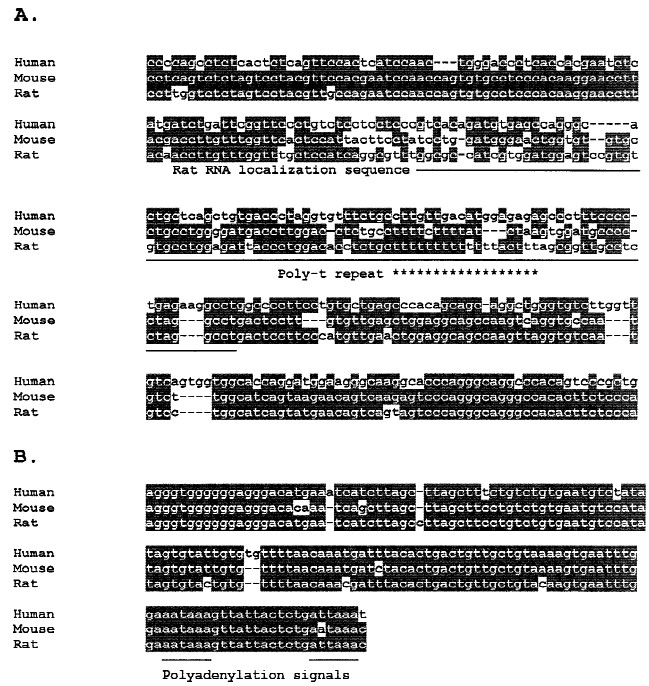

Fig. 5. Comparison of human, mouse, and rat long 3′UTR. **A.** Sequence alignment for the putative mRNA targeting region of the 3′UTR. The rat RNA localization signal corresponds to the 91-nt sequence described previously (Aranda-Abreu et al. 1999). **B.** Alignment of the region of the terminal polyadenylation signal. The location of the polyadenylation signal is based on the location determined from rat cDNA clones. No other polyadenylation signal is present between those shown here and the polyadenylation site shown in Fig. 4A.

given is defined as the total number of nucleotides in all conserved regions within an intron divided by the total length of the intron). The lowest level of sequence conservation is found in I8 (10.4%), I9 (14%), I11 (14.7%), and I12 (3.9%).

The E10 region of MAPT was examined in detail because E10 splicing regulation is altered by some FTDP-17 mutations. Also, in PSP (Vermersch et al. 1994; Schmidt et al. 1996; Chambers et al. 1999), Pick's disease, and CBD (Ksiezak-Reding et al. 1994; Buee-Scherrer et al. 1996), regulation of E10 splicing appears to be altered. However, no specific MAPT mutation or allele is known for PSP or CBD. Also, E10 splicing is developmentally regulated, and E10 usage in the adult CNS differs between mouse and human. Immediately upstream of E10, only about 20 nt of sequence are conserved. In MAPT, starting at nt −21 relative to E10, there is a 60-nt tandem repeat (93% identical) that is not present in mouse. Immediately 3′ to E10, the first 50 nt are partially homologous, but the mouse and human sequences are not

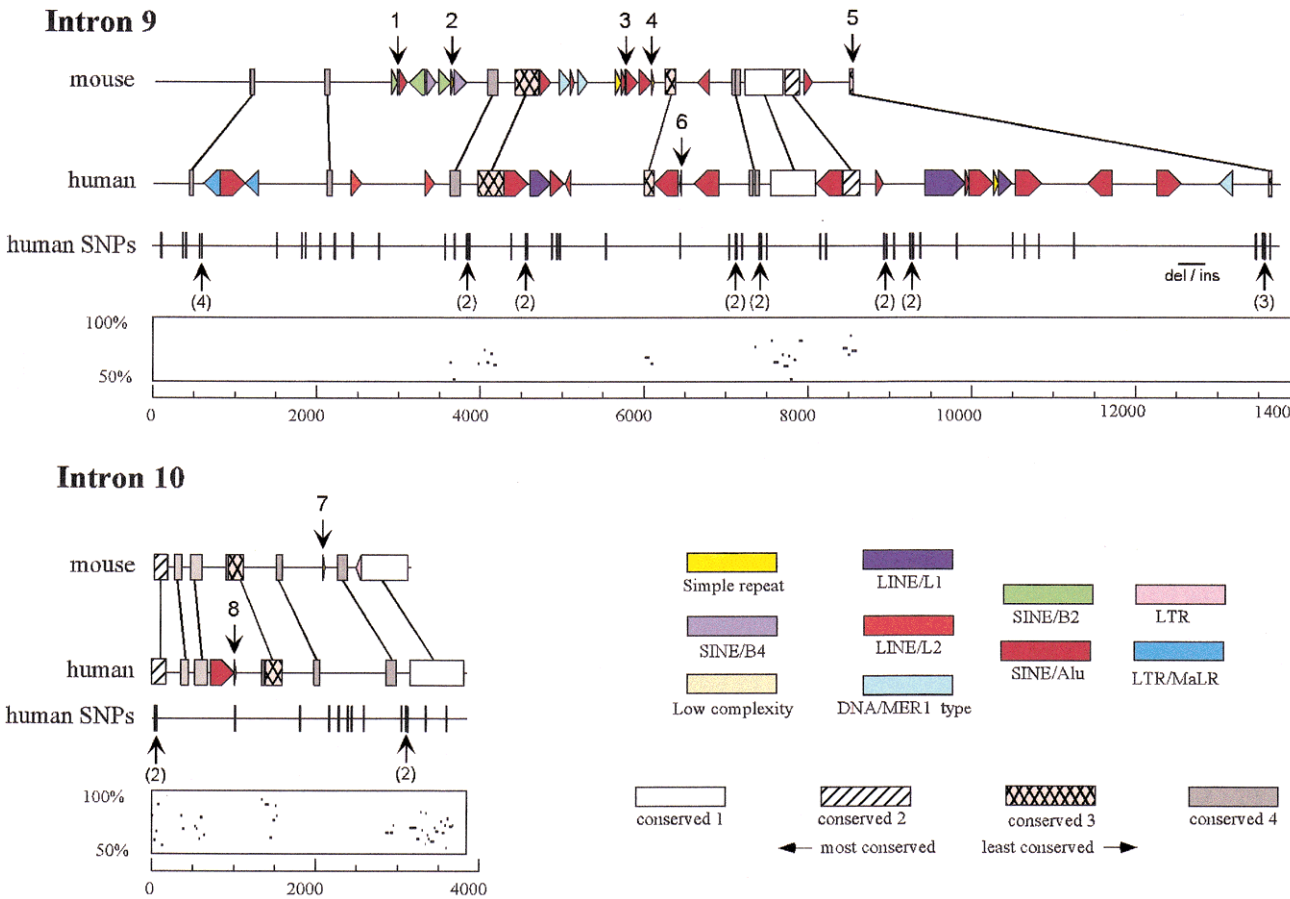

Fig. 6. Repeat and SNP content of MAPT I9 and I10 and sequence comparison of MAPT with *Mtapt.* Different classes of repeat elements are indicated by colored boxes with the repeat orientations shown by the direction of the point. Conserved elements were assigned a Crossmatch score based on a matrix that assigns 3 points to a match, −2 points to a transition mismatch, and −4 points to a transversion mismatch. The total score is the sum of the scores for all aligned pairs, which is dependent on the percentage identity, the length of the conserved region, and the nature of the mismatches. Conserved elements are boxes where the conservation level is indicated by the calculated Crossmatch scores between MAPT and *Mtapt:* conserved 1, score >160; conserved 2, 120 < score <160; conserved 3, 80 < score < 120; conserved 4, 40 < score < 80. Small repeat or conserved elements that are difficult to resolve are indicated numerically above the element: 1, LTR; 2,4,6, and 7, simple repeats; 3, SINE/B4; 5, conserved 3; 8, low complexity repeat. Human SNP locations are shown with hatch marks along a bar. When closely spaced SNPs are not distinguishable, the number of SNPs at a given site is given in parentheses beneath the SNP location. PIP plots are for comparison of MAPT and *Mtapt.*

functionally equivalent in splicing (D'Souza and Schellenberg 2000). The amount of conserved sequence is similar for I9 and I10, even though I9 is substantially longer. I9 and I10 contain 10 and 8 conserved islands representing 1336 nt and 1353 nt, respectively. In I9, the most highly conserved segment is a 467-nt sequence that is 71% identical, with 19 gaps (average gap size is 5.1 nt) and non-identical nt being 80 transitions and 55 transversions. This sequence (white box in I9; Fig. 6) is part of a cluster of conserved blocks. This cluster in mouse is much closer to the 3′ splice site compared with the location of these sequences in the human intron. In I10, the most highly conserved element is 562 nt in length, 70% identical, with 31 gaps (average gap size is 3.5 nt), and the non-identical nucleotides are 94 transitions and 75 transversions. The location of this sequence is similar in human and mouse.

*Functional analysis of MAPT conserved elements in RNA splicing.*To determine the role of conserved sequences on the regulation of E10 splicing, a tau minigene was constructed (Fig. 7) and tested for splicing in COS-7 cells (Fig. 8). The minigene construct RSV/9-10-11 contains human E9, E10, and E11 separated by shortened I9 and I10 segments. The I9 and I10 segments used are sequences immediately flanking E9, E10, and E11. When the entire minigene was transfected into COS-7 cells, E10 was retained in 73% of the transcripts (Fig. 8). The 640 nt of I10 adjacent to E11 in RSV/9-10-11 are from a highly conserved region of I10 (Fig. 6). Interestingly, the 5′ end of this conserved sequence contains within a 132 base sequence, 8 CTG motifs, two of which are in tandem. In RNA transcripts, CUG-repeats in introns and in 3′UTR sequences are potential regulatory sequences and can bind to CUG-binding proteins resulting in altered RNA processing (Timchenko et al. 1996; Philips et al. 1998). However, when these eight CTG motifs were removed from RSV/9-10-11 by a 98-base deletion (construct AAH3), E10 inclusion was not altered (72% for AAH3 versus 73% for RSV/9-10-11; Fig. 8). Construct AAH3 was further modified to address the role of the intron sequences immediately adjacent to E10. A 567-nt I10 sequence immediately flanking E10 is conserved between human and mouse, while the 190-nt I9 segment is not (Fig. 6). Shortening only the adjacent I10 sequence from 190 to 51 bases in AH3 shows no change in E10 inclusion. Shortening only the adjacent I9 sequence from 567 to 33 bases in S1AH3 causes an increase in E10 inclusion to 80% (Fig. 8). Since the non-conserved 60 nt tandem repeat sequence present immediately 5′ to E10 is removed in S1AH3, the increase in E10 splicing suggests that the tandem repeat sequence down-regulates E10 usage. However, when both I9 and I10 sequences in S1H3 are shortened to 33 and 51 nt, respectively, E10 inclusion is the same as

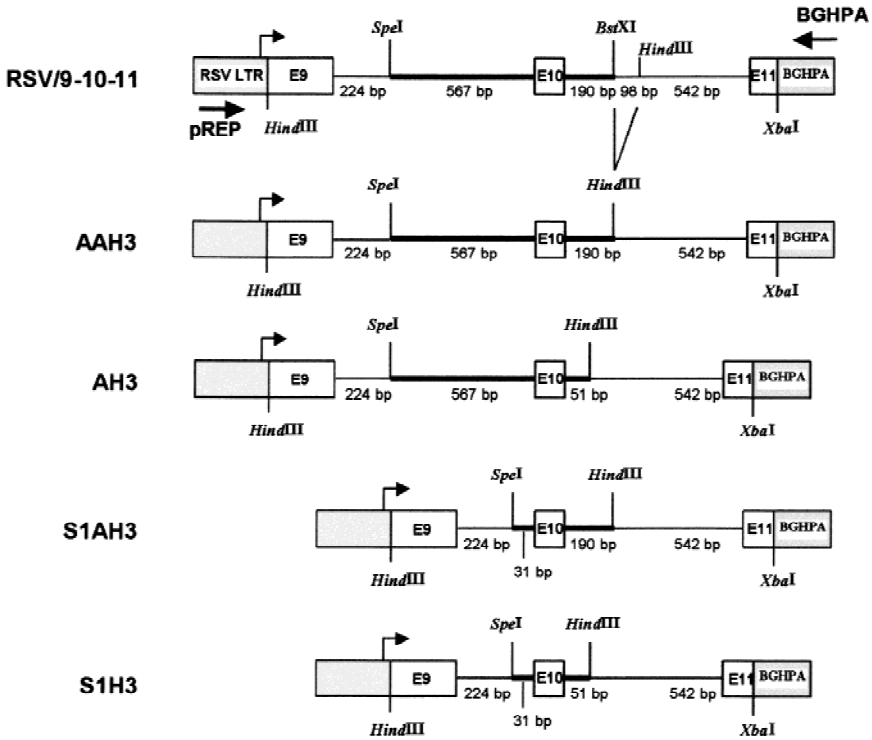

Fig. 7. MAPT minigene constructs for in vitro splicing assays. RSV/9-10-11 was generated by subcloning three PCR-amplified genomic fragments containing tau exons 9, 10, and 11, each with adjacent intronic sequences. The E9 segment includes 209 nt of downstream I9 sequence; the E10 segment contains 567 nt and 190 nt of flanking I9 and I10 sequences, respectively; and the E11 fragment contains 659 nt of adjacent I10 sequence. Minigene construct AAH3 is a 98-nt deletion mutant derived from RSV/9-10-11 by deleting the first 95 nt of the 132 nt human-mouse conserved sequence in I10 at the 5′ end of the E11 segment. Minigene constructs AH3, S1AH3 and S1H3 are modifications of AAH3 where I9 and/or I10 sequences immediately flanking E10 are shortened. Primers used for amplification of spliced transcripts (BGHPA and pREP) are indicated with arrows.

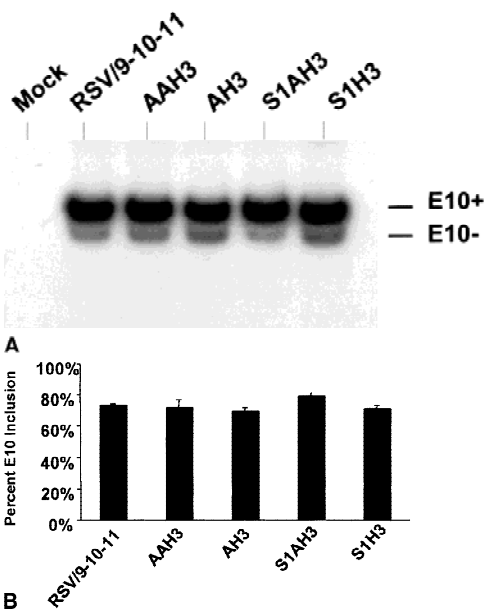

Fig. 8. The role of conserved versus non-conserved intron sequences on MAPT E10 splicing. **A.** Autoradiograph of reverse-transcription-PCR products from splicing assays utilizing the minigene constructs shown in Fig. 7. E10− and E10+ transcripts yield 261-bp and 354-bp fragments, respectively. **B.** Quantitation of E10+ and E10− splicing. Each bar represents the mean of three different transfection experiments, and 100% is the sum of E10+ and E10−. Error bars are the standard deviations. When E10 inclusion for the parent construct (RSV/9-10-11) was compared with the four derivative constructs, only results for S1AH3 were significantly different ($p = 0.000151$). When E10 inclusion for S1AH3 was compared with the other three derivative constructs, there was a significant difference between S1AH3 and AH3 ($p = 0.00045$), AAH3 ($p = 0.00045$), and S1H3 ($p = 0.0011$).

AAH3. The results suggest that removal of a negative sequence in I9 is offset by removal of a positive sequence in I10.

MAPT SNP and insertion/deletion polymorphisms. Polymorphisms in MAPT exons and some intronic polymorphic sites have been described previously (Hutton et al. 1998; Poorkaj et al. 1998; Baker et al. 1999; Bullido et al. 2000). To identify polymorphic sites potentially involved in susceptibility to PSP and other tauopathies, 12 normal control subjects were sequenced for segments of the 5′UTR, I1, I2, and I3, and all of I9, I10, and the complete long 3′UTR. In total, 94 variable sites were identified, with the majority of the polymorphisms occurring in only one or two control subjects (Table 4, Fig. 6). The majority were simple SNPs or 1–2 nt insertion/deletion polymorphisms. However, an insertion/deletion site (± 237 nt) was detected in I9 713 nt from the 5′ end of E10. This polymorphism was not within a conserved sequence. Since all of I9 and I10 were sequenced in a panel of subjects, variability in conserved islands, repeat elements, and non-conserved, non-repeat sequences could be compared. Polymorphic sites occurred at 1/668 nt and 1/451nt of conserved island sequence for I9 and I10, respectively. In contrast, polymorphic sites in non-repeat non-conserved sequences were more frequent at 1/182 nt and 1/200 nt for I9 and I10, respectively. For I9, polymorphic sites were found at a rate of 1/500 nt of repeat element sequence. For I10, which has only two repeats, no polymorphisms were observed in the repeat elements.

Discussion

Genomic analysis of the genes encoding human and mouse tau leads to the following conclusions. First, MAPT is not part of a cluster of related genes, as the flanking genes (KIAA1267 and CRFR) show no homology to MAPT. As all three genes are expressed in the CNS, there may be regional *cis*-acting sequences

Table 4. Intronic and non-coding nucleotide changes in MAPT. Polymorphisms within MAPT.

MAPT Site	61D6 Location	nt Change	MAPT Site	61D6 Location	nt Change	MAPT Site	Location	nt Change
5′UTR	75,730[a]	A to G	I9	114,658	T to C	I10	123,832[a]	G to A
I1	85,172	T to C	I9	114,933[c]	G to A	I10	123,855[a]	A to C
I2	85,836	C to T	I9	115,022[c]	A to T	I10	124,807	C to A
I2	87,627[a]	C to G	I9	115,125[c]	G to A	I10	125,598	C to T
I2	87,628[a]	T to G	I9	115,607	C to T	I10	125,956	C to T
I2	87,651[a]	A to G	I9	116,505	GT deletion	I10	126,070	C to G
I3	87,885	A to G	I9	117,104	A to G	I10	126,183	A to C
I3	91,677[a]	T to A	I9	117,260	T insertion	I10	126,231	T to C
I3	91,686[a]	A to T	I9	117,264	G to A	I10	126,374	T to C
I4	91,890[a]	A insertion	I9	117,308	G to A	I10	126,835	C to G
I4	100,374	T to C	I9	117,471[a]	G insertion	I10	126,889	C to T
I8	109,779	G to A	I9	117,502	T to C	I10	126,893	C to T
I9	110,173	G to A	I9	117,567	T to C	I10	127,127	G to C
I9	110,475	G to C	I9	118,221[c]	A to G	I10	127,383[a]	G to A
I9	110,515[a]	C to T	I9	118,289	A to T	I11	127,447	G to A
I9	110,621	A to G	I9	118,996	T deletion	I11	127,624	A to G
I9	110,626	G to A	I9	119,021	G to A	I11	127,759[a]	G to A
I9	110,653	A to G	I9	119,121	G to A	I12	127,815	A to G
I9	110,659	G to A	I9	119,337	T to C	I12	136,950[a]	A to G
I9	111,575	T to C	I9	119,363	C to T	I12	136,960	C to T
I9	111,877	T to C	I9	119,442	G to A	3′ UTR	137,607[e]	T to C
I9	111,941	A to G	I9	119,885[d]	A to G	3′ UTR	137,649[e]	AAT deletion
I9	112,103	T to C	I9	120,568[d]	T to C	3′ UTR	137,822[e]	T insertion
I9	112,280	T to C	I9	120,711[c]	C to T	3′ UTR	137,894[e]	C insertion
I9	112,506[b]	A to G	I9	120,879[c]	G to C	3′ UTR	138,494[e]	CT insertion
I9	112705	A to G	I9	121,317	T deletion	3′ UTR	138,727[e]	T deletion
I9	113,630	G to A	I9	122,761–122,996	237 nt deletion	3′ UTR	138,787[e]	CA deletion
I9	113,748	G to T	I9	123,535	C to A	3′ UTR	138,910[e]	A to C
I9	113,890	G to A	I9	123,543	G to A	3′ UTR	138,978[e]	T to C
I9	113,891	T to A	I9	123,604	G to A	3′ UTR	139,342[e]	C to T
I9	114,392[c]	G to A	I9	123,653	C to A			
I9	114,656	G to A	I9	123,664	G to T			

[a] Located in a region conserved between mouse and human.
[b] Located in LTR/MaLR repeat.
[c] Located in a SINE/Alu element.
[d] Located in a LINE/LI element.
[e] Located in the 3′UTR which is primarily conserved between mouse and man.

possibly affecting chromatin structure, permitting these genes to be coordinately expressed. However, no evidence directly supports this hypothesis. Second, there is no evidence for additional exons not previously described for either MAPT or *Mtapt.* Also, the sequence comparison indicates that the alternative splicing of E14 in mouse but not human is probably caused by differences in the sequences equivalent to the 3′splice site for mouse E14, or the upstream splicing branch point (Fig. 4B). Third, there are substantial regions of sequence conservation between MAPT and *Mtapt*, providing evidence for numerous *cis*-acting regulatory sequences that control tau gene expression and alternative splicing.

Conserved sequences in non-coding regions potentially regulate gene expression, RNA splicing, chromatin structure, nuclear matrix attachment, or chromosome function (e.g. replication, cell division, etc.). A recent comparison of the noncoding regions of 77 orthologous mouse and human gene pairs found that blocks of >60% identity covered, on average, 36% of regions 5′ of the genes, 50% of the 5′ UTR's, 23% of the introns, and 56% of the 3′ UTRs (Jareborg et al. 1999). This conservation pattern is consistent with the results for comparison of MAPT and *Mtapt* reported here, where conserved regions include an approximately 2-kb island that is 1 kb upstream of the promoter, the promoter region and flanking sequences, substantial portions of I0, some intronic regions, the 3′UTR, and about 1 kb of sequence 3′ to the apparent end of transcription (Figs 3, 6). These conserved sequences, or specific sites within these sequences, potentially interact with *trans*-acting proteins to regulate gene expression and RNA splicing. As noted by others (Duret and Bucher 1997), DNA- and RNA-binding proteins typically interact with relatively short 5 to 25-nt sequences. Yet, most of the conserved blocks observed here and for other genes are longer, with some MAPT blocks extending up to 6 kb (Fig. 3). Several explanations are possible. First, secondary DNA or RNA structures involving hundreds to thousands of nucleotides could be important for gene regulation, though examples of such regulatory structures are limited. Second, each conserved region may contain multiple *trans*-acting factor binding sites. A clustering of binding sites could explain conserved sequences of several hundred nucleotides such as those found in I9 and I10 (Fig. 6). An example of such a clustering of regulatory sites is the recognition and regulation of alternatively spliced exons, a process that requires numerous discrete *cis*-acting elements within and flanking a single exon. For example, regulation of MAPT E10 alternative splicing involves at least five distinct exon splicing silencer and enhancer elements within the 93-nt exon along with two or more elements in I10 immediately flanking E10 (Hutton et al. 1998; D'Souza et al. 1999; D'Souza and Schellenberg 2000). Functional analysis of human tau I9, I10, and I11 sequences (Fig. 8, D'Souza and Schellenberg 2000) confirms the presence of multiple intronic splicing regulatory elements that control E10 inclusion. Since at least some if not all of these regulatory elements interact to control splicing, all of these sites may be simultaneously occupied during the process leading to exon definition. Another well-characterized example of the regulatory significance of extensive conserved sequences is the locus control region (LCR) of the β-globin gene cluster (Li et al. 1999). Extensive analysis of this 16-kb LCR indicates that normal regulation of the β-globin gene cluster requires the presence of at least four conserved non-coding sequence blocks (50–70% conserved), each co-localizing with a unique DNase hypersensitive site. Individual conserved blocks span up to 2 kb and can contain multiple *trans*-acting factor binding sites (Hardison et al. 1997). Thus, the extensive tracks of conserved sequences in MAPT (Fig. 3) may reflect elements where

multiple proteins bind to regulate gene expression and RNA splicing both during development and in the adult organism.

Comparison of MAPT and *Mtapt* should help to identify the similarities and differences in how humans and mice regulate alternative splicing of E10. Both species completely exclude E10 early in development. In the adult mouse, E10 is present in all transcripts, while in adult humans, E10 is present in approximately 50% of the transcripts. This splicing difference between mouse and human could be due to a difference in either the complement or function of *trans*-acting factors present in brain. Alternatively, the difference could be caused by sequence differences between MAPT and *Mtapt.* Preliminary transgenic animal work suggests both explanations are correct. When the entire human gene cloned as BAC 61D6 (Fig. 2) is used as a transgene, in the adult animals, E10 is included in approximately 15–20% of transcripts from the human gene. In the same animals, transcripts from the mouse gene are 100% $E10^+$ [Poorkaj and Schellenberg, unpublished data; see also Grover et al. (1999)]. Thus the human MAPT sequence dictates that the human E10 is alternatively spliced in the adult mouse. However, since human E10 is not included to the same extent in mouse compared with human, *trans*-acting factors must also differ between the two species.

The sequence elements that control developmental regulation of E10 inclusion are unknown. Alternative splicing is typically controlled by sequences in the introns and splice sites directly flanking alternatively spliced exons. Thus, candidate E10 developmental regulatory signals are the conserved sequences in I9 and I10 (Fig. 6). Differences between human and mouse E10 adult splicing could be owing to specific sequence differences in the conserved segments. Alternatively, the position of these conserved elements relative to the exons could alter their regulatory function. Specifically, in mouse I9, the paired conserved elements near the 3′ end of the intron (white box and hatched black and white box, Fig. 6) have been split in humans by the insertion of a LINE element, and the mouse elements are much closer to E10 than in mouse.

Interpretation of orthologous sequence comparisons is complicated by the fact that evolutionary changes in a regulatory sequence can be compensated for by evolutionary changes in another regulatory element. For example, when E10, flanked by short regions of I9 and I10 and heterologous exons, is assayed for splicing, E10 inclusion is roughly equivalent when either MAPT or *Mtapt* sequences are used (45% and 34%, respectively, D'Souza and Schellenberg 2000). Splicing in this system is regulated in part by an intron splicing silencer (ISS) located in the first 53 nt after E10. Because of specific sequence differences, the human ISS is a much weaker inhibitor than the mouse ISS. In contrast, within E10, there is an exon-splicing silencer (ESS) that differs between MAPT and *Mtapt* by 1 nt (E10 position 57). The result is that the human ESS is stronger than the mouse ESS (D'Souza and Schellenberg 2000). Thus, for the human gene, the weaker ISS off-sets the stronger ESS to yield near-equivalent splicing ratios indicating that in this case, evolutionary changes at one location are offset by changes at another element.

The 3′UTR for the tau genes is highly conserved, as is the case for numerous other genes (Tournier-Lasserve et al. 1989; Lipman 1997; Jareborg et al. 1999). For tau, the long 3′UTR presumably contains a binding site for a protein that transports MAPT mRNA to the axonal hillock in neurons, while transcripts with the short 3′UTR remain in the cytoplasm (Wang et al. 1993; Thurston et al. 1997). For the rat, the 3′UTR site responsible for transport is less than 91 nt in length and thus does not explain why the remainder of the 3′UTR is conserved. The 3′UTR also may contain sequences to regulate transcript stability. A unique feature of MAPT/*Mtapt* is that the conservation of the 3′UTR extends over 2 kb past the end of the longest transcript. Note that for KIAA1267, the terminal exon that contains the 3′UTR is conserved, but the region of identity stops at the end of the exon (E14K; Fig. 3). The conserved block 3′ to the end of E13T may be important possibly in determining termination of transcription or in regulating some more distant aspect of tau gene expression.

Polymorphism analysis of human I9 and I10 was undertaken to identify potential sites contributing to susceptibility to PSP. A number of polymorphisms within MAPT show a genetic association with PSP (Conrad et al. 1997; Baker et al. 1999). However, since polymorphic sites across MAPT are in linkage disequilibrium, it is not possible to determine which sites contain the true susceptibility allele. Because the PSP populations are not strongly familial, presumably susceptibility is caused by a relatively common allele with low penetrance rather than a rare mutation. Because in PSP, in affected regions of the brain, there appears to be elevated 4R tau produced from E10-containing transcripts (Vermersch et al. 1994; Schmidt et al. 1996), the susceptibility allele may affect regulation of E10 splicing (Chambers et al. 1999). The polymorphisms detected in I9 and I10 are candidates for PSP susceptibility alleles. The comparative sequence analysis performed here permits classification of these polymorphisms as occurring in repeat elements, in conserved and in non-conserved sequences. Obviously, the latter two categories are more likely to be involved in splicing regulation, and these are better candidate PSP susceptibility sites. Functional analysis of these polymorphic sites in splicing assays will be needed to determine which site and allele confers susceptibility to PSP.

Comparison of MAPT and *Mtapt* identifies a large number of conserved sequences that are potential regulatory elements, because DNA and RNA binding motifs within these elements are difficult to identify unambiguously. To completely understand how MAPT is regulated, functional analysis combined with binding assays of *trans*-acting factors is needed.

Acknowledgments. Supported by National Institute on Aging Grant R01-AG11762 (GDS), and PO10135316 (GDS) a Merit Award from the Department of Veterans Affairs (GDS), and an Eloise H. Troxel Memorial grant from the Society for Progressive Supranuclear Palsy (PP). We thank Elaine Loomis and Leojean Anderson for technical assistance.

References

Andreadis A, Brown WM, Kosik KS (1992) Structure and novel exons of the human-tau gene. Biochemistry 31, 10626–10633

Andreadis A, Wagner BK, Broderick JA, Kosik KS (1996) A tau promoter region without neuronal specificity. J Neurochem 66, 2257–2263

Aranda-Abreu GE, Behar L, Chung S, Furneaux H, Ginzburg I (1999) Embryonic lethal abnormal vision-like RNA-binding proteins regulate neurite outgrowth and tau expression in PC12 cells. J Neurosci 19, 6907–6917

Baker M, Litvan I, Houlden H, Adamson J, Dickson D et al. (1999) Association of an extended haplotype in the tau gene with progressive supranuclear palsy. Hum Mol Genet 8, 711–715

Behar L, Marx R, Sadot E, Barg J, Ginzburg I (1995) Cis-acting signals and transacting proteins are involved in tau mRNA targeting into neurites of differentiating neuronal cells. Int J Dev Neurosci 13, 113–127

Boyne LJ, Tessler A, Murray M, Fischer I (1995) Distribution of big tau in the central nervous system of the adult and the developing rat. J Comp Neurol 358, 279–293

Buee-Scherrer V, Hof PR, Buee L, Leveugle B, Vermersch P et al. (1996) Hyperphosphorylated tau proteins differentiate corticobasal degeneration and Pick's disease. Acta Neuropathol 91, 351–359

Bullido MJ, Aldudo J, Frank A, Coria F, Avila J, Valdivieso F (2000) A polymorphism in the tau gene associated with risk for Alzheimer's disease. Neurosci Lett 278, 49–52

Butner KA, Kirschner MW (1991) Tau protein binding to microtubules through a flexible array of distributed weak sites. J Cell Biol 115, 717–730

Caceres A, Kosik KS (1990) Inhibition or neurite polarity by tau antisense oligonucleotides in primary cerebellar neurons. Nature 343, 461–463

Caceres A, Potrebic S, Kosik KS (1991) The effect of tau antisense oli-

gonucleotides on neurite formation of cultured cerebellar macroneurons. J Neurosci 11, 1515–1523
Chambers CB, Lee JM, Troncoso JC, Reich S, Muma NA (1999) Overexpression of four-repeat tau mRNA isoforms in progressive supranuclear palsy but not in Alzheimer's disease. Ann Neurol 46, 325–332
Chen R, Lewis KA, Perrin MH, Vale WW (1993) Expression cloning of a human corticotropin-releasing-factor receptor. Proc Natl Acad Sci USA 90, 8967–8971
Clark LN, Poorkaj P, Wszolek Z, Geschwind DH, Nasreddine ZS et al. (1998) Pathogenic implications of mutations in the tau gene in pallidoponto-nigral degeneration and related neurodegenerative disorders linked to chromosome 17. Proc Natl Acad Sci USA 95, 13103–13107
Collet J, Fehrat L, Pollard H, Depouplana LR, Charton G et al. (1997) Developmentally regulated alternative splicing of mRNAs encoding N-terminal tau variants in the rat hippocampus: structural and functional implications. Eur J Neurosci 9, 2723–2733
Conrad C, Andreadis A, Trojanowski J, Dickson D, Kang D et al. (1997) Genetic evidence for the involvement of tau in progressive supranuclear palsy. Ann Neurol 47, 277–281
D'Souza I, Schellenberg GD (2000) Determinants of 4 repeat tau expression: coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion. J Biol Chem 275, 17700–17709
D'Souza I, Poorkaj P, Hong M, Nochlin D, Lee VMY et al. (1999) Missense and silent tau gene mutations cause front temporal dementia with parkinsonism–chromosome 17 type by affecting multiple alternative RNA splicing regulatory elements. Proc Natl Acad Sci USA 96, 5598–5603
Duret L, Bucher P (1997) Searching for regulatory elements in human non-coding sequences. Curr Opin Struct Biol 7, 399–406
Georgieff I, Liem RKH, Mellado W, Nunez J, Shelanski ML (1991) High molecular weight tau: preferential localization in the peripheral nervous system. J Cell Sci 100, 55–60
Georgieff I, Liem RKM, Couchie D, Mavilia C, Nunez J et al. (1993) Expression of high molecular weight tau in the central and peripheral nervous system. J Cell Sci 105, 729–737
Goedert M, Wischik CM, Crowther RA, Walker JE, Klug A (1988) Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. Proc Natl Acad Sci USA 85, 4051–4055
Goedert M, Spillantini MG, Jakes R, Rutherford D, Crowther RA (1989) Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. Neuron 3, 519–526
Grover A, Houlden H, Baker M, Adamson J, Lewist J et al. (1999) 5′ Splice mutations in *tau* associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10. J Biol Chem 274, 15134–15143
Gu YJ, Oyama F, Ihara Y (1996) Tau is widely expressed in rat tissues. J Neurochem 67, 1235–1244
Gustke N, Trinczek B, Biernat J, Mandelkow EM, Mandelkow E (1994) Domains of tau protein and interactions with microtubules. Biochemistry 33, 9511–9522
Hardison R, Slightom JL, Gumucio DL, Goodman M, Stojanovic N, Miller W (1997) Locus control regions of mammalian β-globin gene clusters; combining phylogenetic analysis and experimental results to gain functional insights. Gene 205, 73–94
Higgins JJ, Litvan I, Pho LT, Li W, Nee LE (1998) Progressive supranuclear gaze palsy is in linkage disequilibrium with the tau and not the alpha-synuclein gene. Neurology 50, 270–273
Himmler A (1989) Structure of the bovine tau gene: alternatively spliced transcripts generate a protein family. Mol Cell Biol 9, 1389–1396
Himmler A, Drechsel D, Kirschner MW, Martin DW (1989) Tau consists of a set of proteins with repeated C-terminal microtubule-binding domains and variable N-terminal domains. Mol Cell Biol 9, 1381–1388
Hong M, Zhukareva V, Vogelsberg-Ragaglia V, Wszolek Z, Reed L et al. (1998) Mutation-specific functional impairments in distinct Tau isoforms of hereditary FTDP-17. Science 282, 1914–1917
Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S et al. (1998) Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17. Nature 393, 702–705
Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human genes. Genome Res 9, 815–824
Kosik KS, Orecchio LD, Bakalis S, Neve RL (1989) Developmentally regulated expression of specific tau sequences. Neuron 2, 1389–1397
Ksiezak-Reding H, Morgan K, Mattiace LA, Davies P, Liu W-K et al. (1994) Ultrastructure and biochemical composition of paired helical filaments in corticobasal degeneration. Am J Pathol 145, 1496–1508
Lee G, Cowan N, Kirschner M (1988) The primary structure and heterogeneity of tau protein from mouse brain. Science 239, 285–288
Lendon CL, Lynch T, Norton J, Mckeel DW, Busfield F et al. (1998) Hereditary dysphasic disinhibition dementia: a frontotemporal dementia linked to 17q21-22. Neurology 50, 1546–1555
Levy-Lahad E, Poorkaj P, Wang K, Fu YH, Oshima J et al. (1996) Genomic structure and expression of STM2, the chromosome 1 familial Alzheimer's disease gene. Genomics 34, 198–204. doi:10.10061/geno.1996.0266.
Li Q, Harju S, Peterson KR (1999) Locus control regions coming of age at a decade plus. Trends Genet 15, 403–408
Lipman DJ (1997) Making (anti)sense of non-codong sequence conservation. Nucleic Acids Res 25, 3580–3583
Litman P, Barg J, Gipzburg I (1994) Microtubules are involved in the localization of Tau mRNA in primary neuronal cell cultures. Neuron 13, 1463–1474
Litman P, Barg J, Rindzoonski L, Ginzburg I (1993) Subcellular localization of tau mRNA in differentiating neuronal cell culture: implications for neuronal polarity. Neuron 10, 627–638
Litman P, Behar L, Elisha Z, Yisraeli JK, Ginzburg I (1996) Exogenous tau RNA is localized in oocytes: possible evidence for evolutionary conservation of localization mechanisms. Dev Biol 176, 86–94. doi:10.1006/dbio.1996.9992.
LoPresti P, Szuchet S, Papasozomenos SC, Zinkowski RP, Binder LI (1995) Functional implications for the microtubule-associated protein tau: localization in oligodendrocytes. Proc Natl Acad Sci USA 92, 10369–10373
Mavilia C, Couchie D, Mattei MG, Nivez MP, Nunez J (1993) High and low molecular weight tau proteins are differentially expressed from a single gene. J Neurochem 61, 1073–1081
Mavilia C, Couchie D, Nunez J (1994) Diversity of high-molecular weight tau proteins in different regions of the central nervous system. J Neurochem 63, 2300–2306
Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu JR et al. (1989) The distribution of interspersed repetitive DNA sequences in the human genome. Genomics 4, 273–289
Nagase T, Ishikawa K, Kikuno R, Hirosawa M, Nomura N, Ohara O (1999) Prediction of the coding sequences of unidentified human genes. XV. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res 6, 337–345
Oliva R, Tolosa E, Ezquerra M, Molinuevo JL, Valldeoriola F et al. (1998) Significant changes in the tau A0 and A3 alleles in progressive supranuclear palsy and improved genotyping by silver detection. Arch Neurol 55, 1122–1124
Philips AV, Timchenko LT, Cooper TA (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. Science 280, 737–740
Poorkaj P, Bird TD, Wijsman E, Nemens E, Garruto RM et al. (1998) Tau is a candidate gene for chromosome 17 frontotemporal dementia. Ann Neurol 43, 815–825
Sadot E, Marx R, Barg J, Behar L, Ginzburg I (1994) Complete sequence of 3′-untranslated region of tau from rat central nervous system. J Mol Biol 241, 325–331. doi:10.1006/jmbi.1994.1508
Sadot E, Heicklenklein A, Barg J, Lazarovici P, Ginzburg I (1996) Identification of a tau promoter region mediating tissue-specific-regulated expression in PC12 cells. J Mol Biol 256, 805–812. doi:10.1006/jmbi.1996.0126
Schmidt ML, Huang R, Martin JA, Henley J, Mawaldewan M et al. (1996) Neurofibrillary tangles in progressive supranuclear palsy contain the same tau epitopes identified in Alzheimer's disease PHF tau. J Neuropathol Exp Neurol 55, 534–539
Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C et al. (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10, 577–586
Smit AFA, Green P. RepeatMasker. http://www.genome.washington.edu/RM/RepeatMasker.html. 1996–1997.
Spillantini MG, Murrell JR, Goedert M, Farlow MR, Klug A, Ghetti B (1998) Mutation in the tau gene in familial multiple system tauopathy with presenile dementia. Proc Natl Acad Sci USA 95, 7737–7741
Thurston VC, Pena P, Pestell R, Binder LI (1997) Nucleolar localization of

the microtubule-associated protein tau in neuroblastomas using sense and anti-sense transfection strategies. Cell Motil Cytoskel 38, 100–110

Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datae KV et al. (1996) Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res 24, 4407–4414

Tournier-Lasserve E, Odenwald WF, Garbern J, Trojanowski J, Lazzatini RA (1989) Remarkable intron and exon sequence conservation in human and mouse homologous genes. Mol Cell Biol 9, 2273–2278

Vanier MT, Neuville P, Michalik L, Launay JF (1998) Expression of specific tau exons in normal and tumoral pancreatic acinar cells. J Cell Sci 111, 1419–1432

Vermersch P, Robitaille Y, Bernier L, Wattez A, Gauvreau D, Delacourte A (1994) Biochemical mapping of neurofibrillary degeneration in a case of progressive supranuclear palsy: evidence for general cortical involvement. Acta Neuropathol 87, 572–577

Wang Y, Loomis PA, Zinkowski RP, Binder LI (1993) A novel tau transcript in cultured human neuroblastoma cells expressing nuclear tau. J Cell Biol 121, 257–267

Wei M-L, Andreadis A (1998) Splicing of a regulated exon reveals additional complexity in the axonal microtubule-associated protein tau. J Neurochem 70, 1346–1356

Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174, 95–102