Sequence of an expressed human β -tubulin gene containing ten *Alu* family members

Mary Gwo-Shu Lee, Cynthia Loomis and Nicholas J. Cowan

Department of Biochemistry, New York University School of Medicine, 550 First Avenue, New York, NY 10016, USA

Received 14 May 1984; Revised and Accepted 19 June 1984

ABSTRACT

The complete sequence of a functionally expressed human β -tubulin gene (5 β) is presented. The amino acid sequence encoded by this gene constitutes a distinct isotype, differing from a previously described human β -tubulin sequence at 21 positions throughout the polypeptide chain. The β -tubulin coding sequence in 5 β is interrupted by three intervening sequences of 1014, 117 and 4826 nucleotides. The largest of these contains ten members of the *Alu* family of middle repetitive sequences. Together, these regions account for sixty percent of this intervening sequence. Two of the *Alu* elements are juxtaposed head to tail, and share the same flanking direct repeat. The ten *Alu* sequences are substantially divergent, both from each other and from an *Alu* consensus sequence, and several contain deletions of up to half the entire sequence.

INTRODUCTION

With the exception of anucleate red blood cells, all eucaryotic cells contain microtubules which occur as long filamentous structures of about 24 nm in diameter. The major subunit proteins of microtubules are the α - and β -tubulins, each with a molecular weight of about 55,000 daltons. These proteins, in association with others (the microtubule associated proteins) form complex assemblies that are central to eucaryotic cell function: the mitotic spindle, centrioles, cilia, flagella and elements of the cytoskeleton all contain microtubular structures (1).

In mammals, the genes encoding α - and β -tubulin constitute two distinct multigene families, each containing about 15-20 members (2). In humans, where the β -tubulin multigene family has been the subject of intensive study, the majority of β -tubulin sequences have been shown to be intronless pseudogenes derived by integration of cDNA copies into host germ-line chromosomes (3-5). Expressed β -tubulin genes therefore account for only a small fraction of the complete multigene family. Because of the high nucleotide sequence homology between expressed and non-expressed genes, the precise number of expressed sequences has been difficult to ascertain.

Nonetheless, by a variety of criteria, it has been possible to identify two expressed human β -tubulin genes (5,6). Here we present the complete sequence of one of these genes. The features of the β -tubulin isotype encoded by this gene are discussed, as well as the implications of the unexpected presence of ten Alu family members within a single 4.8 kb intervening sequence.

METHODS

Subcloning, DNA Sequencing and Data Handling.

Regions of the bacteriophage λ recombinant containing the 5 β gene (7) were subcloned into the bacterial plasmid pUC9. Plasmid inserts were sequenced by the dideoxy chain terminator procedure (8), either by random cloning of sonicated fragments (9) or insertion of restriction fragments trimmed by digestion with BaI 31 into the polylinker contained in the sequencing vectors M13 mp8,9,10,11. Data handling was by the computer programs of Staden (10). With very few exceptions, where short regions yielded an unambiguous sequence, the entire sequence of 5 β was determined in both orientations.

RNA Preparation and Analysis.

PolyA⁺ mRNA was prepared from the following cell lines by methods described previously (12): HeLa cells, a human myeloma cell line CM 1500A (11), human embryonic kidney (HEK) (34), melanoma RPMI 7272 (generously provided by Dr. D. Rifkin, NYU Medical Center), retinoblastoma Y79 (36), glioma 132 INI (37) and neuroblastoma [SK-N-SH (38), CHP 126 and CHP 134 (39)]. RNA blot transfer analysis was carried out using 1.5% agarose gels containing formaldehyde as described (13). The blots were prehybridized and hybridized overnight in 6 x SSC, 5 x Denhardt's, 20 mM Na phosphate, pH 7.0 at 68° using a probe (probe B, ref. 6) derived from the 3' untranslated region of 5 β and labelled with ³²P by nick translation (35). Following hybridization, blots were washed to a final stringency of 0.2 x SSC, 68°.

RESULTS

Comparison of Two Human β -Tubulin Isoforms

Previous work in this laboratory has resulted in the isolation of the majority of human β -tubulin sequences from recombinant bacteriophage genomic libraries (7). One cloned fragment containing a β -tubulin sequence was shown by electron microscopic heteroduplex analysis to contain four coding regions and three intervening sequences. This gene, termed 5 β , is expressed because 1) A 3' untranslated region probe used in a genomic Southern blot experiment

performed at high stringency detected restriction fragments that correspond precisely in size with those predicted on the basis of restriction mapping of the cloned 5 β gene, with only one additional (and faint) band visible per restriction digest; 2) The 3' untranslated region subclone detects a 2.6 kb mRNA species that translates in a cell-free system to yield authentic β -tubulin; and 3) A short β -tubulin cDNA clone, D β -2, has been isolated that bears exact homology with a region of coding sequence within 5 β (6).

The complete sequence of the β -tubulin peptide encoded by 5 β is shown in Fig. 1, and compared with the corresponding sequence of M40, a human β -tubulin gene expressed in several different human cell lines (5). The data reveals a divergence of 16.5% at the nucleic acid level, mostly as a consequence of silent changes in the third codon position that do not affect the amino acid sequence. Nonetheless, there are 21 amino acid differences between the two isotypes scattered throughout the length of the polypeptide. Because of these differences, the polypeptide encoded by 5 β is less acidic than that encoded by M40 (5) by two charges.

Structure of the 5 β Gene: Ten Alu Family Members Within a 4.8 kb Intervening Sequence

Complete sequence analysis of 5 β (Fig. 1) confirmed the overall structure of the gene as determined by electron microscopic heteroduplex analysis using a chicken β -tubulin cDNA probe (7). No rearrangement of the genomic sequence containing the 5 β gene can have occurred during the construction and/or amplification of the λ bacteriophage library, since two overlapping recombinants containing identical restriction sites within the region of overlap were isolated. Furthermore, the subcloned fragment that includes the 5 β gene contained the expected number and location of restriction sites. Therefore, the large number of direct and inverted repeat sequences contained within the 5 β gene (see below) did not result in any detectable rearrangements during the cloning procedures prior to sequence determination. In addition, because the sequence was determined from successive overlapping clones generated by Bal 31 digestion, no region of repeated sequence could have been missed. The 5 β gene consists of four coding regions, interrupted by three intervening sequences (Fig. 2). The coding region blocks encompass amino acids 1-19, 20-55, 56-92, and 93-444, a configuration identical to that found in other expressed β -tubulin genes (5,14). The size of the coding blocks has therefore been evolutionarily conserved, both within and across species boundaries. The intervening sequences in 5 β contain 1014, 117 and 4826 nucleotides, in each case flanked

[illegible]

ATGGTGCCCATGCCATCCAGCCTGAGCGATGGTGTGAGACCCATCTCTAAAAAGATCAATAAAACAGGGATTCTAAATGCTATTGAATGTGAGGC
 TCAGGGTGGGCACGGTGGCTCACACCTGTAATCCACAGCACTTTGGAGGGCCGAGCGGGTGGATCACCTGAGGTGGGAGTTTGAGACCCAGCTGACATA
 CATGGAGAAACCCATCTCTACTAAAAATAAAAAATTAGCCAGGCATGGTGGACATGCTGTAATCCAGCTACCTGGGAGGCTGAGGCAAGAGAAATG
 CTGAACGGCGAGAGGTGTGGTGAAGCAAGATCATGACAGTGCACCCAGCGTGGGCAACAGAGTGAACCTTCGCTCCAAAAAATTTTCT
 CAAGCTTTCTGGCTCCAGCTGGCTGCTATGTCACCCAGCTGGCTCCACCGCTTCAATATCCAGCATGTAAGACGTGCTAAATGACAGGACCTGTCCT
 GCCTCTGTAATCTTTCCGCAACTCGCTCTTACTTCCATTTTACAGGTGAGAAAAAGAGGCACAGAGAGCTTTAGGTCACTTCAGAAAGTCACACT
 GCAAGCCACTGGGGCTGTCAGGAGCCAGAGGAGAGGCTGACTACTAGGTCCACGCTTTTTTTTTTTTTTTTGGAGCGGAGTCCCGTTGCCCAGGCTGGA
 GTGACGTGACGCAACCTCCACCTCCAGATTCAGCGATTCTCCTGCTCAGAGTAGCTGGGACTACAGGCATGCATACCCACACCTAGCTAATTTTTT
 D GTATTTTAAATAGTGAAGGGTTTCGCATGTTGGCCAGGCTGGTCTCTATTCTGACCTCTCAGGTGACTCTGCTGCTCTGCTTCCCAAGGTGCTG
 GGATTACAGGCATGAGCAACGCGCTGGTCTGAGGCCACTACTTTTTAATGCTCCTCCACCCCTGGCCACCCAGAGGCCATAGAAGTCTCTGGCCCA
 ACTGTATAGTCCCTCTGGTCTCCTGAAATCACTCACACCATTTGGCCAATCTCATTTTCACTTTGATTTCACCCAGGCATTTCCCAAGGACCTATGATGG
 CCACGGCCACAAAATCGCAATCAATTAAGTGTGTTGAAAGTCTTCCAACAGATCAGAAAGGGGAAGTCACTTGTCGAAGTCAACAGTTTGTCTTGTCAC
 GTCCCGAGAAATCTCGAGTCTAAACCACCAAGCTCGGGAGGAGGAGGGGGGTGACCAAGTGGCCCTTCCTGAGGATGCAGCATTTTGAAGTGAAGACC
 TGAAGGAAGTGAAGAAAAAGTCAAGTGAATACCTGGGGGTTGCACTTCAAGCAGCCAAGCCACTAGTGCAAAAGGCCCAAGGCAAGGACCACTGG
 AAAGCTGGAGTAACAGCTGCGGCTGAAGCAGAGTGAAGGAGGGGATGGGCAAGTGGCGAGGCTTGTCGCTGCAAGGAGGATTTGGGCTTGACCCCA
 GGTAGGTGGGAGCCTGAGAGGCTGAGGAGGAGGAGGAGGACCTGACTTACTTGCTCTCAGGTGTCTCTGTTGGGATGCTGAGGAGGAGACAGT
 GGGTGACTAAGGGTAGGAGCCAGGGACAGGACAGAGGAATTGAATTTATCCAGGAGGGTATGCTGGAGCTGAACCTAAGGTGGAGACAGATGAGGG
 GTGAAGAAAAACGTTAGCTTGTTTACAGTAATACCGGCCAATATCTTTTTTTTATTCATTTTTTTTTTTTTTTTGGAGCAAGGCTCACTCTGTGCGCC
 E AGGCTGGAGTGCAGTGGCGCATCTCGGCTCACTGAAGCTCCACCTCCTGGGTTACGCGCATTCCTGCTCAGCCTCCCGAGTAGCTGAGACTACAG
 GCACCCACCCACATCCAGTAATTTTTTCTATTTTTAGTAGAGCGGGGCTCACCCTGTTAGCGGGATAGCTTGATCTCCTGACCTCGCTATCCG
 CCGGCTCGGCTCCCAAGTGTGGGTTTACAGCGTAGGCAACACGCTGGCCAAATATCTCTTTTTTTTTTTTTTTTTTTTTTGGAGCGAGGCTT
 GCTGTGCCCCAGGCTGGATGCAGTGGCAGCATCTCGGCTCACTGCAGGCTCAGCTCCCGAGTAGCTGGGACTACAGGCGCCCGACCTCGCCCGGCTAAT
 F TTTTGTATTTTTCAGTAGAGACAGGGTTTCACTGTGTAGCCAGGATGGTCTCGATCTCTAACCTCGCAATCGCCCGCTCGGCTCCCTAAGTGGCTG
 GATTACAGGCTAGGCTCAGGCTCGGCTGGCCAAATATCTTAATCTGATTTCTTTTCCGGGAGCTTGCAATCCAGAGTGAGCTTGCAGACATAAATCATGA
 TTAACAAGAAAAACAGGCCAAGCATGGCTGCGAGCCTGTAATCCAGCACTTTAGGAGCATGGCAGGTGTATGCTCCAGCTCAGGAGTTCAAGAACAG
 CTTGGGCAACATGTGCAAAACCCACCTCTACAAAAAAGATACAAAAATTAGCCAGATCTGGTAGCATAAGGCTGTAGTCCAGCTACTCAAAAGGCTG
 G GGGCAGGAGGATCACTTGACCTGGGAAGTCAAGGCTCAGTGAGGCTCATGTTGCACTCTGATCCAGCCTGGGCAACATACCAAGACCTTGCTGAAAA
 TGATATTATTATAATAATGATGTAAATATTAATCTTTTATTGTATTATTATTGTTTTTAGACAGAGGCTTGCTGTCTCCAGGCTGCAATG
 GTGTGATCATAGCTCAGCTGAGGCTCCAACTCCTGGGCTCAAGTAAGTCTCCACCTGAGGCCCTTAGTAATCTGGTACTACAGGTGTGCCACCACTG
 H ACTGGCTAATTTATATATATATATATATATATTTTTTTTTTTTTTTTTTTTTTTTTTTTGTAGAGACAGGGCTCAGTAGGGCGCCAGGCTA
 AGTCTCCAGCTGCTGGGCTCAAGAACCTCCTCCACCTCTGCTCCCAAGTGTAGGATTACAGCATGAGCAGTGTGCCAGCTTGGTTTTTTTTTT
 AGGCAAGGGGGGGGGAGTGTGTTGTGTATGTGTGTTTTGAGATGGAGTCTCACTCCGTCACCCAGGCTGGATGCAAGTGGCTCAATCTGGCTCCC
 I TGCAGCTCTCTCCAGGTTCAAGCAATCTCTGCTCAGCTCCCGAGTAGCAGGAGCTACGATGCTCACCAGCACTGGCTATAATTTGTTAT
 TATTTATTATTATTATTATTATTATTATTATTGAGACGGAGCTCGCTGTTGGCCAGGCTGAGTGCAGTGGGCAATGTGCTGCTCAGTCCCAA
 J GCTTCGCTCTCTGGGTTACGCGATTCTCTGCTAGGCTCCCGAGTAGCTAGGACTACAGGCGCCGCAACCTCCCGCTAATTTTTTTTGTAAATTT
 TAGTAGAGCGGGTTTACCCTGTAGGCAAGTGGTCTCGATCTCTGACCTCGTATCGCCCACTCGGCTCCCAAGTGTGGGATTACAGGTG
 TGAGCCACCGCAACCGGCTATTTGTTATTATTTTGTAGTACTAGTATTCACTACTAGAGGCCCTCCAGGTGGTGGAGAGCAGAGTCAAGGGTTGG

100
 Gly Gln Ser Gly Ala Gly Asn Asn Trp Ala Lys Gly His
 GGCTCCCCATAACTGCATGTGTGTCCTCCCTCTCCCTCTAG GC CAA TCC GGA GCC GGC AAC AAC TGG GCA AAG GGG CAC
 G T G A T
 110
 Tyr Thr Glu Gly Ala Glu Leu Val Asp Ala Val Leu Asp Val Val Arg Lys Glu Ala Glu Ser Cys Asp Cys Leu
 TAC ACG GAG GGC GCA GAG CTG GTG GAC GCT GTC CTG GAC GTA GTC CGG AAG GAG GCC GAG AGC TGC GAC TGC CTT
 A C T T T G A
 Ser
 120
 140
 Gln Gly Phe Gln Leu Thr His Ser Leu Gly Gly Thr Gly Ser Gly Met Gly Thr Leu Leu Ile Ser Lys Ile
 CAG GGC TTC CAG CTG ACC CAC TCG CTG GGG GGT GGG ACG GGG TCC GGA ATG GGC ACG CTG CTC ATC AGT AAG ATG
 A C G C A C T
 150
 160
 Arg Glu Glu Phe Pro Asp Arg Ile Met Asn Thr Phe Ser Val Val Pro Ser Pro Lys Val Ser Asp Thr Val Val
 CGC GAG GAG TTC CCA GAC CGC ATC ATG AAC ACC TTC AGC GTG GTG CCC TCG CCC AAA GTG TCA GAC ACG GTG GTG
 A A A T T T T
 Tyr

```

                                190                                200
Glu Pro Tyr Asn Ala Thr Leu Ser Val His Gln Leu Val Glu Asn Thr Asp Glu Thr Cys Ile Asp Asn Glu
GAG CCC TAC AAT GCC ACG CTG TCT GTG CAC CAG CTG GTG GAG AAT ACG GAT GAG ACC TAC TGC ATC GAC AAC GAG
                                C C C C T T A T
                                210                                220                                230
Ala Leu Tyr Asp Ile Cys Phe Arg Thr Leu Lys Leu Thr Thr Pro Thr Tyr Gly Asp Leu Asn His Leu Val Ser
GCA CTC TAC GAC ATC TGT TTC CGC ACC CTC AAG CTG ACC ACC CCC ACC TAC GGG GAC CTC AAC CAC CTG GTG TCG
                                T T C T G Arg A A
                                240                                250
Ala Thr Met Ser Gly Val Thr Thr Cys Leu Arg Phe Pro Gly Gln Leu Asn Ala Asp Leu Arg Lys Leu Ala Val
GCC ACC ATG AGC GGC GTC ACC ACC TGC CAC CTG CGC TTC CCG GGC CAG CTG AAC GCC GAC CTC CGC AAG CTG GCC GTC
G GAG T T C T T C T T C T T C T T C T T C
Gly Glu Cys
                                260                                270                                280
Asn Met Val Pro Phe Pro Arg Leu His Phe Phe Met Pro Ala Phe Ala Pro Leu Thr Ser Arg Gly Ser Gln Gln
AAC ATG GTC CCC TTT CCT CGC CTG CAC TTC TTG ATG CCC GCG TTC GCA CCC CTG ACC AGC CGG GGC AGC CAG CAG
                                C A T C T T T GC Gly
                                290                                300
Tyr Arg Gly Leu Thr Val Pro Glu Leu Thr Gln Gln Met Phe Asp Ala Lys Asn Met Met Ala Ala Cys Asp Pro
TAC CGG GGC CTG ACG GTG CCC GAG CTC ACC CAG CAG ATG TTC GAT GCC AAG AAC ATG ATG GGC GGC TGC GAC CCG
T A CT C A G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C G C
Ala Asp Val
                                310                                320                                330
Arg His Gly Arg Tyr Leu Thr Val Ala Ala Val Phe Arg Gly Arg Met Ser Met Lys Glu Val Asp Glu Gln Met
CGC CAC GGC CGC A TAC CTG ACC GTG GCC GCC GTG TTC CCG GGC CGC ATG TCC ATG AAG GAG GTG C GAG CAG ATG
A C
                                340                                350
Leu Ser Val Gln Ser Lys Asn Ser Ser Tyr Phe Val Glu Trp Ile Pro Asn Asn Val Thr Ala Val Cys Asp
CTG AGC GTG CAG AGC AAG AAC AGC AGC TAC TTC GTG GAG TGG ATC CCC AAC AAC GTG AAG ACG GCC GTG TGC GAC
T A Asn Asn
                                360                                370                                380
Ile Pro Pro Arg Gly Leu Lys Met Ala Val Thr Phe Ile Gly Asn Ser Thr Ala Ile Gln Glu Leu Phe Lys Arg
ATC CGC CCC CGC GGC CTG AAG ATG GCC GTG ACC TTC ATC GGC AAC AGC ACG GCC ATC CAG GAG CTG TTC AAG CGC
A T T C A C T T A
                                390                                400
Ile Ser Glu Gln Phe Thr Ala Met Phe Arg Arg Lys Ala Phe Leu His Trp Tyr Thr Gly Glu Gly Met Asp Glu
ATC TCC GAG CAG TTC ACG GCC ATG TTC CGG CGC AAG GCC TTC CAC TGG TAC ACG GGC GAG GGC ATG GAC GAG
G T
                                410                                420                                430
Met Glu Phe Thr Glu Ala Glu Ser Asn Met Asn Asp Leu Val Ser Glu Tyr Gln Gln Tyr Gln Asp Ala Thr Ala
ATG GAG TTC ACC GAG GCC GAG AGC AAC ATG AAT GAC CTG GTA TCT GAG T CAG CAG TAC CAG GAC GCC ACC GAC CCG
T C C
                                440
*** Glu Gln Gly Glu Phe Glu Glu Glu Ala Glu Glu Glu Val Ala TAGGCTGCTCCCATCGCTTCCACCTGCTCCCTCGAGGC
GAG CAG GGC GAG TTC GAG GAG GCG GAG GAG GAG GTG GCC CC ***
GAA G AG T GT C A Ala
Glu Glu Glu Asp Gly

```

TTCTGACCTTTGATCGCTAGGCCCGCCATCTCTGAACCTAGAGCCCGCTTCCCTCCAAGGCTGACTCCCGCTGACCTAACAAATTACCTTTGGA
GCTCGCTTACCTCTGGCTACTTCATCTCCGACCTGGCTCCCGCTTGGAGCCCTAATTTATCTTTAAACCCCTTGAGCTCTTCCAACCTTGACATTCCTCA
GGAGGAGCCCGCTTACCCCTTCTCACTCTGGAACCGCACCTTTAACTTTGAGACCTTCCCTTACCCCTGACTTCTGCTTCACTTTGACCTCTGCTG
CCCCATGAATCCCATTTTACCTCTAGACCTATAAGTTCTGGTTTATGTTTGGACCCCTCCCTCTGAGCTGCACCTTCCCGCTGACCTTGCTCACCCTTTAA
CCCCCACCCTGAGCCCCAGCTCTACCTCTGACCCCAACTTCTCTTTGAATCTCTGAATCCCTCTGACTCCAATCTCTCTTCAACCTCTATGAGTCC
ATTTTACTTCTACACCTGCAAGTCTGGTTTATATTGAGCCCTCCCTCCGAGCTGCAGTTACCTTTGACCTTGCTCTACCTTTACCCCCACCCCTT
CACAGCTCAGCTCTACCTCTGACCCCACTTCTCTGTTCCACAGGCCCATGCTCCCTGCTCAGCTCCCTCAGCCCTGCGGACCTTAG
CTTATCTGGGAGAGAAACAGGCTGCTGAGGAGAGAGGTCACCCCTACCTCCCTCCCGCTTGCTGCTCTACCCCTCAATAAATAAATTA
ATGTGTCTGATGTTCTGCGGAATCCCTTTTCTCTCTACAGCAA

by the splice signal dinucleotides AG and GT. Unexpectedly, the largest (third) intervening sequence contains ten sequences belonging to the middle repetitive Alu family (31) (Figs. 1 and 2). Since each Alu member encompasses approximately 300 nucleotides, these Alu elements account for about 60% of the total length of this intervening sequence. Three of the Alu family members (B, C and G, Fig. 2) are in the same transcriptional orientation as the 5 β gene; the remainder (A, D, E, F, H, I, and J, Fig. 2) are in the opposite transcriptional sense. With three exceptions (G, H and I), all the Alu sequences are flanked by direct repeats, varying in length from four (C) to thirteen (B) nucleotides (Table 1). In one case, where two Alu family members (E and F) are directly juxtaposed, the direct repeats flanking each member are shared. Otherwise, the direct repeats flanking the various Alu family members bear no sequence relationship to one another.

The extent of homology of the ten Alu family members with respect to a consensus human Alu sequence taken either from Alu elements selected at random (15, 31) or derived from the sequence data presented here (Fig. 1) is shown in Fig. 4. The two consensus sequences are in excellent agreement. However, the range of divergence of any given Alu family member from either consensus sequence varies considerably, from 10 to 28 percent (Table 1). In the least homologous case (G), the divergence is evident throughout the length of the Alu sequence, and there is no evidence of flanking direct repeats. With one exception (G), all the Alu family members include the characteristic A-rich sequence at their 3' ends; in one case (J), this sequence consists of the repeated trinucleotide (AAT)₁₀.

A curious feature of Alu family member H is that it differs from the consensus sequence in the central A-rich region. In the consensus, this region is A₅TACA₅, whereas in Alu family member H, the corresponding region contains the sequence A₃₄(TA)₁₀.

Restricted Expression of the 5 β Gene

In a previous study, a 3' untranslated region probe derived from 5 β was

Figure 1. Sequence of an expressed human β -tubulin gene, 5 β . The nucleotide and corresponding amino acid sequences of the 5 β β -tubulin gene are shown, and compared (within the coding regions) with the sequence of a second human β -tubulin isotype, M40 (5). Where no nucleotide base or amino acid is shown, homology with 5 β is preserved. Asterisks have been introduced in either sequence to denote the absence of a corresponding amino acid in one or the other isotype. Ten Alu family members (A-J) that occur within the third intervening sequence are underlined; their flanking direct repeats are doubly underlined. Two overlapping polyA addition signals (AATAAA) are also underlined.

Figure 2. Diagram outlining the structure of the 5 β human B-tubulin gene. Coding blocks are drawn as solid bars; the 3' untranslated region is hatched. Arrows show the location and transcriptional orientation of ten Alu family members (A-J) within the third intervening sequence. Note that one family member (I) lacks sequences corresponding to the 5' half of the consensus Alu family sequence (see Fig. 4).

used to detect a 2.6 kb mRNA present in HeLa cells, albeit at low levels (5). Because of the difficulty in obtaining human tissues suitable for the preparation of undegraded mRNA, we assayed for the presence of 5 β gene transcripts in RNA blot transfer experiments using polyA⁺ mRNA prepared from a variety of transformed human cell lines maintained in culture, including those derived from neuroblastoma (3 independent cell lines),

Table 1
Features of Ten Alu Family Members Located
in the Third Intervening Sequence of 5 β

Designation (see Fig. 2)	Direct repeat	% divergence from consensus <u>Alu</u> sequence*	Remarks
A	TCCTC	20	-
B	TAAACAGGGATT	21	-
C	CTCA	19	-
D	AGG(T)CCAC	13	Contains deletions of 6 and 15 bps
E	AATATTCT	10	Shares direct repeat with F
F	AATATTCT	11	Shares direct repeat with E
G	-	28	-
H	-	23	Midsection sequence consists of A ₃₄ (TA) ₁₀
I	-	14	Truncated - 5' half missing
J	TTGTTATTATT	10	3' end consists of (AAT) ₁₀

*Each substitution, insertion or deletion has been counted as a single event, even if an insertion/deletion involves several bp's.

Figure 3. Restricted expression of 5 β . PolyA⁺ RNA prepared from HeLa cells (A) or human fetal brain (B) was fractionated on a 1.5% agarose gel containing formaldehyde (13), transferred to nitrocellulose and the blot hybridized with a subcloned probe containing 5 β 3' untranslated region sequences (probe B, ref. 6) labelled with ³²P by nick-translation (35).

glioma, kidney, melanoma, fibroblast, hepatoma and myeloma. No signal was detected in any case. However, a sample of human fetal brain mRNA (generously provided by Dr. P. Dobner, University of Massachusetts Medical School) gave a conspicuous 2.6 kb RNA signal (Fig. 3). We conclude, therefore, that the 5 β gene is expressed in developing human brain tissue.

DISCUSSION

The human β -tubulin multigene family contains about 15-20 extensively homologous members that cannot be distinguished by hybridization using β -tubulin coding region probes (5). The presence of multiple copies of processed pseudogenes within the multigene family (3-5) has made the number of functionally expressed human β -tubulin genes hard to define, largely because of the difficulty of undertaking an exhaustive search for sequences that satisfy the required criteria for expressed genes. Nonetheless, the task is important both to establish the number of mammalian β -tubulin isotypes, and to determine whether co-expression of multiple tubulin isotypes within a single cell type results in the assembly of microtubules consisting of two distinct β -tubulin gene products.

The data presented here define two distinct human β -tubulin isotypes that differ in amino acid sequence at 21 positions, an overall divergence of 5 percent. Whether these substitutions are of significance in terms of their possible functional influence on the assembled microtubule is a matter of

speculation. Though we have shown that both isotypes are indeed expressed within a single cell type (6), it is not clear whether both can coexist within the same kind of microtubule. However, our sequence data suggests that this question might be addressed by generating isotype-specific antisera using peptide haptens chosen from those regions of the β -tubulin sequences that show maximum divergence.

The size of the mRNA transcribed from 5 β is about 2.6 kb, and results from polyadenylation at a site approximately 830 bp's downstream from the termination codon. Although this transcript is clearly identifiable in human fetal brain mRNA and in some sublines of HeLa cells (Fig. 3), no RNA sequences homologous to the 3' untranslated region of 5 β could be detected in polyA⁺ mRNA prepared from eight different human cell lines maintained in culture. One possible explanation is that 5 β is restricted in its expression to an early developmental stage. Alternatively, it is possible that expression of 5 β could have been lost, either as a consequence of transformation or during serial passage of the cell lines.

The size and number of exons in 5 β is identical to that found in the human β -tubulin gene M40 (5) and, indeed, in chicken β -tubulin genes (14). Thus, the four-exon structure in the β -tubulin genes of vertebrate organisms seems to have been conserved both within and across species boundaries. In contrast, the intervening sequences are highly variable both in size and in sequence content. A curious feature of the largest intervening sequence of 5 β is that it contains ten members of the middle repetitive Alu family. These elements, about 300 nucleotides in length, are present in human DNA at approximately 3×10^5 copies per haploid genome (15), and are capable of transposition (16-18) probably via an RNA intermediate (20,21) transcribed from a Pol III promotor contained within the Alu sequence itself (21). The involvement of an RNA intermediate and the insertion of cDNA copies into a staggered host-chromosomal break is therefore similar, if not identical, to the mechanism involved in the generation of β -tubulin (and other) processed pseudogenes (5, 19, 40).

There is no evidence that processed pseudogenes are transcribed, and they are probably therefore non-functional. On the other hand, the literature is

Figure 4. Sequence comparison of human Alu family members within 5 β . The sequence of ten Alu family members, A-J (Fig. 2) is compared with a consensus sequence taken from the literature (31) (consensus 1) and with a consensus sequence derived from the data shown in Fig. 1 (consensus 2). Open spaces denote homology; deletions (*) and insertions (Δ) have been introduced so as to maintain maximum homology in the figure.

replete with speculation as to the possible function of Alu family members and their transcription products. Suggestions include their involvement as origins of DNA replication (22), mediators of recombination (23) or chromosomal rearrangement (24), modulators of chromatin structure (25) and stabilizers of gene structure that prevent or inhibit gene conversion (26). Convincing experimental evidence for any of these conjectures is lacking. It seems certain, however, that at least some Alu members are transcribed *in vivo* (27), sometimes colinearly with sequences that are transcribed by Pol II (28). It is not clear whether some or all of the ten Alu family members present within the 5 β Pol II transcription unit are themselves transcribed. Though Alu sequences have been found within other Pol II transcription units (29,30) including a human β -tubulin gene (5), the very high density of Alu sequences within 5 β is unprecedented. The random distribution of Alu family members in the human genome would entail an average spacing of about one member per 8 kb (calculated on the basis of 2.5×10^9 bp's per haploid genome/ 3×10^5 Alu copies). Even if Alu family members are restricted in their distribution to 50% of the genome--a situation that applies to the distribution of most short-period interspersed repeats in eucaryotic DNAs--then the average spacing between individual Alu elements would be 4 kb. The clustering of ten Alu family members within a 4.8 kb intervening sequence therefore differs dramatically from any theoretical average. The 56 kb human globin gene locus, for example, contains seven Alu family members with an average spacing of 8 kb (32,33). The factors that govern the integration of RNA-mediated transposable elements are unknown, but it is possible that the third intervening sequence of 5 β contains sequences that are particularly favorable for the entry of exogenous DNA.

Two of the Alu family members (E and F, Fig. 3) are juxtaposed in a head to tail fashion, sharing the same flanking direct repeat. Presumably these sequences either entered the genome at the identical target site, or arose by a duplication event precisely spanning the Alu element itself. Whatever their origin, the two abutting Alu members are significantly divergent from one another. Indeed, inspection of the sequence of all ten Alu elements and comparison of this data with a consensus sequence taken either from the literature (15) or compiled from the data presented here (Fig. 4) shows significant divergence among the various members, ranging from 10 to 28 percent. The differences with respect to the consensus sequence are not restricted to base substitutions: several changes apparently involved deletions of up to half of the entire Alu sequence (Fig. 4). This

observation suggests to us that, in common with pseudogenes generated via RNA intermediates (3-5, 19,40), a significant portion of Alu family members may be functionless, subject to evolutionary drift, and may, with time, become unrecognizable as they diverge and disappear into the genomic background.

ACKNOWLEDGEMENTS

This work was supported by grants from the NIH (GM 31740) and the Muscular Dystrophy Association.

REFERENCES

1. Kirschner, M.W. (1978) *Int. Rev. Cytol.* **54**, 1-71.
2. Cleveland, D.W., Lopata, M.A., McDonald, R.J., Cowan, N.J., Rutter, W.J. and Kirschner, M.W. (1980) *Cell* **20**, 95-105.
3. Wilde, C.D., Crowther, C.E., Cripe, T.P., Gwo-Shu Lee, M. and Cowan, N.J. (1982) *Nature* **292**, 83-84.
4. Wilde, C.D., Crowther, C.E. and Cowan, N.J. (1982) *Science* **217**, 549-552.
5. Gwo-Shu Lee, M., Lewis, S.A., Wilde, C.D. and Cowan, N.J. (1983) *Cell* **33**, 477-487.
6. Hall, J.L., Dudley, L., Dobner, P.R., Lewis, S.A. and Cowan, N.J. (1983) *Mol. Cell. Biol.* **3**, 854-862.
7. Cowan, N.J., Wilde, C.D., Chow, L.T. and Wefald, F.C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4877-4881.
8. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B. (1980) *J. Mol. Biol.* **143**, 161-178.
9. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963-3965.
10. Staden, R. (1980) *Nucl. Acids Res.* **8**, 3673-3694.
11. Croce, C.M., Shander, M., Martins, J., Ciccurel, L., D'Acona, G.G., Dolby, T.W. and Koprowski, H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3216-3419.
12. Cleveland, D.W., Lopata, M.A., Sherline, P. and Kirschner, M.W. (1981) *Cell* **25**, 537-546.
13. Boedtker, H. (1971) *Biochim. Biophys. Acta* **240**, 448-453.
14. Sullivan, K. and Cleveland, D.W. (1984) Submitted to *J. Cell Biol.*
15. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) *J. Mol. Biol.* **132**, 289-306.
16. Bell, G.I., Pictet, R. and Rutter, W.J. (1980) *Nucleic Acids Res.* **8**, 4091-4109.
17. Grimaldi, G. and Singer, M.F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1497-1500.
18. Kominami, R., Muramatsu, M. and Moriwaki, K. (1983) *Nature* **301**, 87-91.
19. Van Arsdel, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, M. and Gesteland, R.F. (1981) *Cell* **26**, 11-17.
20. Jagadasweeran, P., Forget, B.G. and Weissman, S.M. (1981) *Cell* **26**, 141-149.
21. Elder, J.J., Pan, J., Duncan, C.H. and Weissman, S.M. (1981) *Nucleic Acids Res.* **9**, 1171-1189.
22. Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Choudary, P.V., Biro, P.A., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. and Schmid, C.W. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1398-1402.
23. Jagadasweeran, P., Tuan, D., Forget, B.G. and Weissman, S.M. (1982) *Nature* **296**, 469-470.

24. Calabretta, B., Robberson, D.K., Barrera-Saldana, H.A., Lambrou, T.P. and Saunders, G.F. (1982) *Nature* **298**, 219-225.
25. Duncan, C.H., Jagadasweeran, P., Wang, R.C.C. and Weissman, S.M. (1981) *Gene* **13**, 185-196.
26. Hess, J.F., Fox, M., Schmid, C.W. and Shen, C-K.J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5970-5974.
27. Allen, M. and Paul, J. (1984) *Nucleic Acids Res.* **12**, 1193-1200.
28. Allen, M., Lanyar, W.G. and Paul, J. (1983) *Cell* **35**, 187-197.
29. Della Favera, R., Gelman, E.P., Gallo, R.C. and Wong-Stall, F. (1981) *Nature* **292**, 31-35.
30. Page, G.S., Smith, S. and Goodman, H.M. (1981) *Nucleic Acids Res.* **9**, 2087-2104.
31. Jelinek, W.R. and Schmid, C.W. (1982) *Ann. Rev. Biochem* **51**, 813-844.
32. Fritsch, E.F., Lawn, R.M. and Maniatis, T. (1980) *Cell* **19**, 959-972.
33. Coggins, L.W., Grindley, G.J., Vass, J.K., Slater, A.A., Montague, P., Stinson, M.A. and Paul, J. (1980) *Nucleic Acids Res.* **8**, 3319-3333.
34. Harrison, T., Graham, F. and Williams, J. (1977) *Virology* **77**, 319-327.
35. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* **113**, 237-251.
36. Reid, T.W., Albert, D.M., Robson, A.S., Russell, P., Craft, J., Chu, E.W., Tralka, T.S. and Wilcox, J.L. (1979) *J. Natl. Cancer. Inst.* **53**, 347-360.
37. Leichtling, B.H., Drofar, A.H., Ortmann, R. and Perkins, J.P. (1976) *J. Cyclic Nucleotide Res.* **2**, 89-98.
38. Beidler, J.L., Helson, L. and Spengler, B.A. (1973) *Cancer Res.* **33**, 2643-2652.
39. Schlesinger, H.R., Gerson, J.M., Moorhead, P.S., Maguire, H. and Hammeter, K. (1976) *Cancer Res.* **36**, 3094-3100.
40. Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) *Cell* **32**, 461-472.