Brain-specific genes have identifier sequences in their introns

(cDNA/small cytoplasmic RNA/consensus sequence/gene control/repetitive sequence)

ROBERT J. MILNER, FLOYD E. BLOOM, CARY LAI, RICHARD A. LERNER, AND J. GREGOR SUTCLIFFE

Research Institute of Scripps Clinic, La Jolla, CA 92037

ABSTRACT   The 82-nucleotide identifier (ID) sequence is present in the rat genome in $1-1.5 \times 10^5$ copies and in cDNA clones of precursors of brain-specific mRNAs. One brain-specific gene contains more than one ID sequence in its introns. There is an excess of ID sequences to brain genes, and some ID sequences appear to have been inserted as mobile elements into other genetic locations. Therefore, brain genes contain ID sequences in their introns, but not all ID sequences are located in brain gene introns. A brain ID consensus sequence has been obtained by comparing 8 ID nucleotide sequences.

We recently identified several cDNA clones derived from rat brain poly(A)$^+$ RNA that hybridized to an inappropriately small 160-nucleotide cytoplasmic RNA (BC1) that was present in rat brain but not in liver or kidney (1). Subsequently, BC1 was found in pituitary and peripheral nervous tissue but not in several non-neural tissues (2). The sequences of two of the cDNA clones were analyzed and found to be homologous for 82 nucleotides and otherwise generally nonhomologous. The homologous sequence was responsible for the hybridization of the cDNA clones to the small brain-specific RNA species and was also located in an intron of a rat growth-hormone gene (1, 3). We suggested that the 82-nucleotide sequences, which we called identifier (ID) sequences, might be representative of elements that control tissue-specific gene expression and may occupy noncoding regions of primary eukaryotic transcripts. Indeed, subsequent studies indicate that 62% of brain transcripts synthesized *in vitro* contain ID-homologous sequences, whereas few, if any, transcripts from other tissues contain brain ID sequences (2).

For such an element to be involved in a substantial portion of neural cell-specific gene expression, several requirements should be fulfilled. (*i*) Brain-specific genes should contain ID sequences; (*ii*) the rat genome should contain enough copies of the ID sequence to account for a substantial portion of brain genes; (*iii*) many of these ID sequences should be linked to brain genes; (*iv*) ID sequences should be largely absent from genes not exclusively expressed in brain. Criteria *i* and *iii* will always both be true if and only if brain genes and ID sequences are isomorphic. In this paper, we provide the nucleotide sequences of two more ID sequence-containing rat brain cDNA clones, demonstrate that regions of some of these clones hybridize to mature mRNA species present in brain but not in liver or kidney, and show that at least one such clone is likely to have been derived from a precursor of a brain-specific mRNA. Furthermore, by hybridization of rat genomic clones to ID sequences we have estimated that there are $1-1.5 \times 10^5$ copies of ID sequences in the rat genome. Genomic clones corresponding to a cDNA clone of a brain-specific mRNA have been isolated. Although the cDNA clone does not itself contain any ID sequences, the corresponding genomic clone hybridizes with ID probes. We
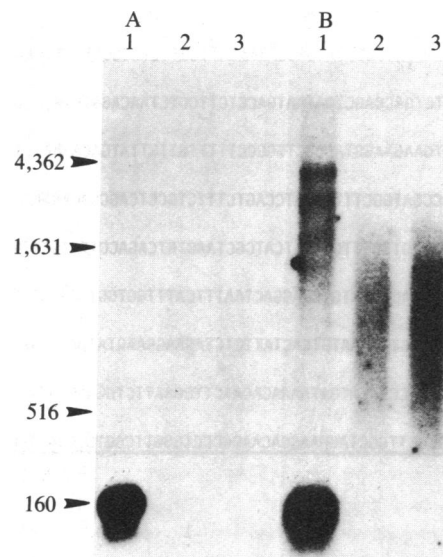
FIG. 1.   RNA blot analysis of ID cDNA clones. Poly(A)$^+$ mRNA (2 μg) from rat brain (lanes 1), liver (lanes 2), or kidney (lanes 3) were subjected to electrophoresis on 1.5% agarose gels in 1 M formaldehyde, blotted onto nitrocellulose, and hybridized with ^{32}P-labeled p1B308 (*A*) or ^{32}P-labeled p1B337 (*B*) (1, 4). The positions of DNA size standards (linear pBR322 and *Hin*fI restriction fragments of pBR322) and the position of BC1 RNA (160 nucleotides) are indicated at the left.

have mapped the ID sequences to each of two introns. Our experiments show that ID sequences are abundant, are dispersed through the genome, are correlated with brain genes, are found in introns, and therefore could be involved in brain-specific gene control.

EXPERIMENTAL

We originally defined the ID sequence as a region within rat brain cDNA clones that hybridized to the 160-nucleotide brain-specific (BC1) RNA molecule and had a common 82-nucleotide sequence. We have determined the sequence of two more brain cDNA clones that contain ID sequences, clones p1B308 [639 base pairs (bp)] and p1B337 (1,016 bp). Each of these was picked up on our RNA blot screen of brain cDNA clones (4) because it hybridized to a small target in brain but not in liver or kidney poly(A)$^+$ RNA (Fig. 1). Regions of the nucleotide sequences of the cDNA inserts were found to be similar to the previously determined ID sequences (Fig. 2 *A* and *B*). The ID sequences contained in these clones and flanking sequences are compared to the corresponding regions of clones p1B224 and p2A120, defined previously (1), and the rat growth-hormone gene (3) in Fig. 2C. Also included in the comparison are homologous sequences that have appeared in three recently published rat

A

CTGCAG$_{13}$TGTTTTGTTTGTTTTTTTGTTTTTTCATCTTTTTTAAACAATGGATTCCTTTCTTCCTGGAATATCTGGCATTTAATAGTCTGAGAAAAAATTCTTCAGGAGGTTGTTATCT

TTAAGACTGGGGATTGGCATAGCCATGAAAGTCACGTAACCTCCCTAGACACTGCTTTCCCACCTATAAAATAAGGATAATGATCGCTCCGATTTGGATCTACGGTGTCTCCAAAATGCC

CATAAGTCCCAGGCTGTAGTGCTATTGGGCAGGTAGGGCCTAAAGGAAGCCTAAAGGAAAAAGGGATGTTAGGACCGGCCCCTCCCCGCCCCTTTCTTTGCTTCCGATCCACACCAATGT

TAGAACAACCTTCCTCTAGCTCTCACCACGGTGTGAGGCCCAAAGCAACAGAGCTGAAACCATGAGCGATGATTTCCTTCTTCCTCAGGAGTGGATTGTGTCAGGCATTTGCTCTAGCAG

CGGAAAGTTGGGTTAACACTATAAGACCTAACAGAGTCAAAACAGCCACATGC<u>GGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGGAAGCGCAAGGCCCTGGGTTCGGTCCC</u>

<u>CAGCTCCCCCCAAAAAAAAAAAC</u>$_7$TGCAG

B

CTGCAG$_{12}$TTCCTCACCACCAGAAGATTCCTCCTCCGCTTCCTTCAACCATTTAATAAATGGCTCTGCTTTGACACGAATCTCTTTGGCAAGTTCTTTTGAGACATATTTCTTAGAGGCC

TTTTCTGACCAGCTGATAATGACCTCTTCCTCTAACAGGTCTGCATCATACATCTCCTTCAAGATATGTGGAATCTTGGAGATCAACTGAGCTTGATGCATTGCTACCACACATTCCAAA

CCATGAAGAAGGTACCGCTGGGCCTTTTTGTTGTTATGACAAAATCTTAAAAAATGGCGCCTGTATTTCTTGATTTGCTCTCTTATCTTCTCATCAAAGAGAACTTCTGTCAAAACGAGA

GGGCCCATGGCTTTTACATCCAGTCTTTCTGCCTCAGCCACAATGTCTTTATCAGATGAATCAATAATGCCCTCTTCTTTCTTTTTCTTAACAAAATCAAACAGGATGTTAACACGCTCT

TCTACAGTTCTTTCCAAATCATCGCTAAGTGTCAGACCTTTTGCATGGTCACTGATTTCATCCATTCTGCGCCTTTGAGCTTCCTCAGTTGTATCCTCCCCCCAATCATCATCTTCCTCT

TCTTCCACAGCATGTGGAGGACTAATTTCATTTGGTGGTGGAGGTGGTGGTGTCTCACTGGTGGATACAGAGCCATTTTCCTTGTCCTTGCCCTTTCTATTTTTCTTTTCTTTCTCTTTC

TTTCCTGTACCAATGTCACTATTCTCTACAAGGAAGTAAGAGTTAGCTTACTTAGTCTCATGGTTTATGTATTACTATCAGACTTAAATTCCACCCTCTCACCATCGTAGTTCTCTGCAG

AAAAAGCCAACATGATAAAACACAACTTCGAATTCTGGAGGGAGTGGCTGGTAAACAATTAAGAGGACTTGTTCCTGCATAAGAAGGAAGTT<u>CAGGGGTTGGGGATTTCGCTCAGTGGTA</u>

<u>GAGCGCTTGCCTAGGAAGCACAAGGCCCTGGGTTCGGTCCCCAGTTCCAAAAAAAAAAAAAAAAAC</u>$_{12}$TGCAG

C

```
                    1        10        20        30        40        50        60        70        80        90
                    .        .         .         .         .         .         .         .         .         .
        p1B224  TTAAAGATCTGGTACAGGGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTACCTAGGAAGCGCAAGGCCCTGGGTTCGGTCCCCAGCTCCGAAAAAAAGAACCAAA
        p2A120  TTAAAAAGGTGAGCAGTGGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTACCTAGGAAGCACAAGGCCCTGGGTTCGGTCCCCAGCTCCGAAAAAAAAAAAAAA
        p1B308  TCAAAACAGCCACATGCGGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGGAAGCACAAGGCCCTGGGTTCGGTCCCCAGCTCCCCCCAAAAAAAAAA
        p1B337  CATAAGAAGGAAGTTCAGGGGTTGGGGATTTCGCTCAGTGGTAGAGCGCTTGCCTAGGAAGCACAAGGCCCTGGGTTCGGTCCCCAGTTCCAAAAAAAAAAAAAA
        GH      TAAAAAAAAACAAAACAGGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTACCTAGGAAGCGCAAGGCCCTGGGTTCGGTCCCCAGCTCCGAAAAAAAAACCAAA
        R drel  TCAAGACCAAGCGTACCGGGGCTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGGAAGCGAGAGGCCCTGGGTTTGGTCCCCAGCTCCGAAAAAAAGAACCAAA
        U16-6   GTTTAGAATTGGCCTAGGGGGTTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGGAAGCGCAAGCCCCTCGGTTCGGTCCCCAGCTCCGAAAAAAAGAACCAAA
        HSV     TAAAACATAGCGTTTTGGGGGTTGGGGATT TAGCTCAGTGA TAGAGCTCTTGCCTAGCACGC--AA-GCCCTGGGTTCGGTCCCCAGCTCTGAAAAAAAGGAAAGAA
                                                                                             Ʌ
                                                                                             CC
```
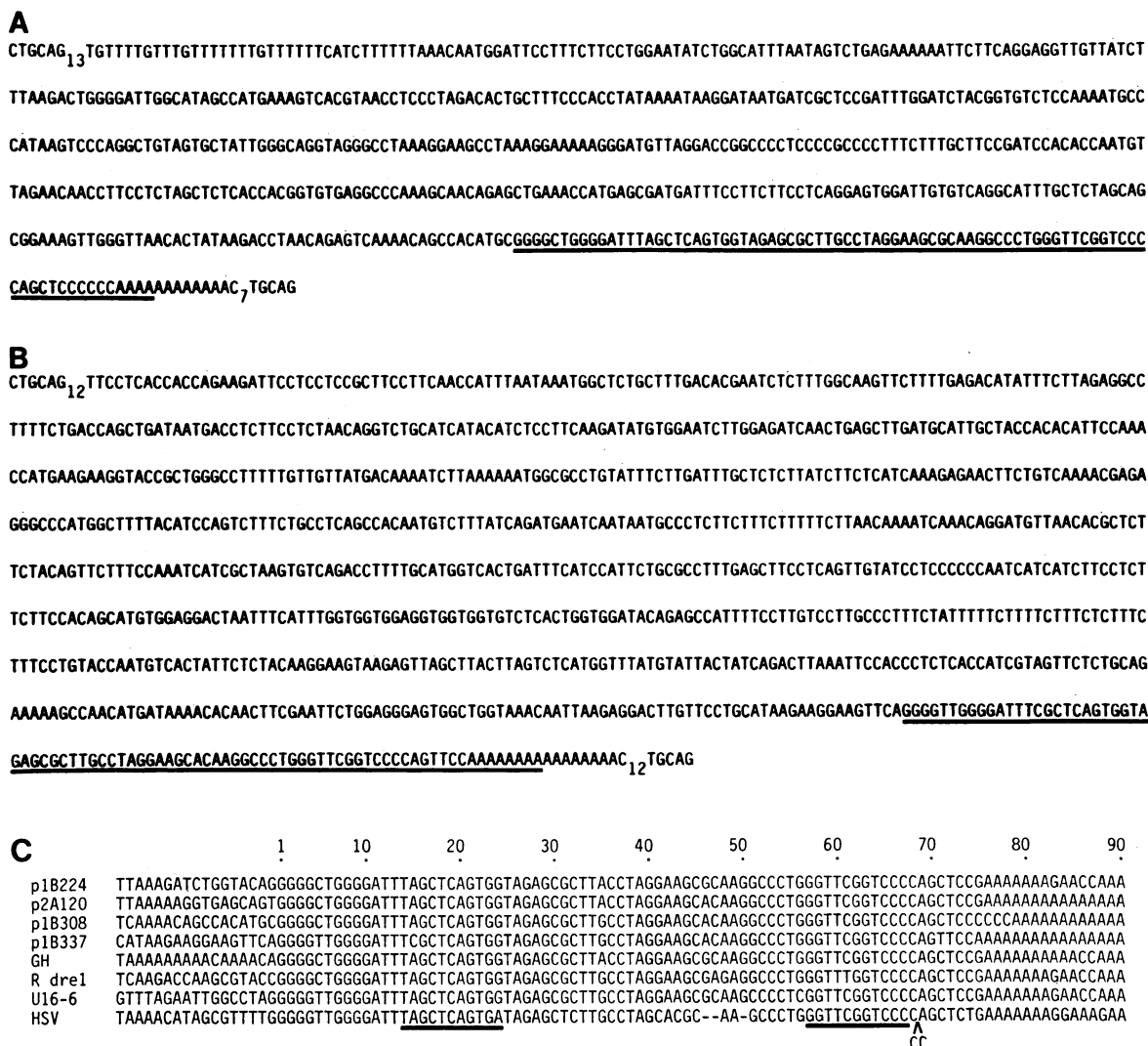
FIG. 2. Nucleotide sequences of ID clones. (*A*) p1B308, (*B*) p1B337. The nucleotide sequence (5' → 3') of the strand of the cDNA insert corresponding to the plus strand of BC1 RNA is shown in each case and the ID core sequence is underlined. For p1B308 the 5' terminal is proximal to the *Eco*RI site of pBR322; for p1B337 the 5' terminal is distal to the *Eco*RI site of pBR322. (*C*) Comparison of ID sequences from p1B224 and p2A120 (1), p1B308, and p1B337, growth hormone (GH; see ref. 3), a tubulin pseudogene (R dre1; see ref. 5), a rat U1 gene (U16-6; see ref. 6), and HSV (7). Three gaps and a C-C insertion were required to align the HSV sequence. Regions homologous to *Pol* III promotor consensus sequences (8) are underlined.

sequences, one just upstream from the *ras* gene of the rat Harvey sarcoma virus (HSV) (7), one in a putative pseudogene of rat α-tubulin (5), and one in a rat U1 RNA gene (6).

ID Sequences Are Represented 10^5 Times in the Genome. The occurrence of the ID sequence in 5 out of 163 different randomly selected cDNA molecules suggested that ID sequences might exist in multiple copies in the genome. Furthermore, in Southern blot analyses, ID clone p1B224 hybridized as a smear across restriction fragments of rat genomic DNA of all molecular sizes, indicating that the rat genome contains many different copies of the ID sequence (Fig. 3*B*). To estimate the number of copies of the ID sequence in the rat genome, we therefore probed a rat genomic library cloned as partial *Eco*RI fragments in Charon 4a by Sargent *et al.* (9). We picked 450 random discrete plaques, suspended each in 0.1 ml of phage buffer, spotted 1 μl of each phage suspension on a freshly plated bacterial lawn, prepared blots of the matrix of resulting plaques, and probed with p1B224. Of the independent phage isolates, 198 out of 450 (43%) hybridized to the ID probe. The Sargent library contains the rat genome as 10- to 20-kilobase (kb) fragments: because the rat genome size is 3.2 × 10^6 kb, the entire

genome would be contained in 2.15 × 10^5 separate clones. Because 43% of randomly picked clones hybridize to the ID sequence (non-ID containing fragments of p1B224 hybridize to single bands on Southern blots, as discussed below, so all hybridizations in this experiment can be considered due to ID), there are at least 9.2 × 10^4 ID copies in the genome. If we assume that ID sequences occur randomly, then 43% of these fragments will contain 2 copies of ID, and 43% of these will contain 3 copies, and so on. This brings the calculation to 1.5 × 10^5 copies of ID per genome. A group of previously described repetitive elements, the *Alu* family (recently reviewed in ref. 10) also hybridize to small cellular RNAs. However, *Alu* sequences do not hybridize to tissue-specific RNAs and are much more prevalent (3–5 × 10^5 copies per genome) than ID sequences and therefore are distinct.

ID Sequences Are Linked to Brain Genes. Clearly, ID sequences are multiply represented in the genome. What about the non-ID portions of our several cDNA clones? Using the restriction maps derived from the nucleic acid sequences of four of the ID-containing cDNA clones, we isolated restriction fragments (Fig. 3*A*) that did not contain ID sequences for use as probes in a Southern blot analysis of rat liver geno-
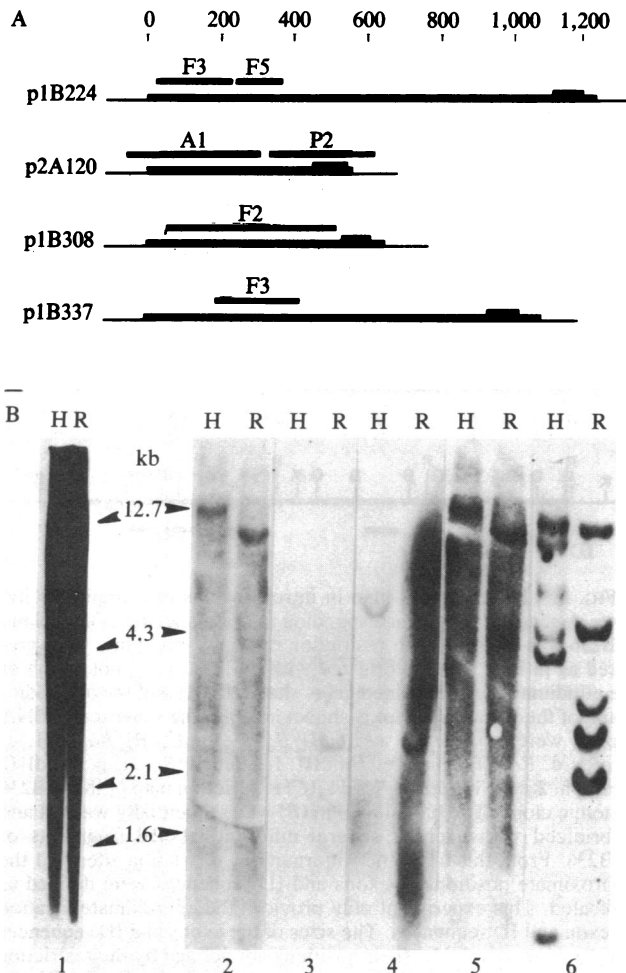
Biochemistry: Milner *et al.*

Proc. Natl. Acad. Sci. USA 81 (1984) 715

FIG. 4. RNA blot analysis of RNAs corresponding to ID-containing clones. (*A*) Clone p2A120 fragment A1. (*B*) Clone p1B224 fragment F5. (*C*) Clone p1B224 fragment F3. (*D*) Clone p1B337 fragment F3. RNA samples from brain (lanes 1), liver (lanes 2), and kidney (lanes 3) were treated as described in Fig. 1 and hybridized with restriction fragments derived as indicated in Fig. 3*A*.

FIG. 3. (*A*) The structure of the rat brain ID containing cDNA clones. The cDNA insert of each clone is indicated by the thick line; surrounding pBR322 sequences are indicated by a thin line. The thicker block at the right-hand end of each insert indicates the position of the ID sequence. Labeled bar above each clone diagram indicates the position of restriction fragments used in Southern and RNA blot analyses. The scale is in bp. (*B*) Southern blot analysis of rat liver DNA; 20 μg of Sprague–Dawley rat liver DNA (4) digested with *Hind*III (H) or *Eco*RI (R) was subjected to electrophoresis on 0.8% agarose gels, blotted to nitrocellulose, and hybridized with ^{32}P-labeled restriction fragments derived as indicated in *A*. 1 and 2, clone p2A120 fragments P2 and A1, respectively; 3 and 4, clone p1B224 fragments F3 and F5, respectively; 5, clone p1B308 fragment F2; 6, clone p1B337 fragment F3.

mic DNA (Fig. 3*B*). The non-ID regions of each clone except p1B337 hybridized to different single restriction bands for each enzyme tested and therefore are probably present in single-copy genes. The p1B337 non-ID fragments each hybridize to many bands so these represent a multigene family.

Are the genes corresponding to the ID-containing clones expressed only in brain? We used the non-ID restriction fragments as probes in RNA blot analysis of brain, liver, and kidney poly(A)$^+$ RNA (Fig. 4). Certain fragments from p2A120 and p1B224 hybridized to brain-specific poly(A)$^+$ RNA: for example, fragment A1 of p2A120 hybridized to a brain-specific mRNA of ≈3,000 nucleotides, and fragment F5 of p1B224 hybridized to another brain-specific mRNA target of ≈3,000 nucleotides and also to a less prevalent brain-specific 1,600-nucleotide species. p1B337 hybridizes to at least five mRNAs in brain, liver, and kidney preparations (Fig. 4), but because there are eight (or more) bands in Southern blot experiments (Fig. 3), the gene family defined by p1B337 may include a brain-specific gene that is not high-
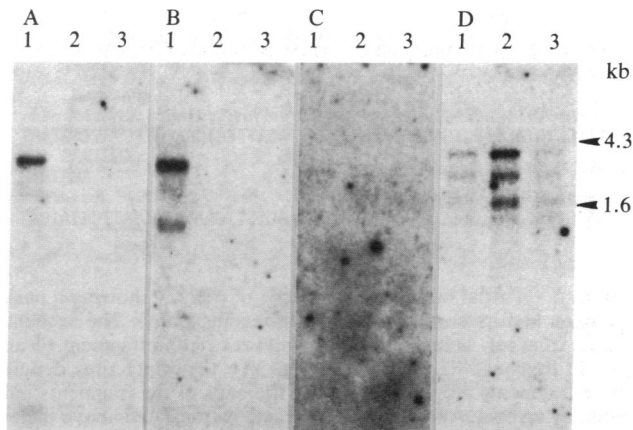
ly transcribed. p1B308 fragments do not appear to hybridize to any brain, liver, or kidney mRNA species. Therefore, p1B308 may be a clone of a rarer brain mRNA. We know from other studies (4) that a substantial portion of brain mRNAs are expressed at less than a few copies per cell. Most such rare mRNAs are likely to be brain specific. Thus, definitely clones p2A120 and p1B224, probably clone p1B308, and possibly clone p1B337 correspond to brain-specific genes.

ID Sequences Are Contained in Precursors of Brain-Specific mRNAs. We had previously suggested that the ID containing cDNA clones were derived from incompletely processed mRNA molecules (1). Definition of the mature cytoplasmic RNAs corresponding to several ID clones now allows us to test this possibility. For example, although fragment F5 of p1B224 detects brain-specific mRNA molecules, fragment F3 of p1B224 does not clearly hybridize to any RNA target (Fig. 4). (F3 may exhibit slight hybridization to a 3,000-nucleotide brain-specific species. We do not know if this is due to F3 sequences or to slight contamination of F3 with other p1B224 fragments.) Both fragments F3 and F5, however, detect genomic DNA fragments of identical sizes in Southern blots (Fig. 3). Because fragments F5 and F3 are immediately adjacent in the p1B224 sequences, fragment F3 must therefore correspond to intron sequences; F5 contains largely exon sequences, which are expressed in mature mRNA. Indeed, an open-reading frame (Fig. 5) runs through most of F5 until reaching a terminator TAA triplet near its 3' end. To use the open-reading frame for translating p1B224, sequences upstream from F5 must be removed. Potential splice acceptor and donor sites are indicated in Fig. 5. The p2A120 fragment A1 sequence does not contain long open-reading frames and, therefore, could contain both introns and exons or could correspond to a noncoding mRNA region.

Because p1B224 represents an mRNA precursor containing both intron and exon sequences, it seems very likely that the p1B224 ID sequence shares the same fate as upstream fragment F3—namely, splicing. Both F3 and ID might each reside in an intron of the brain-specific gene corresponding to cDNA clone p1B224. It is very clear that the cytoplasmic poly(A)$^+$ RNA population from which we made cDNA contains mRNA precursors. Because 62% of the transcripts of brain nuclear extracts contain an ID sequence (2) and only 3% of our cDNA clones made from total poly(A)$^+$ cytoplasmic RNA contain such a sequence, it seems that the ID se-
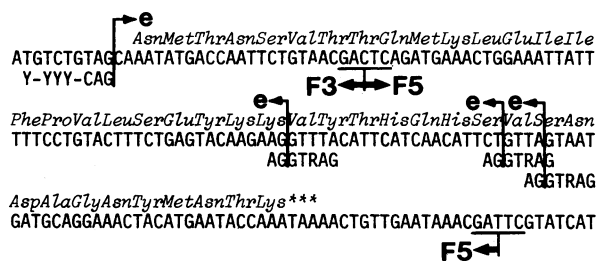
```
                ┌─●e
                │ AsnMetThrAsnSerValThrThrGlnMetLysLeuGluIleIle
     ATGTCTGTAGCAAATATGACCAATTCTGTAACGACTCAGATGAAACTGGAAATTATT
     Y-YYY-CAG
                              F3◄──►F5
                                                ●◄─●◄─
      PheProValLeuSerGluTyrLysLysValTyrThrHisGlnHisSerValSerAsn
     TTTCCTGTACTTTCTGAGTACAAGAAGGTTTACATTCATCAACATTCTGTTAGTAAT
                            AGGTRAG                AGGTRAG
                                                   AGGTRAG

      AspAlaGlyAsnTyrMetAsnThrLys***
     GATGCAGGAAACTACATGAATACCAAATAAAACTGTTGAATAAACGATTCGTATCAT
                                       F5◄─┤
```

FIG. 5. Partial nucleotide sequence of p1B224 showing a possible exon and its corresponding open-reading frame. The sequence (taken from ref. 1) includes all of *Hin*fI restriction fragment F5 and part of fragment F3 (mapped in Fig. 3A): the *Hin*fI sites defining these fragments are underlined and the ends of the fragments indicated by arrows. An open-reading frame is translated above the nucleotide sequence. Potential splice donor and acceptor sites (11) are indicated by their appropriate consensus sequence: the 5' and 3' terminals of a possible exon (e) generated by these sites are indicated by rightward and leftward pointing arrows, respectively.

quences are indeed present in an unstable portion of brain RNA.

Introns of Brain-Specific Genes Contain ID Sequences. The data thus far presented suggest that ID sequences are found in precursors of brain mRNAs and probably occupy intron regions of the corresponding genes. Our previous studies (4) indicate that some 30,000 different mRNAs are expressed in rat brain. There cannot, therefore, be a one-to-one relationship between ID sequences and brain-specific genes because ID sequences are in several-fold excess. However, all or at least most brain-specific genes could still contain ID sequences. To approach this question, we isolated genomic clones corresponding to cDNA clones that we had previously defined as brain specific (4), that did not contain ID sequences, and that corresponded to single-copy genes. One such cDNA clone is p1B236, whose nucleotide sequence we have determined and whose protein-encoding region we have identified using antibodies to chemically synthesized peptides corresponding to an amino acid sequence deduced from the nucleotide sequence (12). Both the 1B236 mRNA and its protein product are brain specific. We have isolated a 15-kb genomic clone corresponding to 1B236 and have probed blots of single and double restriction digests of the genomic clone (Fig. 6A) with ^{32}P-labeled fragments of the cDNA insert. The analysis of the data allowed us to construct a restriction map of the genomic clone and to align four intron and five exon regions approximately to that restriction map (Fig. 6C). The same blot that was used to generate this map was further hybridized with an ID sequence probe (Fig. 6B): two regions of the 1B236 genomic clone contain ID sequences (indicated in Fig. 6C). One of these is clearly in an intron; the other maps just 5' to sequences that are contained in the cDNA clone, but because this clone lacked the 5' terminal 1,000 nucleotides of the mRNA, this second ID sequence is almost certainly contained in another 1B236 intron.

DISCUSSION

We have defined an 82-nucleotide ID core sequence by comparing the sequences of four independent rat brain cDNA clones with four other ID sequences. The ID sequence was previously shown to share homology with a cytoplasmic poly(A)$^+$ RNA molecule of 160 nucleotides (BC1 RNA) present exclusively in neural tissue (1, 2). We have now shown that the ID sequence is (*i*) present in cDNA clones representing precursors of brain-specific mRNA molecules, (*ii*) probably present in some of the introns of those mRNA precursors, and (*iii*) represented about 1.5×10^5 times in the rat genome, a number higher than estimates of the number of
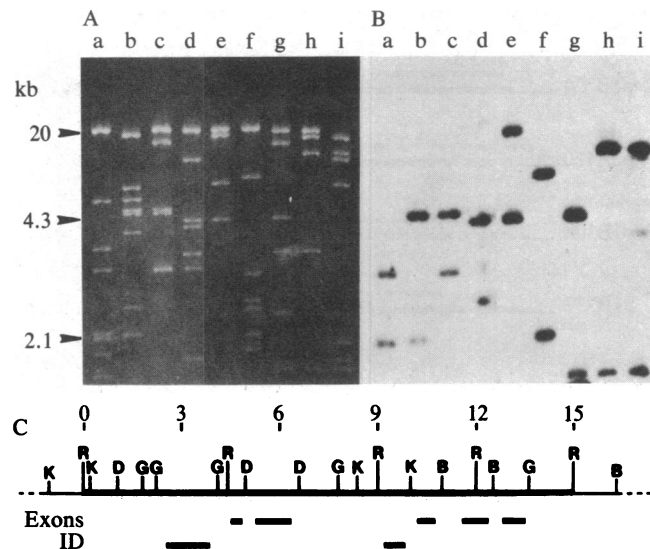

FIG. 6. ID sequences map in intron regions of a brain-specific gene. A genomic clone corresponding to brain-specific cDNA clone p1B236 was digested with restriction enzymes, and blots were prepared as in Fig. 3 and hybridized with p1B224. The photograph of the ethidium bromide-stained gel is shown in (A) and the autoradiograph of the hybridized blot is shown in (B) to the same scale. DNA digests were as follows: a, *Eco*RI/*Bgl* II; b, *Eco*RI/*Bam* HI; c, *Eco*RI; d, *Eco*RI/*Kpn* I; e, *Hin*dIII; f, *Hin*dIII/*Bgl* II; g, *Hin*dIII/*Kpn* I; h, *Kpn* I; i, *Bam*HI/*Kpn* I. (C) Restriction map of the p1B236 genomic clone. The blot shown in (B) was sequentially washed and hybridized with each of several different labeled fragments of p1B236. From this and other information, restriction sites and the approximate positions of exons and ID sequences were defined as indicated. This experiment only provides the approximate location of exon and ID sequences. The sizes of the exons and ID sequences are not shown to scale: their positions are defined by the restriction sites bounding the fragments that contain them. R, *Eco*RI; D, *Hin*dIII; K, *Kpn* I; B, *Bam*HI; G, *Bgl* II.

brain-specific mRNA species. (*iv*) Furthermore, when a rat genomic phage library is screened, clones selected because they contain a particular brain-specific gene also contain ID sequences. For the 1B236 gene, an ID sequence is found in each of two introns, but not in two other introns of the gene. DNA sequence analysis will further detail the relationship of the ID sequences to exon regions.

By considering the several examples in Fig. 2C, we can now sharpen our definition of the brain ID sequence. We have made the tacit assumption that all of these sequences are related. Indeed, we have shown that the HSV-ID analogue, which is the most divergent sequence of the group, hybridizes to BC1 RNA but not to a similar target in liver or kidney RNA. The consensus ID sequence consists of an 82-nucleotide core sequence in which 62 positions are invariant in the eight examples, 15 positions are common in seven out of eight cases, 2 positions are common in six out of eight cases, and the other 3 positions are exclusively either purines or pyrimidines (Fig. 2C). Three clustered gaps and two adjacent insertions are required to align the HSV sequence with the other seven ID sequences. The consensus core sequence terminates with eight adenine residues at the 3' end, which are followed by an oligo(dA)-rich region that is mostly homologous in the various isolates. Position 83 is a purine and positions 86 and 87 are A-A or C-C. As we previously noted, we believe the adenine-rich region to be responsible for the frequent representation of such clones in our oligo(dT)-primed random brain cDNA collection. This adenine-rich region may well be in internal rather than in 3' regions of the RNA molecules copied to produce our cDNA clones. On the 5' side of the core, 10 of 17 positions favor some particu-

lar nucleotides: of special note is the sequence T-X-A-A-A in positions −17 to −13. In addition, the consensus ID sequence contains regions homologous to consensus RNA polymerase III initiation sequences (Fig. 2C)—indeed, ID clones act as *Pol* III promoters *in vitro* (2). Therefore the −17 to −13 sequence may be analogous to a similar sequence (A-A-A-A-G-T) observed the same distance 5′ to 5 *Pol* III transcription units (13, 14).

As previously noted for p1B224 and p2A120, each rat brain cDNA sequence has some of the characteristics of intron sequences, such as long strings of repeated short oligonucleotides or homopolymeric purine or pyrimidine tracts. Although neither new sequence contains an extensive open-reading frame in the ID orientation, p1B337 does contain a 253 triplet open-reading frame on the opposite strand. Given that our clones were constructed by oligo(dT) primer extension, this open-reading frame could not be used in the RNA molecule we copied, but it could possibly function in a complementary transcript.

The evidence that some ID sequences are located in brain-gene introns now seems quite convincing. At least two are found in introns of the 1B236 gene that encodes a brain-specific protein. ID sequences in some of our cDNA clones seem to be in introns, as judged by the hybridization patterns of adjacent regions to mature mRNA and by the nucleotide sequences of the clones. In this respect, the finding that three of the four cDNA clones containing an ID sequence each hybridize to a single genomic fragment in each of two restriction digests argues that these cDNA clones are probably colinear with the genomic DNA and, hence, probably span introns. ID sequences probably reside in the introns of the $\approx 3 \times 10^4$ brain specific genes. This is in accord with the finding that ID sequences are contained in 62% of the transcripts made *in vitro* from brain chromatin (2). They therefore represent a repetitive sequence marking brain-specific genes and are likely to be involved in the control of tissue-specific gene expression. Britten and Davidson (15) were the first to propose such a model to suggest a function for the highly repeated sequences they discovered by DNA complexity measurements.

ID sequences could operate at the level of transcription (see ref. 2) or, less likely, at the level of mRNA maturation, perhaps directing the splicing of hnRNA precursors in brain or providing a recognition sequence for brain-specific proteins. In this regard, it is of interest that the 82-nucleotide ID core structure (Fig. 2C) contains self-complementary sequences at positions 1–13 and 62–74. Annealing these regions would generate a stem–loop structure (Fig. 7). The stem is 13 nucleotide pairs long, including G·U base pairs. Although there are several differences between the eight sequences compared in Fig. 2C, variations within the proposed stem region only interchange G·U and G·C base pairs, except for the discontinuity in the HSV-*ras* gene, which we have indicated as a bulge. This stem–loop structure would be expected to form in all transcripts (*Pol* II or *Pol* III) of brain-specific genes and could act as a nucleation site for RNA
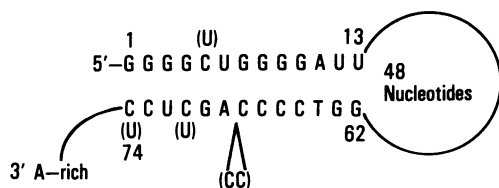
folding, leading either to splicing or to protein recognition. However, because ID sequences are only transcribed in neural tissue (2), models involving tissue-specific splicing seem superfluous.

Clearly, there are more ID sequences in the genome than brain genes, and this confounds a strict correlation. We expect that some cDNAs picked up in our screen (4) will be copies of glial-specific mRNAs and that the corresponding genes will not necessarily contain ID sequences. Some genes, such as 1B236, may contain multiple ID copies, but many ID sequences are likely to be nonfunctional in brain-gene expression, such as the ID sequence found in the tubulin pseudogene (5), growth-hormone gene (3), or more recently on the rat prolactin gene (16). It was suggested by Lemischka and Sharp (5) that the ID sequence may be a mobile genetic element. This could provide the flexibility required to coordinately control 3×10^4 different brain transcripts as the function of the brain expanded during recent mammalian evolution and would also generate a battery of nonfunctional genomic ID copies. If transcription of ID sequences by *Pol* III is an important aspect of gene control, then ID sequences that transpose to irrelevant, nonlethal sites may not be functional *Pol* III units because they lack some positional information.

It has already been shown that splicing is a necessary step in the expression of at least some eukaryotic genes (17). The present studies show that introns may have specific functional significance in gene expression. One wonders if introns carry other functions.

FIG. 7. Possible stem–loop structure for the ID sequences. A consensus sequence is used: variant residues are shown in parentheses. The C-C insert used to align the HSV sequence (Fig. 2C) is also indicated.

1. Sutcliffe, J. G., Milner, R. J., Bloom, F. E. & Lerner, R. A. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4942–4946.
2. Sutcliffe, J. G., Milner, R. J., Gottesfeld, J. M. & Lerner, R. A. (1983) *Nature (London)*, in press.
3. Barta, A., Richards, R. I., Baxter, J. D. & Shine, J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4867–4871.
4. Milner, R. J. & Sutcliffe, J. G. (1983) *Nucleic Acids Res.* **11**, 5497–5520.
5. Lemischka, I. & Sharp, P. A. (1982) *Nature (London)* **300**, 330–335.
6. Watanabe-Nagasu, N., Itoh, Y., Tani, T., Okano, K., Koga, N., Okada, N. & Ohshima, Y. (1983) *Nucleic Acids Res.* **11**, 1791–1801.
7. Dhar, R., Ellie, R. W., Shih, T. Y., Oroszlan, S., Shapiro, B., Maizel, J., Lowy, D. & Scolnick, E. (1982) *Science* **217**, 934–937.
8. Galli, G., Hofstetter, H. & Birnstiel, M. (1981) *Nature (London)* **294**, 626–631.
9. Sargent, T. D., Wu, J., Sela-Trepat, J. M., Wallace, R. B., Reyes, A. A. & Bonner, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3256–3260.
10. Jelinek, W. R. & Schmid, C. W. (1982) *Annu. Rev. Biochem.* **51**, 813–844.
11. Sharp, P. A. (1981) *Cell* **23**, 643–646.
12. Sutcliffe, J. G., Milner, R. J., Shinnick, T. M. & Bloom, F. E. (1983) *Cell* **33**, 671–682.
13. Korn, L. J. & Brown, D. D. (1978) *Cell* **15**, 1145–1156.
14. Peterson, R. C., Doering, J. L. & Brown, D. D. (1980) *Cell* **20**, 131–141.
15. Britten, R. J. & Davidson, E. M. (1969) *Science* **165**, 349–358.
16. Schuler, L. A., Weber, J. L. & Gorski, J. (1983) *Nature (London)* **305**, 159–160.
17. Gruss, P., Lai, C.-J., Dhar, R. & Khoury, G. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4317–4321.