Rodent BC1 RNA gene as a master gene for ID element amplification

(short interspersed repetitive DNA element/retrotransposons/gene amplification/evolution)

JOOMYEONG KIM*, JOHN A. MARTIGNETTI†, M. RICHARD SHEN*, JÜRGEN BROSIUS†, AND PRESCOTT DEININGER*‡§

*Department of Biochemistry and Molecular Biology, Neuroscience Center of Excellence, Stanley S. Scott Cancer Center, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112; †Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, LA 70121; and ‡Fishberg Research Center for Neurobiology Mount Sinai School of Medicine, New York, NY 10029

Communicated by Anthony Cerami, January 3, 1994 (received for review November 22, 1993)

ABSTRACT ID elements are short interspersed repetitive DNA elements (SINEs) which have amplified in rodent genomes via retroposition, a process involving an RNA intermediate. BC1, an abundant ID-related transcript, is transcribed from a conserved, single-copy gene in rodents. The gene encoding BC1 RNA represents one of the earliest and possibly the first ID-containing sequence. Comparison of consensus sequences of each rodent ID with its corresponding BC1 RNA gene showed that the variations of BC1 RNA within rodents corresponded to specific changes within the ID consensus sequence for each rodent species. This supports the hypothesis that the BC1 gene is a master gene responsible for the amplification and evolution of ID elements. The rat ID family consists of at least four subfamilies, with the oldest subfamily having been derived from the BC1 RNA. The other three subfamilies appear to have been derived from a new master gene(s), which has been responsible for the large increase in ID element copy number within the rat genome. We have found that the guinea pig genome contains two copies of the BC1 gene, apparently the result of a DNA-mediated duplication event. Both of these guinea pig BC1 genes have a conserved TATA-like element in the 5' flanking region and have contributed to guinea pig ID amplifications.

ID elements (originally termed R.dre.1) are a major class of short interspersed repetitive DNA elements (SINEs) found in rodent genomes (1). They are 85–105 bp long and are composed of a core region that averages 75 bp and a 10- to 40-bp oligo(dA) tail. The core region of the ID element shares sequence identity with alanine tRNA (2), indicating that ID elements were ancestrally derived from a tRNA gene. ID elements, like all SINEs, are flanked by direct-repeat sequences and contain internal promoter elements for RNA polymerase III transcription (3). The amplification of ID elements is thought to occur by retroposition (4) of an RNA polymerase III-derived transcript. The copy number of ID elements within the genomes of various rodents varies by at least 2–3 orders of magnitude (5), indicating that ID elements have different amplification rates in these species. Rat has the largest ID copy number of any rodent examined, with 130,000 copies, suggesting that this amplification has been recently accelerated specifically in the rat genome (6).

Studies have indicated that the vast majority of human *Alu* repetitive elements, and many other mammalian SINEs as well, are incapable of retroposition (reviewed in ref. 7). Amplification at any given time appears to be dominated by a limited number of master genes, with perhaps a single gene being disproportionately responsible for the overall evolution of a SINE family. The reasons that most copies are incapable

of retroposition are not completely understood. However, a number of potentially important factors have been discussed (reviewed in ref. 8). These include the relative transcriptional silence of most of the SINE copies and potential processing of the primary transcripts. The detailed structure of the RNA being generated from different copies may also influence the efficiency of reverse transcription of the element. Factors affecting transcriptional activity may include mutations within the individual elements (9, 10) or the sequence context into which each element inserts (i.e., upstream elements or factors controlled by general chromatin domains).

Three different ID transcripts, determined by size, are reported in rat tissues: BC1, BC2, and T3 RNAs (3, 5, 11). BC1 is the dominant form in neuronal cells, whereas BC1 and BC2 can be found at modest levels in a wide range of tissues and T3 RNA is detected preferentially in testes (11). BC1 RNA is a homogeneous transcript derived from a single-copy gene in rat (12). BC1 RNAs of other rodent species also have similar unique sequences, indicating that the BC1 gene has been conserved in rodent evolution (13). Its unique expression pattern in neuronal cells and its conservation in evolution strongly suggest that BC1 RNA gene has been “exapted” (14, 15) into a functional role in the rodents. The transcriptional activity of the BC1 RNA gene seen in all rodents, including in germ-line cells (H. Tiedge and J.B., unpublished work), also supports the possibility that it serves as a master gene for ID family evolution. We therefore have carried out an evolutionary analysis of the BC1 gene and ID repeats in various rodent genomes to test this hypothesis.

METHODS

Guinea Pig Genomic Library Construction, Screening, and Sequencing. A guinea pig genomic library containing DNA fragments from partial *Sau3A1* digestion (13) was initially screened with a 75-bp hybridization probe corresponding to the ID portion of the guinea pig BC1 gene (see Fig. 2) at low stringency (at 37°C in 50% deionized formamide/5× Denhardt's solution/0.1% NaDodSO₄ with denatured salmon sperm DNA at 100 µg/ml) and confirmed by hybridization with the same region of the rat BC1 gene. Positive clones were purified to homogeneity by multiple rounds of plaque hybridization. Purified λ DNA was digested to completion with *Sau3A1* and ligated into the *Bam*HI site of phagemid M13mp19. The ligation mixtures were transformed into *Escherichia coli* XL-1 Blue cells (Stratagene) and M13 plaque lifts were screened with the probe used in the initial screening. Single-stranded DNA from M13 subclones from each λ clone was sequenced by standard dideoxy procedures with a Sequenase kit (United States Biochemical). Second strands

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: SINE, short interspersed repetitive DNA element.
§To whom reprint requests should be addressed.

like element has been demonstrated for other RNA polymerase III-transcribed genes (17–19). Analysis of the coding region of the two genes shows 10 mismatches between them (Fig. 2).

Comparison of the four guinea pig ID elements with the two BC1 genes shows a high level of sequence similarity in the ID portion of the gene and the A-rich region (Fig. 2). Guinea pig-specific changes seen in the BC1 genes (regions 46–48 and 94–100) are all found in the four retroposed ID elements, suggesting that these ID elements were derived from the BC1 genes. Interestingly, only one of the four ID elements, ID1, is more closely related to BC1-1 whereas the other three are related to BC1-2. This implies that both genes are involved in ID element amplifications and that the BC1-2 gene may be the more active of the two in terms of amplification. Both BC1 genes have one base not found in any of its ID copies; A⁵¹ in BC1-1 and G¹¹ in BC1-2. These are likely to represent new mutations in the BC1 genes, generated after the amplification of the ID copies.

Mouse ID Elements. With the criteria described in *Methods*, 3 mouse and 62 rat ID elements were found in the EMBL databank containing about 7225 rodent sequences (48% mouse, 48% rat, and 4% guinea pig and hamster), whereas no hamster or guinea pig ID elements were found. This ratio is in agreement with the ID copy-number difference between the mouse and rat genomes (5) and the relative database sizes. We have also included a fourth mouse ID whose sequence we determined previously. The database search also identified a number of more divergent sequences with sequence similarity to the ID repeats. These more divergent sequences were generally more similar to the alanine tRNA than to the ID repeats, suggesting that they represented an independently evolving group of SINEs or even an extensive group of tRNA pseudogenes, and therefore were not considered further in this analysis.

The four mouse ID elements are also flanked by direct-repeat sequences indicating that they are all products of retroposition events. These mouse elements show high sequence similarity in the ID core region (positions 1 to ≈ 76), which helps to determine the consensus base at each position except one, position 73 (Fig. 3). The consensus base at each position of the mouse ID element is identical to the base at the same position of the mouse BC1 RNA gene. Compared with other rodent BC1 RNA genes, the mouse BC1 RNA has two specific changes, T⁷⁴ and G⁷⁶, which are seen in all the mouse ID elements in the database. Additionally, the pMP11 ID element has extensive sequence identity throughout its A-rich region with the A-rich region of the BC1 RNA gene.

GP BC1-1
 GP BC1-2
 RAT BC1

GP BC1-1
 GP BC1-2
 RAT BC1

FIG. 1. Schematic representation of two guinea pig (GP) BC1 genes. The boxes indicate sequence similarities between the two guinea pig BC1 genes, and the solid lines indicate regions with no obvious sequence similarities. The shaded boxes indicate the coding region for the BC1 RNA. The broken line between the boxes for the BC1-2 gene represents a deletion relative to the BC1-1 gene. Comparison of the flanking region of the two guinea pig BC1 genes with rat BC1 shows some sequence similarities, with the potential TATA-like sequence underlined and in bold type. The transcription start site of the BC1 gene, based on the rat BC1 transcript (13), is marked as +1. The regions inside the parentheses represent the BC1 coding region. X, base not present in the rat BC1.

FIG. 2. Alignment of four guinea pig (GP) ID elements with two BC1 genes. Sequences are compared with the BC1-1 sequence. Dashes indicate deletions, and the star in the A-rich region indicates the addition of an A residue. The superscript^a refers to a longer poly(dA) tail, A₂₁CA₈CA₈, in this region. The sequences inside parentheses at the end of individual sequences represent the direct repeats flanking the inserted ID elements, and the underlined bases denote the mismatched positions between 5' direct-repeat and 3' direct-repeat sequences (G to A; G to T). Positions in bold type are those which best demonstrate the relationship between the individual ID repeats and the two BC1 genes.

Rat ID Elements. Through multiple alignment of 62 rat ID elements (data not shown, provided upon request), four diagnostic base changes have been identified which allow us to divide rat ID elements into four subfamilies (Fig. 4). Each subfamily has a different average sequence divergence relative to their consensus, indicating that each is of a different average evolutionary age, as has been seen for other SINE subfamilies (6, 7, 20).

Among the rat subfamilies, type 1 shows the most sequence divergence (4.9%), suggesting that it is the oldest subfamily. The consensus base at each position of type 1 is identical to the same position of the rat BC1 gene. However, most members of the type 1 subfamily show sequence heterogeneity near positions 29–32. Considering the high similarity in all other positions, the origin of this heterogeneity is not clear, although it does include CpG sites, which are known to mutate much more rapidly than other positions (21). Alternatively, small RNAs often undergo base modification (22) that could lead to inaccurate copying during the retroposition process. A more trivial explanation would be sequencing errors due to sequence compressions that occur in this region. The rat BC1 gene has two specific base changes, C⁷⁴ and A⁷⁶, when compared with the mouse BC1 gene, whereas the five rat ID type 1 elements are identical to

the rat BC1 at the 74th position. Interestingly, two rat ID elements have the same base as the rat BC1 at the 76th position, while the other three rat ID elements have a G at that position as seen in the mouse BC1 gene. This difference is thought to reflect the evolutionary change of BC1 because the likely progenitor sequence of the two rodent BC1 RNAs would be CCGG in that region (positions 73–76).

The three other rat ID subfamilies, types 2–4, have one common base change, C⁶⁷, which is not found in other rodent ID elements or BC1 RNAs. When compared with type 1, these three subfamilies show a higher degree of sequence similarity, indicating that they were formed more recently. However, subfamilies 2–4 are thought to have been formed at slightly different evolutionary times, on the basis of sequence divergence. The average age of the subfamilies is estimated to be 9.8 million years (Myr) for type 1, 5.4 Myr for type 2, 3.6 Myr for type 3, and 2.8 Myr for type 4, by assuming that the neutral mutation rate for rodent is 0.5%/Myr (23). When these three subfamilies are placed in order of formation, the diagnostic changes of each subfamily accumulate progressively, indicating that they have been sequentially derived from each other. Some members of the three subfamilies also have a GAACC sequence embedded at approximately positions 83–87 in the A-rich region. Only

FIG. 3. Alignment of mouse ID elements and rat subfamily type 1 ID elements with mouse and rat BC1 RNA gene. Each ID element is named by its EMBL database name. Numbers in parentheses indicate the position of the ID element in the individual sequence. The pMP11 sequence is one we have determined from an unknown mouse locus. The RNDBIPG sequence (R. Freneau, P. Gray, R. Einstein, and J.B., unpublished work) is an ID element upstream from a rat diazepam binding inhibitor (DBI) pseudogene. Sequences are presented relative to the mouse BC1 RNA sequence. Dashes indicate deletion, and the star indicates the addition of an A residue. Positions which differentiate the mouse and rat BC1 genes are in bold type.

FIG. 4. Alignment of four consensus sequences of rat subfamilies with rat BC1 RNA. Each subfamily is named in sequential order, type 1 through type 4, based on the sequence divergence within the subfamily. Subfamilies 2–4 have a unique sequence motif, GAACC (underlined), in the oligo-(dA) tail. This GAACC motif is not found in all members of each subfamily. In fact, the lowercase cc in the type 2 subfamily sequence is not present in the majority of the subfamily members. This suggests that there may be even more subfamily divisions, although we favor the explanation that the sequences without the GAACC motif are from loss of this sequence in individual elements due to well-known instabilities in the A-rich region. A progressively larger fraction of the older subfamily members have lost this sequence, consistent with this latter explanation.

10% of type 2 members contain this sequence motif, compared with 50% of type 3 and 75% of type 4. Some of the sequences may be missing the GAACC motif because the reverse transcription during the retroposition process was primed 5' to the GAACC sequence in the RNA. However, it seems likely that much of the heterogeneity at this position in the different subfamilies is the result of loss of this region in some copies because of the high rate of variability in A-rich regions of SINEs (24). This would result in the observed increased loss of the GAACC sequence in the older subfamilies.

DISCUSSION

Most studies of SINEs have demonstrated that there is not a specific mechanism for the removal of these elements (reviewed in ref. 1). Thus, the low copy number of ID elements in guinea pig is almost certainly due to a very low amplification rate relative to the other rodents. The presence of BC1 genes in guinea pigs, which have only 100–200 ID copies, suggests that it was the first ID-containing gene, although we cannot absolutely rule out the presence of a small number of ID-related sequences that existed prior to the BC1 gene. It seems likely that the BC1 gene was generated through a retroposition process (12, 13), based on the close correspondence of the oligo(dA)-rich region at 3' end of its ID sequence, in a position corresponding closely to the end of the mature alanine tRNA. The lack of direct repeats flanking the ID portion of the BC1 gene is then consistent with its older evolutionary age and resulting divergence of these direct repeats. Alternatively, it is possible that the BC1 RNA gene arose by a mechanism other than retroposition.

Guinea pig, mouse, and rat ID elements have species-specific sequence differences that correlate with the sequence variation of BC1 RNA within the same species (Fig. 5). The close coevolution of the BC1 RNA gene with the ID repeats demonstrates that the BC1 gene is a master gene (7) which has influenced the evolution and amplification of the ID family in most rodent genomes studied and of at least one of the ID subfamilies within the rat genome. The only alternative explanation for this coevolution would be through a homogenization mechanism, such as gene conversion, and studies to date suggest that SINE families are not subject to a significant level of gene conversion (1). As an ID master gene, it is possible that the BC1 gene has directly made the majority of ID elements. Alternatively, BC1 could represent a master gene that has controlled the evolution of ID elements without having made the majority of the elements. If a small proportion of ID elements were capable of amplification activity for a limited period of time after their formation, they would not have time to diverge significantly from the BC1 gene sequence prior to losing their amplification capability.

Thus, these copies would then closely reflect the BC1 gene sequence and would not form an obviously independently evolving subfamily (25).

There are two BC1 genes in the guinea pig genome, one of which has been formed by a DNA-mediated duplication, as evidenced by duplication of extensive flanking sequences with no obvious direct repeats. Although expression of the BC1-2 gene has not been demonstrated, the conserved upstream region –30 to –23 suggests the transcriptional potential of this gene. The finding that three out of four ID

FIG. 5. Evolutionary comparison of the BC1 RNA gene and ID repeats. An approximate evolutionary tree for the rodents is shown with the nucleotides at the diagnostic positions (positions 5, 51, 67, 72, 74, and 76) for the BC1 RNA gene and the ID repeats in the respective species shown. The rat has five sequences shown for ID, the first two are two variant sequences of type 1 and the other three are type 2–4 sequences. Two BC1 RNA gene sequences are shown for hamster, with the upper one representing Chinese hamster and the lower one representing Syrian hamster. Both the BC1 genes and the subfamilies of ID that they are apparently responsible for are shown. Question marks represent positions that could not be assigned definitively. The BC1 sequences to the lower left represent the most likely BC1 sequence at those particular points in rodent evolution, based on the sequences in the modern-day rodents. Position 67 is boxed, as is position 5 in rat type 4 ID elements, because these are the positions in which evolution of the BC1 gene is inconsistent with that gene being a master for the rat ID types 2–4.

elements are closer in sequence to the BC1-2 gene suggests the possibility that in the germ line, the BC1-2 gene is preferentially transcribed, despite the observation that 10 out of 10 guinea pig brain BC1 RNA cDNAs have been derived from the BC1-1 gene rather than the BC1-2 gene (13).

Although guinea pig has two BC1 genes, the amplification capability (100–200 copies) is much lower than that seen for ID elements derived from the other rodent BC1 genes (1000–10,000 copies). There are several possible explanations for this variable amplification rate. The concentration of BC1 RNA in germ-line cells of rat and mouse may be higher than that of guinea pig and hamster, which could result in the different copy number within the various rodent genomes. Additionally, factors such as the sequence difference in the A-rich region or the 3' end of the transcript could also contribute. Compared with other rodent BC1 RNAs, the A-rich region of guinea pig BC1 genes is shorter and more punctuated with other bases, possibly making the A-rich region a less effective template for self-priming by the U residues present at the 3' end of BC1 RNA.

We have classified rat ID elements into four subfamilies. The type 1 subfamily appears to be directly derived from the BC1 gene, whereas the other three subfamilies appear to be derived from another master gene(s). The comparable copy numbers of rat type 1 and mouse ID elements indicate that amplifications derived from these two rodent BC1 genes have occurred at a similar rate. This is consistent with the levels of nucleotide divergence between rat type 1 and mouse ID elements. The ID sequences of types 2–4 have changes that differ from the current BC1 sequence, as well as from predictions of previous BC1 gene sequences as estimated by sequences of other rodent BC1 genes (Fig. 5). Therefore, it seems that the large accumulation of ID elements in the rat genome is likely to be the result of one or more new master gene(s) for these new subfamilies, formed after the divergence of rat and mouse.

One feature of the ID repeat that differs somewhat from most other SINEs is that the 3' end of the repeat does not consist strictly of variations of length and sequence in the A-rich region but, instead, has very specific sequences embedded within it. The GAACC motif found in most of the rat subfamilies is one such example (Fig. 4). However, the guinea pig ID elements (Fig. 2) and one mouse ID example (pMP11, Fig. 3) also have other sequence motifs that correspond well with sequences present in their respective BC1 transcripts. It has been suggested that the A-rich 3' region of SINEs may arise through the polyadenylation of SINE transcripts, rather than from coding by a gene (26). The 3' ends of ID elements suggest that for at least a significant portion of the ID repeats, the A-rich 3' region is part of the primary structure of the master gene(s) (27).

Our rat ID data strongly support the presence of at least a second master gene. Therefore, it is clear that new master genes can form for SINEs, although this appears to be an extremely rare process. When a retroposition event occurs, the duplicate includes only the repetitive DNA portion of the gene and does not include flanking sequences. It is likely that these sequences play an important role in the activity of a retroposon, in terms of both transcription and the reverse transcription process. Any new master gene(s) formed by retroposition would have to integrate into a genomic region that provided favorable sequences for these processes (reviewed in ref. 8). On the other hand, duplication of a gene

region through a DNA-mediated mechanism would duplicate the entire flanking region and would be likely to create an active gene. The finding of a second BC1 gene in the guinea pig genome that is apparently an active master gene made through a DNA amplification event demonstrates that DNA duplication events are a mechanism for generating new master genes. Once a DNA duplication event occurs, one copy might be free to mutate and change its expression pattern, potentially resulting in more active germ-line transcription and retroposition.

We thank Xuemin Xie for preliminary screening of the guinea pig genomic library. We thank David Kass, Susan Morrow, Mark Batzer, and Tamim Shaikh for critical comments during the studies and on the manuscript. This work was supported by Public Health Service Grants HG00340 and GM45668 to P.D. and MH38819 to J.B.

1. Deininger, P. L. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 619–636.
2. Daniels, G. R. & Deininger, P. L. (1985) *Nature (London)* **317**, 819–822.
3. Sutcliffe, J. G., Milner, R. J., Gottesfeld, J. M. & Lerner, R. A. (1984) *Nature (London)* **308**, 237–241.
4. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55**, 631–661.
5. Sapienza, C. & St-Jacques, B. (1986) *Nature (London)* **319**, 418–420.
6. Deininger, P. L. & Daniels, G. R. (1986) *Trends Genet.* **2**, 76–80.
7. Deininger, P. L., Batzer, M. A., Hutchison, C. A., III & Edgell, M. H. (1992) *Trends Genet.* **8**, 307–311.
8. Schmid, C. W. & Maraia, R. (1992) *Curr. Opin. Genet. Dev.* **2**, 874–882.
9. Daniels, G. R. & Deininger, P. (1991) *Nucleic Acids Res.* **19**, 1649–1656.
10. Liu, W. M., Leeftang, E. P. & Schmid, C. W. (1992) *Biochim. Biophys. Acta* **1132**, 306–308.
11. McKinnon, R. D., Danielson, P., Brow, M. A. D., Bloom, F. E. & Sutcliffe, J. G. (1987) *Mol. Cell. Biol.* **7**, 2148–2154.
12. DeChiara, T. M. & Brosius, J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2624–2628.
13. Martignetti, J. & Brosius, J. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 9698–9702.
14. Brosius, J. & Gould, S. J. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10706–10710.
15. Brosius, J. (1991) *Science* **251**, 753.
16. Anzai, K., Kobayashi, S., Suehiro, Y. & Goto, S. (1987) *Mol. Brain Res.* **2**, 43–49.
17. Dahlberg, J. E. & Lund, E. (1991) *Science* **254**, 1462–1463.
18. Willis, I. M. (1993) *Eur. J. Biochem.* **212**, 1–11.
19. Carbon, P. & Krol, A. (1991) *EMBO J.* **10**, 599–606.
20. Labuda, D., Sinnett, D., Richer, C., Deragon, J. M. & Striker, G. (1991) *J. Mol. Evol.* **32**, 405–414.
21. Labuda, D. & Striker, G. (1989) *Nucleic Acids Res.* **17**, 2477–2491.
22. Sakamoto, K. & Okada, N. (1985) *Nucleic Acids Res.* **13**, 7195–7206.
23. Li, W.-H., Gojobori, T. & Nei, M. (1981) *Nature (London)* **292**, 237–239.
24. Economou, E. P., Bergen, A. W., Warren, A. C. & Antonarakis, S. E. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2951–2954.
25. Deininger, P. & Batzer, M. A. (1993) *Evol. Biol.* **27**, 157–196.
26. Matera, A. G., Hellman, U., Hintz, M. F. & Schmid, C. W. (1990) *Nucleic Acids Res.* **18**, 6019–6023.
27. Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Desselle, T. D., Hoppens, C. L. & Deininger, P. L. (1990) *Nucleic Acids Res.* **18**, 6793–6798.