Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 599–601, January 1995
Genetics

Active *gypsy*/Ty3 retrotransposons or retroviruses in *Caenorhabditis elegans*

(nematode/transcription/reverse transcription/mobile elements)

ROY J. BRITTEN

Division of Biology, California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625

Contributed by Roy J. Britten, September 30, 1994

ABSTRACT **A *gypsy*/Ty3-class retrotransposon (*Cer1*) is integrated in the DNA of *Caenorhabditis elegans* chromosome III. It is 8865 nt in length and has 492-nt long terminal repeats that are identical in DNA sequence. There is an exceptionally long (6819 nt) open reading frame uninterrupted by frameshift mutations in the period since the insertion, which must therefore have been rather recent. Alignment with other *gypsy*-class elements and with retroviruses indicates that an *env* gene occupies the 3′ 1.2 kb of the open reading frame. A search through GenBank has uncovered two additional *gypsy*-class elements from *C. elegans* that are very closely related in DNA sequence to this insert and are transcribed. Since *gypsy* of *Drosophila* has been shown to be an infectious element, it is possible that retrovirus-like *gypsy* elements are active in *C. elegans*.**

The sequence of 2.2 megabases of *Caenorhabditis elegans* was recently published (1). Within that sequence there is a *gypsy*-class mobile element inserted at about position 1140 in figure 1 of that reference (1). The existence of a reverse transcriptase gene was recognized, and the gene was described in GenBank entry CELF44E2 as being related to a yeast reverse transcriptase gene. However, the element as a whole was not noticed, probably because the long terminal repeats (LTRs) occur on different clones. Comparison of the two clones showed a pair of 492-nt LTRs that match perfectly in sequence and a duplicated insertion site 3 nt long.

This *gypsy*-class element is likely a very recent insert in the *C. elegans* genome because of the sequence identity of the LTRs and because of the presence of an unusually long open reading frame of 6819 nt, since with time such inserts are subject to base substitutions, insertions, and deletions that cause frameshift mutations and introduce stop codons. The element is 8865 nt long and the nontranslatable sequence amounts to 723 nt just downstream from the 5′ LTR and 337 nt just upstream from the 3′ LTR. Such untranslatable regions are typical of active *gypsy*-like mobile elements. The similarity of the reverse transcriptase protein sequence and the gene organization to that of the original *gypsy* of *Drosophila melanogaster* and to many other *gypsy* elements identifies the *C. elegans* element as a member of the *gypsy*/Ty3-class of mobile elements. Following the practice for naming such elements, the element is named *Cer1* (the first letters of the genus and species names, *r* for retrotransposon, and a number for the first example in that species).

Table 1 shows that the *gypsy*/Ty3 class forms a coherent group on the basis of the reverse transcriptase amino acid sequence and that the *Cer1* element is in this group. All of the elements on this list above the cauliflower mosaic virus are recognized members of the *gypsy*/Ty3 class. Alignment was determined with CLUSTALV (2), and all subsets of elements with reverse transcriptase genes that are closely related to each other are represented by a single example, always the one most closely related to the *C. elegans* sequence. For example, all of the retroviruses are represented by a cat endogenous retrovirus related to leukemia viruses. Many of the viral reverse transcriptase sequences are more distant, though some foamy virus sequences are not much more distant than the cat endogenous retrovirus sequence.

Fig. 1 shows an alignment of *Cer1* with *Drosophila gypsy* and other elements to identify the coding regions and to show the similarity of gene organization. *Drosophila gypsy* has recently been shown to act as a retrovirus (3, 4). The feature that suggests that *Cer1* is a potential retrovirus is the probable *env* gene of about 1.2 kb. There is even a short region of recognizable amino acid sequence identity (18% in 76 amino acids) near the 5′ end of the *Cer1* and *Drosophila gypsy env* genes, which is remarkable, since, in general, *env* genes evolve rapidly, and the evolutionary distance between nematodes and *Drosophila* is large. To check the alignment of sequences in Fig. 1, the amino acid sequence resulting from translation of the open reading frame of the *Cer1* DNA sequence was divided into 200-amino acid segments, and the DNA sequences of GenBank were searched with each segment. Table 2 lists all segments that showed significant relationship to sequences in GenBank, the elements that gave the highest score, and the probability of their occurrence by chance according to the TBLASTN (5) program.

A search with the 5′ end of the open reading frame turned up two *C. elegans*-expressed sequence tags (GenBank CELK022H1F and CELK010B3F, Y. Kohara, H. Mitsuki, A. Nishigaki, T. Motohashi, A. Sugimoto, and H. Tabara) that had recently been listed. These short sequences are the approximate 5′ end of transcripts present in a population of the nematodes at various stages of growth and development. These transcripts (identified here as Cer2x and Cer3x) match the sequence of *Cer1* almost perfectly over their full length. They match segments of sequence near the beginning of the open reading frame of *Cer1* (at nucleotide 1219) as follows, with all numbers representing nucleotide positions in the *Cer1* DNA sequence. Cer2x matches perfectly from 1231 to 1374, from 1588 to 1704 with one mismatch, and perfectly from 1750 to 1846. Thus, Cer2x is probably transcribed from a close relative of *Cer1* that has two deletions (1375–1587 and 1705–1749). Cer3x starts 117 nt from the beginning of the open reading frame at 1335 and matches *Cer1* perfectly to 1374. The next match is from 1588 to 1704, with the same single mismatch as Cer2x, and then Cer3x matches *Cer1* perfectly from 1783 to 1881. The match between Cer3x and *Cer1* is less perfect from 1881 to 2014, differing by seven insertions of single bases in Cer3x. It thus appears that Cer3x is a transcript from another close relative of *Cer1* that shares the same first deletion (1375–1587) and the beginning of the second deletion (1705), but this deletion is larger, extending to 1782. The main parts of these comparisons are straightforward, and the elements from which they are transcribed are identified as *Cer2* and

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: LTR, long terminal repeat.

Table 1. Number of amino acid sequence differences for the most conserved 400-amino acid region of the reverse transcriptase gene

```
 CELPAR3  C. elegans plasmid pAR3 (1)
  |
62 -BMMAG  Bombyx mori retrotransposon mag
   |
63 63 -DMIS297  D. melanogaster 297
      |
63 65 60 -CFT1RTPOS  Cladosporum fulvum CfT-1 LTR-retrotransposon
         |
66 67 61 65 -YSCTY31A  Saccharomyces cerevisiae retrotransposon Ty3-1
            |
66 64 40 60 61 -TRNTED  Autographa californica TED
               |
66 67 53 65 64 56 -DROGYPF1A  D. melanogaster gypsy
                  |
65 66 62 60 63 63 65 -LHDEL  Lilium henryi del transposon
                     |
66 67 60 67 64 59 65 66 -DMRTMGD1  D. melanogaster mgd1 retrotransposon
                        |
68 67 65 57 62 64 69 65 70 -YSPRTPTFL  Saccharomyces pombe retrotransposon Tf1
                           |
67 64 66 66 68 68 70 69 68 68 -SUTRVLEP  Tripneustes gratilla Tgr1
                              |
68 70 66 67 66 63 68 67 69 66 71 -DMMC3DM2  Drosophila microcopia Dm-2
                                 |
68 68 62 51 64 63 65 61 68 58 73 65 -FSOGAGPOL  Fusarium oxysporum
                                    |
69 70 65 69 67 65 72 69 72 67 72 61 68 -DMBLPP  D. melanogaster
                                       |
70 70 64 70 70 67 71 68 68 69 75 70 68 73 -DVULYSS  Drosophila virilis Ulysses
                                          |
70 74 69 74 67 70 73 75 72 71 75 73 72 73 76 -MCAMADCRI  Cauliflower mosaic virus
                                             |
72 76 78 76 79 77 76 77 80 78 80 74 76 76 78 79 -FSECE1  Cat ECE1 endogenous virus
```

The order has been set by the divergence from *Cer1*, the topmost entry. Shown are representatives of the known groups of *gypsy*/Ty3-class retrotransposons and two viruses, represented in each case by the member of the group that is most closely related to *Cer1* in this region. The GenBank names are shown as well as the species that harbor the elements.

Cer3, but longer and completely accurate sequences may change some details, such as the mismatches at the extreme 3′ end of Cer3x. The deletions in *Cer2* and *Cer3* compared to *Cer1* do not interrupt the open reading frame, but not much else can be said because the function of putative genes in the very 5′ region of *Cer1* is unknown. Usually the 5′-most region of the

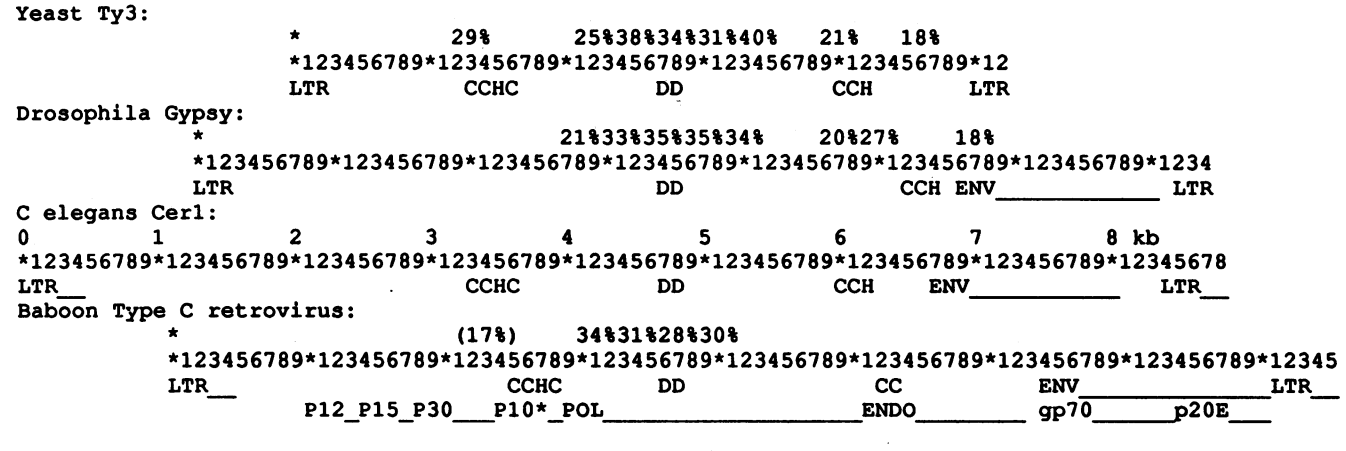

FIG. 1. Alignment of reverse transcriptase-dependent LTR elements. The top three sequences are members of the *gypsy*/Ty3 class and the bottom sequence is a mammalian retrovirus. For each sequence the middle line is made up of digits as a distance marker, 100 nt per character. The bottom line symbolizes regions that are identifiable for alignment. CCHC is a zinc finger nucleic acid binding site of which *Cer1* has three, one similar to retroviral CCHC; DD represents the absolutely conserved Asp–Asp near the center of the reverse transcriptase gene, and the elements have been aligned to it for the figure; CCH or CC represents a nucleic acid binding site in the endonuclease/integrase gene; ENV represents the envelope gene, which is absent or much reduced in the yeast element. The top line shows the percent amino acid sequence identity between the three other elements and *Cer1*. In all cases the absence of a percentage means that there was no significant sequence relationship recognizable by FASTA for a 100-amino acid probe. In all but one case, the alignment is just that shown in the figure, and the exception (17%) in parentheses is between position about 2200 in the baboon virus and position about 1700 in the probable *gag* gene of *Cer1*. The bottom-most line encodes gene identifications for baboon C-type virus. GenBank locus names are YSTCY31A, DROGYPFIA, CELPAR3+CELF44E2, and BABGPE, respectively, and references are available there.

Table 2. Significant similarities between 200-amino acid segments of the *Cer1* open reading frame and other GenBank sequences

Nt position	Source or type of similarity (GenBank nos.)	*P*
1222	Open reading frame starts (ATG)	
2900–3119	Clusters of Asn show homology to DNA binding proteins	
3120–3300	Three $CX_2CX_5HX_3C$ zinc finger sequences are recognizable	
4141–4740	*Saccharomyces cerevisiae* retrotransposon Ty*3* (YSCTY31A)	$7.6e^{-38}$
4741–5340	*Bombyx mori* retrotransposon *mag* (BMMAG)	$7.8e^{-32}$
5341–5940	*Autographa californica* (TRNTED)	$2.4e^{-12}$
5941–6540	*Drosophila melanogaster 412*-like (DRO8DC5Z)	$1.8e^{-14}$
8037	End of open reading frame (first stop codon)	

The translation of the open reading frame was divided into 200-amino acid segments except for the terminal segments, which were 160 and 99 amino acids. All regions not listed above showed no statistically significant sequence similarities to elements from other species in a search with TBLASTN on the complete nonredundant bank of DNA sequences. The number on the left is the position in the DNA sequence starting at the 5′ LTR. The numbers on the right are the Poisson probability of the relationship's occurring by chance as calculated by TBLASTN (5). The upper two regions are clearly recognizable, but it is not practical to calculate their probabilities of occurrence by chance.

open reading frames of *gypsy* elements and retroviruses is identified as the *gag* gene. However, it is clear from Fig. 1 that the excess length of *Cer1* compared with other *gypsy* elements or retroviruses is at the 5′ end, and there might be other genes in this region of *Cer1*, which would be the genes suffering the observed deletions.

The conclusion from the expressed sequences Cer2x and Cer3x and the presence of *Cer1* integrated into chromosome III is that *C. elegans* contains at least three *gypsy* elements, two of which are transcribed at their 5′ ends. Judging from these short transcripts, the transcribed elements obviously need not be full length, but it is a good guess that they are, since they are transcribed from the 5′ part of the open reading frame and match *Cer1* so accurately in sequence. Since *Cer1* is very likely a recent insert in the *C. elegans* genome, it is either an active element or the recent product of an active element. The two transcribed sequences may also be the product of active elements. The likelihood that *Cer1* is an infectious retrovirus in nematodes depends for the moment on the evidence that it contains an ≈1.2-kb *env* gene that aligns with other *env* genes.

The observation of *gypsy* elements in *C. elegans* DNA is another example of the wide range of occurrence of *gypsy* sequences. We have recently observed examples in the tunicate *Ciona intestinalis* and the herring *Clupea pallasi* (6). The list of organisms containing *gypsy* sequences, including these examples and those in Table 1, now includes deuterostomes (echinoderms, tunicate, bony fish), protostomes (many insects, but most protostomes have not been tested), six or more fungi, two plants (lily and pine tree), and nematode. Missing from this list are all of the vertebrates above fish. Many of these higher vertebrates harbor retroviruses, and the mobile elements most closely related to retroviruses are the *gypsy* elements (6). Only among fish do both *gypsy* elements and retroviruses occur, but both have yet to be found in the same individual or species. It appears that at some level among the vertebrates, *gypsy* elements evolved into retroviruses, and, for some reason, the parental *gypsy* lineage did not survive. The broad distribution of *gypsy* sequences among living species could be due, in part, to horizontal transfer, but there is no direct evidence of transfer among widely divergent groups. The large divergence shown in Table 1 is consistent with vertical transfer whether the rate of evolution of the *gypsy* DNAs was equal to or faster than that of their host's genomes. Whether *gypsies* and their descendants are of universal occurrence is a fascinating question for the future.

This work was supported by National Institutes of Health grants.

1. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al.* (1994) *Nature (London)* **368,** 32–38.
2. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8,** 189–191.
3. Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Prud'homme, N. & Bucheton, A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1285–1289.
4. Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. & Corces, V. G. (1994) *Genes Dev.* **8,** 2046–2057.
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
6. Britten, R. J., McCormack, T. J., Mears, T. L. & Davidson, E. H. (1995) *J. Mol. Evol.* **40,** 13–24.