Transposable element Tc1 of Caenorhabditis elegans recognizes specific target sequences for integration

SITE SPECIFICITY/TRANSPOSON/UNC-22/NEMATODE

TRANSPUSON

Mori, Ikue; Benian, Guy M.; Moerman, Donald G.; Waterston, Robert H.

Communicated by Donald P. Shreffler, October 12, 1987

ABSTRACT The frequency of movement of Tc1, a 1.6-kilobase transposable element in the nematode Caenorhabditis elegans, is under genetic control, and Tc1 insertion sites are widely but nonrandomly distributed. The unusually high frequency of insertions at multiple sites in the gene unc-22 suggested that this gene might be particularly rich in preferred target sites. To discover the features of Tc1 target sites, we have sequenced the sites of seven independent Tc1 transpositions into unc-22 and three other sites. Our comparison of these and two other sites from the literature indicates that in all cases Tc1 integrates at the dinucleotide T-A when it is flanked both 5' and 3' by particular preferred nucleotides. Our analysis revealed the following consensus target for Tc1 integration: G-A-K-A-T-T-G-T, in which K = G or T. This target site sequence specificity has implications both for the mechanism of Tc1 transposition and the use of Tc1 in cloning genes by transposon-tagging.

A significant fraction of eukaryotic genomes consists of repetitive sequence families. Very often these repetitive sequences have structural features common to transposable elements, which have the ability under appropriate conditions to move in the genome and thereby to cause mutations. The insertion-site specificity is known for several prokaryotic elements (1-4) and has been inferred for some eukaryotic elements (5-8). The features of the insertion site have implications for the mechanism of transposition and possibly for genome evolution of the organisms that harbor the elements. More practically, insertion-site specificity may affect the usefulness of the transposon for cloning genes by insertional mutagenesis (9, 10).

In the nematode species Caenorhabditis elegans, the Tc1 transposable element was initially identified as a repetitive sequence (11-14). Sequencing of one member showed that the element is 1.6 kilobases (kb) in length, has 54-base-pair (bp) perfect inverted repeats at its ends, and has a large open reading frame (11). This element is present from 30 copies to >300 copies in various strains (13, 14). More recently, a strain of C. elegans called Bergerac or BO was found to exhibit a high frequency of spontaneous unstable mutations (15) caused by the insertion of Tc1 (9, 10, 16). The frequency of insertional mutagenesis varies among genes. One of the most frequently targeted genes is unc-22, a muscle-affecting gene (9, 15), where the frequency of insertion is about 10^-4.

The Tc1 insertion sites of 15 unc-22 mutations are scattered into at least 12 different sites, which span a 30-kb stretch of the unc-22 region (17). On the other hand, the mutation frequency of unc-54, a gene encoding the major myosin heavy chain, is <10^-5 in the same or related backgrounds (16). The difference in gene size alone (about 3-fold) does not explain the difference in mutation frequency.

One plausible explanation for the variation of Tc1 insertional mutagenesis among genes is that Tc1 has a preferred target sequence for insertion and the number of target sequences varies among genes. The wide distribution of Tc1 insertion sites in unc-22 suggests that the site specificity for insertion may not be extremely rigorous. The unc-22 gene may also contain more such target sequences than other genes. To test these hypotheses, we have sequenced Tc1 insertion sites of 7 unc-22 mutations and 3 other sites unrelated to unc-22. The comparison between these 10 sites and 2 other sites clearly demonstrates that Tc1 has a relaxed target site specificity, which probably controls in part the transpositional behavior of Tc1 in the Bergerac (BO) background.

MATERIALS AND METHODS

Cloning and Sequencing of the Tc1 Insertion Sites. Nematode strains containing unc-22::Tc1 insertions (st139::Tc1, st140::Tc1, st141::Tc1, and st185::Tc1) were grown for DNA extraction. Total genomic DNAs were purified by the standard methods (18, 19). Genomic DNAs were digested with Sst I and fractionated on agarose gels. The fragments including Tc1-inserted unc-22 fragments were recovered from the gel and ligated into phage A2001 (20). These partial genomic libraries were screened by plaque hybridization (21) with a nick-translated (22) Tc1 probe as well as appropriate unc-22 probes, and the desired Sst I fragments containing unc-22::Tc1 insertions were recloned into pUC19 (23). A phage clone containing the unc-22(st192::Tc1) insertion was provided by J. Kiff in our laboratory. The Sst I fragments containing unc-22(st2136::Tc1) and unc-22(st1376::Tc1) (9), and the HindIII fragment containing Tc1 insertion for stPI, an unc-22-linked dimorphism due to a recent transposition (D.G.M., unpublished data), were previously cloned into pBR322 (24). A clone containing the lin-12 (n137e1979::Tc1) insertion site was provided by I. Greenwald (10). A clone containing the Tc1 insertion in unc-15 was provided by S. Rioux of this laboratory, and the sequence was determined by him as part of these studies. To sequence from the Tc1 into the flanking unique sequence, a fragment containing only one Tc1 inverted repeat was subcloned into phage vectors M13mp18 or M13mp19 (23) in an orientation such that a synthetic oligonucleotide complementary to the Tc1 inverted repeat could be used as a primer in dideoxy chain-termination DNA sequencing (25). The M13 universal primer (Amersham) was also used in some cases. Sequences were analyzed by using the Beckman Microgenie program (Beckman).

Computer Analysis of the Insertion Sequences. We derived a simple computer program that evaluated all possible strand...
RESULTS AND DISCUSSION

To begin our studies of Tc1 insertion, we analyzed 7 independent unc-22 mutations (Fig. 1). The stl41::Tc1 and stl92::Tc1 alleles, though isolated independently, have Tc1 inserted at precisely the same site and, as determined by restriction mapping, in the same orientation. Besides these 7 unc-22 alleles, we also sequenced an insert in lin-12 (10), an unc-22-linked dimorphism, stp1 (D.G.M., unpublished data), and an insertion into unc-15, the structural gene for paramyosin (ref. 26; H. Kagawa, personal communication). The sequences surrounding these 10 Tc1 insertions along with two sites sequenced by Rosenzweig et al. (11, 12) are shown in Fig. 2 Left.

We have searched these sequences for common features. At each of the 12 sites, Tc1 is inserted at a T-A dinucleotide. As indicated previously (11, 12), this T-A may be duplicated upon insertion and is presumably an obligate part of the insertion site. When the flanking sequences were analyzed, certain preferred bases at the same positions relative to the Tc1 insert were obvious. However, the palindromic nature of the T-A dinucleotide, and in some cases the 5′ and 3′ flanking sequences (e.g., A-T-G-T-A-C-A-T (stl36::Tc1) and A-T-A-T (stl92::Tc1)) introduced an ambiguity into our analysis—namely, that the entire sequence might be compared with the other sites in deriving the best consensus sequence from our data set. Therefore, we carried out a computer analysis of all possible 2048 strand comparisons for the region −5 to +5 relative to the Tc1 insert (see Materials and Methods). The configuration of strands in Fig. 2 Left reveals preferred bases at both 5′ and 3′ of the T-A dinucleotide, and this configuration gives the highest score (824) of all combinations of strands. The derived consensus is given in Fig. 2 Right. It should be pointed out that the lowest score among 2048 comparisons of our data set was 752, which was still significantly better than the scores from 2000 sets of 12 computer-generated 10-bp sequences, even when a T-A dinucleotide was fixed at the −1 and +1 positions and the A+T bias of the nematode genome (64%; ref. 18) was used for the rest of the positions (mean ± SD = 674.75 ± 19.63; P < 0.001). This confirms in fact both the palindromic nature of the sites and the importance of the positions flanking the T-A dinucleotide determining the insertion sites.

The consensus sequence has several interesting features. It is both A+T-rich and palindromic. The exceptional positions to both these trends are at −5, −3, and +3. Given our sample size, we cannot be certain if these positions are an important part of the site, and, if so, that our analysis has revealed the truly preferred base at these exceptional positions. Nevertheless, the palindromic nature of the site is striking and is seen strongly in 8 of the 12 sequences, in which either 6 or 8 of 10 positions are matched as inverted repeats (Fig. 2 Left, lines 1, 3, 5, 7, 8, and 10–12). Even in the remaining four sequences, 4 of 10 positions have palindromic-forming bases. Following the analogy drawn for transposon Tnl0 (1), target recognition by Tc1 can be related conceptually to a type II restriction enzyme, which cleaves symmetrically or asymmetrically about the middle of a palindromic sequence. Most likely Tc1 duplicates only the central T-A dinucleotide upon insertion, although even this duplication remains in doubt because of ambiguity about the precise ends of the Tc1 element (11, 12). Thus, Tc1 does not duplicate the entire palindromic target, unlike the other elements with inverted terminal repeats that recognize specific target sequences (Tnl0, insertion sequence IS5, and P element; refs. 1, 2, and 5, respectively). In any case, a consequence of Tc1 insertion is that Tc1 destroys the consensus site upon insertion, thereby eliminating a target in the genome, rather than creating a second one.

Another interesting feature of the consensus sequence is the alternation of purines and pyrimidines at seven of eight positions, from −4 to +4 (Fig. 2 Right). This alternating motif is not an absolute requirement for insertion; however, some sites show a perfect alternation throughout the region [e.g., stl36::Tc1 and lin-12(nl37e1979::Tc1)], but others do not (e.g., stl40::Tc1). The alternating purine/pyrimidine tracts are intriguing because such tracts are prone to form the alternative left-handed, or Z-DNA, conformation (for review, see ref. 27). It is not clear whether the purine/pyrimidine stretch of sequence around the Tc1 insertion is sufficient to permit B-to-Z transition. However, it is interesting to note that, like Tc1, the Drosophila retrotransposons gypsy and 17.6 elements strongly prefer sites consisting of alternating purines and pyrimidines (6, 7).

Our conclusion that Tc1 has a preferred target site is buttressed by the sequence of Tc1 insertions into the unc-54 gene by D. Eide and P. Anderson (personal communication). Of 11 insertions analyzed, 7 are at precisely the same site in the unc-54 gene. By using our consensus matrix shown in Fig. 2 Right, the score given for this site from position −5 to position +4 is the highest of any of the four unc-54 integration sites that they identified. We then searched the entire 6-kb unc-54 coding sequence (28) for the potential insertion sites using our consensus matrix. The

Fig. 1. Restriction map of the seven Tc1 insertion sites in unc-22 sequenced in this study. The orientation of the Tc1 insertions (shown by arrows above the triangles) was determined by either probing RNA transfer blots with an SP6-Tc1 probe (17) or by determining the location of an asymmetric Cla I site within Tc1 (ref. 9; this study). The sequences span ~30 kb in the unc-22 region. The insertion sites of stl36::Tc1 and stl37::Tc1 are 55 bp apart; those, as well as the sites of stl40::Tc1 and stl92::Tc1, are in open reading frames of unc-22 (data not shown). The site of stl139::Tc1 is at or near the 5′ end of unc-22, and the site of stl185::Tc1 is near the 3′ end of unc-22 (data not shown). A 3.4-kb Bgl II fragment sequenced (J. Kiff, D.G.M. and R.H.W., unpublished results) is shown as a bold line. The orientation and the minimal extent of the unc-22 transcript are also shown (data are from ref. 17).
stringency was set such that there should be T, A, and T at positions −1, +1, and +4, respectively, and the score for nine positions should be equal to or higher than the lowest score of our data set, which was obtained for the unc-22(st139::Tcl) site. We found nine sites that include both multiply targeted sites. Of these nine sites, the site with the highest score corresponds to the "hot" site with 7 insertions. Thus, Tcl clearly does have a preference in target site selection, and our consensus is indeed significant in predicting Tcl insertion sites.

Can the difference in frequency of Tcl insertion into different genes be simply explained by its target site preference? Although we do not have the entire unc-22 sequence, we searched the 3.4 kb of the unc-22 coding sequence (J. Kiff, D.G.M. and R.H.W., unpublished results; Fig. 1) available for analysis. The stringency used for the search was the same as in the unc-54 search. This 3.4 kb of unc-22 has 16 matching sites, including the site for st141::Tcl and st192::Tcl as the site with the highest score among these 16 sites. Assuming that there are approximately 14 kb of unc-22 coding sequence (17) and that the remainder of unc-22 resembles the 3.4 kb fragment analyzed, unc-22 would contain about 7 times as many potential sites as unc-54, which could partially account for the greater than 10-fold difference in mutation frequency between the two genes (15, 16). At present we cannot say what additional bias might be introduced by the variations in match to the consensus sequence. One implication of our results is that genes with an abundance of the preferred target site may be readily tagged with Tcl, whereas other genes may be appreciably more difficult to tag.

We thank I. Greenwald for the lin-12::Tcl clone; S. Rioux, H. Kagawa, D. Eide and P. Anderson, and J. Kiff for allowing us to cite their unpublished results; and K. Kondo for advice on sequencing. We also thank D. Berg and D. Dykhuisen for their comments on this manuscript. This work was supported by a grant GM23883 from the National Institutes of Health to R.H.W.