Chromosoma (1994) 103:215–224

CHROMOSOMA
© Springer-Verlag 1994

Structure of the *Drosophila* HeT-A transposon: a retrotransposon-like element forming telomeres

Olga Danilevskaya[1,2], Franchot Slot[1], Maria Pavlova[2], Mary-Lou Pardue[1]

[1] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2] Institute of Molecular Genetics, Russian Academy of Sciences, 123182 Moscow, Russia

Received: 21 February 1994 / Accepted: 2 March 1994

Abstract. Telomeres of *Drosophila* appear to be very different from those of other organisms. A transposable element, HeT-A, plays a major role in forming telomeres and may be the sole structural element, since telomerase-generated repeats are not found. HeT-A transposes only to chromosome ends. It appears to be a retrotransposon but has novel structural features, which may be related to its telomere functions. A consensus sequence from cloned HeT-A elements defines an element of ~6 kb. The coding region has retrotransposon-like overlapping open reading frames (ORFs) with a −1 frameshift in a sequence resembling the frameshift region of the mammalian HIV-1 retrovirus. Both the HeT-A ORFs contain motifs suggesting RNA binding. HeT-A-specific features include a long non-coding region, 3′ of the ORFs, which makes up about half of the element. This region has a regular array of imperfect sequence repeats and ends with oligo(A), marking the end of the element and suggesting a polyadenylated RNA transposition intermediate. This 3′ repeat region may have a structural role in heterochromatin. The most distal part of each complete HeT-A on the chromosome, the region 5′ of the ORFs, has unusual conserved features, which might produce a terminal structure for the chromosome.

Introduction

It is surprising that *Drosophila melanogaster*, the organism from which the concept of the telomere originated (Muller 1938), appears to lack the short G-rich repeats found at the ends of chromosomes in other eukaryotes (Zakian 1989). Drosophila seems to have a novel solution for forming chromosome ends, a telomere-specific transposable element, HeT-A (Biessmann et al. 1992b).

Edited by: P.B. Moens

Correspondence to: M.L. Pardue

Gen bank accession numbers: Lambda T-A, U06947; 23Zn-1, U06920; 4568, U06921; 9D4; X68130

These elements are present in multiple copies on unbroken chromosome ends (Valgeirsdottir et al. 1990) and have transposed onto broken chromosome ends on several occasions (Biessmann et al. 1990b, 1992b; Traverse and Pardue 1988). A remarkable feature of HeT-A is that transposition is strictly limited to chromosome ends; none of the *D. melanogaster* stocks that have been studied have any evidence of HeT-A sequences in interior sites on the chromosome arms (Valgeirsdottir et al. 1990).

The initial characterization of the HeT-A element was based on sequence from two elements, each transposed to the broken end of a chromosome. (These were elements 394 and 473, transposed to the X chromosomes in the RT394 and RT473 stocks, respectively.) The sequences from the transposed elements were compared with sequence of a fragment (lambda T-A) containing four elements cloned from an established telomere (Biessmann et al. 1992b). The consensus sequence from this set of elements gave the first clues about the nature of HeT-A. The two broken chromosomes onto which HeT-A elements had transposed, are referred to as "healed" chromosomes: however, we do not know whether this healing is the complete reestablishment of a telomere or the first step in a more complex process. The ends of both the healed chromosomes undergo change (Biessmann et al. 1990b). Biessmann et al. (1992a) have recently reported following individual chromosomes from the two healed stocks for 17 generations. They found that ends receded at about the rate expected if the terminal RNA primer was not replaced during DNA replication. In addition, ~1% of the ends had new HeT-A elements added to the end of the existing elements in each generation. Similar processes seem to have been involved in forming established telomeres since three of the HeT-A elements in lambda T-A are truncated at the point at which they join the next element (Valgeirsdottir et al. 1990; Biessmann et al. 1992b). The multiple HeT-A elements on established telomeres, along with other sequences, form a complex region that is only beginning to be unraveled.

At the time of the original study it was known that the elements on the two healed chromosomes were not complete. However, since both elements were attached to the chromosome with the same polarity, the two clones contained equivalent parts of the two elements and the two sequences were very similar. Alignment of the sequences from the 394 and 473 elements gave a consensus sequence of >2 kb. When this consensus was compared with the sequence of lambda T-A, it was seen that three of the lambda T-A elements, T-A2, T-A3, and T-A4, were shorter versions of this sequence, each ending at the start of another copy of the consensus. One of the lambda T-A elements, T-A1, was at least as long as the consensus and could not be further defined since we had no other sequence for comparison at the time. To understand the nature of the *Drosophila* telomere it is necessary first to know the structure of the complete HeT-A element. We have sought other sequence comparisons to define the rest of the element and have now concluded that T-A1 is a complete element.

Our study of healed chromosomes showed clearly that HeT-A is a transposon. HeT-A attachment to the broken ends could not have involved homologous recombination since there was no sequence similarity between HeT-A and either broken chromosome end. Nor did the sequence of the broken end of RT394 have any similarity to the sequence at the break in RT473. Thus, there was no evidence that the DNA sequence at or near a break played any role in whether the break would be healed. In contrast to the differences in the chromosomal sequences, the sequences of the two transposed HeT-A elements were very similar, especially at the end where they joined the chromosome. Both elements were joined to the chromosome end by oligo (A): A_4 for element 394 and A_{11} for element 473.

The oligo(A) segments at the junction between an element and the chromosome suggest that HeT-A transposes through an RNA intermediate and therefore is a retrotransposon. The oligo(A) junction with the chromosome is one of the prime distinguishing features of a major class of retrotransposons, the poly(A)-type or non-LTR retroposons (see Boeke and Corces 1989; Xiong and Eickbush 1990). Studies of the mechanism of transposition of one poly(A)-type retrotransposon, the *Bombyx mori* R2Bm element, show that reverse transcription of the R2Bm RNA is primed by the 3′ hydroxyl of the nicked DNA at the insertion site (Luan et al. 1993). This mechanism provides an explanation for the observation that all elements of this class are joined to the chromosome by an oligo(A) since the insertion begins by reverse transcription of the 3′ end of the RNA. The mechanism also helps to explain a second marked feature of this class, the large numbers of elements that are variably truncated at the 5′ends. The 5′ end of the insertion would be determined by the point at which the reverse transcript joined the other end of the broken DNA. The variability of the 5′ ends suggests that the requirements for making the 5′ junction are not stringent. Our early study showed that HeT-A had two of the characteristics of the poly(A)-type retrotransposons: an oligo(A) junction with chromosomal DNA and variable 5′ truncation.

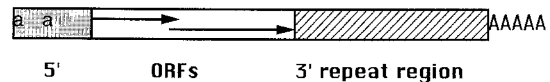

Fig. 1. Diagram of the consensus HeT-A element. The entire ~ 6 kb element is shown. The *gray box* indicates the 5′ structural region. Each *a* marks one of the clusters of A-rich repeats. The *white area* indicates the coding region. *Arrows* mark overlapping open reading frames (ORFs). *Diagonal stripes* mark the 3′ repeat region. *AAAAA* indicates the presumed poly(A) tail of the RNA transposition intermediate

However it still is not clear that truncation of HeT-A, which attaches to ends, occurs in the same way as truncation of elements that insert into chromosomes.

Retrotransposable elements, including RNA viruses, are thought to have arisen from an ancestor encoding both a *gag*-like nucleic acid-binding protein and reverse transcriptase (see Xiong and Eickbush 1990). Although, in theory, the functions of these proteins can be supplied in *trans*, the elements that have been studied have encoded their own proteins. The original sequence for HeT-A had no coding region. Therefore one important question to be asked by extending the sequence of the element was whether there was sequence coding for proteins typical of retrotransposons.

Our approach to determining the overall structure of HeT-A has been to align sequences from four cloned DNA fragments. The clones were from four different genomic libraries made from diverse *Drosophila melanogaster*, differing in location of origin of the library (Russia and the USA) and in the fly stock (Oregon R and Canton S). Three of the four clones had been sequenced solely because of their location on telomeres. The fourth clone (23Zn) was selected because it hybridized to a probe for the 3′ repeat region of HeT-A. We emphasize the diversity and the independent selection of these clones because our definition of the intact element is based on the similarity of their sequences and on the evidence that all overlap consistently and specifically to form a consensus.

The consensus derived from the four sequences appears to define the complete HeT-A element (Fig. 1). The element is approximately 6 kb and can be divided into three regions, a coding region with some characteristics typical of retrotransposons, and 5′ and 3′ regions, each with characteristics that seem to be unique to HeT-A. These 5′ and 3′ regions may be relevant to the role of HeT-A at the telomere.

Materials and methods

Clones. All clones were *D. melanogaster* DNA. Lambda T-A was from a lambdaphage library of Canton S DNA and identified by its telomere hybridization (Valgeirsdottir et al. 1990). Element T-A1 is nucleotides (nt) 2030-7358 of this sequence. 23Zn was isolated from a lambdaphage library of Oregon R DNA by hybridization with sequences of the HeT-A 3′ repeat (Danilevskaya et al. 1992). Element 23Zn-1 is the only fragment of the 23Zn clone that has been completely sequenced. Dm4568 was from a library of Oregon R DNA in the plasmid pBR322 and was selected for its telomere

hybridization (Danilevskaya et al. 1992). 9D4 was cloned from a recently healed telomere in the RT473 stock (Biessmann et al. 1992a).

Sequencing. Additions to published sequences were done as described by Danilevskaya et al. (1992).

Translation in rabbit reticulocyte lysates. A 4.5 kb EagI fragment of the 23Zn clone was inserted at the NotI site of the BluescriptII KS vector (Stratagene). Transcription in vitro with T7 polymerase (Promega) produced an RNA with a 235 nt leader. The HeT-A initiation codon is the first AUG in this RNA and is in a favorable context. The plasmid was opened with restriction enzymes at different sites to produce transcripts of different lengths, allowing the size of the translation product to be correlated with the length of coding region in each RNA. RNAs were translated in a rabbit reticulocyte lysate (Promega) according to the manufacturer's directions.

Production of HeT-A products in Escherichia coli. The open reading frames (ORFs) from clone 23Zn-1 (nt 1-3905) were inserted at the NcoI-BamHI site of the pET-15b vector (Novagene), yielding plasmid p23-ORF1+2, which should express the complete HeT-A ORF with no attached bacterial protein sequence. p23-ORF1 was derived from p23-ORF1+2 by a deletion removing the DNA between 23Zn-1 nt 1237 and the BamHI site in the vector. p23-ORF1 expressed ORF1. The ORFs were expressed in *E. coli* according to Novagene protocols.

Sequence analyses. Analyses were done with programs from the Wisconsin Genetics Computer Group (Devereux et al. 1984) and with the Multalin program (Corpet 1988).

Results

Definition of a complete HeT-A element and its structure

As defined by the aligned sequences, HeT-A is an element of ~ 6 kb with three distinct regions: (1) a 5′ structural region; (2) a protein coding region; and (3) a 3′ repeat region (Fig. 2). The characterization deduced from the sequence comparisons is supported by our discovery of an RNA transcript that contains sequence of all three regions and is approximately the same size as the two longest elements (O. Danilevskaya, F. Slot, K. Traverse, N. Wogan and M.L. Pardue, in preparation). Because each of the DNA sequences requires several small gaps to fit the alignment, and a gap could indicate either a deletion or an insertion. Until we know with certainty the reasons for the gaps we cannot determine a more exact length for the consensus DNA sequence. The RNA was sized only by gel migration. For single stranded RNA of this size, the nucleotide sequence could influence gel migration. Therefore, we can only give approximate sizes for both the consensus DNA and the RNA.

The structure of the HeT-A element derived from the consensus sequence is that expected of a poly(A)-type retrotransposon. The protein coding sequence is on the DNA strand that has the oligo(A) at the 3′ end, as it would be on a poly(A)$^+$ mRNA. In addition, all of the HeT-A RNA that we have detected is the equivalent of this strand.

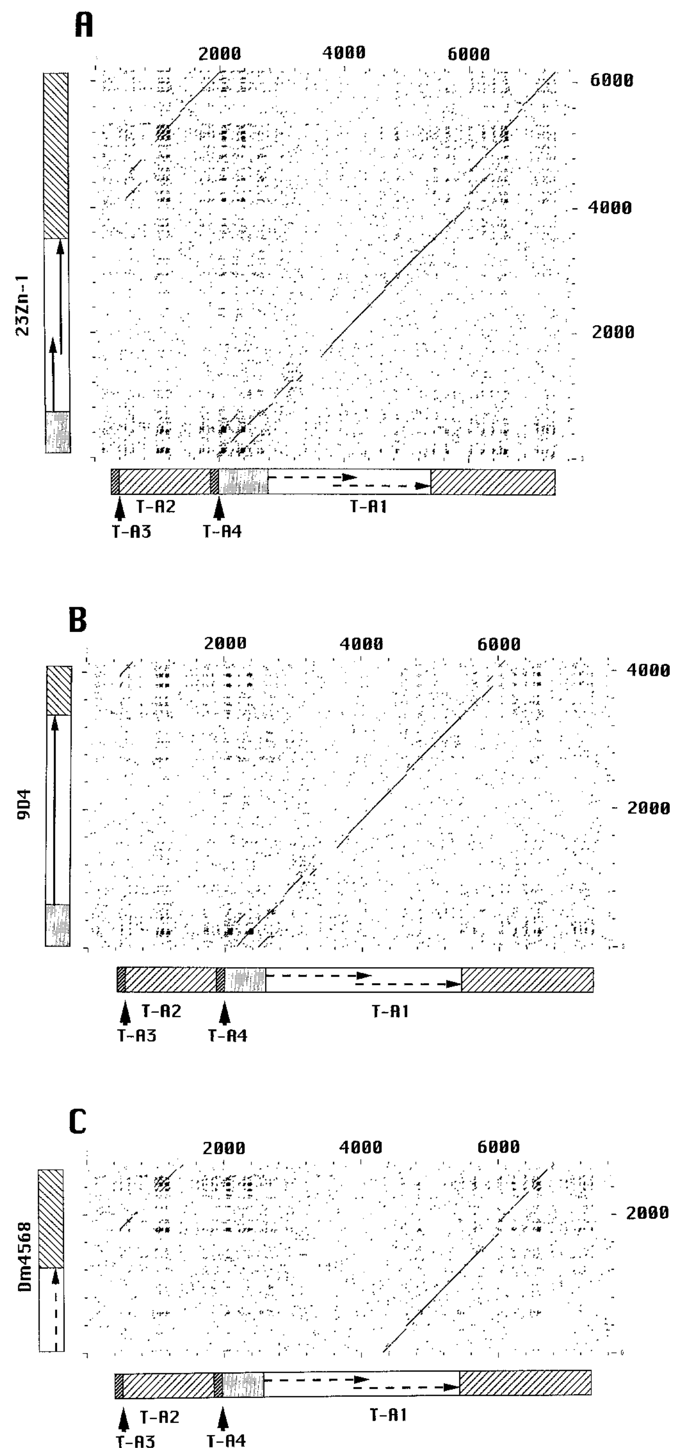

Fig. 2A–C. Dotplots comparing the DNA sequence of a fragment from an established telomere (lambda T-A, containing elements T-A1, T-A2, T-A3, and T-A4) with sequences from three other elements. The lambda T-A element is compared with: A 23Zn-1, B 9D4, C Dm4568. Axes depict HeT-A elements, showing the regions of each element (see Fig. 1 for explanation of symbols). *Arrows* indicate ORFs. *Dashed arrows* indicate pseudo ORFs. Strand shown is the DNA equivalent of an RNA transposition intermediate, shown 5′->3′. Comparisons were done with a window of 21 and a stringency of 14. Elements 23Zn-1 and T-A1 are believed to be complete elements. Dm4568 and 9D4 were truncated at both ends by cloning. T-A2, T-A3 and T-A4 are truncated by the attachment of another HeT-A element. The 3′ repeat region of T-A3 and T-A4 is shown *darkened* for visibility

Table 1. Pairwise sequence comparisons of HeT-A ORFs

Sequences compared	No. identical bases / No. bases compared	% Identity	No. gaps
Region 1–886[a]			
23Zn / 9D4	823 / 877	94	1
Region 887–1515			
23Zn / 9D4	612 / 630	97	1
23Zn / T-A1	512 / 596	86	6
9D4 / T-A1	513 / 596	86	7
Region 1516–2766			
23Zn / 9D4	1189 / 1249	95	0
23Zn / 4568	1060 / 1244	85	6
23Zn / T-A1	1029 / 1203	86	14
9D4 / 4568	1075 / 1245	86	6
9D4 / T-A1	1028 / 1199	86	16
T-A1 / 4568	1040 / 1198	87	15

Sequences in the alignment were compared by counting only bases aligned with a base in the other sequence being compared. % Identity represents the fraction of identical sequences out of the total number of sequences compared. No. gaps is the number of gaps on one or more bases introduced to give the alignment used for the calculations.

[a] Numbers used are based on the 23Zn-1 ORF sequence

The protein-coding region

The consensus sequence shows that HeT-A has a protein-coding region composed of two overlapping open reading frames (ORFs) just upstream of the 3′ repeat region (Fig. 1). The HeT-A ORFs were initially defined as an independent telomere element and named the T element (Danilevskaya et al. 1992). The sequence published in that original report was from the element that we have now completely sequenced and named 23Zn-1. Soon afterwards, Biessmann et al. (1992a), found a nearly identical ORF sequence as part of a second transposition onto the healed 394 chromosome. Part of the DNA of this second addition was cloned and named 9D4. The sequence alignments that we used in our attempt to define the complete HeT-A element included 23Zn-1 and 9D4, as well as lambda T-A, and pDm4568, a clone isolated on the basis of its telomere localization (Danilevskaya et al. 1992). All four sequences fit easily to a consensus, indicating that the T element is actually part of the HeT-A element, rather than an independent element. (Two of the elements, 9D4 and Dm4568 were truncated by cloning and are therefore incomplete at both ends).

Two of the elements in our alignment, 23Zn-1 and 9D4, have ORFs that appear to be functional and would produce nearly identical products. These ORFs are 95% identical with only two gaps (Table 1). In contrast, the equivalent regions in T-A1 and Dm4568 have multiple stop codons and thus are probably non-functional pseudo-ORFs. For the first 890 nucleotides (all numbering based on the 23Zn-1 sequence beginning with the initiation codon) the T-A1 ORF is so badly rearranged that the sequence cannot be fitted to the consensus. Beyond that point the T-A1 sequence shows much stronger conservation and can be aligned with the two functional ORFs to give >85% sequence identity with only a few gaps. Our clone of the Dm4568 sequence does not begin until nucleotide 1521 (of the 23Zn-1 sequence). Dm4568 shows approximately the same identity to the functional ORFs as T-A1 but requires fewer gaps for alignment.

It is interesting that the ORFs seem to fall into two groups, based on the degree of sequence conservation; the two apparently functional ORFs are 95% identical while the two pseudo ORFs show 85%–87% identity with each other as well as with both 23Zn-1 and 9D4. Thus, with the exception of the comparison between 23Zn-1 and 9D4, pairwise comparisons between the coding regions of these elements show approximately the same level of identity as was seen in the 3′ repeat region (Biessmann et al. 1992 b).

The two presumed functional ORFs would produce nearly identical products of 921 amino acids (23Zn-1) or 918 amino acids (9D4). The size difference is due to a region of three amino acids (Thr–Lys–Leu), which is either an insertion (in 23Zn-1) or a deletion (of 9D4) near the amino-terminus. Unfortunately, the three amino acids are in a part of the sequence where the T-A1 element is too diverged for comparison, and the Dm4568 sequence does not extend into this region. Comparison of the 23Zn-1 and 9D4 sequences suggests that the coding potential is being conserved; half of the nucleotide substitutions result in synonymous codons.

```
23Zn-1  C C C A A T A A A A A A . T T A T G C G C G A C
23Zn-2  * * * * * * * * * * * * . * * * * * * * * * * *
9D4     * * * * * * * * * * * * A * * * * * * * * * * *
T-A1    * * . * * * * * * * * * . * G * * * * * * * * *
17B3    * * * * * C * * * * G * . A * * * * T * * * * *
```

Fig. 3. Sequence of the region of the −1 frameshift from the ORFs of five HeT-A elements. The ORF from element 9D4 has one more nucleotide than any of the other sequences. Only a −1 frameshift in this region would permit the other elements to produce a polypeptide with strong sequence identity to that of 9D4. This small region contains several nucleotide changes in 17B3 and in the T-A1 pseudoORF. The 23Zn-2 sequence is from a partially sequenced element that lies upstream of 23Zn-1 on the cloned DNA. The 17B3 sequence is from unpublished work by H. Biessmann, B. Kasravi, T. Bui, G. Fujiwara, L. Champion, and J. Mason. *Asterisks* indicate sequence identity with the 23Zn-1 sequence. Nucleotide changes are indicated. *Dots* indicate gaps introduced for alignment

The only other significant difference between the two presumed functional ORFs is that the 23Zn-1 sequence requires a −1 frameshift after amino acid 382 to yield a full length peptide while the 9D4 sequence does not. (Therefore the ORF in 9D4 is equivalent to the overlapping ORF1+ORF2 in 23Zn-1.) The frameshift is conserved in the T-A1 sequence, in a second element (partially sequenced) on the 23Zn clone, and in one more element that has recently been sequenced by Biessmann, Mason, and collaborators (H. Biessmann, B. Kasravi, T. Bui, G. Fujiwara, L. Champion, and J. Mason, in preparation). The Dm4568 sequence does not extend into the frameshift region. A comparison of the five sequences available for this frameshift region shows several nucleotide changes within the small region (see Fig. 3) but only in the 9D4 sequence do the changes eliminate the need for the frameshift. It is interesting that a region of 25% proline occurs roughly 50 amino acids after the frameshift. This level of proline is reminiscent of the proline-rich activation domains in transcriptions factors (Mitchell and Tjian 1989). Such a region could also act as a hinge between two domains.

Overlapping ORFs are found frequently in retrotransposons and retroviruses but are rare in cellular genes. Thus this feature of HeT-A is retrotransposon-like. HeT-A shows a second resemblance to retro elements in the three zinc finger motifs found in ORF2. This zinc finger region has significant similarity to the *gag* proteins, which characterize retro elements (Danilevskaya et al. 1992; Biessmann et al. 1992a). We have not been able to find evidence that HeT-A encodes a reverse transcriptase, the other protein typically encoded by retro elements.

We have noted, in ORF1, a second conserved motif, $D(X)_4E(X)_3(G,C)XT(I,V)$, that may also have an RNA-binding function (Fig. 4). This motif has not been reported in retrotransposons but it is conserved in GTP-dependent elongation factors. It is interesting that the HeT-A ORF shares more than the minimal motif with the *Drosophila* elongation factor 2 (Grinblat et al. 1989). The motif does not seem to be part of the mechanisms for

```
23Zn-1     D E R K Q E E R P C T T I

9D4        D E R K Q E E R P C T T I

DmEF2      D t R K d E q e r C i T I


motif      D x x x x E x x x G x T I
                             C   V
```

Fig. 4. Comparison of sequences from elements 23Zn-1 and 9D4 with the GTP-binding elongation factor signature. This motif is thought to be involved in the interaction of elongation factors with either the ribosome or peptidyl tRNA. Its significance in HeT-A is unclear. Sequence from the *Drosophila* elongation factor 2 (Grinblat et al. 1989) is also shown (*DmEF2*). The motif at the *bottom* is the generalized consensus from the PROSITE dictionary (Bairoch 1993). The two HeT-A elements have identical amino acid sequences in this region and share more amino acids with DmEF2 than are fixed in the general consensus

GTP binding. Instead, it has been suggested that, in the elongation factors, this sequence might interact either with the ribosome or in the peptidyl tRNA. Its significance to HeT-A is unclear but the possibility of RNA binding is intriguing.

The sequence alignments offer strong evidence that ribosomes make a −1 frameshift in 23Zn-1 at the AAAAAA sequence (nucleotides 1152-1157 of the coding region) since shifting in this region maximizes the amino acid identity to the non-frameshifted 9D4 sequence. This region of the HeT-A ORF is very similar to the frameshift regions of some vertebrate retroviruses, such as HIV-1 (reviewed by Weiss et al. 1989). The HIV-1 frameshift occurs on a hexamer of U, rather than A, as in HeT-A. The two frameshift regions are similar in that there is a stop codon in the new reading frame immediately 5′ to the shift point and both have a long overlap between ORF1 and ORF2 after the shift point (279 nucleotides for HeT-A and 241 nucleotides for HIV-1).

In retroviruses the sequence 3′ of a frameshifting sequence can often be folded into pseudoknots or stem-loops. These 3′ sequences can affect the efficiency of frameshifting but the effect appears to depend greatly on the sequence context. For instance, the 3′ sequences are not required for HIV-1 frameshifting (Wilson et al. 1988). The HeT-A sequence does not show obvious potential to form a 3′ secondary structure.

Translation of HeT-A RNA in heterologous systems

The sequence analyses provide a strong argument that HeT-A sequences such as 23Zn-1 can undergo a translational frameshift. This theoretical argument is supported by experiments showing that the 23Zn-1 ORFs can be translated both in vitro and in *E. coli* to yield a full size polypeptide. Vertebrate retroviral frameshifts have been studied in vitro using rabbit reticulocyte lysates, which appear to give a reasonable reproduction of in vivo translation efficiency. Although the A_6 and U_6 frameshifting sequences have been reported only from vertebrate retroviruses, *E. coli*, in vitro translation systems also frameshift on these retroviral RNAs (Weiss et al. 1989). Our studies show that the *Drosophila* HeT-A ORFs appear to be frameshifted with surprising efficiency by both mammalian and bacterial ribosomes (Fig. 5).

To analyze translation by mammalian ribosomes we placed the ORF sequence under the control of a T7 polymerase promoter. Plasmids carrying this construct were linearized and transcribed in vitro with T7 polymerase. By using several different restriction enzymes, we linearized plasmids at different sites. Run-off transcription yielded RNAs encoding polypeptides with predictable size differences. Each transcript was translated in a rabbit reticulocyte lysate and in each case the longest product was of the size expected if the appropriate frameshift had been made (Fig. 5A). The shortest RNA ended in the region where the two ORFs overlap; therefore it was the only RNA that could have been translated to completion without frameshifting. Thus, the shortest polypeptide would be the same size, ~ Mr 53000, whether or not

Fig. 5A, B. The HeT-A frameshift is efficiently accomplished in heterologous systems. A Autoradiogram of [^{35}S]-methionine-labeled products translated from variously truncated transcripts of the HeT-A ORF by a rabbit reticulocyte lysate. In each case the most abundant product is of the size expected if the RNA were completely translated. Transcripts were made by T7 polymerase. Truncation was accomplished by cleaving cloned DNA with appropriate restriction enzymes. Lane 1, RNA of 2675 nucleotides (nt) produces a polypeptide ~ Mr 100000; 2, RNA of 1789 nt produces a polypeptide of ~ Mr 72000; 3, RNA of ~ 1325 nt produces a polypeptide of ~ Mr 53000. The product of ORF1 should be ~ Mr 53000, if no frameshift occurs. Note that there is no major accumulation of a product of this size in lanes 1 or 2. This suggests that ribosomes are shifting to the second reading frame very efficiently. B Antibody blot of total protein from *Escherichia coli* carrying two HeT-A ORF constructs transcribed from an inducible *E. coli* promoter. Products are detected by antiserum against the product of ORF1. Lane 1, protein from *E. coli* before induction of HeT-A transcription, showing no HeT-A products in uninduced bacteria. Lane 2, products of construct containing entire 23Zn-1 coding region with ORF1+ORF2 after induction of transcription. The longest product is ~ Mr 100000, as expected from frameshifting. There are also various shorter products but none are of the size expected from termination at the ORF1 stop codon. Lane 3, products of induced bacteria expressing only ORF1. The major polypeptide detected by the antibody is of the size expected from the ORF1 sequence. This ORF1 polypeptide was used to raise the antibodies used for the blot. Sizes in kilodaltons are indicated

the frameshift was made. This smallest polypeptide also serves as a measure of the size of the polypeptide produced by ribosomes that translate only ORF1. Since frameshifting is expected to be inefficient, it is surprising that a product of ~ Mr 53000 does not accumulate in the translation products of the other RNAs. All the translations yield a spectrum of shorter polypeptides but this is not uncommon with such in vitro reactions. Both ribosome "early quitters" and RNA degradation can produce shorter products. The failure to see a preferential accumulation of ORF1-sized polypeptides suggests that the frameshift is unusually efficient.

We have evidence that the HeT-A frameshift can also be made by *E. coli* ribosomes (Fig. 5B). A construct, p23-ORF1+2, which contains the entire HeT-A ORF from 23Zn-1 cloned for expression in *E. coli*, leads to expression of a HeT-A polypeptide of ~ Mr 100000. Since the HeT-A ORF can yield the large polypeptide only by the −1 frameshift, it appears that the *E. coli* ribosomes have used the HeT-A frameshift signal.

The 3′ repeat region

The non-coding region 3′ of the ORFs makes up nearly half of the element. It is the region characterized in our initial study (Biessmann et al. 1992). The sequences in the present study give results entirely consistent with the earlier characterization. There is no significant ORF in this region. Instead the sequence has a regular pattern of imperfect repeats that are most easily seen by comparing any two elements (as in Fig. 2). This 3′ repeat region is one of the features that distinguish HeT-A from known transposable elements. Other transposable elements contain very little sequence not involved in coding for protein. We have suggested that the 3′ repeat region sequence has a role in the chromatin structure of heterochromatin (Biessmann et al. 1992b; Danilevskaya et al. 1993). Very closely related sequences, which are tandem

```
        <-------------
23Zn-1  CCACCAGCAAAGTAAAAAAAAAAAAAAAAAAAATAAATAAATAAAATAAATTAAACAA.T
        ************** *** **** **** ***  **** ** **** ***** ****** *
T-A1    CCACCAGCAAAGTTAAACAAAATAAAATAAAT.TAAACAATTAAATTAAATAAAACAAAT
        <--------------

23Zn-1  TAACTAAATAATTAAA.TAACTAAAATTAATAATATAATCCGTTCGC...TTGCCAAAGA
        ***  *********** *** *  * **** ** ********   *     **  ****
T-A1    TAAACAAATAATTAAACTAAATTTATTTAACAACATAATCCGCGCAAAATTTTTCAAATT

23Zn-1  CTCTCACGCGCATAACTAATTAAAATCGATTTTCAAGTTGACAAATAAATGGTTTAAAAT
         ** ***  *** ** *  *     *  *****                * *  ***** **
T-A1    TTC.CACCTGCAAAATTTTTC....TTTATTTT..............ATTTTTTTAATAT

23Zn-1  TGTCCTCAGGCTGCAAAGAAAAGCCGCGGCAACAATAAACATTTAGTGACACGCGAAA.G
        *    ** *** ******  ** *    ** ************** ****  ***** *
T-A1    TCGT.TCCGGCCGCAAAGTTAAAC....GCGACAATAAACATTTAACGACAAACGAAAAG

23Zn-1  CGAACATTTGATTAGTGTAATACTTGTGCAAAC.CGACAAGCTGCCGCCATAACAAAACG
        ****** **          *************  ************************** *
T-A1    CGAACAGTT..........ATACTTGTGCAAAAACGACAAGCTGCCGCCATAACAAAAAG

23Zn-1  GAGACGAAGAATCATAAAGAACAAAAGCTAAATCCACCAGCATAGCAAAAAT..AAATTA
        ** * ****** ** *** *** **** ************** **  *** *  ***  *
T-A1    GAAATGAAGAACCACAAATAACTAAAGTTAAATCCACCAGCAAAGTTAAAGTTCAAAGCA

23Zn-1  ACAAATAAAATAAAAGCAAATTTAAATAACATAATAAATTAAACTTATTTAATAAACCAA
        *   **********       ***** **    ***** * **** *** ******* **
T-A-1   AGCTATAAAATAAA.......TTAAACAA....ATAAACTGAACTAATTAAATAAACTAA
```

Fig. 6. Aligned sequences from the 5′ ends of 23Zn-1 and T-A1. *Arrows* mark the beginning of the adjacent HeT-A element in the cloned DNA. The oligo(A) segments of the flanking elements would lie just to the right of the end of each arrow. Because the 5′ region is very A-rich and apparently not rigidly conserved, we cannot unambiguously mark the limit of either oligo(A) tail. Sequences were aligned by the Multalin program (Corpet 1988) with a gap penalty of 2. If gaps>3 are ignored, the sequences are 77% identical in this region

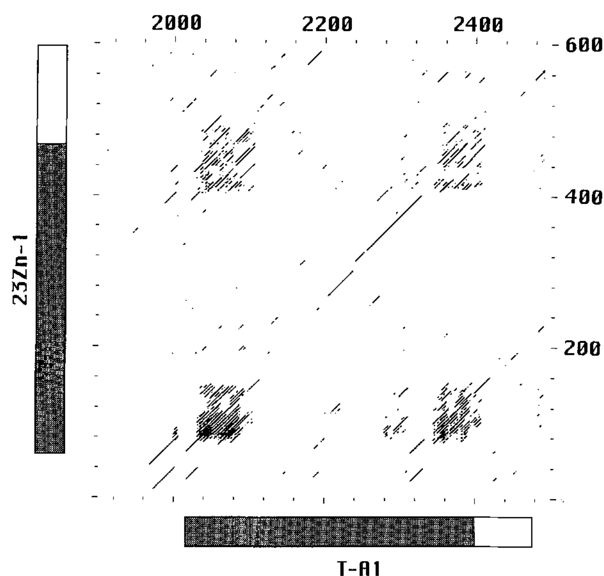

Fig. 7. Dotplot comparing the 5′ structural region of 23Zn-1 with that of T-A1. A very regular pattern of overlapping sequence repeats is detected. This is an enlargement of the regions from Fig. 1 (nucleotides 1–600 of 23Zn-1 and 1900–2500 of lambda T-A). Alignment was with a window of 21 and a stringency of 14. The *bars* on each axis mark the part of the sequence belonging to 23Zn-1 and T-A1. *Filled bars* mark parts corresponding to sequence in Fig. 6

repeats including part of the HeT-A 3′ repeat region, occur in another heterochromatin context; non-telomeric regions of the *Y* chromosome (Danilevskaya et al. 1993). Thus, it is possible that these residual repeats are a motif for heterochromatin, analogous to the sequence motifs indicating functions in proteins.

The 5′ structural region of HeT-A

The alignments suggest that HeT-A has a conserved sequence structure 5′ of the ORF. As can be seen in the dotplots, the most prominent feature of this region is two very precisely spaced clusters of overlapping repeats. When 23Zn-1 is compared with T-A1 (Fig. 7 and also Fig. 2), the repeats can be visualized as two nearly square clusters with a spacing of ~ 220 nucleotides in each sequence. The 9D4 sequence stops in the middle of this region but includes one copy of the repeat. The 4568 clone does not extend this far. In both the 23Zn-1 and the T-A1 sequences these 5′ repeats immediately precede the oligo(A) junction of another HeT-A. In fact the sequences surrounding those junctions are nearly identical (Figs. 6 and 7) although 23Zn-1 was recovered from Oregon R DNA and T-A1 was from Canton S DNA. This similarity in the 5′-most regions suggests that these two elements may define the intact 5′ end of HeT-A. This is the first time that we have observed two elements terminated at the same site (see "Note added in proof"). Studies of HeT-A elements joined to various sites within other elements or terminal deletions in the *yellow* gene have shown no sequence requirement for HeT-A transposition (Biessmann et al. 1990b, 1992a).

Obviously, the coincidence of the T-A1 and 23Zn-1 ends could be due to chance; however the sequence identity in the region of the 5′ end is almost as strong as it is in most of the rest of the element. The alignment is better in this region than in ORF1. We think it likely that the coincidence of the two ends indicates the preferred site for 5′ addition onto the element. We have recently identified an HeT-A RNA (O. Danilevskaya, F. Slot, K. Traverse, N.C. Wogan and M.L. Pardue, in preparation). Its length suggests that the T-A1 and 23Zn-1 elements are full size.

What is conserved in the 5′ region is not so much a defined sequence as two precisely spaced A-rich regions with the A-residues concentrated on one strand. The spacing is most easily seen in the dotplot shown in Fig. 7. This strong predominance of A-residues on the single strand causes the A-rich regions to appear in the dotplot as clusters of overlapping repeats but other analyses show that the repeats are very diverged. The precision of the spacing of the regions suggests that these sequences might form a structure, most likely by specific protein attachment. There are several possible ways in which a 5′ structure on HeT-A might help to explain why HeT-A transposes only to chromosome ends. For example, if the element has a 5′ structure that is blocked and cannot bind to other sequences (with the possible exception of other HeT-A elements), that property would ensure that HeT-A elements form ends (see Discussion).

Discussion

The HeT-A element appears to be a retrotransposon with unusual characteristics

Studies of broken *Drosophila* chromosomes, both the terminally deleted chromosomes isolated by Mason et al. (1984) and the spontaneously broken C(1)A ring chromosome (Traverse and Pardue 1988), have shown that HeT-A elements can transpose. Failure to find any trace of the elements in non-telomeric regions of any *Drosophila* stock (Valgeirsdottir et al. 1990; our unpublished results) is evidence that the transposition is very specifically limited. These findings raise two important questions. How does HeT-A transpose? Why does it transpose only to ends? The sequence of the elements provides several clues that will help answer these questions.

How does HeT-A transpose? The original suggestion that HeT-A was a retrotransposon (Biessmann et al. 1992b) is strengthened by several pieces of evidence. HeT-A has a coding region with characteristics usually associated with retrotransposons and retroviruses. In addition, this coding region is on the DNA strand that should represent the presumed poly(A)$^{+}$ RNA. A cellular RNA of the appropriate size and strandedness has been detected (O. Danilevskaya, F. Slot, K. Traverse, N. Wogan and M.L. Pardue, in preparation). Thus the evidence strongly suggests that HeT-A transposes via an RNA intermediate.

Why does HeT-A transpose only to chromosome ends? This characteristic sets HeT-A apart from all

known transposable elements. The two most obvious explanations for the specific transposition of HeT-A to chromosome ends are: (1) the element may be able to attach only to chromosome ends, or (2) the element may form an end whenever it attaches to the chromosome. In the second case it is presumed that if the element attaches to a break in the chromosome, it will form an end, preventing reattachment of the distal piece of chromosome. The distal fragment would be lost, resulting in aneuploidy and probably cell death. Sequence analyses of HeT-A and HeT-A transposition sites do not distinguish between the two explanations but do give some insight into the question.

The telomere-specific transposition of HeT-A does not appear to depend on the DNA sequence of the target since none of the sites of HeT-A addition have had obvious similarities (Biessmann et al. 1990b, 1992a). This suggests that, if transposition is limited because HeT-A can only attach to ends, the element must be recognizing a structural feature, probably a feature of chromatin rather than DNA alone. On the other hand, the HeT-A element has at least two regions that are not found in other retrotransposons, the 5′ structural region and the 3′ repeat region. One or both of the unique features might have a role in the specificity of the transposition, possibly by forming a structure that limits attachment at the end.

The coding region of HeT-A has retrotransposon-like characteristics

Frameshifts such as that required by the 23Zn-1 ORF are commonly found in retrotransposons and retroviruses. The sequence in the HeT-A frameshift region shows structures not found in any overlapping reading frames reported from *Drosophila* retrotransposons; however this region is very similar to the frameshift region found between the *gag* and *pol* sequences in the human retrovirus, HIV-1. The HeT-A and HIV-1 frameshift regions are alike in having a mononucleotide hexamer, a stop codon immediately 5′ of the start of the new reading frame, and a relatively long overlap of the two reading frames in the 3′ direction (279 nt in HeT-A and 241 in HIV-1). Both HeT-A and HIV-1 appear to be able to undergo frameshift without the assistance of a pseudoknot or stem loop downstream of the frameshift region (Wilson et al. 1988).

Surprisingly, it seems that the non-frameshifting sequence of element 9D4 is the atypical HeT-A ORF. Four other HeT-A ORFs have been sequenced in this region and all of them would require the same frameshift that is needed to translate the 23Zn ORF. The predominance of ORFs with the frameshift argues that it is functional. Although we do not yet know what products the HeT-A ORFs produce in *Drosophila* cells, we know that HeT-A RNA can direct frameshifting in other situations. Our experiments show that both rabbit reticulocyte lysates and *E. coli* cells are capable of translating the entire 23Zn-1 ORF, apparently frameshifting with high efficiency. The ability of these heterologous systems to translate the *Drosophila* sequences is not unexpected since the HeT-A signals appear very similar to those of mammalian retroviruses. *E. coli* ribosomes have also been shown to frameshift appropriately on mammalian retroviral RNAs (Weiss et al. 1989).

It is generally thought that frameshifting is a device for regulating the amount of product from the second reading frame. This is an appealing explanation for the retroviral *gap-pol* sequences because the RNA-binding *gag* product of the first ORF should be needed in larger amounts than the polymerase coded by the second ORF. Very little is known about possible functions for the products of the HeT-A ORFs. However, HeT-A sequences with similarity to *gag* proteins appear after the frameshift, rather than before as in the retrovirus, suggesting that in HeT-A the *gag*-like polypeptide would be less abundant of the two ORF products. The function of the ORF1 product is unknown so we cannot speculate on whether it might be needed in greater abundance. If the frameshift is regulating amounts of product, the loss of the frameshift in element 9D4 may be a deleterious mutation.

Does the 3′ repeat region have a role in limiting transposition to chromosome ends?

All of the HeT-A elements that have been studied have portions of the 3′ repeat region. Surprisingly, segments of sequence with significant homology to this region are also found separate from HeT-A elements (Danilevskaya et al. 1992). These HeT-A related sequences are joined with non-related sequences to form units of tandem repeats in several clusters on the *Y* chromosome. Neither the HeT-A coding region nor the 5′ structural region has any similarity to these tandem clusters. In addition, sequence analysis argues that transposition of the HeT-A-related sequences to the *Y* chromosome was not by retroposition. There are several clusters of these HeT-A-related tandem repeats on *Y* chromosome. These repeat clusters are not telomeric. Thus, the sequence of the 3′ repeat region can be transposed to internal chromosome regions and is not likely to be responsible for limiting HeT-A transposition to ends. On the other hand, the repeat clusters are on the *Y* chromosome, which, like the telomeres, is heterochromatic. Thus the sequence of the 3′ repeat region seems to be limited to heterochromatin, suggesting that it may have a role in chromatin structure.

Does the 5′ structural region have a role in limiting transposition to chromosome ends?

If the telomere specificity of HeT-A is due to a tendency to form some type of a terminal structure, the 5′ end of the element might be responsible since it is at the distal end of each complete element. For that reason it is interesting that this region has very regularly spaced A-rich clusters. The conserved regions might form some type of 5′ structure, probably through specific protein binding. We speculate that the structure might be specialized for the addition of other HeT-A elements but not able to

bind other sequences. The preferred addition site for HeT-A would produce linked elements such as seen in the 23Zn-1 and T-A1 ends. The failure to bind other sequences would produce ends.

Although HeT-A appears to have a preferred 5′ addition site, it is clearly able to transpose onto different sequences when those sequences are at broken chromosome ends, as seen with the healed telomeres. If the 5′ end of HeT-A is specialized for attachment, transposition to random broken ends might be expected to be relatively inefficient. The large number of unhealed telomeres in Mason's stocks with terminally deleted *X* chromosomes (Biessmann et al. 1990a) are consistent with the idea that healing of broken ends is not a very efficient process.

It is interesting that the 5′ structural region also bears some resemblance to parts of the 3′ repeat region that distinguishes the HeT-A element. These similarities raise the possibility that recombination might occur between the 5′ and 3′ repeats, deleting the ORF regions and producing truncated elements.

How is the Drosophila telomere related to telomeres in other organisms?

Telomeres containing simple G-rich repeats are nearly ubiquitous. They have been found in organisms as diverse as protozoa, yeast, mammals, and plants (Zakian 1989). It is hard to imagine that the *Drosophila* telomere is not derived from the same ancestral element as telomeres of other organisms. Therefore, the *Drosophila* telomere seems more likely to be a modification of the general telomere than a radical reinvention.

At the moment we can suggest two possible ways in which the HeT-A mechanism might have been derived from the standard telomere. One possibility is suggested by the formal similarity between telomerase plus its template RNA (Blackburn 1990) and reverse transcriptase plus an RNA transposition intermediate. Perhaps the putative HeT-A RNA transposition intermediate is simply a very complex template for a telomerase. An alternative possibility arises from the observation that HeT-A sequences resemble telomere-associated sequences of other organisms in size and complexity. Perhaps HeT-A evolved from telomere-associated sequences.

An evolutionary relationship between HeT-A and telomere-related sequences in other organisms is attractive in light of recent studies on the *Saccharomyces cerevisiae* mutant, *est-1* (*ever shorter telomeres*). These studies have shown that a major family of telomere associated sequences, Y′, can serve as a backup system for telomeres (Lundblad and Blackburn 1993). Yeast lacking *est-1* function cannot maintain telomerase repeats and cultures become senescent as the chromosomes gradually shorten. All of the revertants found in these cultures have amplified Y′, or partial Y′, elements on the chromosome ends. The amplified elements are dynamic and are rapidly decreased when the *est-1* function is put back into the cells. Thus, the Y′ elements are reproducibly serving as a backup for a defective telomerase system. This evidence that a telomere-associated sequence can serve as an alternative telomere is particularly intriguing since both Y′ and HeT-A have characteristics of retrotransposons, although the Y′ rescue of *est-1* mutants seems to have occurred by homologous recombination rather than transposition. In addition, D. Gottschlung (personal communication) finds that Y′ elements increase the position effect that he finds at yeast telomeres. The 3′ repeat region sequence of HeT-A suggests that it has a role in heterochromatin structure and therefore possibly in position effect, although neither role has been tested. These similarities between HeT-A and Y′ seem worth noting; however, we have not been able to find any supporting sequence similarity in the two elements.

Acknowledgements. This work has been supported by grants from the American Cancer Society and the National Science Foundation. Sequence analysis was done through the Whitaker College Computing Facility of the Massachusetts Institute of Technology. We thank Ranee Mehra for assistance in plasmid construction. We thank the members of the Pardue laboratory and Ky Lowenhaupt for fruitful discussions. We are grateful to H. Biessmann and D. Gottschlung for sharing unpublished results.

References

Bairoch A (1993) The PROSITE dictionary of sites and patterns in proteins, its current status. Nucleic Acids Res 21: 3097–3103

Biessmann H, Carter SB, Mason JM (1990a) Chromosome ends in *Drosophila* without telomeric DNA sequences. Proc Natl Acad Sci USA 87: 1758–1761

Biessmann H, Mason JM, Ferry K, d'Hulst M, Valgeirsdottir K, Traverse KL, Pardue ML (1990b) Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in *Drosophila*. Cell 61: 663–673

Biessmann H, Champion LE, O'Hair M, Ikenaga K, Kasravi B, Mason JM (1992a) Frequent transposition of *Drosophila melanogaster* HeT-A elements to receding chromosome ends. EMBO J 1: 4459–4469

Biessmann H, Mason JM, Ferry K, D'Hulst M, Valgersdottir K, Lofsky A, Chin C, Ginther B, Levis RW, Pardue M-L (1992b) HeT-A, a transposable element specifically involved in "healing" broken chromosome ends in *Drosophila*. Mol Cell Biol 12: 3910–3918

Boeke JD, Corces VG (1989) Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43: 403–434

Blackburn EH (1990) Telomeres and their synthesis. Science 249: 489–490

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16: 10881–10890

Danilevskaya ON, Petrov DA, Pavlova MA, Koga A, Kurenova EV, Hartl DL (1992) A repetitive DNA element, associated with telomeric sequences in *Drosophila melanogaster*, contains open reading frames. Chromosoma 102: 32–40

Danilevskaya O, Lofsky A, Kurenova EV, Pardue M-L (1993) The Y chromosome of *Drosophila melanogaster* contains a distinctive subclass of HeT-A-related repeats. Genetics 134: 531–543

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395

Grinblat Y, Brown NH, Kafatos FC (1989) Isolation and characterization of the *Drosophila* translation elongation factor gene. Nucleic Acids Res 17: 7303–7314

Luan DD, Korman MM, Jacubczak JL, Eickbush TM (1993) Reverse transcription of R2Bm RNA is primed by a nick at the

chromosomal target bite: a mechanism for non-LTR retrotransposition. Cell 72: 595–605

Lundblad V, Blackburn EH (1993) An alternative pathway for yeast telomere maintenance rescues est1-senescence. Cell 73: 347–360

Mason JM, Strobel E, Green MM (1984) mu-2: mutator gene in *Drosophila* that potentiates the induction of terminal deficiencies. Proc Natl Acad Sci USA 81: 6090–6094

Mitchell PJ, Tjian R (1989) Transcriptional regulation in mammalian cells by DNA binding proteins. Science 245: 371–378

Muller HJ (1938) The remaking of chromosomes. Collect Net 13: 181–195

Traverse KL, Pardue ML (1988) A spontaneously opened ring chromosome of *Drosophila melanogaster* has acquired HeT DNA sequences at both new telomeres. Proc Natl Acad Sci USA 85: 8116–8120

Valgeirsdottir K, Traverse KL, Pardue ML (1990) HeT DNA: a family of mosaic repeated sequences specific for heterochromatin in *Drosophila melanogaster*. Proc Natl Acad Sci USA 87: 7998–8002

Weiss RB, Dunn DM, Shuh M, Atkins JF, Gesteland RF (1989) *E. coli* ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. New Biol 1: 159–169

Wilson W, Braddock M, Adams SE, Rathjen PD, Kingsman SM, Kingsman AJ (1988) HIV expression strategies, ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. Cell 55: 1159–1169

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based on their reverse transcriptases. EMBO J 9: 3353–3362

Zakian VA (1989) Structure and function of telomeres. Annu Rev Genet 23: 579–604

Note added in proof

We have sequenced the first 255 nucleotides of the 5′ end of the HeT-A element that joins the 3′ end of 23 Zn-1. The end is nearly identical to 23 Zn-1 and has only 3 changes in this region.