J. Mol. Biol. (1995) 245, 34–42

JMB

Polymorphic SINEs in Chironomids with DNA Derived from the R2 Insertion Site

Hong He, Carlos Rovira, Shirlei Recco-Pimentel, Ching Liao and Jan-Erik Edström*

Department of Molecular Genetics, University of Lund Sölvegatan 29, S-22362 Lund Sweden

*Corresponding author

A short interspersed repeat (SINE) in the two sibling species *Chironomus pallidivittatus* and *Chironomus tentans* is described. It is present at many sites in the genome and is surrounded by 10 to 14 bp target site duplications. It consists of two sequence modules in different numbers and variable order relative to each other and often has large inversions of different sizes at one end. One of the modules contains pol III promoter consensus sequences. This SINE, nevertheless, is likely to have been dependent on an outside promoter for its formation. It is therefore interesting that both modules start with a 22 bp region with striking similarity to the R2 insertion site in the preribosomal gene of insects. We suggest that this type of SINE, termed Cp1, was formed after a series of events among which the first step was the retroposition of a tRNA gene into the R2 site in the preribosomal gene by the R2 coded protein. The final step is likely to have been due to retroposition from this site.

Keywords: SINEs; pol III; tRNA; R2 integration site; *Chironomus*

Introduction

SINEs are short interspersed repetitive elements believed usually to be formed by reverse transcription of pol III transcripts, integrated without pronounced target sequence dependence (Rogers, 1985; Weiner *et al.*, 1986; Deininger, 1989; Okada, 1991; Eickbush, 1992 for reviews). The *Alu* sequences in primates and the rodent B1 elements are derivatives of 7SL RNA genes (Ullu & Tschudi, 1984) and the majority of the remainder have an origin in tRNA genes (Okada, 1991). Nevertheless SINEs may be complex like the Galago type II retroposon (Daniels & Deininger, 1983*a*,*b*) derived from both tRNA and 7SL RNA genes and they may contain sequence elements of other origin like an A or A + T-rich 3' end. Other blocks may also exist like the 31 bp insertion in the second half of the *Alu* elements (Deininger *et al.*, 1981) or segments in the rodent type 2 *Alu* family (Haynes & Jelinek, 1981) and the rabbit C family (Hardison & Printz, 1985). It has been suggested that recombination and extension of pol III transcription units could create such complex units (Sakamoto & Okada, 1985). Another factor may be that SINEs have a tendency to insert into 3' regions of pre-existing elements (Rogers, 1985).

SINEs belong to the non-LTR class of retrotransposons (without long terminal repeats). Like retropseudogenes derived from pol II transcription they depend on enzyme produced elsewhere for their transposition. It was shown for a non-LTR retrotransposon, the R2 element, that the transcript is identified at the 3' end by its own gene product (Luan *et al.*, 1993). This protein also finds the specific insertion site in the rRNA gene, cleaves DNA strands at the integration site and catalyses reverse transcription. It was proposed that non-LTR retroelements without open reading frames like SINEs use the transposition machinery of other non-LTR elements (Eickbush, 1992).

Here we provide evidence that the above proposal applies for a specific type of SINE identified in chironomid insects: *Chironomus pallidivittatus* and the sibling species *Chironomus tentans*. This SINE, which we term Cp1, contains two different, partially identical modules, one of which shows significant similarity with tRNA genes, including pol III promoter consensus sequences. Cp1 is variable in design. The two modules can appear in different combinations within the element, one end of which is inverted in the majority of cases. Sequences hybridizing to Cp1 components are present at many sites, including the centromeres, outside of which they are also found in degenerate forms. Each

Present address: S. Recco-Pimentel, Department of Cell Biology, Universidade Estadual de Campinas, 13081 Campinas, S. P. Brazil.

Abbreviations used: SINE, short interspersed repeat; pol III, polymerase III; LTR, long terminal repeat.

0022-2836/95/010034–09 $08.00/0

module has an initial segment of 22 bp, with pronounced similarity to the insertion site for the R2 element (Roiha *et al.*, 1981). Although Cp1 contains promoter consensus sequences for pol III, it is unlikely that pol III was the only polymerase responsible for its formation. We suggest that Cp1 was formed in several steps at the R2 integration site after retroposition of tRNA genes with the aid of the R2 protein and that it was dependent for its final transposition away from the R2 site on an outside preribosomal promoter.

Results

The centromeric clone pCp627

Figure 1A shows the arrangement of units in pCp627 (Rovira *et al.*, 1993) containing two tandem 155 bp centromeric repeats surrounded by non-repeated DNA. It is likely that in genomic DNA the number of repeats is much larger and that repeat elimination has occurred. The two repeats are followed by a short sequence, constituting the start of a third repeat. The 155 bp repeats are flanked to the left by the 61 bp long *A* segment followed by the 193 bp long *B* module and to the right by an inverted *A* segment. The *AB* region is subcloned as pCp254 and used as a probe to identify clones described in the present paper.

It is possible that the *AB* region represents the right end of the Cp1 element extending between 155 bp tandem arrays and that the inverted *A* segment is the left end of another, similar element inserted in the same way. If so the start of a third repeat could be a target site duplication. The results to be presented here support this interpretation.

Structure of different Cp1 elements

Two clones, pCp116 (EMBL accession no. X79503) and pCp413 (EMBL accession no. X79505), were isolated from a *Hin*dIII digest of *C. pallidivittatus* larval DNA ligated into λNM1149. The arrangement in the inserts of Cp1 elements, which are surrounded on both sides by non-centromeric DNA, is shown in Figure 1B and C. The insertions are flanked by target site duplications. In these and other clones shown in Figure 1 two sequence modules can be distinguished, designated the *SCA* (Figure 1G) (EMBL accession no. X79539) and the *B* module (Figure 1H) (EMBL accession no. X79540), both of which start with 22 bp sequences with only one base difference. The *SCA* module consists of the *S*, *C* and *A* segments between which *Eco*RI sites are present. The *A* segment and the *B* module are also components of the *AB* region in pCp627. It can be seen in Figure 1B and C that there are inversions at the left end of these Cp1 elements, in which modules that are truncated at their initial parts extend in diverging orientations. The *SCA* module reaches the left ends of the elements with its tip in both cases. In pCp116 the truncated *B* module extending to the right is followed by complete *SCA* and *B* modules. In pCp413 the *B* module bordering the inversion is followed by another, complete *B* module.

The Cp1 element in pCp116 illustrates a possible design of an intracentromeric unit. If one assumes that similar units are integrated between tandem arrays of 155 bp centromeric repeats, this would lead to a sequence similar to the one represented by pCp627 in Figure 1A.

We also isolated a clone, λCt2 (EMBL accession no. X79506), from an EMBL 3 library of DNA from the sibling species *C. tentans*. The insert of 16.7 kb was restricted with *Eco*RI. The region containing sequences hybridizing to pCp254 (Figure 1D) was distributed between one 1.6 kb fragment (right end) and one 2.1 kb fragment (left end) which seemed to adjoin each other as demonstrated by restriction mapping. This was confirmed by PCR (see Genomic arrangements). Sequencing showed, starting from the left (in the right part of the 1.6 kb fragment) first a *B* module then an *S* part of the *SCA* module. The left end of the 2.1 kb fragment starts with the *A* part (54 bp if the entire *Eco*RI site is included) and is then joined to non-centromeric DNA. All parts of this Cp1 element are in the forward orientation and it is flanked by target site duplications. Obviously subcloning resulted in the loss of the small *C* segment (31 bp) of the *SCA* module and this element, consequently, is likely to consist of a complete *B* module followed by a complete *SCA* module in λCt2 and in the genome. There is also, between the left target duplication and the start of the B module, a short sequence of 11 bp similar to the end of a *SCA* module.

Fragments of Cp1 elements

One clone, pCp284 (EMBL accession no. X79504) was obtained by PCR amplification of genomic DNA from *C. pallidivittatus*. The forward primer, 5' GTGCTGCCAGCTAAGCGAGCGAATG 3', with a *Pst*I cloning site was made to correspond to the beginning of the *B* module. The reverse primer, 5' GTTGAGCTCTTCGTAGTAGCTTAGGC 3', with a *Sac*I cloning site, was made to fit the end of the *C* segment in the inverse orientation. Due to the similarity of the beginnings of the *B* and *SCA* modules the use of the forward primer led to the cloning of a product starting with a *SCA* module, and followed by the beginning of another *SCA* module (up to the primer in the *C* segment; Figure 1E). This clone shows that a *SCA* module can directly follow another *SCA* module, which is not demonstrated by any other cloned element. We do not believe this duplication is a PCR artifact. This is because the downstream primer is within the *C* segment; nevertheless the upstream *SCA* module is complete. This conclusion agrees with the fact that the complete *SCA* module has a termination that conforms with the sequence of internal modules (see Beginnings and ends of Cp1 modules).

Another fragment of a Cp1 element was obtained from the genomic *Hin*dIII library of *C. pallidivittatus*.

It contained a complete *B* module followed by a part of the *S* segment extending to a *Hin*dIII site. The *B* module is connected by the sequence TATTAT to a truncated inverted *S* segment at its left end. The inverted *S* segment can be followed to an upstream *Hin*dIII site. This clone is designated pCp125 (Figure 1F; EMBL accession no. X79507).

Genomic arrangements

To investigate whether duplications and inversions seen in Cp1 elements are cloning artifacts rather than representing the genomic arrangements, we first determined the size of the element in pCp116 (Figure 1B) and λCt2 (Figure 1D) in the clones and in genomic DNA. For this purpose we constructed oligonucleotides matching sequences in flanking DNA and carried out PCR, followed by gel electrophoresis of the amplified products. In both cases we obtained the expected bands, identical in size for genomic and cloned DNA (Figure 2). Further analysis was done with the element in Figure 1B, which has both a duplication and an inversion. An oligonucleotide primer in inverse orientation was constructed for the upstream part of the *B* module which was used in combination with the upstream flanking primer. Another primer was constructed representing the *S* segment in the forward direction, which was used together with the reverse downstream flanking primer. In both cases bands of the expected size, identical for cloned and genomic DNA, were obtained (Figure 2). This experiment, therefore, probes in a simple way the overall structure of this Cp1 element and shows that two of its modules have similar orientation and location in both cloned and genomic DNA.

Beginnings and ends of Cp1 modules

Consensus sequences for *B* and *SCA* in Figure 3A show that the modules start with an almost identical 22 bp sequence: CTCTCTAAGCGAGC (G,A)AATGCTT, the G in the variable position being characteristic for the *B* module consensus and the A for the *SCA* module consensus. These sequences, in turn, can be aligned to the R2 insertion site in the 28 S part of the preribosomal gene (Figure 3B). There is some sequence similarity between the two modules in the subsequent 40 bp (Figure 3B) but none with the preribosomal gene.

The ends of both modules differ somewhat in flanking and internal positions. They are most complete and regular when preceding other modules (Figure 4). When bordering a target site duplication or the centromeric repeat the ends are shortened and more irregular. Their termini show mutual similarities (Figure 3B).

Structure of inversions

The inversions in pCp116 (Figure 1B) and pCp413 (Figure 1C) are similar in design. The inversion points are situated between a *B* module in forward and a *SCA* module in reverse orientation. The truncations that accompany the inversions eliminate the upstream parts of the modules. The shortened *B* modules are similar in size whereas the eliminations

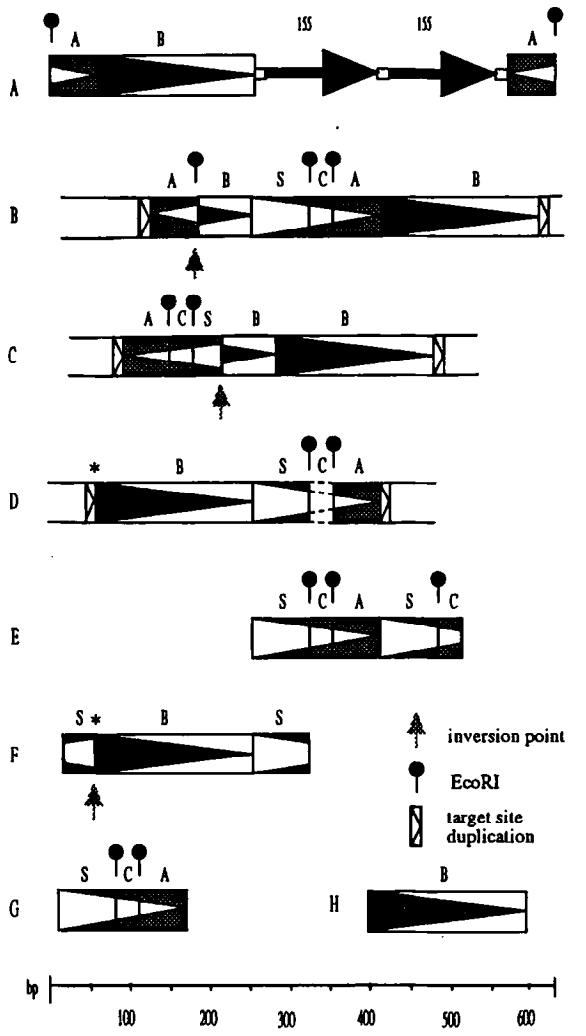

Figure 1. Arrangement of the *SCA* and *B* modules in centromeric DNA and extracentromeric Cp1 elements. Schematic representation of pCp627, which is a centromeric clone containing two 155 bp tandem repeats and surrounded by components of the *AB* region (A). The first 14 bp of each repeat, also present as the start of a third repeat are indicated by white boxes. The element cloned in pCp116 is shown in B and the element in pCp413 in C. A Cp1 element from *C. tentans*, present in the λCt2 clone, is shown in D. Here part of a module has been lost during cloning, shown with a dashed contour. A PCR product, pCp284, with one complete *SCA* module followed by the *SC* part of a second module is shown in E. In F is a Cp1 fragment from pCp125, containing an inversion point between a truncated, inverted *SCA* module and a forward *B* module followed by the first part of a forward *SCA* module. The *SCA* module is shown in G and the *B* module in H. Asterisks in D and F indicate that 3' end fragments of the *SCA* module precede the *B* module, AACAATATTAT in D and TATTAT in F.

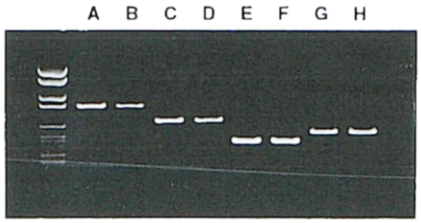

Figure 2. Amplification by PCR of cloned and genomic DNA containing Cp1 elements. Amplification products of pCp116 DNA are shown in A, C and E, and amplifications with identical primers applied to genomic DNA in B, D and F. Amplification product of λCt2 in G and of genomic DNA with identical primers in H. Four different primers were used in the 6 first amplifications (A to F): Pa, Pb, Pc and Pd. Pa is a 23 nt forward primer, starting 381 bp upstream of the start of the 5' flanking target site duplication: 5'TCTAGAATCAGTGGTATCCTGAG3'. Pb is a 23 nt reverse primer, the 5' end of which is 125 bp downstream of the 3' end of the downstream target site duplication: 5'CGAAAACAGGAGTATTGTCAGTC3'. Pc is a 22 nt reverse primer representing the *B* module in its middle part (pos. 60 to 81 in Figure 3A): 5'TGCCATTGGCATTGG CAGAATG3'. Pd is a 23 nt forward primer representing the *S* segment of the *SCA* module (pos. 36 to 58 in Figure 3A): 5'ATAAGTCGTTCTGTTAGCAGAGC3'. The combination Pa-Pb was used in A and B. This gives a 1022 bp product according to sequencing data in agreement with the electrophoretic migration for both substrates. The combination Pa-Pc was used in C and D and should give a product of 771 bp, in both cases in agreement with electrophoretic results. Finally, in E and F the combination Pd-Pb was used, which should give a product of 470 bp, again in agreement with both electrophoretic separations. According to Figure 1B the products of the 2 latter amplifications overlap, with 219 bp according to sequencing data. For amplifications in G and H 2 primers were used, Pe and Pf. Pe is a 20 nt forward primer flanking the Cp1 element in λCt2, starting 108 bp upstream of the 5' end of a 14 bp target site duplication: 5'TGCTTGGCACTAAAT CTGCG3'. Pf is a 20 nt reverse primer the 5' end of which is 67 bp downstream of the 3' end of the 14 bp target site duplication: 5'TTGTGAGTAGTGAATTGATG3'. Amplification with the 2 primers should result in a 593 bp long DNA. A similar product is obtained with both the λCt2 DNA (G) and genomic DNA (H) as substrate. Separations were run in 1% (w/v) agarose gel with size marker VI (Boehringer), the 10 visible bands of which (from top to bottom) are: 2176, 1766, 1230, 1033, 653, 517, 453, 394, 298 and 234/220 bp.

```
A

B module
      CTCTCTAAGC GAGCGAATGC TTTACAACCG TTGAAAAAAG
      AGTGGTTCAT ATATATTGGC ATTCTGCCAA TGCCAATGGC
      AAAGCCCGAT AGCTCAGTGG TCTGAGCACT TGACCGGCAA
      TCGAGAGGTG CGAGGTTCGA TTCCCGCTCG GGAAGAGTCA
      TTGGGTGAAT TACTTTTTTT TCAACTTTTT CTTTTAAACT
      ATTTTTT

SCA module
      CTCTCTAAGC GAGCAAATGC TTATTCAAAA TGAGAATAAG
      TCGTTCTGTT AGCAGAGCAA GCTTAAGCTT GATAGCAGTT
      CTCGAATTCA TCAAAAAGTA GCCTAAGCTA CTACGAATTC
      AGACAAAACA GAACATCAGA CTTAACTTCT CATATATTCC
      CTTTGAAAAT ATTAT

B

      5'      5910        5920
rDNA  CTCTCTTAAG GTAGCCAAAT GCCT
B     ------*--- CG---*G--- --T-TA*CAA CCGTTGAAAA
SCA   ------*--- CG---*---- --T-ATT--- **AA---G--

B     AAGAGTGGTT CATATATATT GGCATTCTÔT TTCCTTT*AA
SCA   T*A---C--- -TGT--GCAG A---AG--Ô- -C-----G--

          3'
B     ACTATTT        Ô = 120bp in B, 92 bp in SCA
SCA   -A----AT
```

Figure 3. Consensus sequences of the *B* and *SCA* modules are shown in A. The versions at the 3' ends characteristic of internal modules are given. Alignment of the initial parts of the *B* and *SCA* modules with the R2 integration site in preribosomal gene DNA (using the sequence given by Roiha *et al.* (1981) and Tautz *et al.* (1988), the numbering of the pre-rRNA gene according to the latter authors) is shown in B. The subsequent segments of the *B* and *SCA* modules are then aligned with each other. After regions that are 120 bp and 92 bp, respectively, within which there are no stretches of mutual similarities, the 3' ends are compared with each other. Again, the 3' end versions of internal modules are given. Asterisks show deletions and dashes identities.

have left *SCA* modules of different size. The inverted *A* segment in pCp116 could start at any of the TTC positions within the *Eco*RI site or an adjoining G but the location is ambiguous since the TTCG sequence could, alternatively, be the start of the *B* module fragment. The same G starts the *B* module fragment in pCp413. The fact that an inversion point is possibly present within an *Eco*RI site is a coincidence since this restriction enzyme was not used for cloning. A similar inversion pattern would also be compatible with the structure of the insert in pCp627 (Figure 1A).

The inversion in pCp125 (Figure 1F) differs in principle from the other two. The cloning procedure allows only a part of the element, including the inversion, to be seen, between *Hin*dIII sites in two *S* segments. The inverted *S* segment starts 18 bp downstream of the beginning of the module and can be followed to a *Hin*dIII site. It joins a complete *B* module in forward orientation. The sequence TATTAT, characteristic of the end of an internal *SCA* module, is, however, interposed between the two elements (Figure 3B). Thus, in this case the inversion is localized between fragments of two *SCA* modules.

The inversions do not occur in all Cp1 elements, nor do they engage both ends of the elements like in the foldback elements (Potter *et al.*, 1980) and are therefore unlikely to be required for transposition.

Target site duplications

If one assumes that the *AB* region and the inverted *A* segment in pCp627 represent the ends of an element inserted between tandem repeats, it will be surrounded by 14 bp duplications, AAAGCTTT-TATTTT, one of which would also constitute the start of a third repeat. The last 4 bp could alternatively be part of the inverted *A* segment that forms the left end of the pCp627 clone, and the real length may be shorter, down to 10 bp.

For pCp116 the target site duplication is 13 bp, TAAACATAACTAA, but the 5'-proximal T in the distal target site might be the end of the *B* module and the corresponding T in the proximal site could be part of surrounding genomic DNA and show agreement by chance. It is possible, therefore, that the duplication is only 12 bp. Also pCp413 has a target site duplication of 12 bp, TGTTATTAAGGA.

In the case of the 14 bp duplication surrounding the Cp1 element in λCt2, ATAAAACAATTTAA, the start position of the downstream site is ambiguous since the first 3 bp could also be the end of the *A* segment. A chance agreement with more than 1 bp in the DNA preceding the upstream duplication appears unlikely and a size of 13 to 14 bp is therefore most probable.

In conclusion, the target size duplications are 10 to 14 bp, a size within the range typical for SINEs. If, however, only one size is possible it has to be 12 bp. As is usually the case for SINEs the duplication consists of A + T-rich DNA, here with one to three G·C base-pairs, reflecting that retroposition usually occurs into A + T-rich DNA (Rogers, 1985).

Origin of modules

Computer search (Altschul *et al.*, 1990) showed significant similarities between the *B* module and several tRNA genes. Best agreements were obtained with two sequences from *Drosophila melanogaster* (Figure 5), particularly with respect to similarities within putative A and B promoter boxes and termination signals as well as the distances between these elements. These were the Lys-tRNA-5 gene corresponding to the region 29A1-2 (Defranco *et al.*, 1982) and the Gly-tRNA cluster from region 35B (Meng *et al.*, 1988). The putative A and B boxes in a pol III promoter in Figure 5 agree with established consensus sequences (Murphy & Baralle, 1983). There are two pol III termination sequences (Bogenhagen *et al.*, 1980) at the end of the *B* module.

No pol III promoter boxes were found in the *SCA* module in any orientation, nor were there any significant similarities with other parts of tRNA genes.

Sequence conservation

Sequences in different clones representing the same module are highly similar. The mutation rate from the consensus is 1.6% for five complete *B* modules (A, B, C, D and F in Figure 1). Of the 16

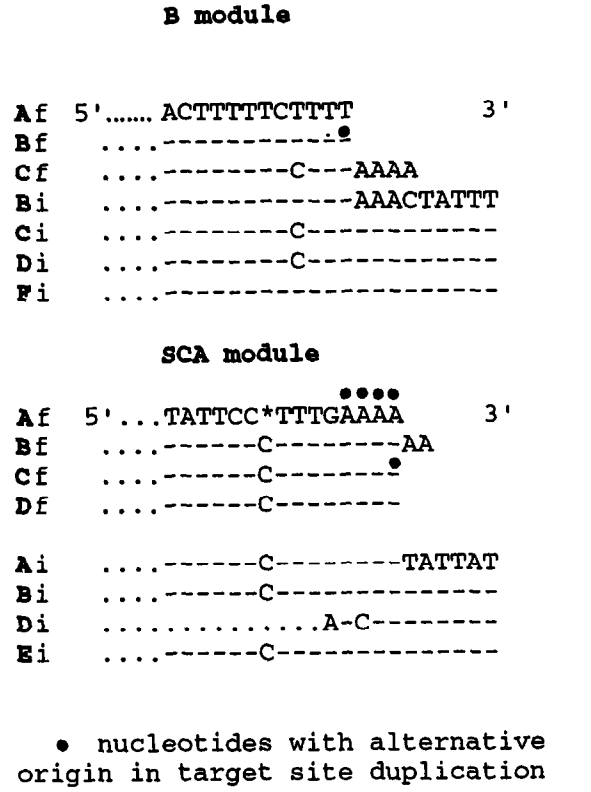

Figure 4. Sequences at the termini of flanking (f) and internal (i) *B* and *SCA* modules. The elements from which the modules originate are designated as in Figure 1. Dashes indicate identities.

mutations 11 are single base deletions or insertions and the remainder single base substitutions. One insertion (for pCp413) changes the A box so as not to agree with the consensus whereas all B boxes conform with the consensus in spite of a deletion (pCp254) and a base substitution (pCp413). The mutations were distributed along the whole module.

Since only two complete *SCA* modules were available partial sequences were also compiled, and used to determine mutation frequencies. There are 17 mutations (eight deletions/insertions, seven single and two double base substitutions) in 991 analysed positions, giving a mutation frequency of 1.7%. The mutations are clustered within one unstable region (positions 13 to 17) having three mutations and within another one with seven mutations in positions 71 to 79. Thus most of the *SCA* module is strikingly well conserved. The 40 bp stretch with partial identity to the *B* module, following the 22 bp start sequence, is particularly stable with only one mutation in 259 analysed positions.

Regions around inversion points are not significantly mutated above background. Regions of 40 bp in the elements shown in Figure 1B, C and F surrounding the inversion points contained altogether only one mutation (double base substitution).

All in all both modules are relatively constant, suggesting either recent formation or a high degree of sequence conservation.

```
B module  GAAAAAAGAG TGGTT***CA TATATATT*G GCATTCTGCC
gly tRNA  -------A-C --T--ATC-- AT---G--C- -A---TCAG-
lys tRNA                        ----** -----T-T-G

                                          A
          AATGCCAATG GCAAAG***C *CCGATAGCT CAGTGGTCTG
          --A-AT-T-T ------***- AT--G-G-T- -------*A-
          CGA---TG-T --C-T-TGC- *-G-G----- ----C-GTA-

                                                  B
          AGCACTTGAC CGGCAATCGA GAGGTGCGAG GTTCGATTCC
          -ATG--C-*- -T--C-CGCG -GC-GC-CG- ----------
          ----T-G--- TTTT----C- AG---CTAG- ----A-G---

                                       stop
          C*GCTCGGGA AGAGTCATTG GGTGAATTAC TTTTTTTTCA
          -*-GC--**- T-CAAA--A- TT-TTTAA-- --------TT
          -T------** *-C--A-CAT CA-TTT--T- CAC-----*-

           stop
          ACTTTTT
          -T--C-A
          -------
```

Figure 5. Sequence of part of the *B* module (top) compared with 2 tRNA gene sequences. A and B boxes are indicated like termination signals. Dashes show identities and asterisks deletions.

Degenerating elements

Several clones hybridizing to pCp254 were isolated from our *Hin*dIII library in λNM1149, containing more or less degenerate segments of different sizes (Table 1). None of these sequences were connected to centromeric tandem repeats. In no case could target site duplications be identified. The 12 clones containing these sequences were screened from a library sample that gave three non-degenerate clones (pCp116, pCp413 and pCp125), suggesting that most elements complementary to pCp254 are degenerate. It is thus of some interest that the elements are either well conserved or highly mutated.

In situ hybridization

Double colour *in situ* hybridizations were done to reveal the distribution of Cp1 elements relative to centromeres. Figure 6A shows the localization of the insert in pCp116 to the right arm of chromosome I and Figure 6B the insert from λCt2 to a region in the left arm of chromosome 2. The origin of the third complete Cp1 element in pCp413 could not be determined since surrounding sequences hybridized to repetitive, interspersed DNA. pCp254 was used as a probe in Figure 6C and shows the widespread distribution of complementary sequences and also their relatively high concentrations in centromeres.

Table 1

DNA segments in *C. pallidivittatus* hybridizing in *AB region of pCp627*

Length (bp)	% Identity with *AB* region
157	87
119	67
107	83
096	81
063	65
056	57
045	60
043	65
040	65
038	74
029	69
028	68

Discussion

We describe a new type of SINE, called Cp1, which is modular in design with variable internal arrangement of the two constituent modules and with 10 to 14 bp target site duplications. Some of the elements have inverted segments at their left ends in which the inversions are accompanied by eliminations of variable extent from the 5' ends of the modules bordering the inversion points. The number of elements is not easily titrated, because of the presence in the genome of numerous degenerating elements. On the basis of *in situ* hybridization pictures one can surmise that there are several dozens of regions with sequences more or less identical to the modules within Cp1, distributed throughout the whole genome but particularly concentrated to the centromeres.

Each module starts with 22 bp long, almost identical sequences. These 22 bp boxes can be aligned to a 24 bp region in the 28 S part of the preribosomal gene from insects. This region contains the R2 non-LTR retroposon insertion site. It is highly conserved and several insect species use one and the same size (Jakubzak *et al.*, 1991), in *Bombyx mori* generated by a 2 bp staggered cut (Luan *et al.*, 1993), located 10–11 positions downstream of the left end of the rRNA region in Figure 3B. The presence of these boxes in both Cp1 modules suggests a previous state in their generation, in which they were associated with the R2 insertion sites.

One of the modules, *B*, contains distinct putative pol III promoter boxes and termination signals and shows significant similarity to tRNA genes from *D. melanogaster*. The origin of the second module, *SCA*, is not known. It could, however, derive from a tRNA gene like the *B* module. We speculate that two originally identical units diverged in evolution once they became part of the same retroposon. Some support for this is a region of about 40 bp following the 22 bp box, showing partial mutual identity like the 3'-ends. This alternative would to some extent parallel the diverging evolution of two 7SL-derived components in the *Alu* element (Deininger *et al.*, 1981). An alternative is that the *SCA* element is more complex and contains components of different origin in addition to the 22 bp rDNA.

A variable module arrangement like the present one has to our knowledge not previously been found in any retroposon. This feature may have a causal relation to the 22 bp rDNA boxes at the beginning of each module, which may have offered recombination hotspots during the formation of these elements.

Although Cp1 falls within the definition for SINEs insofar as it seems to contain pol III promoter elements (Deininger, 1989), it is not a typical SINE.

Thus it is not likely that the final Cp1, although containing pol III promoter consensus sequences, is generated by pol III action. The transcription unit probably terminates within the element, which is usually not the case (Duncan *et al.*, 1981; Fuhrman *et al.*, 1981). It also seems unlikely that the *SCA* module would be transcribed from an inside pol III promoter irrespective of whether it lies upstream or downstream of a *B* module. Finally pol III would not be expected to transcribe the beginning of the *B* module. All this suggests that Cp1 has been dependent for its formation on an outside promoter which allowed it to be transcribed irrespective of its variable internal structure and the presence of pol III stop signals. Only a few non-LTR retroposons with specific integration sites are known and their occurrence may reflect the need for external promoters for retrotransposition (Eickbush, 1992). The R2, like the R1 integration sites (Roiha *et al.*, 1981) in the 28 S region of the preribosomal gene belong to this category. It is against this background that the presence of the 22 bp rDNA boxes becomes significant.

We suggest that the Cp1 elements have been formed at this site and made use of the pre-rRNA promoter for their retroposition. This could have occurred in a sequence of events (Figure 7) in which the first step was the misidentification of the 3' end of the tRNA transcript for an R2 transcript by the R2 protein and insertion of the resulting tRNA retrogene into the R2 insertion site (Figure 7A, B). In this context or in a subsequent step the insert became surrounded on both sides by repeats similar to the 22 bp boxes, e.g. by gene conversions induced from rRNA genes without insertions (Figure 7C). (This could have happened even if the proximal rDNA border was duplicated (Roiha *et al.*, 1981; Burke *et al.*, 1987).) This in turn could promote recombination by unequal crossing over or circle formation and re-insertion (Figure 7D). This would result in two tandem tRNA retrogenes inserted between and separated by 22 bp repeats. This double unit could then have evolved into two regions with different functions, one of which, the *B* module, retained the pol III internal promoters, whereas the other one, the *SCA* module, lost most tRNA gene features (Figure 7E). A few mutations in the 22 bp rDNA box (Figure 3B) would prevent recombinations with unoccupied R2 sites in preribosomal genes. New combinations of modules and longer arrays would be created by further recombinations. During the recombinations inversions combined with truncations of the 5' parts of the modules could have occurred. In retrotranspositions studied here 22 bp repeats are not present where the 3' ends of the modules join surrounding DNA, meaning that the final transposition does not encompass a hypothetical distal 22 bp repeat and a few terminal nucleotides belonging to the adjoining modules (Figure 7F, G). Isolation of pre-Cp1 elements in the R2 region of *Chironomus*, should they still exist, would be required to verify central parts of this model.

The present results provide the first factual

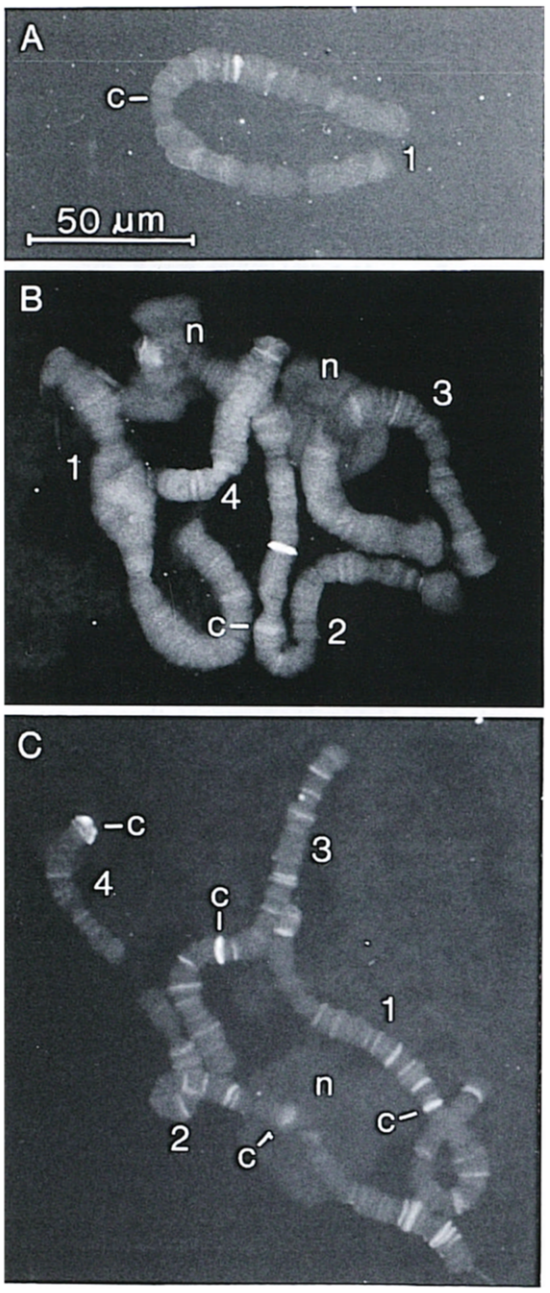

Figure 6. *In situ* hybridization to localize pCp116 in A shows hybridization to the right arm of salivary gland chromosome 1 of *C. pallidivittatus*. The insert of the λCt2 clone is hybridized to a squash of *C. tentans* salivary gland chromosomes in B, showing hybridization to the left arm of chromosome 2. The *AB* region in pCp627 is hybridized in C to *C. pallidivittatus* chromosomes. In all cases probes were labelled with biotin and immunological detection was with avidin-FITC. The localizations of centromeres in A and C were determined by simultaneous hybridization with digoxigenin-labelled 155 bp centromeric repeat and immunological detection with rhodamine-labelled antibody. The main hybridization signal in A and B comes from genomic sequences adjoining the Cp1 element, which gives weak cross hybridization to many other regions. Chromosome numbers are given, *c* stands for centromeres and *n* for nucleoli.

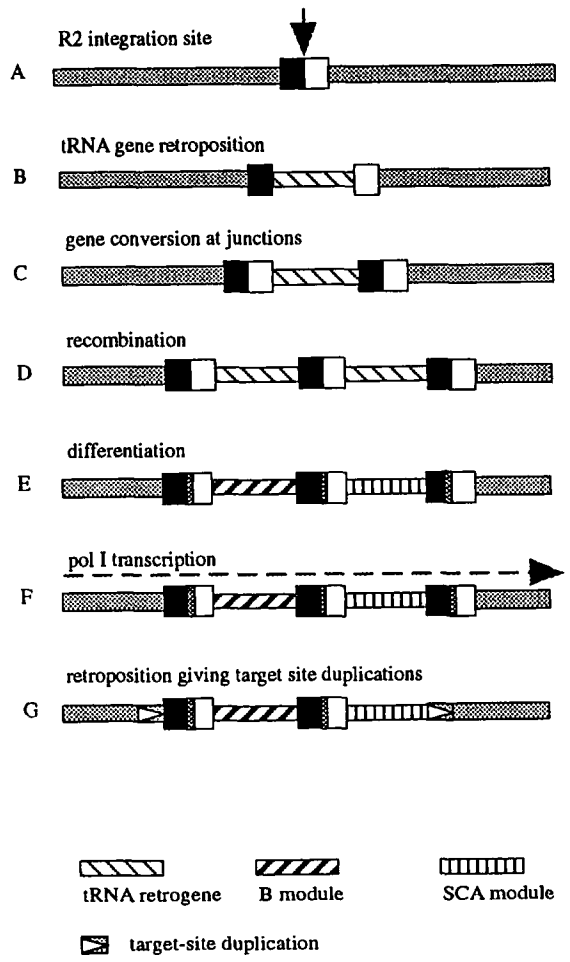

Figure 7. A model for the formation of Cp1 elements. The R2 integration site is shown in A into which a tRNA transcript becomes reverse transcribed and integrated (B). Gene conversion creates regions of 22 bp identity before and after the inserted gene in C. Recombination creates units with 2 tRNA gene units in D. Evolutionary differentiation creates *SCA* and *B* modules and some mutations in the central parts of the 22 bp regions in E. Continued recombination creates polymorphism and inverted forms. Transcription induced from the outside pol I promoter (F) and retrotransposition creates Cp1 elements surrounded by 10 to 14 bp target site duplications in G.

support for predictions that non-LTR retroelements like SINEs, not coding for enzyme needed for transposition, are using the protein coded by other non-LTR elements (Eickbush, 1992). This conclusion is possible because of the highly sequence-specific insertion of the R2 non-LTR retrotransposon and the recovery of the sequence around the insertion site in slightly modified form in both Cp1 modules.

Materials and Methods

Cloning and sequencing

Genomic DNA from *C. pallidivittatus* was restricted with *Hin*dIII and ligated into *Hin*dIII restricted λNM1149 (Murray, 1983). The library was screened with a subclone to pCp627, which is a clone that originates in one of the centromeres of *C. pallidivittatus* (Rovira *et al.*, 1993). The subclone is pCp254, which contains G + C-rich sequences (region *AB* in Figure 1A) upstream of 155 bp repeats. The same kind of screening was done with a *Sau*3A partial digest of *C. tentans* genomic DNA cloned into EMBL3. All sequences were determined with the dideoxy chain termination method.

In situ hybridization

Chromosome squashes were hybridized in 4 × SSC (SSC is 0.15 M NaCl, 0.015 M Na acetate, pH 7.0) at 58°C overnight with biotin or digoxigenin labelled probe. For biotin labelling about 100 ng of linear DNA was denatured at 100°C for 10 minutes, followed by incubation with dATP + dCTP + dGTP, 1 μl 0.5 mM solution of each and 1.6 μl of a 1:1 mixture of 0.5 mM dTTP and 0.5 mM biotin-16-dUTP (Boehringer), 2 μl hexanucleotide mixture (Boehringer) and 1 μl Klenow fragment, 2 units/μl, labelling grade (Boehringer) at 37°C for at least two hours. Hybridized probe was detected in a three-step procedure with avidin-FITC, biotinylated anti-avidin IgG and avidin-FITC (Scherthan *et al.*, 1992). Labelling with digoxigenin-dUTP (Boehringer) was done according to the protocol of the manufacturer and detection with rhodamine-labelled anti-digoxigenin (Boehringer) at a concentration of 1:250.

Acknowledgements

The work was supported by grants from the Swedish Cancer Society and the Philip-Sörensen Foundation. S.R.-P. was supported by the Brasilian Research Council (grant no. 2039997-89-1). We are indebted to Erwin R. Schmidt, University of Mainz for a gift of the EMBL3 library of *C. tentans* DNA and to Thomas Hankeln, University of Mainz for valuable suggestions.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bogenhagen, D. F., Sakonju, S. & Brown, D. D. (1980). A control region in the center of the 5S RNA gene directs specific initiation of transcription. II. The 3' border of the region. *Cell*, **19**, 27–35.

Burke, W. D., Calalang, C. C. & Eickbush, T. H. (1987). The site-specific ribosomal insertion element type II of *Bombyx mori* (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol. Cell. Biol.* **7**, 2221–2230.

Daniels, G. R. & Deininger, P. L. (1983*a*). Repeat sequence families derived from mammalian tRNA genes. *Nature (London)*, **317**, 819–822.

Daniels, G. R. & Deininger, P. L. (1983*b*). A second major class of Alu family repeated DNA sequences in a primate genome. *Nucl. Acids Res.* **11**, 7595–7610.

Defranco, D., Burke, K. B., Hayashi, S., Tener, G. M., Miller, R. C. Jr & Soell, D. (1982). Genes for Lys-tRNA-5 from *Drosophila melanogaster*. *Nucl. Acids Res.* **10**, 5799–5808.

Deininger, P. L. (1989). SINEs: short interspersed repeated DNA elements in higher eukaryotes. In *Mobile DNA* (Berg, D. H. & Howe, M. M., eds), pp. 619–636, American Society of Microbiology, Washington, DC.

Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. & Schmid, C. W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.* **151**, 17–33.

Duncan, C. H., Jagadeeswaran, P., Wang, R. R. C. & Weissman, S. M. (1981). Structural analysis of templates and RNA polymerase III transcripts of Alu family sequences interspersed among the human b-like globin genes. *Gene*, **13**, 185–196.

Eickbush, T. H. (1992). Transposing without ends: the non-LTR retrotransposable elements. *New Biol.* **4**, 430–440.

Fuhrman, S. A., Deininger, P. L., LaPorte, P., Friedmann, T. & Geiduschek, E. P. (1981). Analysis of transcription of the human Alu family ubiquitous repeating element by eukaryotic RNA polymerase III. *Nucl. Acids Res.* **9**, 6439–6456.

Hardison, R. C. & Printz, R. (1985). Variability with the rabbit C repeats and sequences shared with other SINEs. *Nucl. Acids Res.* **13**, 1073–1088.

Haynes, S. R. & Jelinek, W. R. (1981). Low molecular weight RNAs transcribed in vitro by RNA polymerase III from Alu-type dispersed repeats in Chinese hamster DNA are also found in vivo. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 6130–6134.

Jakubzak, J. L., Burke, W. D. & Eickbush, T. H. (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 3295–3299.

Luan, D. D., Korman, M. H., Jakubzak, J. L. & Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.

Meng, Y. B., Stevens, R. D., Chia, W., McGill, S. & Ashburner, M. (1988). Five glycyl tRNA genes within the noc gene complex of *Drosophila melanogaster*. *Nucl. Acids Res.* **16**, 7189.

Murphy, M. H. & Baralle, F. E. (1983). Directed semisynthetic point mutational analysis of an RNA polymerase III promoter. *Nucl. Acids Res.* **11**, 7695–7700.

Murray, N. E. (1983). Phage Lambda and molecular cloning. In *The Bacteriophage Lambda* (Hendrix, R., Weisberg, R., Stahl, F. & Roberts, J., eds), vol. 2, pp. 395–432, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Okada, N. (1991). SINEs. *Curr. Opin. Genet. Develop.* **1**, 498–504.

Potter, S., Truett, M., Phillips, M. & Maher, A. (1980). Eucaryotic transposable genetic elements with inverted terminal repeats. *Cell*, **20**, 639–647.

Rogers, J. H. (1985). The origin and evolution of retroposons. *Int. Rev. Cytol.* **93**, 187–279.

Roiha, H., Miller, J. R., Woods, L. C. & Glover, D. M. (1981). Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in *D. melanogaster*. *Nature (London)*, **290**, 749–753.

Rovira, C., Beermann, W. & Edström, J.-E. (1993). A repetitive DNA sequence associated with the centromeres of *Chironomus pallidivittatus*. *Nucl. Acids Res.* **21**, 1775–1781.

Sakamoto, K. & Okada, N. (1985). Rodent type 2 Alu family, rat Identifier sequence, rabbit C family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. *J. Mol. Evol.* **22**, 134–140.

Scherthan, H., Köhler, M., Vogt, P. von Malsch, K. & Schweizer, D. (1992). Chromosomal in situ hybridization with double-labeled DNA: signal amplification at the probe level. *Cytogenet. Cell Genet.* **60**, 4–7.

Tautz, D., Hancock, J. M., Webb, D. A., Tautz, C. & Dover, G. A. (1988). Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol. Biol. Evol.* **5**, 366–376.

Ullu, E. & Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature (London)*, **312**, 171–172.

Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55**, 631–661.

Edited by J. Karn

(Received 22 August 1994; accepted 29 September 1994)