Birth of a Retroposon: The Twin SINE Family from the Vector Mosquito Culex pipiens May Have Originated from a Dimeric tRNA Precursor

Cédric Feschotte, Nicolas Fourrier, Isabelle Desmons, and Claude Mouchès

Laboratoire Ecologie Moléculaire et Faculté Sciences et Techniques Côte-Basque, Université de Pau et des Pays de l’Adour, Pau, France

SINEs are short interspersed repetitive elements found in many eukaryotic genomes and are believed to propagate by retroposition. Almost all SINEs reported to date have a composite structure made of a 5′ tRNA-related region followed by a tRNA-unrelated region. Here, we describe a new type of tRNA-derived SINEs from the genome of the mosquito Culex pipiens. These elements, called Twins, are ~220 bp long and reiterated at approximately 500 copies per haploid genome. Twins have a unique structure compared with other tRNA-SINEs described so far. They consist of two tRNA-Alu-related regions separated by a 39-bp spacer. Other tRNA-unrelated sequences include a 5′ leader that precedes the left tRNA-like unit and a short trailer located downstream of the right tRNA-like region. This 3′ trailer is a 10-bp sequence that is ended by a TTTT motif and followed by a polyA tract of variable length. The right tRNA-like unit also contains a 16-bp sequence which is absent in the left one and appears to be located in the ancestral anticodon stem precisely at a position expected for a nuclear tRNA intron. According to this singular structure, we hypothesize that the Twin SINE family originated from an unprocessed polymerase III transcript containing two tRNA sequences. We suggest that some peculiar properties acquired by this dicistronic transcript, such as a polyA tail and a 3′ stem-loop secondary structure, promote its retroposition by increasing its chances of being recognized by a reverse transcriptase encoded elsewhere in the C. pipiens genome.

Introduction

Retroposons are DNA sequences generated by the reverse transcription of RNA and reintegrated into the genome (Weiner, Deininger, and Efstratiadis 1986). This process is widely spread among eukaryotes (Weiner, Deininger, and Efstratiadis 1986; Xiong and Eickbush 1990; Malik, Burke, and Eickbush 1999), so retroposons, being represented by retrogenes as well as short and long interspersed elements (SINEs and LINEs), often represent a large fraction of their genomes. For example, over 30% of the human genome is made of retroposed sequences which have accumulated over a long evolutionary period (Smit 1999).

Retroposons have long been considered selfish DNA, but a growing number of examples indicate that some of them can play major roles in genome evolution. They can mediate chromosome rearrangements (Brosius 1991; Schmid 1998), provide or define regulatory domains for gene expression (McDonald 1995; Britten 1996; Willoughby, Vilalta, and Oshima 2000), give rise to new genes or new gene regions (Brosius 1991, 1999; Long, Wang, and Zhang 1999), or even assume a cellular function (Pardue et al. 1996; Schmid 1998). Thus, retroposition, being an important mediator of genomic plasticity, has emerged as a major evolutionary force.

Retrogenes are retroposed derived from a messenger RNA transcript (Weiner, Deininger, and Efstratiadis 1986). They are usually found in low copy numbers and are generally nonfunctional because they lack their original regulatory elements. Therefore, they are doomed to degenerate by neutral drift unless they integrate near sequences which can promote their transcription (Weiner, Deininger, and Efstratiadis 1986; Brosius 1991) or they become part of a new gene (Brosius 1999; Long, Wang, and Zhang 1999).

SINEs define another group of retroposons, which are 100–400-bp sequences derived from small structural RNA genes transcribed by RNA polymerase III (pol III) (Deininger 1989; Okada 1991). Consequently, unlike retrogenes, reintegrated SINE copies retain their own internal promoter (A and B boxes) and can potentially give rise to new transcripts capable of further retroposition (Deininger 1989; Schmid 1998; Weiner 2000). Consequently, SINE families can be represented in very high copy numbers in genomes.

One of the most prolific SINE families, the primate Alu family, is present in up to one million copies in the human genome (Smit 1999). Most Alus are about 300 bp long and are composed of two imperfect monomeric repeats. The original monomers were derived from 7SL RNA, one of the components of the signal recognition particle (Ullu and Tschudi 1984; Quentin 1992). Since the dimerization of the ancestral Alu element, the two monomers diverged, and only the left monomer has retained a functional pol III promoter (Deininger 1989; Schmid and Maraia 1992; Schmid 1998). Like most retroposed sequences, Alus are ended by a polyA stretch and flanked by target site duplications, reflecting integration at a staggered DNA break (Weiner, Deininger, and Efstratiadis 1986).

With the exception of the primate Alu and rodent B1 elements, all SINEs described to date are related to tRNAs (Okada 1991; Shedlock and Okada 2000). tRNA SINEs share three distinct regions: a 5′ tRNA-related region containing the internal pol III promoter, a tRNA-unrelated region, and a 3′ tail which is AT-rich or com-
posed of simple repeats (Okada 1991). While 7SL-derived SINEs are found only in primate and rodent genomes, tRNA-derived SINEs have been described in a wide range of organisms, including vertebrates, invertebrates, plants, and fungi (reviewed in Shedlock and Okada 2000).

Since SINEs lack coding capacity, it is obvious that their retroposition depends on reverse transcriptase produced elsewhere in the genome. Several lines of evidence suggest that SINEs may have borrowed the retrotransposition machinery of autonomous LINES, which can code for reverse transcriptase (RT) and endonuclease activities. Indeed, the 3′ ends of several tRNA-derived SINEs share sequence homology with the 3′ end of a LINE present in the same organism (Oshshima et al. 1996; Okada et al. 1997; Gilbert and Labuda 1999; Ogawa et al. 1999). Hence, the LINE-encoded RT might be able to recognize the 3′ end of the SINE transcript and initiate cDNA synthesis. The 3′ end of an Alu does not share significant sequence similarity with any LINE identified so far. Nevertheless, Alu flanking sequences share homology with the target motif recognized and cleaved by the human L1 LINE endonuclease, which suggests an intimate relationship between Alu and L1 (Boeke 1997; Jurka 1997).

Here, we report the characterization of a SINE family named Twin from the vector mosquito Culex pipiens. High sequence conservation between Twin copies, as well as their distribution among culicine mosquitoes, suggests a relatively recent amplification history for this SINE family. Interestingly, the structure of Twin defines a new type of SINE, sharing two tRNA-related regions separated by a 39-bp spacer and followed by a short polyA tract. Based on primary- and secondary-sequence analysis, we propose a scenario for the origin of this new type of SINEs involving reverse transcription of a dimeric tRNA precursor.

Materials and Methods
Mosquito Strains and Genomic DNAs

The first Twin-Cp1 copy was identified in a λ clone previously isolated from a genomic library of the Tem- R strain of C. pipiens (California). The Twin-Cp2 element was isolated from the MSE strain of C. pipiens (France). All other copies were from the Ravenna strain of C. pipiens (Italy). For Southern and PCR experiments, we also used genomic DNAs from the C. pipiens strains Idron (collected in the field, south of France), Montpellier (collected in the field, south of France), Frankfurt (collected in the field, Germany), Pro-R, Pat, Willow (California), C. pipiens cells (Taiwan), Culex hortensis, Aedes triseriatus cells (Trois Rivières, Canada), A. albopictus Oahu 71 (Hawaii), Aedes aegypti Hanói (Vietnam), Anopheles stephensi (obtained from the MNHN, Paris), Toxorynchites emboinensis (Polynésie, ORSTOM), and nonculicid dipterans Drosophila melanogaster (Canton strain) and Ceratitis capitata (collected in the field, Italy). Total genomic DNA was prepared from adult insects as described previously (Mouchès et al. 1986).

Southern Blot Hybridization of Genomic DNA

Aliquots of 10 μg of genomic DNA were digested to completion with EcoRI restriction endonuclease. Resulting fragments were separated on 1% agarose gels, transferred to a Nytran membrane (Amersham Pharmacia Biotech, Upsala, Sweden) and hybridized at high stringency (65°C) with radiolabeled probes. Other procedures were as previously described (Mouchès et al. 1990). Twin probes were obtained by PCR amplification from a plasmid carrying the Twin-Cp1 copy using primers TP1 and TP2 (see below), gel-purified, and labeled with α32P-dCTP by random priming (Amersham Pharmacia Biotech).

PCR Amplification of Twin-Related Elements in Several Culex Species

Genomic DNA (~10 ng) from various C. pipiens strains and several insect species were subjected to PCR amplification using a pair of Twin internal primers (TP1: 5′-CCGAGCTWCCGTGGCCGTGA-3′; TP2: 5′-TCCCCGTACGAGMATCGAAGACT-3′). PCR reactions were performed according to standard procedures, and cycling conditions were as follows: 5 min at 94°C, followed by 30 cycles of 45 s at 94°C, 90 s at 55°C, and 60 s at 72°C, followed by a final 10-min elongation at 72°C. PCR products were analyzed on agarose gels, and those related to Twin were identified by Southern hybridization using a Twin-Cp1 probe.

Isolation of Additional Twin Copies from a C. pipiens Genomic Library and Estimation of Twin Copy Number

A library was prepared by complete EcoRI digestion of genomic DNA from the Ravenna strain of C. pipiens and ligation into a λ-gt11 cloning vector (Stratagene, La Jolla, Calif.). About 20,000 recombinant phages were plated and screened using a Twin-Cp1 probe. Prehybridization, hybridization, and washing were carried out at 65°C as previously described (Mouchès et al. 1990). After a first round of screening, a large number of positives were obtained. Several positive plaques were plugged in SM buffer and amplified, and each was used as a template for PCR amplification with primers for the arms of the λ-gt11 vector. PCR parameters were the same as those described above except that the annealing temperature was reduced to 54°C and the elongation time was increased to 2 min 30 s. PCR products containing Twin elements were identified by Southern hybridization with a Twin-Cp1 probe, gel-purified, and subcloned into pCR-TOPO plasmid vectors (Invitrogen, Groningen, the Netherlands).

Copy number for Twin elements was estimated based on the ratio of positive plaque plaques to the total number of plaques screened, taking into account the haploid genome size of C. pipiens of 540 Mb (Black and Rai 1988) and an average 4-kb insert size of the genomic library.
Sequence Analysis

Sequence analysis was done by the Eurogentec sequencing department (Seraing, Belgium) with synthetic primers, using an ABI-377 automatic sequencer. Most sequence analysis was done with tools available at the Infobiogen server (http://www.infobiogen.fr). Database searches were performed with BLASTN (Altschul et al. 1997) using default parameters. Multiple-sequence alignments were constructed by CLUSTAL W, version 1.7 (Thompson, Higgins, and Gibson 1994), using default parameters. The ability of twins to form secondary structures was estimated by the M. Zuker DNA and RNA mfold programs, available through the server http://mfold.wustl.edu. We also used the S. Eddy tRNAscan-SE program (Lowe and Eddy 1997) to assess the presence of tRNA-like sequences in twin elements (http://www.genetics.wustl.edu/eddy/tRNAscan-SE).

Results

Discovery of the Twin Family of Repetitive Elements in the C. pipiens Genome

The first twin copy, twin-Cp1, was discovered as a 215-bp sequence inserted into the second intron of the C. pipiens homolog of the Drosophila white gene. The C. pipiens white gene was cloned from a genomic library of the California Tem-R strain (unpublished data). The 215-bp sequence was PCR-amplified and used as a probe in Southern experiments against C. pipiens genomic DNA. A long continuous smear was obtained (data not shown), showing that this insertion sequence belongs to a family of interspersed repetitive elements. By using sequence analysis, we identified a second member of this repeat family, twin-Cp2, in another C. pipiens lambda clone previously isolated from a genomic library of the French MSE strain of C. pipiens. Twin-Cp2 shares 90.7% similarity with the 215-bp insertion sequence found in the white gene, while the DNA flanking the two elements shares no obvious similarity. We conclude that these elements are members of the same family of interspersed repeats from the C. pipiens genome that we called the twin family.

Structure of twin Elements

Sequence analysis of the C. pipiens MSE clone reveals that twin-Cp2 is inserted in a tandem repeat sequence named TRCp. TRCp units are 116 ± 1 bp long and well conserved in sequence, with pairwise identity between units ranging from 84% to 97%. Based on sequence analysis, it is obvious that TRCp4 is the “youngest” tandem unit (not shown). Thus, integration of twin-Cp2 in TRCp4 can be considered a relatively recent event. This insertion allows us to define the boundaries of this twin copy by comparing sequences of the four tandem repeats (fig. 1). The insertion sequence in TRCp4 is 229 bp long and is ended by a 14-bp pure A stretch.

Twin-Cp2 and twin-Cp1 have no coding capacity and no specific terminal sequence arrangements like inverted terminal repeats, which characterize transposons moving via a DNA intermediate. Rather, the polyA tract at the 3’ end of twin-Cp2 is reminiscent of the end of retroposoned DNA sequences (Weiner, Deininger, and Efirstradiis 1986). Therefore, it appeared that twins might belong to a new family of non-LTR retroelements, namely, a SINE or a LINE family.

In order to define the structure of twin elements, we isolated additional copies by screening a C. pipiens genomic library using the 215-bp twin-Cp1 element as a probe. Four positive phage clones were randomly chosen and further characterized. Each genomic clone contained one copy of the repeat family. According to the alignment shown in figure 2, twin elements can be defined as a 217-bp consensus sequence terminated by a TTTT motif and followed by a variable number (0–13) of A residues. One element, twin-Cp6, lacks 124 bp at its 3’ end, and twin-Cp3 and twin-Cp5 are slightly truncated at their 5’ ends. However, sequence of truncated copies is as well conserved as the “full-length” copies. Excluding deleted regions, pairwise similarity between twin copies ranges from 83% to 96%.

Retroposons are frequently surrounded by short direct repeats (~5–20 bp) due to integration at staggered chromosomal breaks (Weiner, Deininger, and Efirstradiis 1986). No obvious target site duplications are recognizable in genomic DNA flanking twin copies. Nevertheless, twin-Cp2 is flanked by the sequence AAAA-CAAAA at its 5’ end, and its 3’ polyA tract is much longer than those of other twin copies. Therefore, part of the polyA tract might represent a 2–8-bp target site duplication as well (see fig. 2). Alternatively, target site duplications could be very short (1–3 bp), or twin elements might not integrate at staggered chromosomal breaks. Otherwise, it is possible that twin copies were frequently integrated via the host recombination machinery. Twin elements analyzed in this study are all surrounded by AT-rich DNA, except twin-Cp5, which is flanked by a 3’ GC-rich sequence (fig. 2). Further analysis revealed that this GC-rich sequence represents one of the terminal inverted repeats of a putative miniature transposable element inserted within the 3’ AT-rich end or immediately downstream of twin-Cp5 (data not shown). It is noteworthy that the six twin elements are all found in genomic regions which are highly enriched in transposable elements (unpublished data).
Copy Number and Distribution of Twin Elements in Dipteran Insects

The copy number of the Twin elements in the C. pipiens genome was estimated by screening a genomic library from the Ravenna strain with Twin-Cp1 as a probe. Based on the ratio of positive plaques to the total number of plaques screened and assuming a haploid genome size for C. pipiens of 540 Mb (Black and Rai 1988), the copy number of Twin elements is \( \sim 500 \) per haploid genome.

We used PCR with two specific internal primers for the Twin family to investigate the presence of related sequences in genomic DNA of several C. pipiens strains and various dipteran species, including Aedes and Anopheles mosquitoes. A single strong band of the expected size (~200 bp) was obtained in all Culex strains analyzed, as well as in the close relative species C. hortensis (fig. 3, upper panel). The identification of PCR products as members of the Twin family was confirmed by hybridization of PCR products with a Twin-Cp1 probe (fig. 3, lower panel). No amplification was detected from dipterans outside the genus Culex. These findings were corroborated by Southern hybridization of total genomic DNA digests from the same insect species and from additional Culex species using the Twin-Cp1 probe. Again, hybridization signals were obtained only for Culex species. Besides, some variations in the banding pattern suggest that several Twin insertions may be polymorphic among C. pipiens strains (data not shown).

Twins Contain Two tRNA-Related Regions

Twin elements have no coding capacity for a protein. However, a computer-assisted search in DNA databases using the Twin consensus sequence as a query revealed that the 5' region (positions 6–78) shares significant nucleotide similarity (56%–67%) with tRNAArg genes from various organisms and with the tRNA-related regions of several SINEs from the AFC family of Cichlidae fishes (Takahashi et al. 1998). Interestingly, sequence similarity between Twin and AFC is not restricted to the pol III promoter boxes, but is even higher in the region located between the two boxes (fig. 4B). This feature does not necessarily imply a phylogenetic relationship between the two SINE families, but suggests that they may be derived from the same species of tRNA, namely, tRNAArg.

In addition, a short region located near the 3' end (positions 174–204 in the consensus), displays up to 85% similarity to the 3' ends of several tRNA genes. Further sequence analysis showed that Twins are indeed dimeric in structure, being broadly composed of two related units separated by a 39-bp sequence (fig. 4A). Both Twin units can be well aligned except for a 16-bp sequence which is absent in the left unit. When this 16-
bp sequence is removed, both units share significant sequence similarity to tRNAArg genes from various organisms and to the 5' tRNA-related region of AFC SINEs (fig. 4B). Accordingly, both tRNA-like regions can be folded into cloverleaf secondary structures similar to those established for tRNAArg (fig. 4C). Strikingly, most of the invariant and semi-invariant residues in the “universal” tRNA structure (according to Sprinzl et al. 1987) are still present in Twin tRNA-like monomers (fig. 4C). Finally, the left unit is still predicted as a tRNA gene by the tRNAScan-SE program using default parameters (score 38.3). Guided by these analyses, we conclude that Twin is a new family of tRNA-derived SINEs containing two tRNAArg-related regions. The left tRNA-like unit spans from position 6 to position 78 in the Twin consensus sequence, and the right one spans from nucleotide 118 to nucleotide 207 (figs. 2 and 4A).

It is noteworthy that the 16-bp insertion sequence found in the right tRNA-like region of Twins is located in the anticodon loop, 1 bp 3' of the putative anticodon (fig. 4B and C), a position identical to those of eukaryotic tRNA introns (Ogden, Lee, and Knapp 1984; Abelson, Trotta, and Li 1998). Moreover, the size of this insertion sequence fits well with those of eukaryal tRNA intervening sequences, which range from 14 to 60 nt (Abelson, Trotta, and Li 1998).

Until recently, introns in tRNA genes were thought to be very rare in higher eukaryotes, since they had been detected only in tRNA genes coding for tRNA Tyr and tRNA Lys (Arends, Kraus, and Beier 1996). However, introns have now been identified in tRNA Met from plants (Akama and Kashihara 1996), in tRNA Lys genes from mollusks (Matsuo et al. 1995), and in a human tRNAArg gene (Bourn et al. 1994). By searching current DNA databases, we found three human tRNAArg genes that contain an intron as well as several tRNAArg genes without introns. Introns all are located 1 nt downstream of the anticodon, range from 14 to 18 bp, and are highly variable in sequence (fig. 5). Interestingly, the only conserved nucleotide is the first G residue, which is also the first nucleotide of the 16-bp sequence interrupting the left tRNAArg-like region of Twins (fig. 5). Together, these data strongly suggest that the right tRNA-like region of Twins may have derived from an intron-containing tRNAArg gene. Furthermore, this implies that the two tRNA-related regions are derived from two distinct tRNAArg cistrons.

Discussion

Twin Is a Novel SINE Family from the Vector Mosquito C. pipiens

We have characterized a family of repetitive DNA elements called Twin from C. pipiens. One member of this family was recently integrated into a copy of a tandem repeat sequence. Analysis of additional copies shows that Twins possess some features that define the SINE class of retroposons, including a short size (~220 bp), the presence of consensus motifs for pol III promoter (A and B boxes), and a 3' polyA tract. We estimated that there were at least 500 Twin copies per haploid genome and we found that this family was present in all C. pipiens strains analyzed, as well as in the closely relative C. hortensis. We were unable to detect any Twin-related element in Aedes species, which are members of the same subfamily, Culicinae. We conclude that the Twin family arose specifically in the lineage leading to the genus Culex (fig. 3).

Consistent with their relatively recent origin, the six Twin copies isolated from the C. pipiens genome share an average sequence divergence of 15%. Assuming that the substitution rate for retroposons is similar to those defined for Drosophila pseudogenes (1.5%/Myr; Petrov et al. 2000), a major amplification of Twin SINEs in the C. pipiens genome may have occurred approximately 10 MYA. Furthermore, several preliminary results indicate that intraspecific dimorphism exists for some Twin insertions among different populations (data not shown), which suggests that Twin amplification might be an ongoing process in some C. pipiens strains. Dimorphic SINE insertions are potentially a rich source of genetic markers for population biology studies, as
was previously illustrated for the SINEs of some vertebrate species (Batzer et al. 1994; Hamada et al. 1998). Given the current recrudescence of mosquito-transmitted diseases, the development of powerful genetic markers is of major importance for a better understanding of the population structure and dynamics of each vector mosquito species in the field, and thus for better control of these insects.

### Origin of the Twin Family of tRNA-Derived SINEs

*Twin* is the first SINE family to be described from the genome of the vector mosquito *C. pipiens*. However, *Twins* are atypical SINEs in terms of their structure, consisting of two related regions, both similar to a tRNA Arg gene, separated by a 39-bp sequence (fig. 4A). Therefore, *Twins* share a dimer-like structure with two
sets of potential pol III promoters (see below). Most SINEs described so far possess a single tRNA-related region located in their 5’ half, while their 3’ half is made up of a tRNA-unrelated region followed by a polyA tail or short tandem repeats (Shedlock and Okada 2000).

Other multimeric SINEs include the primate Alus (Deininger 1989; Quentin 1992), the chironomid insect Cp1 elements (He et al. 1995), and the zebrafish DNA elements (Izsvák et al. 1996). It is believed that all of these elements arose by multimerization of at least two ancestral retroposons through a mechanism that remains unclear. In the case of Cp1, the two tRNA-related modules are tandemly arranged, and both start with a 22-bp sequence strikingly similar to the insertion site of the R2 LINE in the 28 S preribosomal gene. According to this structure, it is hypothesized that Cp1 arose by duplication of an ancestral tRNA retrogene integrated into the R2 insertion site (He et al. 1995). Alu monomers are also tandemly arranged, and it is proposed that the progenitor of the dimeric Alu family is the result of the fusion of a free left monomer (FLAM) with a right monomer (FRAM). Indeed, FLAM and FRAM elements are still present in the genome but are found at lower copy numbers than the dimeric Alu (Quentin 1992). Each monomer originated from an ancestral retroposon (FAM) which has been derived from 7SL RNA (Ullu and Tschudi 1984; Quentin 1992). Consequently, both FLAM and FRAM elements are present in the genome but are found at lower copy numbers than the dimeric Alu (Quentin 1992).

Although we cannot rule out the possibility that such recombinational events lead to the Twin structure, we prefer an alternative scenario for the origin of this SINE family for the following reasons. First, unlike Cp1 and Alu, the two related Twin units are not truly tandemly arranged, since they are separated by a 39-bp sequence. Moreover, this spacer sequence found between the two tRNA-like regions of Twins is not particularly A-rich. Thus, it seems unlikely that it represents a “fossil” of a polyA tail from an ancestral tRNA retrogene.

What is the origin of this 39-bp sequence? According to our hypothesis, it may correspond to the DNA region ancestrally separating two tRNA^{Arg} genes. In other words, we believe that the structure of Twin SINEs reflects the ancient clustered organization of two tRNA^{Arg} genes. It is known that many nuclear tRNA genes are frequently clustered in the same chromosomal region. For example, 10 tRNA genes are clustered within a 1.9-kb chromosomal region in Leishmania tarentolae (Shi, Chen, and Suyama 1994), 4 tRNA^{Arg} genes are found within a 1-kb region of the D. melanogaster genome (GenBank accession number L09196), and a Xenopus laevis tRNA gene cluster contains a tRNA^{Phe} and a tRNA^{Tyr} separated by only 72 bp of DNA (Hosbach, Silberklang, and McCarthy 1980). These genes are organized as individual transcriptional units, since each gene contains its own internal pol III promoter, and a termination signal for pol III (i.e., at least four consecutive T residues) is present in the downstream sequence of each gene. However, this rule has often been found to be broken in yeast. In this organism, two tRNA genes can be cotranscribed into dimeric precursors and then are processed into two mature tRNAs. To date, two examples of such polycistronic tRNA transcripts are known: a Saccharomyces cerevisiae tRNA^{Arg}-tRNA^{Ap} precursor, in which the two genes are separated by a 10-bp spacer (Schmidt et al. 1980), and a Schizosaccharomyces pombe dimeric precursor, which consists of an intron-containing tRNA^{Arg} gene and a tRNA^{Met} gene separated by a 7-bp spacer (Mao, Schmidt, and Soll 1980).

We believe that such a dimeric tRNA precursor could have been produced in C. pitiens as well and might have given rise to the Twin SINE family. Consistent with this hypothesis, the pol III termination motif (four or more T residues) is absent from the Twin 39-bp spacer while being present at the 3’ end of the Twin consensus, downstream of the right tRNA-like unit. Such a motif also agrees with the polyU sequence typical of the 3’ end of pol III transcripts (Bogenhagen and Brown 1981). Therefore, the 10 nt found downstream of the right tRNA-like region may correspond to a “relic” of the 3′ trailer of a tRNA precursor. Similarly, the 5 nt located upstream of the left tRNA-like unit could represent the short 5′ leader of a tRNA precursor. In this regard, the presence of a 16-bp intervening sequence in the ancestral downstream tRNA^{Arg} (fig. 4B and C) is in agreement with previous reports, showing that splicing can be a relatively late event in tRNA maturation and often occurs after end-processing (Bertrand et al. 1998; Wolin and Matera 1999). Taken together, these data are consistent with the idea that Twin SINEs have originated from an unprocessed dimeric pol III transcript containing two related, but distinct, tRNA cistrons.

Nevertheless, we have no indication that such a dimeric precursor could have ever been efficiently processed into functional tRNAs in the mosquito genome. Indeed, such a cotranscriptional event can be viewed as
accidental, possibly resulting from mutations in the termination signal for the upstream tRNA gene. Consequently, many structural features of the aberrant dimeric transcript might have prevented its maturation but, in the same way, could have increased its chances of becoming an efficient template for a reverse transcriptase (see below).

Are Twin SINEs Amplified Through an RNA Intermediate?

Our model for the origin of Twin SINEs involves an ancestral retroposition event of an unprocessed pol III transcript. This event could be considered very unusual, since retroposons are generally derived from fully processed transcripts (Weiner, Deininger, and Efstratiadis 1986), although some exceptions are well known (Weiner, Deininger, and Efstratiadis 1986; Brosius 1999). This also raises the possibility that Twin amplification could have occurred through a DNA intermediate. Yet, several features indicate that, rather, Twins were most likely to be generated by retroposition.

The first step in retroposition is transcription of the entire DNA element by RNA polymerase III. Consequently, retroelements usually contain an internal promoter. While Twins diverge significantly from their ancestral tRNA progenitors, the left tRNA-related region still has well conserved A and B boxes, i.e., a potential internal promoter for RNA polymerase III (fig. 4B). In addition, the polyA termination signal for RNA polymerase III is found at the 3' end of the Twin consensus sequence and nowhere else in the sequence. Thus, it is plausible that Twin source genes could be transcribed by RNA polymerase III. We were able to detect Twin transcripts of the expected size (approximately 220 bp) by Northern blot analysis, showing that Twin is efficiently transcribed in vivo (data not shown). However, additional studies are needed to determine if Twin is actually transcribed by polymerase III.

The second step in retroposition involves recognition of the 3' end of the retroposon RNA by an RT, followed by first-strand cDNA synthesis (Luan et al. 1993; Kazazian and Moran 1998). Because first-strand synthesis is often an incomplete process, many 5' -truncated LINEs and SINEs are reintegrated in the genome (Weiner, Deininger, and Efstratiadis 1986; Luan et al. 1993; Takasaki et al. 1994; Kazazian and Moran 1998). It is noteworthy that two out of the six Twin copies randomly isolated from the C. pipiens genome are slightly truncated at their 5' ends (Twin-Cp3 and Twin-Cp5; fig. 2). This suggests that these copies may be the products of incomplete reverse transcription and, by extension, further supports the hypothesis that Twins are retroposed sequences.

What Is the Source of RT for Twin SINEs?

Most SINEs described so far resemble a fusion product of a tRNA-derived sequence with a tRNA-unrelated sequence. In some cases, the tRNA-unrelated region can be further divided into a 5' part and a 3' part, with the latter being derived from the 3' tail of a LINE (Ohshima et al. 1996; Okada et al. 1997; Ogiwara et al. 1999). In this way, it is thought that SINEs can "hijack" the retropositional machinery of the corresponding LINE.

In the case of Twins and in some other cases, such as those of CHR-1 and CHRS families (Shimamura et al. 1999) or the rodent ID and B2 families (Deininger 1989), the 3' tRNA-unrelated region is so short that it appears unlikely that they share extensive similarity with a LINE tail sequence. The same conclusion can be drawn for the primate Alu and rodent B1 elements, since their sequences are derived exclusively from 7SL RNA. Therefore, if we assume that Twins and these other SINEs transpose by using the enzymatic machinery of a partner LINE, it is obvious that additional factors may influence the propensity of these SINE families to be efficiently and frequently recognized by a LINE-encoded RT.

One key feature is probably the secondary or tertiary structure of the SINE transcript. Such structures not only facilitate recognition of and access to the LINE RT, but may also influence SINE transcript stability and localization, as well as priming of reverse transcription (Sinnett et al. 1991; Schmid and Maraia 1992; Boeke 1997; Mathews 1997; Schmid 1998; Brosius 1999). Interestingly, the single-stranded Twin consensus sequence can be potentially folded into an elaborate secondary structure (data not shown and fig. 6). While the Twin left unit has retained a cloverleaf tRNA-like structure, the right tRNA-related region can form a long stem-loop structure including the 16-bp putative intron relic. We do not know if such a structure exists in vivo, but if so, it might reflect a structural evolution of the Twin transcript leading to efficient retroposition. It is also possible that the inability of the right monomer to form a tRNA-like structure stabilizes the dimeric transcript and increases its propensity for retroposition. Indeed, it was shown that the first step in the maturation of the yeast dimeric transcript is endonucleolytic cleavage between the two tRNA sequences (Mao, Schmidt, and Soll 1980; Schmidt et al. 1980). This cleavage is mediated by RNase P, which recognizes the tRNA structure of the downstream tRNA (Pearson et al. 1985). Therefore, the inability of the Twin downstream unit to form a tRNA-like structure may have provided positive selection for Twin by stabilizing its transcript and increasing its chances of being retroposed (R. Maraia, personal communication).

Another key feature which might increase retroposition efficiency resides in the presence of a polyA tail in the retroposon transcript. Indeed, it was shown that the polyA tail of the human LINE L1 transcript was critical for its retroposition, with the L1 RT interacting with the polyA itself rather than with the 3' untranslated region of the L1 transcript (Maraia et al. 1996; Kazazian and Moran 1998; Moran, DeBerardinis, and Kazazian 1999). More recently, it was also shown that L1 products are able to generate retropseudogenes (Esnault, Maestre, and Heidmann 2000). These findings reveal that there is no primary RNA sequence specificity for L1-mediated retroposition events, which further su-
ports the hypothesis that L1 LINEs are the most probable candidate to mediate Alu retroposition. It is believed that the presence of a polyA tail in Alu RNAs, probably in concert with some structural properties, may greatly increase their chances of being recognized and reverse-transcribed by the L1 enzymatic machinery (Boeke 1997; Schmid 1998; Weiner 2000). In a similar manner, we speculate that acquisition of a polyA tail by the ancestral Twin transcript may have contributed to its reverse transcription. Although polyadenylation of such a putative pol III transcript might be considered aberrant, it has often been reported for several stable RNAs (Yokobori and Pääbo 1997; Li, Pandit, and Deutscher 1998; Komine et al. 2000). Besides, this acquisition is very likely to have taken place at the RNA level, which further argues that Twin SINEs arose by retroposition.

As discussed by Okada et al. (1997), there might be two different type of LINEs, a stringent type and a relaxed type. L1 may belong to the relaxed type of LINEs, for which the 3’ region is not required for retroposition (Kazazian and Moran 1998), and the recognition specificity by RT became relaxed or changed from the 3’ end tail to the polyA stretch (Boeke 1997; Weiner 2000). These can explain why in mammals there are so many SINEs and pseudogenes ending in a polyA stretch. The present report of a SINE family lacking an obvious 3’ tail in the Culex genome provides evidence that some relaxed LINEs may also exist in an insect genome. L1-like elements have been described in a wide range of eukaryotes, ranging from plants to higher vertebrates, and are considered one of the oldest LINE clades (Makrikakis, Burke, and Eickbush 1999). Although to date no L1-like LINEs have been described from C. pipiens, it is very likely that some are present in its genome. Alternatively, it is possible that some LINEs belonging to other clades could encode for an RT that is “relaxed,” i.e., able to recognize the polyA tail of Twin SINEs. For example, Juan-C elements are polyA-ended LINEs reiterated in more than 2,500 homogeneous copies in the genome of C. pipiens (Agarwal et al. 1993). This suggests recent activity for this LINE family, and some recent data revealed that some Juan-C elements are actively transcribed in mosquito cells (unpublished data). Therefore, it would be very interesting to test in vitro whether Juan-C LINE products can mediate Twin SINE reverse transcription.

Supplementary Material

Nucleotide sequences reported in this paper will appear in the GenBank database under accession numbers AF282724–AF282729. A consensus sequence for Twin SINEs was deposited in Rybase Update (available at http://www.girinst.org).

Acknowledgments

We thank P. Blanchard and S. Karama for technical assistance and R. Kuhn for providing the Frankfurt strain of C. pipiens. We are grateful to R. Maraia for critical comments on the manuscript and generous encouragement. We also thank J.-M. Deragon, D. Engelke, N. Gilbert, N. Pourtau, and J. C. Salvado for helpful
discussion and valuable advice. C.F. was supported by a grant from the Ministère de l’Éducation Nationale, de la Recherche et de la Technologie to the University of Paris 6.

LITERATURE CITED


———. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–134.


PIERRE CAPY, reviewing editor

Accepted September 24, 2000