Site-specific ribosomal DNA insertion elements in *Anopheles gambiae* and *A. arabiensis*: nucleotide sequence of gene-element boundaries

Susan M.Paskewitz* and Frank H.Collins

Malaria Branch, Division of Parasitic Diseases, Center for Infectious Diseases, Centers for Disease Control, Public Health Service, US Department of Health and Human Services, Atlanta, GA 30333, USA

Received July 17, 1989; Revised and Accepted September 14, 1989

EMBL accession nos X15296-X15304 (incl.)

ABSTRACT

The nucleotide sequence of the junctions between the 28S ribosomal gene and site-specific insertion elements from two sibling mosquito species, *Anopheles gambiae* and *A. arabiensis*, is reported. In both species, elements insert at the same point within the 28S gene, but this site is 634 basepairs (bp) 3' of the R1 (Type I) insertion site in *Drosophila melanogaster*. The two mosquito elements each have poly A tails and a polyadenylation signal, but the extreme 3' and 5' ends show no other similarity to each other or to any other insertion element. In both mosquito species, identical target site duplications of 17 bp are generated. The sequence TNTCCCTNT found in this duplication is also found in the 14 bp target site duplications that flank R1 elements in *D. melanogaster*. Another sequence in this duplication, GGGATAACT, is very similar to the sequence GGGAGTAACT found in the 24 base sequence required by the *Bombyx mori* R2 endonuclease.

INTRODUCTION

Interruptions of the 28S ribosomal genes have been studied in detail in several insect species. In *Drosophila melanogaster* (1), *D. virilis* (2), *Calliphora erythrocephala* (3), and *Bombyx mori* (4), R1 (Type I) elements insert at exactly the same position in the 28S gene and generate identical 14-basepair (bp) target site duplications (although these are missing in the larger R1 elements of *D. melanogaster*). R2 (Type II) inserts are less widespread, but in *D. melanogaster* and in *B. mori* these elements insert at the same position in the 28S gene, share a short region of homology at their 3' ends, and have poly A tails (4). The available data indicate that both types of elements are transposable (5, 6, 7, 8).

Mosquitoes are classified as lower Diptera (Nematocera) and are only distantly related to higher flies like *Drosophila* and *Calliphora* (Cyclorrhapha). However, restriction mapping of rDNA in *Anopheles gambiae* demonstrated that this species also has insertion elements in the 28S gene (9). In this paper, we report the nucleotide sequence from the 3' and 5' insertion element-gene boundaries of eight interrupted rDNA clones from *A. gambiae* and one from a sibling species, *A. arabiensis*, and from the corresponding region of uninterrupted clones. The data show that the insertions share characteristics of both R1 and R2 elements but are located downstream of R1 and R2 insertion sites. Similarities between target sites probably indicate related mechanisms for insertion of all these elements.

MATERIALS AND METHODS

Lambda clones with and without insertion elements were isolated from EMBL-3 libraries made from genomic DNA partially digested with Sau3A (10). Clones that contained interruptions were identified by restriction analysis. Previous analysis demonstrated that the insertions occur in a 0.8-kilobase (kb) HindIII fragment from the 3' region of the 28S

sequence (9). The 0.8 kb fragment is flanked by two smaller HindIII fragments (9). Initially, two clones (Agr23 from *A. gambiae* and Aar23 from *A. arabiensis*) which had the two small HindIII fragments but did not have the normal 0.8-kb HindIII fragment were selected for further examination. Seven additional clones from *A. gambiae* were later identified using part of Agr23 (pAgr23H4, see Figure 1) as a probe.

The interrupted and uninterrupted HindIII fragments were subcloned into the phagemid Bluescript (Stratagene Cloning Systems; use of trade names is for identification only and does not constitute endorsement by the Public Health Service or the U.S. Department of Health and Human Services). Fragments were sequenced using the dideoxy chain termination method (11) as modified for double-stranded sequencing (12). The sequencing strategy is summarized in Figure 1. The 5' junction was approximately 80 bp from one HindIII site and thus could be sequenced using commercially available primers. The 3' junction was about 740 bp from the other HindIII site and 240 bp from an internal EcoRI site. The EcoRI/SalI fragment, pAgr12a (Figure 1), had already been sequenced for another purpose and provided sequence for generating a synthetic oligonucleotide primer near the EcoRI site. This primer was used to sequence through the 3' junction.

Southern blots of HindIII-digested genomic DNA from single mosquitoes were prepared as described earlier (10). Mosquitoes from two *A. gambiae* colonies (AS-46 and GMMK6) and two *A. arabiensis* colonies (MAN and GMAL) were examined (see reference 10 for origin of the colonies). Blots were probed with pAgr23H5 or with pAgr23H4 (see Figure 1 for locations of subclones).

Aar23 was digested with KpnI and PstI to release a 1.0-kb fragment that was free from rDNA and plasmid sequence (Figure 1). The 1.0-kb band was cut out of a 0.7% agarose gel, and the DNA was separated from the agarose by adsorption to, and subsequent elution from, glass beads using the GeneClean kit (Bio101). The DNA was then labelled with 32 PdCTP to a specific activity of 10^7 cpm/ μ g. The labelled fragment was used to probe a Southern blot of Agr23 that had been digested with EcoRI and SalI. The same fragment was used to probe a HindIII genomic Southern blot.

An 800-bp HindIII/BamHI fragment that contained the region homologous to that surrounding the R1 and R2 insertion sites was subcloned from Agr12. This fragment was sequenced and was used to probe Southern blots of genomic DNA from single mosquitoes that had been double-digested with HindIII and BamHI.

RESULTS

The restriction maps for two of the interrupted lambda clones (Agr23 and Aar23) selected for study and for one uninterrupted clone from each species (Agr12 and Aar9) are shown in Figure 1. The insertion in Agr23 is approximately 8.2-kb long and contains an internal HindIII site. Two subclones that encompass the insert, pAgr23H4 (4.0 kb) and pAgr23H5 (5.0 kb), were produced from this lambda clone. Three of the other seven insertion element clones from this species (Agr2D, Agr1B, and Agr5B) contained two HindIII fragments of the same length. The other four clones (Agr5D, Agr2A, Agr7C, and Agr7B) contained the 4.0-kb HindIII fragment but the 5.0-kb fragment was truncated to lengths that varied from 1.6 kb to 4.5 kb. All of these HindIII fragments were subcloned and the ends were sequenced (see below).

When HindIII genomic digests were probed with pAgr23H4, only 1 band at 4.0 kb was revealed in *A. gambiae* (in addition to the 0.8-kb band from uninterrupted cistrons), but there were very faint bands at approximately 2.0 and 9.0 kb in *A. arabiensis*. Probing

Figure 1. Subclones and sequences used to identify the rDNA insertion-coding sequence junctions of *Anopheles gambiae* and *A. arabiensis*. Agr12, Agr23, Aar9 and Aar23 are lambda clones of rDNA that contain, respectively, i) an uninterrupted 28S rRNA coding sequence from *A. gambiae*, ii) a 28S rRNA coding sequence containing an 8.2-kb insertion from *A. gambiae*, iii) an uninterrupted 28S rRNA coding sequence from *A. arabiensis*, and iv) a 28S rRNA coding sequence containing a 1.7-kb insertion from *A. arabiensis*. The subclones prepared from these clones are indicated on the map. Arrows beneath the subclones indicate the start points, directions and extents of DNA sequencing. Coding regions are shown as solid blocks. Not all restriction sites are displayed. (B = BamHI; E = EcoRI; H = HindIII; K = KpnI, P = PstI, S = SalI; X = XhoI.)

the same digests with pAgr23H5 revealed several bands in *A. gambiae* that varied between the two colonies observed (Figure 2A). A different set of bands hybridized to this clone in *A. arabiensis* and again variation was observed between the two colonies observed.

The insertion in Aar23 is 1.7-kb long and does not contain any HindIII sites. Thus, one 2.5-kb HindIII subclone (0.8 kb + 1.7 kb), pAar23H2.5, was sufficient for sequencing the junctions in this species (see below). When the 1.0-kb KpnI/PstI fragment (which contained no rDNA sequence) was hybridized to HindIII genomic Southern blots, at least 20 bands were identified in *A. arabiensis*, the predominant band being approximately 7.5-kb long (Figure 2B). In *A. gambiae*, only six bands light up; four of these are very faint, but the predominant band is almost the same size as the *A. arabiensis* predominant band. (Three other members of the *A. gambiae* species complex were also examined in this blot. Only *A. quadriannulatus* displayed fragments that hybridized with the KpnI/PstI probe [Figure 2B]. No bands were revealed in the two saltwater species, *A. merus* and *A. melas*.) Hybridization of the same probe to a Southern blot of EcoRI/SalI digested Agr23 revealed a small 0.2-kb band, which suggests some homology between the two insertion elements.

Insertion element-coding sequence junctions

Figure 3 shows the 5' and 3' junction sequences of all 9 clones that contain insertions and compares them to uninterrupted 28S rDNA sequence from the same region in *A. gambiae* and *A. arabiensis*. (The designation 5' or 3' is based on the coding strand of the rDNA rather than the insertion unless otherwise noted.) Figure 3 shows the strand that is complementary to the coding strand (shown in Figure 4), oriented 3' to 5', to highlight features of the insertion element sequence. The junction sequences are identical between the two mosquito species and between all elements. The same 17-bp duplication is found at both ends of all insertions but occurs only once in the uninterrupted rDNA.

At the 5' junction, poly A tails and a polyadenylation signal (AATAAA) occur on one strand. Poly A tails consisted of 14 residues in *A. arabiensis* and 15 residues in the four 8.2-kb insertion elements from *A. gambiae*. Tails from truncated *A. gambiae* elements always contained the sequence AAC at the end and were of varying lengths. The 3' sequences of all eight *A. gambiae* insertion elements were identical for at least 100 bp upstream of the polyadenylation signal. Aside from the poly A tails and polyadenylation signals, no other sequence similarity was detected at either end of the *A. gambiae* and *A. arabiensis* insertion elements. Nor was any homology detected to other sequenced elements found in rDNA of other insects.

Comparison of the insertion site with Drosophila melanogaster 28S rDNA

Because of the high degree of sequence conservation in this region of the 28S gene, the sequence that surrounded the mosquito insertion point could be aligned with 28S sequence of *D. melanogaster* (Figure 4). The alignment shows that the insertion point in *A. gambiae* and *A. arabiensis* is 634 bp 3' of the insertion site for the R1 element in *Drosophila*. The numbering system is based on that for *D. melanogaster* (13).

Beginning at the position corresponding to *Drosophila* 6605, a 9 bp sequence, ACAGGGATA, occurs within the target site duplication in the 2 mosquitoes. The sequence of the complementary, non-coding strand (reading 5' to 3') is very similar to part of the target site duplication for R1 elements of *Drosophila* (Figure 4). In both mosquito and R1 elements, the target site duplication sequence TNTCCCTNT is in the same strand as the coding sequence of the element.

Part of the mosquito target site duplication is very similar to a sequence found in the R2 element target site in *Drosophila* and *Bombyx*. The sequence GGGAGTAACT found 13 bp on the left side of the Type II insertion site is quite similar to the GGGATAACT found 5 bp to the left of the mosquito insertion site.

Since the anopheline elements described here do not insert at the R1 or R2 insertion site, we examined an 800 bp HindIII/BamHI subclone from *A. gambiae* that contains the region homologous to the R1 and R2 insertion sites. Sequence analysis showed that the mosquito 28S sequence in these regions is identical with that in *Drosophila*. HindIII/BamHI genomic Southern blots probed with this 800-bp fragment revealed the 800-bp fragment plus several additional bands that could not be explained by restriction site loss or gain. The additional bands were different between the two mosquito species. Thus, these two

Figure 2. Panel A. HindIII genomic digest of *Anopheles gambiae* (G; lane 1 = AS46 colony, lane 2 = GMMK6 colony) and *A. arabiensis* (A; lane 3 = MAN colony, lane 4 = GMAL colony) probed with pAgr23H5. Panel B. HindIII genomic digest of *Anopheles gambiae* (G) and *A. arabiensis* (A), probed with the 1.0-kb KpnI/PstI fragment from Aar23. Q = *A. quadriannulatus*, another member of the *A. gambiae* cryptic species complex.

rDNA	...	CAATGG	TGTCCCTATTGACCGAA		
Aar23	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAAAAAAAAATTTGCCCCAAAAATAAAAA	
Agr23	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAAAAAAAAATATAGCTCACCTAAAAATACTA	
Agr5D	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAACAATATAGCTCACCTAAAAATACTA	
Agr2A	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAACAATATAGCTCACCTAAAAATACTA	
Agr7C	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAACAATATAGCTCACCTAAAAATACTA	
Agr7B	...	CAATGG	TGTCCCTATTGACCGAA	AAAAAAACAATATAGCTCACCTAAAAATACTA	
rDNA					CACCGG ...
Aar23	— about 1.7 kb —	GTGTAGGTCGTGGGGCAGTA	TGTCCCTATTGACCGAA	CACCGG ...	
Agr23	— about 8.2 kb —	TGCTGGAGTGCAAACGTGA	TGTCCCTATTGACCGAA	CACCGG ...	
Agr5D	— about 6.6 kb —	CACCCCCCAAGTATCCGTTG	TGTCCCTATTGACCGAA	CACCGG ...	
Agr2A	— about 6.1 kb —	TCGTAAGCGTTATCCGGCTA	TGTCCCTATTGACCGAA	CACCGG ...	
Agr7C	— about 4.7 kb —	GACGGAGTAAACGCCGCCGA	TGTCCCTATTGACCGAA	CACCGG ...	
Agr7B	— about 7.6 kb —	CTGGAGTGCAAACGTGTATG	TGTCCCTATTGACCGAA	CACCGG ...	

Figure 3. Sequences of uninterrupted rDNA and of 5' and 3' insertion-gene junctions in *Anopheles gambiae* and *A. arabiensis* (Aar23). Sequence labeled 'rDNA' was identical in both mosquito species. The 17-bp target site duplication is boxed. Only the 3' and 5' ends of the two insertions were sequenced. The interior of the insertions is represented by the dashes. The strands shown are complementary to the coding strand shown in Figure 4 and are displayed 3' to 5' to highlight features of the elements.

mosquitoes are likely to harbor other insertion elements in the 28S genes, and these elements may be more similar to the R1 and R2 elements than are the two described in this paper.

DISCUSSION

Characteristics of both R1 (Type I) and R2 (Type II) rDNA insertion elements of *D. melanogaster* are combined in site-specific rDNA insertion elements found in *A. gambiae* and *A. arabiensis*. The examined insertion elements of these two mosquito species have poly A tails (as do R2 elements) and generate flanking target site duplications (as do R1 elements). The 17-bp target site duplication of the mosquitoes contains one 9-bp sequence, TNTCCCTNT, that is the same as a 9-bp sequence found in the 14-bp target site duplications of R1 insertion elements. A second 9-bp sequence in the mosquito duplication is nearly identical to a 10-bp sequence included in the 24 bp sequence required by the *Bombyx* R2 endonuclease (8). Thus, these sequences in the mosquito may be important to the site-specific insertion of these elements.

R1 and R2 elements were originally identified as interruptions of the rDNA (15) and the designations 'R1' and 'R2' are based on the rDNA insertion sites. Our studies of two elements from anopheline mosquitoes revealed that their insertion site is much closer to the 3' end of the 28S gene than is the R1 insertion site. Furthermore, no obvious similarity with R1 elements is apparent in the approximately 200 bases that have been sequenced at either end of the mosquito insertion elements.

However, the 9-bp sequence found in both the anopheline target site duplication and the R1 target site duplication indicates that similar mechanisms may be responsible for the integration of these elements. Both R1 and R2 elements have been found outside of the rDNA (7, 15) and are known to be non-LTR (long terminal repeat) retroposons (5, 7). Examples of R1 and R2 elements have been sequenced from *Bombyx mori* and each

Figure 4. Alignment of *Anopheles gambiae* and *A. arabiensis* insertion site with *Drosophila melanogaster* 28S gene sequence (12). The sequence covering the insertion sites of both R1 and R2 elements in *Drosophila* is shown in the top 2 lines. The bottom half of the figure compares the sequences flanking the insertion site (in an uninterrupted clone) in *A. gambiae* with the homologous sequence in *D. melanogaster*. The target site duplications in both species are underlined; those sequences which are underlined twice are found in both species (TNTCCCTNT, occurring on non-coding strand in the mosquito). Sequences that form part of a 24-bp duplication at the 5' end of some *Bombyx mori* R2 elements (this is not a target site duplication) are underlined with a broken line. Sequences that occur in this duplication as well as the target site duplication of mosquitoes are identified by bracketing. The sequence is numbered based on the *Drosophila* numbering system (13).

element contains an open reading frame with reverse transcriptase homology (5, 6). Recent findings that non-rDNA R1 elements of *B. mori* insert at sites that share a TGNCCC motif with the target site duplication in rDNA suggest that sequence-specific integrases are responsible for the insertion of R1 elements (7). This shared motif is similar to part of the 9-bp TNTCCCTNT sequence in mosquitoes.

R2 elements insert about 75 bp upstream of the R1 insertion site (5). A short region of sequence similarity was noted between the R2 insertion site and the *Anopheles* target site duplication. A 10-bp sequence found 13 bp 5' to the R2 insertion site is present, with the deletion of 1 G, in the 17-bp *Anopheles* target site duplication. This 10-bp sequence is found in the 24-bp *B. mori* R2 element endonuclease target sequence in the rDNA (5, 8). However, none of the short sequences of target site similarity observed between non-rDNA R2 elements or between non-rDNA and rDNA R2 elements in *Bombyx* (7) are present in or near the 17-bp *Anopheles* duplication. Other than the poly A tails, no sequence similarity was found in the 5' or 3' region of the mosquito elements and R2 elements.

The R2 elements of *Drosophila* and *Bombyx* are inserted so that the poly A tails are located at the 3' end of the strand that would be transcribed in the same direction as the

rDNA (5). This is of interest because the poly A tails of the *Anopheles* elements are located in the reverse orientation. This reversal may be related to the inversion of the putative 9-bp integrase recognition signal found in these two species. It has been suggested that in-frame elements may not require their own promoters for transcription and mobilization, but may rely on read-through transcription of the rDNA (5). The two elements in the *Anopheles* rDNA would be unable to operate by such a mechanism. These elements must be either immobilized in the rDNA or have their own promoters in the appropriate orientation.

The reversed orientation of the elements is reflected in the truncations, which occur at the 3' end if the rDNA is used for the frame of reference, but at the 5' end if the poly A tails are used for reference instead. Truncations occur at the 5' end of the R1 elements in *D. melanogaster* and *Calliphora erythrocephala*, and of the R2 elements in *Drosophila*. The truncated elements from *A. gambiae* have different poly A tail lengths but share the sequence AAC(poly A), while all the 8.2 kb elements have poly A tails of the same length and do not have the AAC sequence. Although the four 8.2 kb elements could be different clones of the same genomic fragment, they probably represent different genomic copies produced by either independent insertion events or amplification by unequal exchange. One pair of truncated clones, Agr2A (6.1 kb) and Agr5D (6.6 kb), share homologous regions at the truncated end. Another truncated clone, Agr7B (7.6 kb), is nearly identical to the predominant 8.2 kb insertion elements at the 5' end. These results may indicate that the truncations occur in a region of repeated units at the 5' end of the element. However, repeat sequences were not observed within the first 200 to 300 bp at the 5' ends of any of the elements.

With respect to the poly A tails and the target site duplications, the anopheline elements are most similar to two families of non-LTR retroposons from *D. melanogaster*, the G and F elements. The G elements usually insert at a specific site in the intergenic spacers of rDNA. These elements have poly A tails and generate target site duplications (16). The F elements generate target site duplications of 8 to 13 nucleotides and have poly A tails of 12 to 30 residues at the 3' end of one strand (17, 18). These elements have been found to interrupt R1 elements in the chromocenter but are not generally associated with rDNA sequences.

Because the anopheline elements share characteristics of all the elements described above, an appropriate designation for them is difficult at this time. Names will not be assigned to the anopheline elements until further information about their structure and sequence is available.

Aside from the target site characteristics and the poly A tails, the elements from *A. gambiae* and *A. arabiensis* are surprisingly dissimilar, given the close evolutionary relationship of these two cryptic species (19). Genomic Southern blot data showed that an internal fragment from the element in Aar23 hybridized to a set of bands that were completely different from those to which pAgr23H4 and pAgr23H5 hybridized. The type of element present in Aar23 appears to be widespread and extremely variable in *A. arabiensis*, but much less so in *A. gambiae*, although the predominant 7.5 kb band appears in both species. Each lane of the Southern blots probed in figure 2 contains DNA extracted from one mosquito, thus the variation between lanes in hybridization intensity reflects either differences in copy number or homology. The second type of element, the Agr23 type, is also present in *A. arabiensis* but, oddly, sequences homologous to the 5.0 kb fragment, which is frequently truncated in *A. gambiae*, are present in *A. arabiensis* at much higher

copy numbers than are fragments which are homologous to the 4.0 kb fragment (which is conserved in the *A. gambiae* elements). Since preliminary sequence data show the 4.0 kb fragment contains an open reading frame with possible reverse transcriptase homology, it will be of great interest to examine an element with homology to Agr23 from *A. arabiensis*, as well as a full-length representative of the Aar23 type of insertion element.

ACKNOWLEDGMENTS

We thank Brian Holloway and the staff of the Center for Infectious Diseases Biotechnology Core Facility for synthesizing the oligonucleotides. This work was supported by U.S. Agency for International Development PASA No. BST-0453-P-HC-2086-07.

*To whom correspondence should be addressed

REFERENCES

1. Dawid, I.B. and Rebbert, M.L. (1981) *Nucleic Acids Res.* 9, 5011–5020.
2. Rae, P.M.M., Kohorn, B.D. and Wade, R.P. (1980) *Nucleic Acids Res.* 8, 3491–3504.
3. Smith, V.L. and Beckingham, K. (1984) *Nucleic Acids Res.* 12, 1707–1724.
4. Eickbush, T.H. and Robins, B. (1985) *EMBO J.* 4, 2281–2285.
5. Burke, W.D., Calalang, C.C. and Eickbush, T.H. (1987) *Mol. Cell. Biol.* 7, 2221–2230.
6. Xiong, Y. and Eickbush, T.H. (1988) *Mol. Cell. Biol.* 8, 114–123.
7. Xiong, Y., Burke, W.D., Jakubczak, J.L. and Eickbush, T.H. (1988) *Nucleic Acids Res.* 16, 10561–10573.
8. Xiong, Y. and Eickbush, T.H. (1988) *Cell* 55, 235–246.
9. McLain, D.K. and Collins, F.H. (1989) *Heredity* 62, 233–242.
10. Collins, F.H., Mendez, M.A., Rasmussen, M.O., Mehaffey, P.C., Besansky, N.J., Finnerty, V. (1987) *Am. J. Trop. Med. Hyg.* 37, 37–41.
11. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
12. Fuller, C.W. (1988) *U.S. Biochemical Corp. Comments* 14, 1–7.
13. Tautz, D., Hancock, J.M., Webb, D.A., Tautz, C., and Dover, G.A. (1988) *Mol. Biol. Evol.* 5, 366–376.
14. Beckingham, K. (1982) In Busch, H. and Rothblum, L. (eds) *The Cell Nucleus*. Academic Press, New York. Vol. X, pp. 205–269.
15. Dawid, I.B. and Botchan, P. (1977) *Proc. Natl. Acad. Sci. USA* 74, 4233–4237.
16. Di Nocera, P.P., Graziani, F. and Lavorgna, G. (1986) *Nucleic Acids Res.* 14, 675–691.
17. Dawid, I.B., Long, E.O., Di Nocera, P.P. and Pardue, M.L. (1981) *Cell* 25, 399–408.
18. Di Nocera, P.P., Digan, M.E. and Dawid, I.B. (1983) *J. Mol. Biol.* 168, 715–727.
19. Davidson, G., Paterson, H.E., Coluzzi, M., Mason, G.F., and Micks, D.W. (1967) In Wright, J.W. and Pal, R. (eds.) *Genetics of Insect Vectors of Disease*. Elsevier Publ. Co., London, pp. 211–250.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.