2524–2537 Nucleic Acids Research, 2002, Vol. 30, No. 11

A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*

Jeffrey D. Wuitschick, Jill A. Gershan, Andrew J. Lochowicz, Shuqiang Li and Kathleen M. Karrer*

Department of Biology, Marquette University, Milwaukee, WI 53201-1881, USA

Received December 31, 2001; Revised March 11, 2002; Accepted March 27, 2002 DDBJ/EMBL/GenBank accession nos+

ABSTRACT

In the ciliated protozoan *Tetrahymena thermophila*, extensive DNA elimination is associated with differentiation of the somatic macronucleus from the germline micronucleus. This study describes the isolation and complete characterization of Tlr elements, a family of approximately 30 micronuclear DNA sequences that are efficiently eliminated from the developing macronucleus. The data indicate that Tlr elements are comprised of an ~22 kb internal region flanked by complex and variable termini. The Tlr internal region is highly conserved among family members and contains 15 open reading frames, some of which resemble genes encoded by transposons and viruses. The Tlr termini appear to be long inverted repeats consisting of (i) a variable region containing multiple direct repeats which differ in number and sequence from element to element and (ii) a conserved terminal 47 bp sequence. Taken together, these results suggest that Tlr elements comprise a novel family of mobile genetic elements that are confined to the *Tetrahymena* germline genome. Possible mechanisms of developmentally programmed Tlr elimination are discussed.

INTRODUCTION

Transposable elements, retroviruses and some DNA viruses proliferate by integrating into the chromosomes of their hosts. These mobile genetic elements are common features of most eukaryotic genomes, often comprising a substantial percentage of the bulk DNA within a cell. Transposable elements alone have been estimated to constitute ~15% of the *Drosophila* genome (1), 35% of the human genome (2) and >50% of the maize genome (3).

The inherent instability of mobile genetic elements represents a fundamental challenge to the integrity of a host genome. Insertions into or near coding regions can result in alteration or disruption of gene expression. In addition, the repetitive nature of most elements can promote non-homologous or ectopic recombination leading to large-scale genomic deletions and rearrangements (reviewed in 4). Thus, their uncontrolled proliferation could drastically disrupt genome order and lead to a rapid decline in host fitness. As a result, the activity of mobile genetic elements is generally restricted by a variety of constraints imposed by the host organism or by the elements themselves (reviewed in 1,5).

The unusual genetic organization of the ciliated protozoa may provide a unique opportunity for effective control of mobile genetic elements. Although unicellular, ciliates exhibit nuclear dualism (reviewed in 6). The holotrichous ciliate *Tetrahymena thermophila* is a typical case. Each cell contains two types of structurally and functionally distinct nuclei: a large somatic macronucleus and a small germline micronucleus. During asexual propagation, *Tetrahymena* divide by vegetative fission. Under these conditions, the macronucleus is responsible for all detectable nuclear gene expression while the micronucleus remains transcriptionally silent. However, during conjugation, the sexual phase of the ciliate life cycle, cells of complementary mating types pair, their micronuclei undergo meiosis and partners reciprocally exchange the resulting haploid gametic pronuclei (reviewed in 7). Following pronuclear fusion, new micronuclei and macronuclei develop from identical mitotic products of the diploid zygotic nucleus and the old macronuclei degenerate in a manner closely resembling apoptosis (8,9).

Although genetically indistinguishable from the micronucleus at the onset of development, the newly formed macronucleus rapidly undergoes dramatic genomic reorganization (reviewed in 7). In the developing macronucleus, the five germline-derived diploid chromosomes of *T.thermophila* are fragmented into 50–250 subchromosomal molecules (10,11). In addition, over 6000 chromosomal segments are removed through reproducible internal deletion events involving (i) elimination of specific stretches of micronucleus-limited DNA and (ii) ligation of the macronucleus-retained flanking sequences (12). These internally eliminated sequences (IESs) range in length from 0.6 to >13 kb (13,14), are highly variable in nucleotide

*To whom correspondence should be addressed. Tel: +1 414 288 1474; Fax: +1 414 288 7357; Email: kathleen.karrer@marquette.edu
Present addresses:
Jill A. Gershan, Platelet Immunology, Blood Research Institute of Southeastern Wisconsin, 8727 Watertown Plank Road, Wauwatosa, WI 53226, USA
Andrew J. Lochowicz, Platelet and Neutrophil Immunology, Blood Center of Southeastern Wisconsin, 638 North 18th Street, Milwaukee, WI 53201, USA

+AF451860–AF451870

sequence (13,15–18) and consist primarily of moderately repetitive DNA elements (15,18–20). Deletion of IESs ultimately results in a loss of ~15% of the germline genome from the mature *Tetrahymena* macronucleus (21). Throughout this process, chromosomes in the micronucleus remain intact and unaltered, thus ensuring a means by which the complete genome may be transferred to future generations.

Although the biological function of programmed DNA elimination in ciliates is presently unclear, one consequence of this unusual process appears to be the efficient removal of mobile genetic elements from the somatic macronucleus. Repetitive transposon-like elements have been identified in the micronuclear genomes of several ciliates (19,22–25). Remarkably, all copies of these elements seem to be eliminated from the developing macronuclei of their respective hosts. The most thoroughly characterized examples of micronucleus-limited mobile genetic elements in ciliates are the repetitive, transposon-like Tec (24,26,27) and TBE1 elements (28–30) of *Euplotes* and *Oxytricha*, respectively. At present, little is known about the mechanism by which Tec and TBE elements are deleted during macronuclear development. However, this process appears to bare striking similarities to the removal of smaller IESs in those ciliates (reviewed in 31).

Tlr elements are a family of approximately 30 germline-limited DNA sequences in *T.thermophila* that structurally resemble mobile genetic elements (14,32). The best characterized Tlr element, Tlr1, contains an 825 bp terminal inverted repeat separated by a long internal region (14). Most or all of Tlr1 is deleted from the developing macronucleus as part of an IES with variable excision boundaries (14,33).

Gershan and Karrer (32) previously reported the isolation of several kilobases of DNA immediately internal to the right side of the Tlr1 terminal inverted repeat. Analysis of this sequence revealed the presence of a 1998 bp open reading frame (ORF) encoding a deduced protein with similarity to retroelement integrases. Micronuclear genomic clones consisting of the corresponding region from four other Tlr family members also contained the putative integrase ORF. All clones were ~95% identical at the nucleotide level, suggesting the internal region of most elements is likely to be highly conserved between Tlr family members.

In this study we describe the isolation and sequence analysis of a complete Tlr composite genome. Tlr elements contain a 22 kb internal conserved region that consists of numerous ORFs, some of which resemble genes encoded by viruses and transposable elements. The Tlr internal region is apparently flanked by complex terminal inverted repeats that are quite variable between family members. These data suggest that Tlr elements are a family of novel mobile genetic elements that are confined to the germline genome in *Tetrahymena*.

MATERIALS AND METHODS

Genomic library screenings

Most Tlr clones were isolated from the previously described pMBR library (34), which was constructed from *T.thermophila* micronuclear DNA partially digested with *Mbo*I and ligated into the plasmid vector pUC19. Clone pMicV-A1 was isolated from a pMicV library constructed from micronuclear DNA completely digested with *Eco*RV and ligated into the plasmid vector pBluescriptSK II. These plasmid libraries, which carry average insert sizes of 6–9 kb, were chosen for use in this study because phage and BAC libraries containing longer segments of *Tetrahymena* DNA are unstable in bacteria.

DNA from the appropriate library (~100 ng) was transformed by electroporation into electrocompetent *Escherichia coli* SURE cells (Stratagene, La Jolla, CA). Transformants were spread onto Luria–Bertani broth (LB) plates containing 50 μg/ml carbenicillin and incubated at 30°C overnight. The resulting colonies were transferred to GeneScreen Plus Colony/Plaque Screen hybridization membranes (NEN Life Science Products, Boston, MA) and lysed by autoclaving. Immobilized DNA was UV crosslinked to the nylon transfer membranes then incubated with appropriate radiolabeled hybridization probes in 1× P buffer (5× P buffer is 1% bovine serum albumin, 1% polyvinylpyrrolidone, 1% Ficoll, 0.25 M Tris pH 7.5, and 0.01 M sodium pyrophosphate) plus 10% dextran sulfate, 1% SDS, 1 M sodium chloride and 0.1 mg/ml salmon sperm DNA at 65°C overnight. Probes were generated as PCR products or restriction fragments from cloned DNA and subsequently labeled with [α-^{32}P]dATP by random priming (Roche Molecular, Indianapolis, IN). Hybridized membranes were washed twice in 2× SSC (20× SSC is 3 M sodium chloride and 0.3 M sodium citrate pH 8.0) at room temperature, 2× SSC plus 1% SDS at 65°C and 0.1× SSC at 58°C. Membranes were exposed to X-Omat *AR* X-ray film (Eastman Kodak, Rochester, NY) and colonies harboring plasmids with positive inserts were selected from the original LB growth plates for further analysis.

DNA sequencing

The Erase-a-Base exonuclease III deletion system (Promega, Madison, WI) was used to generate subclones of most large plasmid inserts obtained from library screens. The resulting constructs, containing progressive unidirectional deletions, were sequenced with universal primers from vector polylinker regions. Gaps were filled using custom-designed oligonucleotides as sequencing primers. In some cases, a series of successively internal custom primers was used to directly sequence intact clones. Sequencing reactions were performed using an ABI Prism dye terminator cycle sequencing ready reaction kit (Perkin Elmer, Boston, MA) with AmpliTaq DNA polymerase and read by an Applied Biosystems sequencer at the University of Wisconsin (Milwaukee, WI) automated DNA sequencing facility.

Sequence assembly and analysis

Multiple sequence reads from individual clones were assembled using GCG Geloverlap followed by GCG Gelassemble. Contigs containing overlapping regions of Tlr elements were aligned with GCG Gap or GCG Fasta and analyzed for sequence continuity. Sequences from all clones were compared with the non-redundant public database using the National Center for Biotechnology Information (NCBI) BLASTn and BLASTx servers (www.ncbi.nlm.nih.gov/BLAST/). Theoretical translations from selected ORFs were compared with the peptide database using NCBI BLASTp. In some cases, NCBI position-specific iterative PSI-BLAST searches were employed to identify subtle similarities to annotated protein entries in the database. Nucleotide and amino acid sequence pile-ups were generated using the Vector

NTI Suite (Informax, Bethesda, MD) multiple sequence alignment function.

Isolation and analysis of genomic DNA

Tetrahymena thermophila laboratory strain CU428, Mpr/Mpr (6-methylpurine-sensitive, VII), was kindly provided by Peter Bruns (Cornell University, Ithaca, NY). Cells were propagated axenically under standard growth conditions. Micro and macronuclei from CU428 cultures were separated by differential centrifugation as described by Gorovsky *et al.* (35). Purified nuclei were lysed in 20 mM Tris–HCl, pH 7.5, 40 mM EDTA and 1% SDS at 50°C for 20 min then incubated with 200 µg/ml proteinase K at 37°C for 1 h. Genomic DNA was subsequently extracted with phenol:chloroform (1:1) and ethanol precipitated. For Southern blot analyses, equal molar amounts of micronuclear and macronuclear DNAs were digested with restriction endonucleases and fractionated by electrophoresis on 0.7% agarose gels. Gels were stained with ethidium bromide to verify the presence of DNA in appropriate lanes. Separated DNA fragments were then transferred by capillary action to GeneScreen Plus hybridization transfer membranes (NEN Life Science Products) and crosslinked to the filters with UV light. Hybridizations with radiolabeled probes and subsequent washes were performed under the same conditions described for colony screenings. In some cases, hybridized blots were stripped by boiling in 0.1× SSC plus 1% SDS for 30 min and then reprobed.

PCR amplification from genomic DNA

Clone AJL1 was fortuitously PCR amplified from *T.thermophila* micronuclear DNA using primers Tlr.left1 (5′-GTTGGAATAGAGAAAGTGATAAAG-3′) and IntA.right1 (5′-AAACCAATCAACAAGCAAAAGG-3′). The resulting PCR product was agarose gel purified according to the QIAquick Gel Extraction procedure (Qiagen, Valencia, CA) and cloned into vector pCR 2.1 using the TA Cloning Kit (Invitrogen, Carlsbad, CA).

Accession numbers of cloned sequences

Accession numbers of the sequences characterized in this study are as follows: pMBR 4C1, AF451860; pMBR 6C1, AF451861; pMBR 8E1, AF451862; pMBR 2, AF451863; pMBR 9, AF451864; pMBR 8, AF451865; pMBR 4D1, AF451866; pMBR 5B1, AF451867; pMicV-A1, AF451868; pMBR 6C1-15A, AF451869; AJL1, AF451870.

RESULTS

Isolation of the Tlr genome

In order to characterize the complete structure of Tlr elements, we performed a ‘chromosome walk’ across the family (from right to left) by successively screening a plasmid library of micronuclear DNA from *T.thermophila* (Materials and Methods). The initial library screen was carried out using the 338 bp LSC1 probe (Fig. 1A) from the left end of Tlr Int B, the longest internal clone obtained in a previous study (32). Hybridizing colonies were relatively abundant, as expected for a target sequence that is moderately repetitive within the genome. Several positive clones with long inserts were selected for further analysis. Those determined by PCR analysis to extend furthest into the unknown internal region of the Tlr genome were sequenced completely. A total of three consecutive library screens were performed using progressively internal PCR products as probes (Fig. 1A). Alignment of sequenced inserts revealed that most overlapping Tlr clones were 90–97% identical at the sequence level, resembling the high degree of nucleotide conservation previously observed in the integrase-like ORF (Tlr ORF 1R) region (32). However, enough nucleotide differences and short insertions/deletions were detected to suggest that overlapping clones probably represent similar regions of different Tlr family members.

Although a high degree of sequence continuity was observed between most imbricated Tlr segments analyzed in this study, two clones were isolated that contain apparent insertions within Tlr internal conserved sequences. Two ~600 bp insertions were detected in pMBR 8 and a related ~1.2 kb insert was found in pMBR 8E1 (Fig. 1A). Each insertion element contains a single ORF encoding a conceptual protein with significant similarity (BLAST expect value = $2e^{-18}$) to HNH intron-homing nucleases.

pMBR 4C1 and pMBR 6C1, the final clones obtained from the chromosome walk, align with 96% identity with the innermost 280 bp of known sequence from the left end of Tlr1 (data not shown). However, this region of nucleotide conservation abruptly terminates when all three sequences diverge (Fig. 1A). No extensive regions of similarity between pMBR 4C1, pMBR 6C1 and Tlr1 could be detected beyond this coordinate. We therefore define this point of divergence as the left boundary of the Tlr internal conserved region. In the case of Tlr1, the terminal inverted repeat is located 32 bp outside the point of divergence (Fig. 1A).

Based on the cumulative distance from the previously established right end of conserved Tlr sequence (32) to the newly determined left boundary, the internal conserved region of Tlr elements is ~22 kb in length. A composite structure of the complete Tlr genome was assembled from a total of 10 overlapping internal clones (Fig. 1B). Most regions were sequenced on at least two different cloned family members. The Tlr genome is 27% G + C, which is similar to the low G + C content (24%) of the overall *T.thermophila* genome (36).

Nucleotide conservation at the right end of Tlr elements extends 986 bp to the right of the previously described integrase-like Tlr ORF 1R (32). Thus, the rightmost 276 bp of the Tlr internal conserved region comprises the inner portion of the Tlr1 terminal inverted repeat (Fig. 1A). In contrast, this 276 bp sequence is not associated with the left end of any cloned element other than Tlr1. Given the small sample size of left-end Tlr clones currently available, it is not possible to determine whether the duplication of this sequence on the left side of Tlr1 is unique among family members. However, such a hypothesis would be consistent with previous Southern blot analyses demonstrating that the innermost portion of the Tlr1 inverted repeat is present in approximately 30 copies within the micronuclear genome (14), equaling about one copy per Tlr element. A detailed analysis of cloned sequences flanking the internal conserved region of Tlr elements is presented in a later section.

Clones used to generate the consensus Tlr genome were analyzed for ORFs. The genetic code of *Tetrahymena* is unusual in that UAA and UAG encode glutamine, making UGA the only stop codon (37). Therefore, ORFs starting with AUG and ending with UGA were detected using a

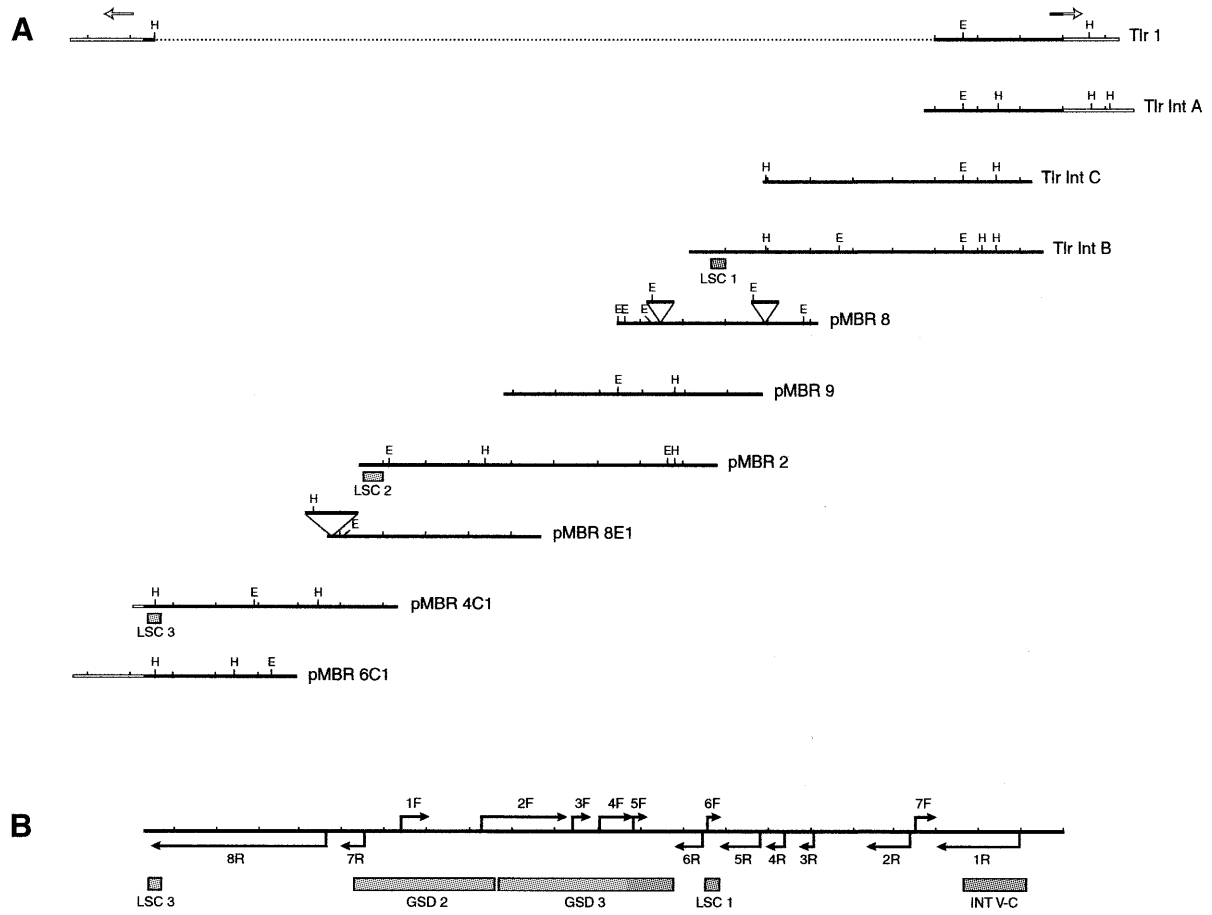

Figure 1. (**A**) Alignment of Tlr clones obtained from consecutive micronuclear library screens. Solid bars indicate sequence within the Tlr internal conserved region. Open bars indicate non-conserved sequence outside the Tlr internal region. Arrows over the Tlr1 element depict the position of its terminal inverted repeat. The dotted line represents uncharacterized Tlr1 sequence. Shaded boxes indicate the locations of probes used in library screens. Apparent insertion elements are depicted as large triangles. The distance between hash marks represents ~1 kb. *Hin*dIII (H) and *Eco*RV (E) sites are indicated. Clones Tlr1 and Tlr Int A–C were isolated and partially characterized by Gershan and Karrer (32). (**B**) Composite diagram of the assembled Tlr internal conserved region. Arrows indicate locations and orientations of major ORFs. Shaded boxes represent positions of probes used in Southern blot analyses.

Tetrahymena translation table. In order to apply additional stringency to this survey, only ORFs >300 nt and present in at least two family members were considered for further investigation. A total of 15 ORFs meeting these criteria were identified in the Tlr genome (Fig. 1B). These ORFs are equally distributed on both strands of the elements. They range in size from 312 to 4218 nt, are predominantly non-overlapping and comprise ~73% of the Tlr internal conserved region.

Amino acid sequence comparisons of deduced proteins encoded by Tlr elements

Tlr ORF 1R was previously reported to encode an integrase-like putative protein (32). BLASTp and iterated PSI-BLAST searches revealed that the conceptual translation products from 4 of the 14 newly described Tlr ORFs showed similarities to protein sequences in the GenBank protein database, most of which are encoded by viruses or transposable elements.

The deduced amino acid sequence of the protein encoded by ORF 6R (Tlr 6Rp) shows similarity to putative Walker-type ATPases from double-stranded (ds)DNA viruses, including phycodnaviruses, iridoviruses and poxviruses (Fig. 2A). Tlr 6Rp and its viral counterparts are roughly similar in length and share several regions of conservation, including apparent Walker A and B motifs (38,39). Although a specific function has not yet been established for this group of viral peptides, the protein encoded by fowlpox virus has been implicated in virus DNA packaging during virion assembly (40).

The ORF 6F predicted polypeptide (Tlr 6Fp) resembles (BLAST expect value = $4e^{-09}$) a small protein of unknown function encoded by the A121R gene from the phycodnavirus *Paramecium bursaria* Chlorella virus (PBCV-1) (41). Figure 2B depicts the 103 amino acid consensus translated sequence of ORF 6F derived from clones Tlr Int B, pMBR 2, pMBR 8 and pMBR 9 aligned with the 97 amino acid conceptual protein from A121R. Several blocks of conserved amino acid residues are distributed along the length of the peptide chains.

ORF 8R, the largest ORF in the Tlr genome, is 4218 nt in length and encodes a putative protein (Tlr 8Rp) of 1405 amino acids. A stretch of approximately 300 amino acids near the C-terminal end of Tlr 8Rp strongly resembles several members of the SF1 helicase superfamily. SF1 helicases comprise a diverse group of related enzymes that mediate the nucleotide triphosphate-dependent unwinding of DNA duplexes, RNA duplexes and DNA/RNA heteroduplexes (42). These proteins are found in prokaryotes and eukaryotes as well as in numerous

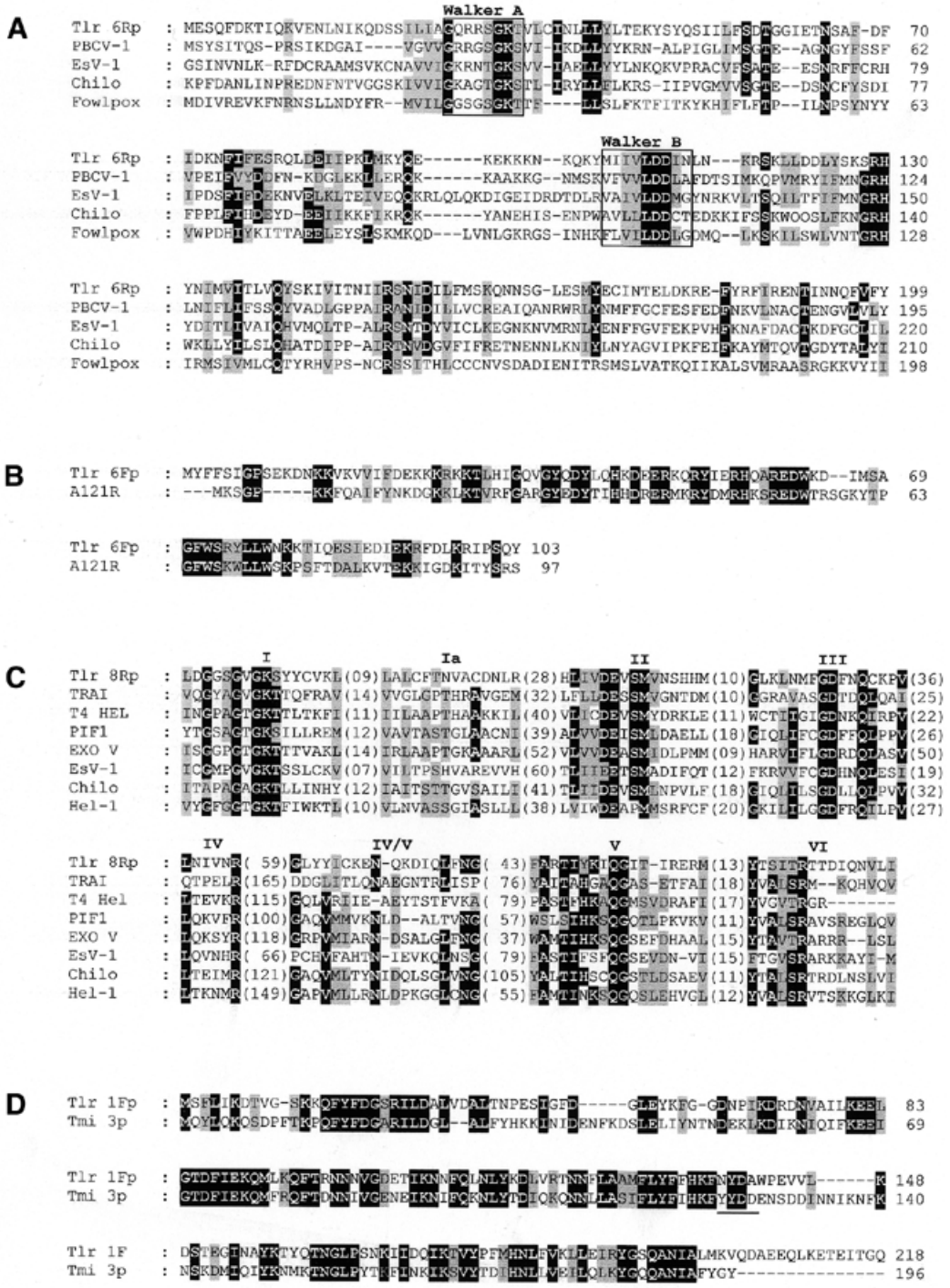

Figure 2. Conceptual amino acid sequences of proteins encoded by selected Tlr ORFs aligned with similar proteins from the GenBank peptide database. Amino acids that are identical in >60% of aligned sequences are shaded in black. Similar amino acids are shaded in gray. (**A**) Alignment of the first 199 amino acids of the Tlr 6Rp consensus sequence with ATPase-like putative proteins from diverse viruses. Walker A and B motifs are boxed. Definitions and accession numbers of compared amino acid sequences are: PBCV-1 A392Rp, NP_048749; EsV-1-26 putative protein, NP_077511; Chilo iridescent virus protein 075L, NP_149538; fowlpox virus FPV197 virion assembly protein, NP_039160. (**B**) Amino acid sequence of Tlr 6Fp aligned with the putative protein of unknown function encoded by the PBCV-1 A121R gene, accession no. NP_048469. (**C**) Conserved helicase domains of Tlr 8Rp from pMBR 4C1 and a set of DNA helicases and putative DNA helicases aligned as described by Kapitonov and Jurka (43). Helicase domain designations are indicated above amino acid sequence blocks. Distances between motifs are indicated in parentheses. Definitions and accession numbers of compared amino acid sequences are: TRAI DNA helicase I from the F plasmid of *E.coli*, P14565; enterobacteria dsDNA phage T4 DNA helicase (T4 Hel), P32270; PIF1 DNA helicase from *Saccharomyces cerevisiae*, P07271; α-subunit of exonuclease V from *E.coli* (EXO V), P04993; EsV-1-29 putative helicase, NP_077514; Chilo iridescent virus 030L putative helicase, AAD48149; Hel-1 putative helicase from *Helitron1* of *Arabidopsis thaliana* (43). (**D**) The first 218 amino acids of the Tlr 1Fp consensus sequence aligned with Tmi 3p, accession no. CAA42575. The YXDD motif from Tmi 3p is underlined.

viruses, including the phycodnavirus *Ectocarpus siliculosus* virus (EsV-1) and the Chilo iridescent virus. SF1 helicases typically contain seven conserved motifs distributed over a region of 200–700 amino acids. These motifs probably comprise most of the helicase catalytic core and their identification within protein sequences has been demonstrated to be a reliable indicator of potential helicase activity (42). Tlr 8Rp contains all seven established SF1 helicase motifs plus a recently described eighth putative SF1 domain (Fig. 2C) (43). Tlr 8Rp exhibits the highest degree of similarity to the DNA helicase subgroup of SF1 helicases. The activity of these enzymes is most commonly associated with replication, repair and recombination of cellular and viral DNA (42). A putative SF1 DNA helicase gene has recently been characterized in an unusual family of eukaryotic transposons, designated helitrons (43). However, other than the common helicase motifs, helitron elements exhibit no structural similarities to Tlr elements.

Tlr ORF 1F encodes a putative 235 amino acid polypeptide (Tlr 1Fp) that is similar (BLAST expect value = e^{-30}) to a 196 amino acid conceptual protein encoded by ORF 3 from Tmi elements (Fig. 2D), another family of moderately repetitive micronucleus-limited elements in *T.thermophila*. Tmi elements have been partially characterized at the sequence level and are thought to structurally resemble RNA-based mobile genetic elements (E. Blackburn and C. Wyman, personal communication). The Tmi ORF 3 conceptual polypeptide (Tmi 3p) was tentatively identified as a reverse transcriptase-like enzyme based primarily on the presence of a central YXDD motif, the most constant region of amino acid conservation between known reverse transcriptases (44). Although Tlr 1Fp is 70% similar to the predicted amino acid sequence of Tmi 3p, none of the four Tlr elements examined contained the YXDD motif (Fig. 2D). Therefore, Tlr ORF 1F is unlikely to encode a protein with reverse transcriptase activity.

Alignment of Tlr ORF 1F with Tmi ORF 3 demonstrated that these sequences are 58% similar at the nucleotide level (data not shown). This information, together with the observed conservation between the predicted amino acid sequences, suggests that the two ORFs may have been derived from a common progenitor. Nucleotide comparisons from flanking regions of both ORFs revealed no detectable similarities. Thus, other than the common ORF, Tlr and Tmi elements do not appear to be closely related.

Conservation of Tlr ORFs

All five Tlr family members previously analyzed showed 98% nucleotide sequence conservation within ORF 1R. Nucleotide changes that do occur in these integrase-like ORFs were demonstrated to be non-random with respect to their position within codons. Of the 196 nucleotide changes detected in aligned Tlr ORF 1Rs, 93% encode identical or similar amino acids, suggesting that these ORFs have evolved under selection to maintain a functional protein (32). In order to determine whether the newly described Tlr ORFs might display a similar degree of selective conservation, we analyzed the distribution of nucleotide changes within ORF 6F and ORF 1F. These ORFs were chosen for analysis because of the availability in both cases of four different Tlr clones containing the complete ORF.

Table 1. Nucleotide changes and amino acid substitutions in ORF 6F from four Tlr family members

Amino acid	Nucleotide position			Total nucleotide changes
	1	2	3	
Identical	2	0	17	19 (56%)
Similar	5	2	3	10 (29%)
Dissimilar	0	3	2	5 (15%)
Total	7	5	22	34

Table 2. Nucleotide changes and amino acid substitutions in ORF 1F from four Tlr family members

Amino acid	Nucleotide position			Total nucleotide changes
	1	2	3	
Identical	3	0	46	49 (51%)
Similar	21	8	11	40 (42%)
Dissimilar	1	4	2	7 (7%)
Total	25	12	59	96

ORF 6Fs from clones Tlr Int B, pMBR 2, pMBR 8 and pMBR 9 (Fig. 1A) were aligned with 97% identity at the nucleotide level (data not shown). Only 34 nucleotide changes were detected among the 1248 bp analyzed. If these changes have accumulated at random, they would be expected to occur at relatively equal frequencies among the three possible nucleotide positions within ORF 6F codons. However, according to χ^2 analysis, the positions of nucleotide changes are non-random (P = <0.001). Of the 34 nucleotide polymorphisms, 65% occur at the third position in codons (Table 1), where they are most likely to result in amino acid conservation. Indeed, 85% of the nucleotide changes in ORF 6F translate into identical or highly similar amino acids as determined by a score of 4 or higher in the Structure–Genetic matrix scoring system (45).

ORF 1Fs from pMBR 2, pMBR 8E1, pMBR 4D1 and pMBR 5B1 (pMBR 4D1 and 5B1 not shown in Fig. 1A) are 95% conserved at the DNA level. Differences among the four sequences are also distributed non-randomly with respect to nucleotide position within codons (P = <0.0001). Of the 96 base changes, 93% encode identical or similar amino acids (Table 2). These data suggest that, like ORF 1R, both ORF 6F and ORF 1F are, or recently were, under selective pressure to preserve functional proteins.

Tlr elements are limited to the germline micronucleus

Previously characterized regions of Tlr elements were demonstrated to be repeated ~30-fold in the micronuclear genome of *T.thermophila* (14,32). No copies were detected in macronuclear DNA, indicating that these sections of Tlr family members undergo efficient developmentally programmed DNA elimination. Thus, it was of interest to determine the nuclear specificity of the remainder of the Tlr genome. Four segments of newly characterized Tlr internal sequence (Fig. 1B) were used to probe Southern blots containing restriction

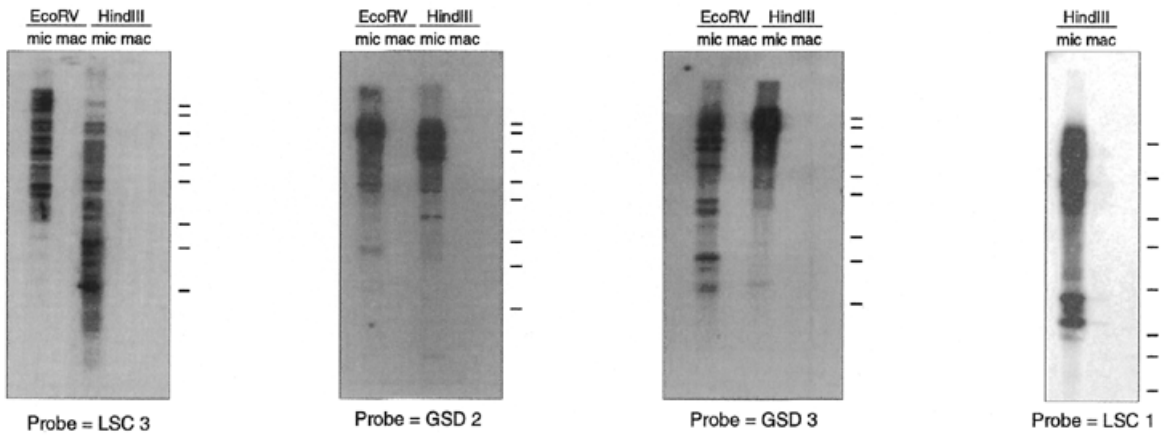

Figure 3. Southern blots of micronuclear (mic) and macronuclear (mac) DNA digested with *Eco*RV or *Hin*dIII and probed with fragments from within the Tlr internal conserved region (Fig. 1B). Bars to the right of each blot indicate mobilities of the 10, 8.0, 6.0, 4.0, 3.0, 2.0, 1.5 and 1.0 kb DNA fragments of the HiLo marker (Minnesota Molecular, Bloomington, MN), run as molecular weight standards.

enzyme-digested nuclear DNAs (Fig. 3). All probes hybridized to multiple fragments in micronuclear DNA but to none in macronuclear DNA, suggesting that all copies of the ~22 kb Tlr internal conserved region are removed in their entirety from the developing macronucleus.

Probe LSC 3 (Fig. 1) hybridized to approximately 30 fragments in lanes containing micronuclear DNA digested with *Eco*RV (Fig. 3). The number of fragments detected in this lane is likely to reflect Tlr copy number since the sequence encompassed by the probe is outside the leftmost *Eco*RV site in the Tlr internal conserved region. This value is in good agreement with previous estimates of Tlr copy number (14,32). LSC 3 spans the leftmost conserved Tlr *Hin*dIII site. Thus, this probe is expected to hybridize to both the collection of left terminal fragments and to conserved internal fragments.

Based on their size and copy number, we calculate that Tlr elements constitute ~0.3% of the 2×10^8 bp *T.thermophila* genome and ~2.2% of the ~3×10^7 bp of micronucleus-limited DNA. The average G + C content of known internally eliminated sequences in *Tetrahymena* is ~19% (46). Thus, at 27% G + C, Tlr elements are considerably more G + C-rich than other micronucleus-limited sequences characterized to date. Since most other micronucleus-limited sequences are presumably non-coding, this disparity could be a consequence of base composition constraints necessary to maintain functional Tlr ORFs. However, it should be noted that no significant differences in G + C content were observed between ORF and non-ORF regions within cloned segments of Tlr family members (data not shown).

Arrays of short direct repeats are commonly found adjacent to the Tlr conserved region

The outer region of the Tlr1 terminal inverted repeat contains two different 19 bp direct repeats, designated 19A and 19B (Fig. 4). Each of these 19mers is tandemly repeated in three perfect copies followed by three degenerate copies (14). Analysis of sequences from right-end clone Tlr Int A and left-end clone pMBR 6C1 (Fig. 1) revealed that these family members also contain arrays of direct repeats outside their internal conserved sequences (Fig. 4). In Tlr Int A, a 26mer and a different 28mer are each repeated twice. Even more striking, the corresponding region of pMBR 6C1 contains a 34mer repeated twice, a 10mer repeated eight times and a 9mer repeated four times. In most cases, the direct repeats within Tlr Int A and pMBR 6C1 are iterated in tandem. However, some are separated by short spacer sequences. All of the direct repeats detected in these family members, including the Tlr1 19A and 19B repeats, are located within 600 bp of the point of sequence divergence from the inner, conserved region. These data indicate that direct repeats may be a common structural feature of many Tlr element termini. No terminal repetition was detected in right-end clone Tlr Int E and left-end pMBR 4C1 (Fig. 4), suggesting that direct repeats may not be associated with the termini of all Tlr elements. However, since Tlr Int E and pMBR 4C1 contain only 425 and 321 bp, respectively, of DNA outside the conserved region, it is possible that direct repeats could be located outside the cloned DNA.

In order to analyze the terminal structure of additional Tlr elements, the ends of two previously uncharacterized Tlr elements were isolated and sequenced. These clones, designated pMicV-A1 and AJL1 (Fig. 4), also contain tandem repeats adjacent to the boundaries of the their internal conserved sequences. Clone pMicV-A1 was obtained from a preliminary screening of the pMicV micronuclear DNA library using the hybridization probe INT V-C (Fig. 1). The genomic library used to isolate this clone was generated from *T.thermophila* micronuclear DNA that was completely digested with *Eco*RV. Since the INT V-C fragment was generated from sequence outside the rightmost *Eco*RV site in the internal conserved region, it is predicted to hybridize to clones containing the right end of Tlr family members. pMicV-A1 is 5790 bp in length and extends rightward, as expected, from the conserved *Eco*RV site within ORF 1F. Like other Tlr elements, it is ~95% similar at the nucleotide level to overlapping clones within the internal conserved region. This nucleotide conservation extends 29 bp farther than Tlr Int A and Tlr Int E, diverging 305 bp into the right half of the Tlr1 inverted repeat. Similar to Tlr1, Tlr Int A and pMBR 6C1, pMicV-A1 contains a series of direct repeats outside the point of divergence, including three different 31mers, a 63mer and a 58mer, each repeated twice (Fig. 4). The final set of direct repeats in pMicV-A1 are located almost 2 kb outside the Tlr conserved sequence.

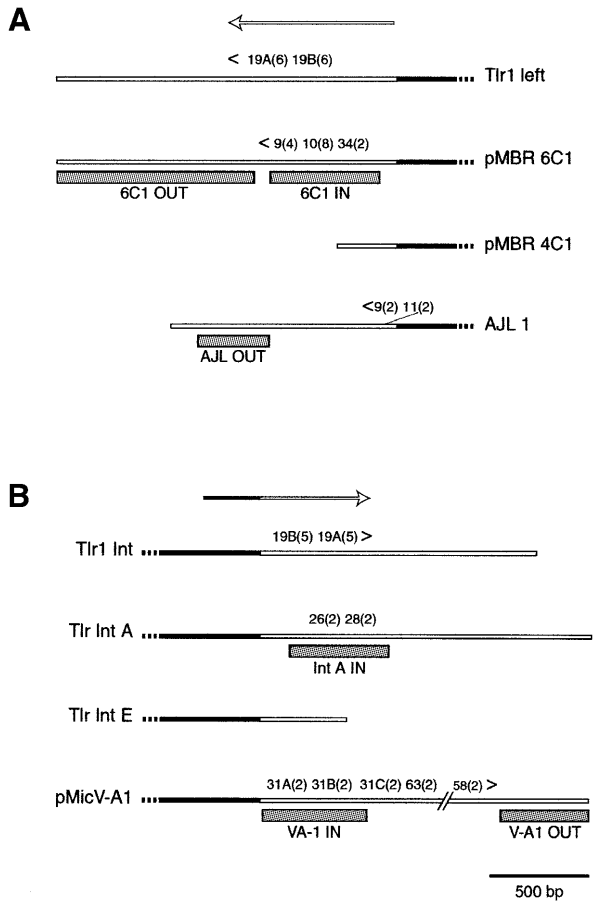

Figure 4. Left (**A**) and right (**B**) terminal regions of Tlr elements. Solid bars indicate sequence within the Tlr internal conserved region. Open bars indicate sequence outside the Tlr internal conserved region. Broken bars indicate continuation of cloned sequence beyond the diagram. Overhead arrows depict the positions of the Tlr1 terminal inverted repeat. Approximate locations of sub-terminal direct repeats are indicated above clones. Numbers outside parentheses indicate repeat unit length. Where an individual clone contains different repeats of the same length, letters are used to differentiate sequences. Numbers inside parentheses indicate the number of iterations. Greater than and less than symbols indicate location and orientation of conserved terminal 47mers. The break in pMicV-A1 indicates ~1 kb of cloned sequence not diagrammed. Shaded boxes indicate positions of probes used in Southern blot analyses. Clone Tlr Int E was isolated and partially characterized by Gershan and Karrer (32).

Tlr1
19Amer
ATTATTTCTTTTTACATTT
ATTATTTCTTTTTACATTT
ATTATTTCTTTTTACATTT
ATgATTTCTTTTTAtATTT
cTTAcTTCTTTTTAtATTT
cTTATTTCTTTTTAtTTTT
19Bmer
TTTCTCATTTTATGAAAAG
TTTCTCATTTTATGAAAAG
TTTCTCATTTTATGAAAAG
TTTCTtATTTTATGAAAAG
TTTCTtATTT-ATcAAAAt
TTTCTtATTT-ATcAAAAt

pMBR 6C1
9mer
TAAAAAGTC
TAAAAAGTC
TAAAA-GTC
TAAAAAGTC
10mer
TTTTTCCTCC
TTTT-CCTCC
TTTT-CCTCC
TTTTT-CTCC
TTTTT-CTCC
TTTT-CCTC-
TTTTTCCTCC
TTTT-CCTCC
34mer
AAACTCCTATTTTCTGAAAAcTCTTATTTTCTAA
AAACTCCTATTTTCTGAAAAtTCTTATTTTCTAA

AJL1
9mer
ACTTTTT-A
ACTTTTTTA
11mer
AACAA-TTTTT
AACAAATTTTT

Tlr Int A
26mer
AATAATAgTAGAATGAAAATTGAAAT
AATAATAtTAGAATGAAAATTGAAAT
28mer
ATTTATAATTTGATTAATATATTTTATG
ATTTATAATTTGATTAATATATTTTATG

pMicV-A1
31Amer
AATTTaAAAATATTTAACATTtAAAAAATTT
AATTTtAAAATATTTAACATTgAAAAAATTT
31Bmer
TAaAACATAAATTTACCATAAAATTATTAgT
TAtAACATAAATTTACCATAAAATTATTAtT
31Cmer
TATTTATTTAATAGATTtTAAAAATAATTTG
TATTTATTTAATAGATTcTAAAAATAATTTG
63mer
TGTTGTATTATTTGAaTTTTTATCTATTCATATTATTTTTCTTAAAACtTTcTCCGCGCTTT
TGTTGTATTATTTGAgTTTTTATCTATTCTTATTATTTTTCTTAAAACaTTtTCCGCGCTTT
58mer
AAACCCCTTTAAATtCAgTTTTCCCTCTTcAAATTTAAATTTATtATGTCCGTTGAAA
AAACCCCTTTAAATgCAtTTTTCCCTCTTtAAATTTAAATTTATgATGTCCGTTGAAA

Figure 5. Nucleotide sequence of direct repeats located within cloned DNA outside the Tlr internal conserved region. Nucleotides that do not match the consensus are shown in lower case.

Clone AJL1 was PCR amplified from *T.thermophila* micronuclear DNA using an outward-directed primer at the left end of the conserved element sequence in combination with a separate oligonucleotide that fortuitously primed near a previously uncharacterized Tlr family member (Materials and Methods). The resulting ~1.5 kb product was cloned and sequenced. Subsequent analysis revealed that AJL1 appears to contain the left boundary of a typical Tlr element (Fig. 4). Two sets of short direct repeats are located ~30 bp outside the point of divergence. The sequences of the Tlr-associated direct repeats diagrammed in Figure 4 are listed in Figure 5.

Tlr-associated direct repeats are multicopy and predominantly micronucleus limited

The Tlr1 19A and 19B tandem repeats were previously determined by Southern blot analyses to be associated with each other and with programmed DNA elimination at six or seven locations in the micronuclear genome, implying that Tlr1 may belong to a sub-family of three or four Tlr elements that contain these 19mers near their termini (14). Thus, it was of interest to determine genomic distribution and copy number of the direct repeats from newly described Tlr elements. Southern blots of *Hin*dIII-digested micronuclear and macronuclear DNA were probed with the direct repeat regions from Tlr Int A (Int A IN), pMBR 6C1 (6C1 IN) and pMicV-A1 (V-A1 IN). The locations of these probes within their respective clones are indicated in Figure 4. Each probe hybridized to multiple fragments in micronuclear DNA (Fig. 6), suggesting that these sequences are repeated elsewhere in the germline genome. Most of the direct repeats are apparently eliminated from the

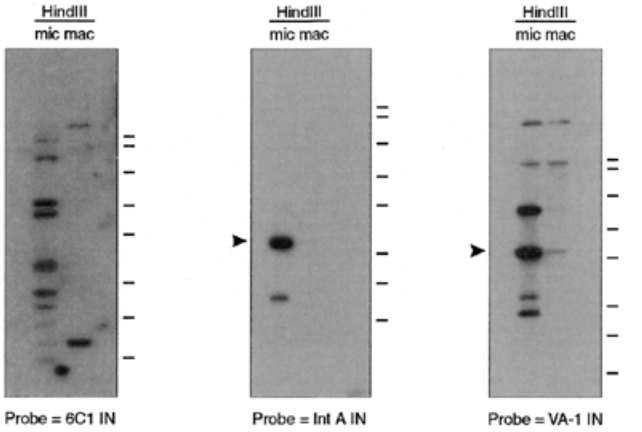

Figure 6. Southern blots of micronuclear (mic) and macronuclear (mac) DNA digested with *Hin*dIII and probed with fragments from direct repeat regions of cloned Tlr elements (Fig. 4A and B). Arrowheads indicate sizes of expected *Hin*dIII fragments based on restriction maps of cloned sequences. The ~3.1 kb minor band in macronuclear DNA that co-migrates with an intensely hybridizing fragment in the micronuclear DNA on the blot probed with VA-1 IN is likely a result of contamination of macronuclear DNA preparations with micronuclear DNA. Bars to the right of each blot indicate mobilities of the 10, 8.0, 6.0, 4.0, 3.0, 2.0, 1.5 and 1.0 kb DNA fragments of the HiLo marker.

macronucleus since the vast majority of hybridizing fragments are absent from macronuclear DNA. These results are consistent with a model in which Tlr elements are divided into sub-families based on the composition of their direct repeat regions. It should, however, be pointed out that the Southern blot experiments do not directly test whether the additional copies of these direct repeat regions are physically linked to other Tlr elements.

Based on restriction maps compiled from cloned sequences, the probes from Tlr Int A and pMicV-A1 are predicted to hybridize to 2.2 and 3.1 kb *Hin*dIII fragments, respectively. In both cases, discrete bands of the expected size were observed exclusively in lanes containing micronuclear DNA (Fig. 6). Thus, the excision boundaries for these two cloned elements are likely to be located outside their direct repeat region. Since pMBR 6C1 does not contain a *Hin*dIII site in the cloned sequence to the left of its internal conserved region, it was not possible to predict the length of the fragment corresponding to this particular element.

Probe 6C1 IN hybridized to ~1.2 and >10 kb *Hin*dIII fragments in macronuclear DNA (Fig. 6), indicating that the macronucleus contains at least two copies of this repeat region. No hybridizing fragments of micronuclear DNA co-migrated with these bands, suggesting that the macronucleus-retained *Hin*dIII fragments containing 6C1 repeat sequences also contain rearrangement junctions. Similarly, the 19A tandem repeats from the left side of the Tlr1 terminal inverted repeat are also located in macronucleus-retained DNA, just outside the predominant Tlr1 excision boundary. All other copies of the Tlr1 19mers are eliminated from the developing macronucleus (14). The weaker intensity of the >10 kb macronuclear *Hin*dIII fragment that hybridized to 6C1 IN is likely due to the lower transfer efficiency of large DNA molecules during the Southern blotting process. The VA-1 IN probe also hybridizes to two major *Hin*dIII fragments in macronuclear DNA (Fig. 6). However, in both cases, micronuclear bands of equal intensity

```
Tlr1     (L): AGAGAATTTACAATCGGAGCATTTTTAGATCAGTACAGGTAAAAAAA
Tlr1     (R): AGAGAATTTACAATCGGAGCATTTTTAGATCAGTACAGGTAAAAAAA
6C1      (L): AGAGAATTTACAATCGGAGCATTTTTAGATCAGTACAGGTAAAAAAA
AJL1     (L): AGAGAATTTACAATCGGAGCATTTTTAGATCAGTA--GGcAgAAAAA
pMicV-A1 (R): AGAGAATTTACAATCGGgGCtTTTTTAGATCAGTA--GGcAAtAAAA
```

Figure 7. Nucleotide sequences of conserved terminal 47mers from cloned Tlr elements. R or L in parentheses indicates whether a sequence is associated with the right or left end of its respective element. All right-end sequences are inverted to facilitate alignment with left-end sequences. Nucleotides represented by lower case letters indicate deviation from consensus.

co-migrate with the macronuclear bands, indicating that these *Hin*dIII fragments do not contain detectable rearrangement junctions.

Other Tlr elements appear to contain terminal inverted repeats

Clones pMBR 6C1, pMicV-A1 and AJL1 each contain a sequence that is almost identical to the outermost 47 bp of the Tlr1 terminal inverted repeat. In all three clones, this sequence is located just outside the final set of tandem repeats (Fig. 4). Figure 7 shows a nucleotide alignment of the terminal 47mers from Tlr clones. In pMBR 6C1 and AJL1, both left-end clones, the 47 bp sequence is oriented in the same direction as on the left side of the Tlr1 inverted repeat (Fig. 4). The right-end clone pMicV-A1 contains an inverted copy of this sequence, similar to the right boundary of the Tlr1 inverted repeat. This strongly suggests that, in addition to Tlr1, other Tlr elements may contain terminal inverted repeats minimally consisting of the conserved 47 bp sequence. Despite close inspection, no sequence similar to the terminal 47mer was detected in Tlr Int A, Tlr Int E or pMBR 4C1. However, since the 47mer from pMicV-A1 is located almost 3 kb outside the Tlr internal conserved region, it is possible that the additional elements may contain this nucleotide motif beyond the cloned sequence.

In addition to the conserved terminal 47mer, the Tlr1 inverted repeat also encompasses the 19A and 19B tandem repeats (14). The data described here suggest a model in which long terminal inverted repeats are a general feature of Tlr elements. According to this model, the Tlr inverted repeats would consist of the conserved terminal 47mer and sub-terminal direct repeat regions that are shared by sub-families of Tlr elements. Such a model would be consistent with the observed genomic copy number of these sub-terminal direct repeat regions.

In order to test this hypothesis, the pMBR micronuclear genomic library was screened with the left-end probe 6C1 IN, which contains the pMBR 6C1 tandem repeats (Fig. 4). Five of the unique clones isolated from this screening were subjected to further molecular analysis. Of these, one clone, designated pMBR 6C1 26A, hybridized to probe LSC 3 (Fig. 1B), indicating that, like pMBR 6C1, it contained sequences from the left end of the Tlr internal conserved region (data not shown). Two of the remaining clones, designated pMBR 6C1 7A and pMBR 6C1 15A, hybridized to probe INT V-C (Fig. 1B), indicating that they contained sequences from the right end of the Tlr internal conserved region (data not shown). The two clones that did not hybridize to internal Tlr probes, pMBR 6C1 3A and pMBR 6C1 16A, were partially sequenced. They also contained sequences from the right end of the Tlr internal conserved region. The left ends of these clones terminate to the right of the Tlr region encompassed by INT V-C,

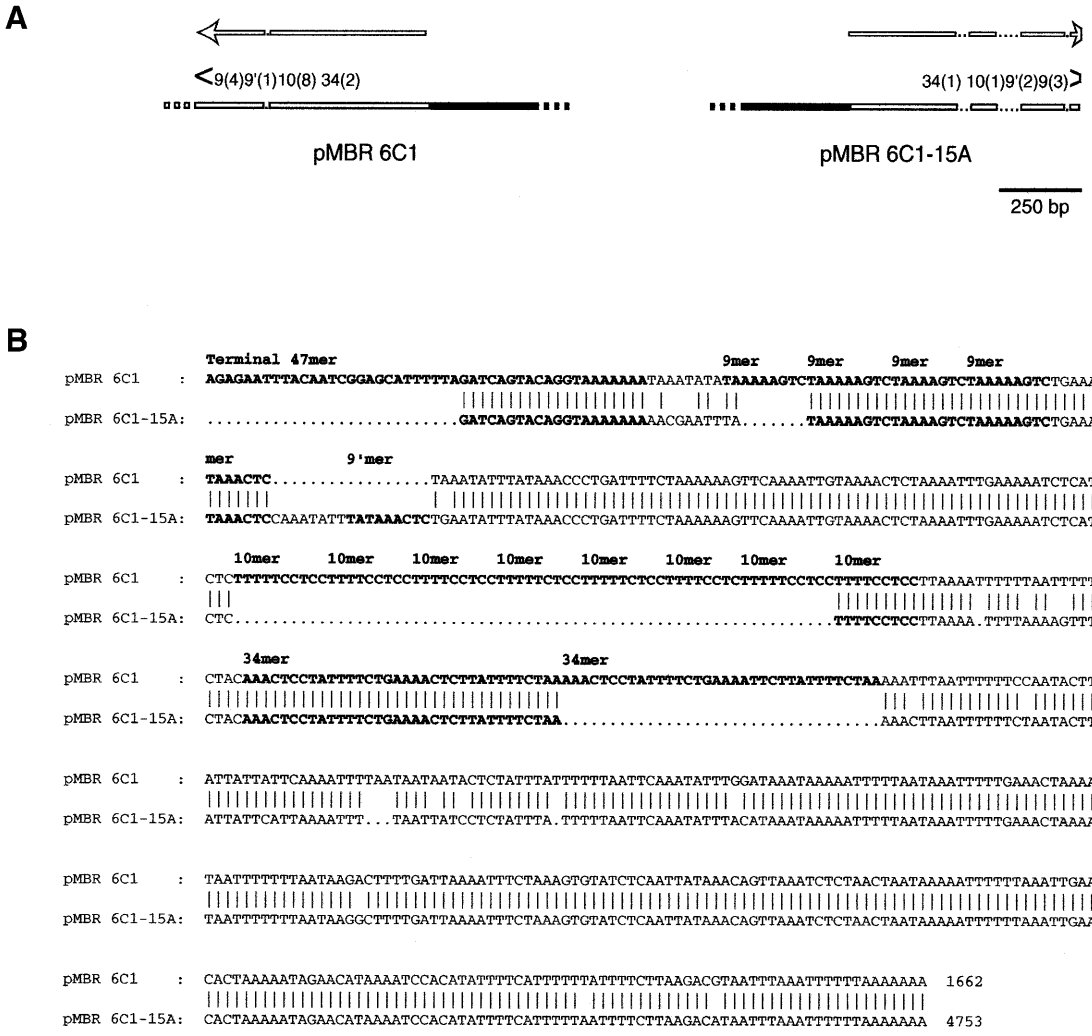

Figure 8. (**A**) Diagram of the sub-terminal regions of left-end clone pMBR 6C1 and right-end clone pMBR 6C1-15A. Solid bars indicate sequence within the Tlr internal conserved region. Open bars indicate sequence outside the Tlr internal conserved region. Overhead arrows depict the positions of the proposed 6C1 terminal inverted repeats. Dotted lines represent gap regions. Broken bars indicate continuation of cloned sequence beyond the diagram. Approximate locations of sub-terminal direct repeats are indicated above the clones. Numbers outside parentheses indicate repeat unit lengths and numbers inside parentheses indicate iterations. Greater than and less than symbols indicate location and orientation of conserved terminal 47mers. The cloned sequence of pMBR 6C1-15A ends within its terminal 47mer, indicated by a truncated arrowhead. (**B**) Gapped nucleotide alignment of the proposed terminal inverted repeat regions from pMBR 6C1 and pMBR 6C1-15A. pMBR 6C1-15A sequence is inverted to facilitate alignment. Terminal 47mer and 9mer, 10mer and 34mer are indicated in bold.

explaining their failure to hybridize to this probe. Thus, the 6C1 sub-terminal repeat region was associated with Tlr family members in all clones analyzed and 6C1-like repeats were found in both left- and right-end clones.

The terminal region of the Tlr element represented in right-end clone pMBR 6C1-15A was sequenced. Subsequent analysis revealed that this clone contains an apparently normal Tlr right boundary, diverging from the Tlr1 sequence 310 bp into the inverted repeat (4 nt beyond the point of pMicV-A1 divergence). As expected, the segment of pMBR 6C1-15A immediately outside this boundary strongly resembles the sub-terminal region of pMBR 6C1 in an inverted orientation (Fig. 8A). The similarity between these sequences begins 38 bp outside the left boundary of the pMBR 6C1 internal conserved region and extends into its terminal 47mer (Fig. 8B). Over this 695 bp region, the two aligned clones are 94% identical to each other with the exception of three short gaps in pMBR 6C1-15A and one short gap in pMBR 6C1. Interestingly, each gap corresponds to the location of the direct repeats in the opposite clone. The three short alignment gaps in pMBR 6C1-15A are at the 34mers, 10mers and 9mers in pMBR 6C1, leaving a single intact copy of the 34mer and 10mer sequences and only three copies of the 9mer sequence in pMBR 6C1-15A. Similarly, the short gap in pMBR 6C1 encompasses the 9′mer that is repeated twice in pMBR 6C1-15A, leaving a single copy of this sequence in pMBR 6C1. The cloned sequence from pMBR 6C1-15A ends at a GATC *Mbo*I recognition site 20 bp into the frequently observed terminal 47mer (Fig. 8B), suggesting that this element also contains the conserved terminal 47mer. Since pMBR 6C1 and pMBR6C1-15A were obtained independently, it is not possible to determine whether they are contiguous in the genome. Nevertheless, these data strongly suggest that, like Tlr1, other Tlr elements terminate in long, complex inverted repeats. Experiments are currently underway to further investigate this possibility.

Sequences flanking most Tlr elements are micronucleus limited

Tlr1 is eliminated from the developing macronucleus as part of an IES with variable excision boundaries (14). In most cell lines examined, the right rearrangement boundary is located ~50 bp outside the right half of the Tlr1 inverted repeat while the left boundary lies within the left half of the inverted repeat, between the 19A and 19B tandem repeats (14,33). Thus, macronucleus-retained DNA flanks Tlr1. As mentioned above, two copies of the 6C1 repeat region were detected in macronuclear DNA, suggesting that at least one member of the proposed 6C1 sub-family may also be located in close proximity to a DNA rearrangement junction (the data presented below show that this family member is different from that in clone pMBR 6C1).

In order to determine whether Tlr family members are generally located near macronucleus-destined sequences, PCR products from cloned regions outside the conserved 47 bp terminal sequence from pMBR 6C1, AJL1 and pMicV-A1 (Fig. 4) were used to probe Southern blots containing restriction enzyme-digested micro and macronuclear DNAs. Sequences flanking all three of these elements hybridized exclusively to micronuclear DNA (Fig. 9), demonstrating that, unlike Tlr1, the Tlr family members analyzed here are embedded within other micronucleus-limited DNA sequences. Since 6C1 OUT and VA1 OUT hybridize to multiple fragments (Fig. 9), the germline-limited sequences these two elements inserted into are themselves repeated.

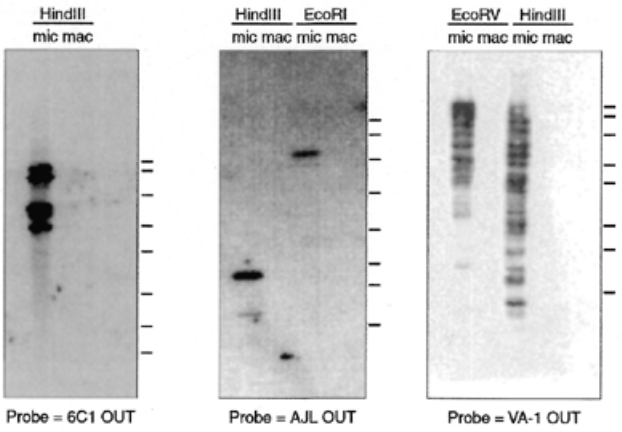

Figure 9. Southern blots of micronuclear (mic) and macronuclear (mac) DNA digested with *Hin*dIII, *Eco*RI or *Eco*RV and probed with DNA fragments from flanking regions of cloned Tlr elements (Fig. 4A and B). Bars to the right of each blot indicate the mobilities of the 10, 8.0, 6.0, 4.0, 3.0, 2.0, 1.5 and 1.0 kb DNA fragments of the HiLo marker.

DISCUSSION

The Tlr family of micronucleus-limited elements is comprised of approximately 30 closely related DNA sequences in the germline genome of *T.thermophila*. The data collected in this survey indicate that Tlr elements consist of a 22 kb internal region flanked by long terminal inverted repeats. The nucleotide sequence of the Tlr internal region is 90–97% conserved between family members and contains 15 predominantly non-overlapping ORFs distributed with equal frequency on both strands of the elements. Several of these ORFs show similarity to genes encoded by known mobile genetic elements.

Any ORF analysis based on sequence inspection alone is likely to be incomplete. In this study, relatively stringent criteria have been set for the designation of ORFs. However, the lack of experimental evidence regarding the presence or absence of introns may have led to overestimates or underestimates in the length and/or number of potential genes within Tlr elements. For example, there is a short 84 amino acid ORF between ORFs 3R and 4R (Fig. 1) that does not meet the length criterion of at least 100 amino acids. The nucleotide sequences of clones Tlr Int B and Tlr Int C are highly conserved in this region and the nucleotide polymorphisms between the two clones are non-random, suggesting that there was selective pressure to maintain an ORF. Thus this region may be a short ORF. There is, however, an equally likely alternative explanation. One significant difference between the two clones is a 1 nt deletion that would result in a frameshift. The polymorphism occurs in a stretch of 54 nt that has a G + C content of only 15%, a characteristic of introns in *Tetrahymena* (46). Joining of the short ORF to ORF 4R at consensus RNA splice sites would produce an in-frame fusion of 78 amino acids of the short ORF with ORF 4R. Resolution of these kinds of issues and finalization of the ORF analysis must await experimental evidence.

In contrast to the highly conserved nature of the 22 kb internal region, Tlr terminal inverted repeats vary dramatically in sequence and length between family members. Despite these striking differences, the proposed Tlr terminal inverted repeats appear to contain several common structural features, including (i) a series of direct repeats that themselves vary in number, length and position among elements and (ii) a terminal ~47 bp sequence that is conserved in most family members analyzed. Blocks of direct repeats from the sub-terminal regions of several Tlr elements were demonstrated by Southern blot analyses to be present at multiple locations mostly within the micronuclear, but not the macronuclear, genome. Although the Southern data do not address the frequency at which additional copies of the direct repeats are linked to Tlr elements, all five clones obtained from a library screening with probe 6C1 IN were found to be associated with Tlr elements. The data are consistent with the previously proposed model that the Tlr family may consist of several smaller sub-families of elements, each distinguished by the composition of their sub-terminal direct repeats (14). Comparison of the nucleotide sequences from the proposed terminal inverted repeat regions of pMBR 6C1 and pMBR 6C1-15A suggests that the copy number of each direct repeat might be variable among elements within a sub-family.

Given Tlr copy number, internal ORFs and terminal inverted repeat structure, it seems likely that these elements comprise a novel family of mobile genetic elements. The presence of a CTCGTG target site duplication flanking the Tlr1 terminal inverted repeat (14) further supports this hypothesis, since short duplications are often a hallmark feature of sequences surrounding the termini of transposons and integrated viruses (1). Unfortunately, Tlr1 is the only family member for which left and right termini of the same element have been isolated. Thus, the frequency at which target site duplications flank

other Tlr elements cannot be determined from the data currently available.

Statistical analyses on the distribution of nucleotide changes within ORFs 1R (32), 1F and 6F suggest that all three of these potential Tlr genes are, or recently were, under selective pressure to maintain functional proteins. Similar observations have been reported for ORFs from Tec and TBE micronucleus-limited transposon-like elements in *Euplotes* and *Oxytricha* (26,30). Our results imply that the proliferation of Tlr elements within the *Tetrahymena* germline genome may have been a relatively recent occurrence. In order to investigate this possibility further, efforts are currently underway to analyze the distribution of Tlr family members in numerous independently isolated *T.thermophila* cell lines.

Whether Tlr elements are currently functional is unknown. Mobility of the elements was not observed during the course of these experiments. However, even if they were fully functional, the apparent stability of Tlr family members within the micronuclear genome might not be entirely surprising since (i) many active mobile genetic elements transpose at extremely low or undetectable frequencies (1) and (ii) the micronucleus is transcriptionally silent throughout most of the *Tetrahymena* life cycle (47–49).

To our knowledge, the structure of Tlr elements does not closely resemble that of any previously characterized mobile genetic element. Thus, it is difficult to accurately predict their mechanism of proliferation. Since the DDE motif-containing putative protein encoded by ORF 1R is most similar to retroviral integrases (32), it is plausible that Tlr elements may transpose through an RNA intermediate. However, as pointed out by Gershan and Karrer (32), the Tlr integrase-like protein lacks the N-terminal HHCC zinc finger domain that is required for retroelement cDNA processing. Furthermore, Tlr elements lack an identifiable reverse transcriptase gene.

Several structural features of Tlr elements are also inconsistent with those of retroelements. At >22 kb in length, Tlr elements are considerably larger than most known retrotransposons and retroviruses. Also, Tlr elements contain terminal inverted repeats, which are common features of DNA-based mobile genetic elements. In contrast, the ends of most retroelements are comprised of long direct repeats or no terminal repeats at all (1). Only a small number of putative retroelements are known to contain terminal inverted repeats (50–52) and Tlr family members bear no additional similarities to these rare elements. Thus, it seems more likely that Tlr elements may transpose through a DNA intermediate. Although unusual, mobile DNA elements with a functional transposase resembling retroviral integrases would not be unprecedented (53). The bacterial IS*30* element, which transposes through a DNA intermediate, has been previously demonstrated to encode a transposase containing retroviral-like integrase active site domains (54). Like the Tlr 1R putative protein, the IS*30* transposase also lacks the N-terminal HHCC integrase domain.

Three of the Tlr ORFs characterized in this study are similar to genes encoded by various dsDNA viruses, including the phycodnaviruses PBCV-1 and EsV-1, as well as iridescent viruses and poxviruses. The functions of these putative proteins, if any, in the life cycle of Tlr elements is currently unknown. However, their inclusion within the Tlr genome is a further indication that Tlr elements may transpose through a DNA intermediate, since the genomes of dsDNA viruses have no known RNA stage. Interestingly, several dsDNA viruses have been reported to contain long terminal inverted repeats comprised of complex arrays of tandem repeats (40,55,56) and thus resemble the apparent structure of Tlr terminal inverted repeats.

These similarities suggest that Tlr elements could be evolutionarily related to eukaryotic dsDNA viruses. This observation is somewhat surprising since, unlike Tlr elements, very few eukaryotic DNA viruses are believed to integrate into their host genomes. A notable exception is EsV-1, which is known to enter a lysogenic phase by integrating at several chromosomal locations within its host, the brown algae *E.siliculosus* (reviewed in 57). Eukaryotic dsDNA viruses typically have extremely large genomes approaching 200–300 kb in length. Although Tlr elements are considerably longer and more complex than most known transposable elements, they are still an order of magnitude smaller than their viral counterparts. One possible interpretation of this observation is that Tlr elements may be an extremely simplified form of a lysogenic DNA virus. However, there is currently no evidence that Tlr elements exist in an extracellular form.

Alternatively, a progenitor Tlr element may have acquired viral ORFs through illegitimate recombination. Given the highly conserved nature of the Tlr genome, such an acquisition would have necessarily occurred before the proliferation of modern Tlr elements in *T.thermophila*. To date, no DNA viruses have been reported to infect ciliates. However, PBCV-1 is known to infect eukaryotic unicellular chlorella-like algae that live endosymbiotically within the ciliate *P.bursaria* (reviewed in 57). In any case, since Tlr ORFs apparently utilize the same unusual genetic code as their host, it seems likely that Tlr elements must have evolved for some time within the genetic background of *Tetrahymena* or a related ciliate species.

One of the most intriguing characteristics of Tlr elements is their confinement to the germline genome in *T.thermophila*. According to Southern blot analyses, all of the approximately 30 Tlr family members detected in the micronucleus are absent from macronuclear DNA, indicating that these elements are included in the ~15% of the germline genome that is deleted during development of the somatic macronucleus. All other putative mobile genetic elements characterized to date in ciliates are also exclusively micronucleus limited, including the Tel-1 (19) and Tmi elements (E. Blackburn and C. Wyman, personal communication) from *Tetrahymena*, as well as the Tec elements of *Euplotes* (23,25,58) and the TBE1 elements of *Oxytricha* (22).

The Tlr family members analyzed in this study appear to be imbedded within longer stretches of micronucleus-limited DNA. This finding represents a significant departure from the impression of Tlr DNA elimination taken only from Tlr1, which terminates near macronuclear rearrangement junctions (14,33). Similarly, sequences flanking Tel-1 elements are also predominantly micronucleus limited (19). In contrast, most Tec and TBE elements are precisely deleted with respect to element DNA boundaries (reviewed in 31). These dissimilar situations suggest that there may be important variations in the molecular pathways that recognize and delete mobile genetic elements from the developing macronuclei of different ciliate species.

At present, the mechanisms controlling programmed DNA elimination in *Tetrahymena* are unclear. It is conceivable that

the Tlr self-encoded transposition machinery could catalyze excision of these elements from the developing macronuclear genome. This possibility might explain the apparent pressure on Tlr ORFs to maintain functional proteins if deletion of Tlr elements from the somatic macronucleus is selectively advantageous to the host, the elements, or both. A similar model has also been suggested to explain the high degree of conservation observed in TBE1 ORFs (30).

However, since poly(A)$^+$ transcripts of the Tlr integrase-like ORF were not detected on northern blots (J. A. Gershan and K. M. Karrer, unpublished data), it seems more likely that Tlr transposition and Tlr programmed elimination are catalyzed by different mechanisms. Analyses of the *cis*-acting DNA signals involved in Tlr1 elimination seem to favor this model. rDNA-based *in vivo* rearrangement assays have demonstrated that the Tlr1 terminal inverted repeat is not required for programmed DNA deletion (J. D. Wuitschick and K. M. Karrer, unpublished observations). In contrast, terminal inverted repeats are thought to be critical structures for the delineation of excision boundaries during transposition of many mobile genetic elements. Additional experiments revealed that macronucleus-retained DNA sequences flanking Tlr1 are involved in controlling the location of its rearrangement junctions (59). This result seems incongruous with the necessity for mobile genetic element sequence requirements to be self-contained. It is, however, consistent with previously reported results demonstrating that *cis*-acting signals for programmed DNA elimination are located in the macronucleus-destined sequences flanking other deleted DNA regions in *Tetrahymena* (60–62). This suggests that Tlr elements may be removed from the developing macronucleus by the same mechanism as other micronucleus-specific sequences. Indeed, the elimination of Tlr1 has been demonstrated to occur during the same ~2 h window of macronuclear development as all other known *Tetrahymena* deletion elements (E. Capowski and K. M. Karrer, unpublished data).

One of several explanations may account for these observations. First, it is possible that during transposition Tlr elements might have preferentially integrated into regions of the micronuclear genome that are predisposed to undergo developmentally regulated DNA elimination. This model would be consistent with our findings that the Tlr elements characterized in this study are apparently embedded within other micronucleus-limited DNA sequences. Preferential transposition of mobile genetic elements into selective chromosomal environments has been well documented in several systems (reviewed in 63). At present, the mechanism by which Tlr elements might target micronucleus-specific sequences is unknown. It is also plausible that transposed Tlr elements within macronucleus-destined sequences may be unstable or extremely deleterious to host fitness. Thus, only the elements integrated within micronucleus-limited DNA would be detectable over time.

Alternatively, it is conceivable that components of the *Tetrahymena* DNA rearrangement machinery might recognize some feature of Tlr elements other than their terminal inverted repeats. Programmed DNA elimination in *T.thermophila* results in the removal of most repetitive sequences from the developing macronucleus. Coyne *et al.* (64) have postulated that deletion of repetitive regions of the genome, such as heterochromatic and centromeric DNA, may be a necessary step in the formation of functional macronuclear chromosomes, which are highly unusual in that they do not undergo condensation during nuclear division. Thus, the elimination of Tlr elements could be a consequence of their copy number.

ACKNOWLEDGEMENTS

We thank Elizabeth Blackburn and Claire Wyman for communicating unpublished results regarding Tmi elements. This work was supported by grant MCB-9974885 from the National Science Foundation to K.K. J.W. was supported in part by fellowships from the Arthur J. Schmitt Foundation, the Richard W. Jobling Trust Fund and the Marquette University Graduate School.

REFERENCES

1. Capy,P., Bazin,C., Higuet,D. and Langin,T. (1998) *Dynamics and Evolution of Transposable Elements*. Landes Bioscience, Austin, TX.
2. Bestor,T.H. (1998) The host defence function of genomic methylation patterns. *Novartis Found. Symp.*, **214**, 187–195.
3. SanMiguel,P., Tikhonov,A., Jin,Y.K., Motchoulskaia,N., Zakharov,D., Melake-Berhan,A., Springer,P.S., Edwards,K.J., Lee,M., Avramova,Z. and Bennetzen,J.L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.
4. Kidwell,M.G. and Lisch,D. (1997) Transposable elements as sources of variation in animals and plants. *Proc. Natl Acad. Sci. USA*, **94**, 7704–7711.
5. Fedoroff,N.V. (1999) Transposable elements as a molecular evolutionary force. *Ann. N. Y. Acad. Sci.*, **870**, 251–264.
6. Prescott,D.M. (1994) The DNA of ciliated protozoa. *Microbiol. Rev.*, **58**, 233–267.
7. Karrer,K.M. (2000) *Tetrahymena* genetics: two nuclei are better than one. *Methods Cell Biol.*, **62**, 127–186.
8. Davis,M.C., Ward,J.G., Herrick,G. and Allis,C.D. (1992) Programmed nuclear death: apoptotic-like degradation of specific nuclei in conjugating *Tetrahymena*. *Dev. Biol.*, **154**, 419–432.
9. Mpoke,S. and Wolfe,J. (1996) DNA digestion and chromatin condensation during nuclear death in *Tetrahymena*. *Exp. Cell Res.*, **225**, 357–365.
10. Altschuler,M.I. and Yao,M.C. (1985) Macronuclear DNA of *Tetrahymena thermophila* exists as defined subchromosomal-sized molecules. *Nucleic Acids Res.*, **13**, 5817–5831.
11. Conover,R.K. and Brunk,C.F. (1986) Macronuclear DNA molecules of *Tetrahymena thermophila*. *Mol. Cell. Biol.*, **6**, 900–905.
12. Yao,M.C., Choi,J., Yokoyama,S., Austerberry,C.F. and Yao,C.H. (1984) DNA elimination in *Tetrahymena*: a developmental process involving extensive breakage and rejoining of DNA at defined sites. *Cell*, **36**, 433–440.
13. Austerberry,C.F. and Yao,M.C. (1988) Sequence structures of two developmentally regulated, alternative DNA deletion junctions in *Tetrahymena thermophila*. *Mol. Cell. Biol.*, **8**, 3947–3950.
14. Wells,J.M., Ellingson,J.L., Catt,D.M., Berger,P.J. and Karrer,K.M. (1994) A small family of elements with long inverted repeats is located near sites of developmentally regulated DNA rearrangement in *Tetrahymena thermophila*. *Mol. Cell. Biol.*, **14**, 5939–5949.
15. Katoh,M., Hirono,M., Takemasa,T., Kimura,M. and Watanabe,Y. (1993) A micronucleus-specific sequence exists in the 5′-upstream region of calmodulin gene in *Tetrahymena thermophila*. *Nucleic Acids Res.*, **21**, 2409–2414.
16. Heinonen,T.Y. and Pearlman,R.E. (1994) A germ line-specific sequence element in an intron in *Tetrahymena thermophila*. *J. Biol. Chem.*, **269**, 17428–17433.
17. Chau,M.F. and Orias,E. (1996) Developmentally programmed DNA rearrangement in *Tetrahymena thermophila*: isolation and sequence characterization of three new alternative deletion systems. *Biol. Cell*, **86**, 111–120.
18. Huvos,P.E., Wu,M. and Gorovsky,M.A. (1998) A developmentally eliminated sequence in the flanking region of the histone H1 gene in *Tetrahymena thermophila* contains short repeats. *J. Eukaryot. Microbiol.*, **45**, 189–197.
19. Cherry,J.M. and Blackburn,E.H. (1985) The internally located telomeric sequences in the germ-line chromosomes of *Tetrahymena* are at the ends of transposon-like elements. *Cell*, **43**, 747–758.

20. White,T.C., el-Gewely,M.R. and Allen,S.L. (1985) Eliminated sequences with different copy numbers clustered in the micronuclear genome of *Tetrahymena thermophila*. *Mol. Gen. Genet.*, **201**, 65–75.
21. Yao,M.C. and Gorovsky,M.A. (1974) Comparison of the sequences of macro- and micronuclear DNA of *Tetrahymena pyriformis*. *Chromosoma*, **48**, 1–18.
22. Herrick,G., Cartinhour,S., Dawson,D., Ang,D., Sheets,R., Lee,A. and Williams,K. (1985) Mobile elements bounded by C4A4 telomeric repeats in *Oxytricha fallax*. *Cell*, **43**, 759–768.
23. Baird,S.E., Fino,G.M., Tausta,S.L. and Klobutcher,L.A. (1989) Micronuclear genome organization in *Euplotes crassus*: a transposonlike element is removed during macronuclear development. *Mol. Cell. Biol.*, **9**, 3793–3807.
24. Jahn,C.L., Krikau,M.F. and Shyman,S. (1989) Developmentally coordinated en masse excision of a highly repetitive element in *E. crassus*. *Cell*, **59**, 1009–1018.
25. Krikau,M.F. and Jahn,C.L. (1991) Tec2, a second transposon-like element demonstrating developmentally programmed excision in *Euplotes crassus*. *Mol. Cell. Biol.*, **11**, 4751–4759.
26. Jahn,C.L., Doktor,S.Z., Frels,J.S., Jaraczewski,J.W. and Krikau,M.F. (1993) Structures of the *Euplotes crassus* Tec1 and Tec2 elements: identification of putative transposase coding regions. *Gene*, **133**, 71–78.
27. Jaraczewski,J.W. and Jahn,C.L. (1993) Elimination of Tec elements involves a novel excision process. *Genes Dev.*, **7**, 95–105.
28. Hunter,D.J., Williams,K., Cartinhour,S. and Herrick,G. (1989) Precise excision of telomere-bearing transposons during *Oxytricha fallax* macronuclear development. *Genes Dev.*, **3**, 2101–2112.
29. Doak,T.G., Witherspoon,D.J., Doerder,F.P., Williams,K. and Herrick,G. (1997) Conserved features of TBE1 transposons in ciliated protozoa. *Genetica*, **101**, 75–86.
30. Witherspoon,D.J., Doak,T.G., Williams,K.R., Seegmiller,A., Seger,J. and Herrick,G. (1997) Selection on the protein-coding genes of the TBE1 family of transposable elements in the ciliates *Oxytricha fallax* and *O. trifallax*. *Mol. Biol. Evol.*, **14**, 696–706.
31. Klobutcher,L.A. and Herrick,G. (1997) Developmental genome reorganization in ciliated protozoa: the transposon link. *Prog. Nucleic Acid Res. Mol. Biol.*, **56**, 1–62.
32. Gershan,J.A. and Karrer,K.M. (2000) A family of developmentally excised DNA elements in *Tetrahymena* is under selective pressure to maintain an open reading frame encoding an integrase-like protein. *Nucleic Acids Res.*, **28**, 4105–4112.
33. Patil,N.S., Hempen,P.M., Udani,R.A. and Karrer,K.M. (1997) Alternate junctions and microheterogeneity of Tlr1, a developmentally regulated DNA rearrangement in *Tetrahymena thermophila*. *J. Eukaryot. Microbiol.*, **44**, 518–522.
34. Rogers,M.B. and Karrer,K.M. (1989) Cloning of *Tetrahymena* genomic sequences whose message abundance is increased during conjugation. *Dev. Biol.*, **131**, 261–268.
35. Gorovsky,M.A., Yao,M.C., Keevert,J.B. and Pleger,G.L. (1975) Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. *Methods Cell Biol.*, **9**, 311–327.
36. Karrer,K.M. (1986) The nuclear DNA of holotrichous ciliates. In Gall,J.G. (ed.), *The Molecular Biology of Ciliated Protozoa*. Acedemic Press, Orlando, FL, pp. 85–110.
37. Horowitz,S. and Gorovsky,M.A. (1985) An unusual genetic code in nuclear genes of *Tetrahymena*. *Proc. Natl Acad. Sci. USA*, **82**, 2452–2455.
38. Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, **1**, 945–951.
39. Gorbalenya,A.E. and Koonin,E.V. (1989) Viral proteins containing the purine NTP-binding sequence pattern. *Nucleic Acids Res.*, **17**, 8413–8440.
40. Afonso,C.L., Tulman,E.R., Lu,Z., Zsak,L., Kutish,G.F. and Rock,D.L. (2000) The genome of fowlpox virus. *J. Virol.*, **74**, 3815–3831.
41. Li,Y., Lu,Z., Burbank,D.E., Kutish,G.F., Rock,D.L. and Van Etten,J.L. (1995) Analysis of 43 kb of the *Chlorella* virus PBCV-1 330-kb genome: map positions 45 to 88. *Virology*, **212**, 134–150.
42. Gorbalenya,A.E. and Koonin,E.V. (1993) Helicases: amino acid sequence comparisons and structure–function relationships. *Curr. Opin. Struct. Biol.*, **3**, 419–429.
43. Kapitonov,V.V. and Jurka,J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl Acad. Sci. USA*, **98**, 8714–8719.
44. Xiong,Y. and Eickbush,T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.*, **9**, 3353–3362.
45. Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1984) Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.
46. Wuitschick,J.D. and Karrer,K.M. (1999) Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. *J. Eukaryot. Microbiol.*, **46**, 239–247.
47. Gorovsky,M.A. and Woodard,J. (1969) Studies on nuclear structure and function in *Tetrahymena pyriformis*. *J. Cell Biol.*, **42**, 673–682.
48. Bruns,P.J. and Brussard,T.B. (1974) Positive selection for mating with functional heterokaryons in *Tetrahymena pyriformis*. *Genetics*, **78**, 831–841.
49. Mayo,K.A. and Orias,E. (1981) Further evidence for lack of gene expression in the *Tetrahymena micronucleus*. *Genetics*, **98**, 747–762.
50. Cappello,J., Handelsman,K., Cohen,S.M. and Lodish,H.F. (1985) Structure and regulated transcription of DIRS-1: an apparent retrotransposon of *Dictyostelium discoideum*. *Cold Spring Harbor Symp. Quant. Biol.*, **50**, 759–767.
51. Ruiz-Perez,V.L., Murillo,F.J. and Torres-Martinez,S. (1996) Prt1, an unusual retrotransposon-like sequence in the fungus *Phycomyces blakesleeanus*. *Mol. Gen. Genet.*, **253**, 324–333.
52. Sharma,R., Bagchi,A., Bhattacharya,A. and Bhattacharya,S. (2001) Characterization of a retrotransposon-like element from *Entamoeba histolytica*. *Mol. Biochem. Parasitol.*, **116**, 45–53.
53. Capy,P., Vitalis,R., Langin,T., Higuet,D. and Bazin,C. (1996) Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? *J. Mol. Evol.*, **42**, 359–368.
54. Dalrymple,B., Caspers,P. and Arber,W. (1984) Nucleotide sequence of the prokaryotic mobile genetic element IS*30*. *EMBO J.*, **3**, 2145–2149.
55. Strasser,P., Zhang,Y.P., Rohozinski,J. and Van Etten,J.L. (1991) The termini of the *chlorella* virus PBCV-1 genome are identical 2.2-kbp inverted repeats. *Virology*, **180**, 763–769.
56. Delaroque,N., Muller,D.G., Bothe,G., Pohl,T., Knippers,R. and Boland,W. (2001) The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology*, **287**, 112–132.
57. Van Etten,J.L. and Meints,R.H. (1999) Giant viruses infecting algae. *Annu. Rev. Microbiol.*, **53**, 447–494.
58. Jahn,C.L., Nilles,L.A. and Krikau,M.F. (1988) Organization of the *Euplotes crassus* micronuclear genome. *J. Protozool.*, **35**, 590–601.
59. Patil,N.S. and Karrer,K.M. (2000) A developmentally regulated deletion element with long terminal repeats has *cis*-acting sequences in the flanking DNA. *Nucleic Acids Res.*, **28**, 1465–1472.
60. Godiska,R., James,C. and Yao,M.C. (1993) A distant 10-bp sequence specifies the boundaries of a programmed DNA deletion in *Tetrahymena*. *Genes Dev.*, **7**, 2357–2365.
61. Chalker,D.L., La Terza,A., Wilson,A., Kroenke,C.D. and Yao,M.C. (1999) Flanking regulatory sequences of the *Tetrahymena* R deletion element determine the boundaries of DNA rearrangement. *Mol. Cell. Biol.*, **19**, 5631–5641.
62. Fillingham,J.S., Bruno,D. and Pearlman,R.E. (2001) *Cis*-acting requirements in flanking DNA for the programmed elimination of mse2.9: a common mechanism for deletion of internal eliminated sequences from the developing macronucleus of *Tetrahymena thermophila*. *Nucleic Acids Res.*, **29**, 488–498.
63. Craig,N.L. (1997) Target site selection in transposition. *Annu. Rev. Biochem.*, **66**, 437–474.
64. Coyne,R.S., Chalker,D.L. and Yao,M.C. (1996) Genome downsizing during ciliate development: nuclear division of labor through chromosome restructuring. *Annu. Rev. Genet.*, **30**, 557–578.