



Non-mammalian c-integrases are encoded by giant transposable elements

Cédric Feschotte and Ellen J. Pritham

Department of Biology, The University of Texas at Arlington, TX 76019, USA

In a recent report in *Trends in Genetics*, Gao and Voytas [1] described a new family of integrase genes that were identified in diverse eukaryotic species, including slime mold, *Caenorhabditis elegans*, *C. briggsae*, zebrafish, *Takifugu rubripes*, *Xiphophorus maculatus*, cow, dog and humans. These genes potentially encode proteins containing ~400 amino acids with homology to retroviral integrases and transposases, including the integrase-like proteins encoded by Tlr, a family of atypical mobile elements with long terminal inverted repeats (TIRs) from the ciliate *Tetrahymena thermophila* [2,3].

Despite the similarity of c-integrases to the Tlr integrases and their presence in multiple copies in some genomes, Gao and Voytas found no evidence linking the c-integrase genes to retroelements [1]. The authors did not exclude the possibility that the c-integrases might be part of an unusual type of mobile element, but instead proposed that the c-integrases are 'host' genes. Two observations supported this hypothesis: (i) an excess of synonymous to nonsynonymous substitutions among c-integrase genes, indicative of purifying selection, and; (ii) the distant relationship of c-integrases to Fob1p, a protein from *Saccharomyces cerevisiae* involved in rDNA metabolism.

Two observations prompted us to investigate the origin of c-integrases. First, we noticed that the ESTs encoding *D. discoideum* c-integrases reported by Gao and Voytas displayed >99% nucleotide identity with the currently active mobile element Tdd-4 (GenBank accession number U57081). Tdd-4 elements essentially consist of the integrase gene flanked by ~125-bp TIRs, a structure reminiscent of DNA transposons [4]. Second, we were intrigued that they found two distinct pairs of c-integrases from *C. elegans* to be part of larger duplicated genomic regions flanked by large inverted-repeats (IRs; see Figure 3 in Ref. [1]). In light of the relationship between *D. discoideum* c-integrases and the TIR-containing Tdd-4 transposons, the presence of IRs flanking the *C. elegans* c-integrase genes might indicate that these genes are part of larger mobile elements.

A hallmark of mobile-element transposition is the duplication of a short genomic sequence at the site of chromosomal integration [5]. These target-site duplications (TSDs) result from the staggered cleavage mediated by integrases and transposases. We examined each flank of the three IR pairs associated with the *C. elegans* c-integrase genes and found that in all three cases, the IRs were immediately flanked by a 6-bp direct

repeat of variable sequence (Table 1 in the supplementary material online). The presence of a short, direct repeat immediately flanking the IRs and its conservation in size is consistent with the hypothesis that this repeat represents the TSD provoked by insertion of large (7–17 kb) transposable elements (TEs) delimited by TIRs and carrying the c-integrases. We referred to these giant repeats as *Maverick* elements.

Gao and Voytas also reported multiple copies of c-integrases from the genomes of the nematode *C. briggsae* and from the zebrafish *Danio rerio*. We compared the flanking genomic regions of highly similar c-integrases from both of these genomes and found that, as in *C. elegans*, these genes lie within larger interspersed repeats delimited by long TIRs. In addition, 16 out of 18 complete elements (i.e. with both TIRs) were flanked by a 6-bp direct repeat, representing putative TSDs (Table 1 in the supplementary material online). These findings indicate that, in both nematodes and zebrafish, c-integrases were duplicated and propagated as *Maverick* elements. Within a genome, *Maverick* elements can be grouped into families with moderate copy number (three to ~40) and a high level of nucleotide sequence identity among copies of the same family (92–99%; Table 1 supplementary material online). This is indicative of recent propagation of *Mavericks* in these genomes.

To demonstrate whether the 6-bp duplications that flank *Maverick* elements are created following insertion of these elements, we searched for conserved paralogous insertion sites of *Maverick* elements that might be devoid of the insertion. We identified such sites in the genomes of *C. briggsae* and *D. rerio*. These sites occur in multiple copies in these genomes because they are part of other repeat families (Figure 1 in the supplementary material online; data not shown). In both cases, comparison of the repeat copies with and without the *Maverick* insertion showed that the 6-bp sequence was present in only one copy in the pre-integration site and confirmed that it was duplicated following insertion of the *Maverick* element (Figure 1 in the supplementary material online). Together, these data provide evidence for the past mobility of *Maverick* elements within the respective host genomes and indicate that their insertion creates a TSD of conserved length, a hallmark of TE integration.

However, we found no evidence that the mammalian c-integrases are currently part of *Maverick* elements. It is possible that these have evolved from *Maverick* elements into stationary genes with a cellular function. However,

Corresponding author: Feschotte, C. (cedric@uta.edu).

their phylogenetic relationship to other c-integrases is unclear [1] and apparently they have been lost from the rodent lineage (data not shown). *Maverick* elements from nematodes and zebrafish share similarities to each other and to the ciliate Tlr elements [2,3], both in terms of structure (large size, long TIRs, 6-bp TSD) and coding capacity [integrase-like open reading frame (ORF) and a set of additional conserved ORFs that we will describe elsewhere]. How these elements propagate is at present unknown, but they presumably define a novel group of eukaryotic mobile elements.

Acknowledgements

We thank Dan Voytas and Xiang Gao for critical reading and suggestions on an earlier version of this article. This work was supported by funds from the Department of Biology, University of Texas at Arlington.

Supplementary data

Supplementary data associated with this article can be found at [doi:10.1016/j.tig.2005.07.007](https://doi.org/10.1016/j.tig.2005.07.007)

References

- 1 Gao, X. and Voytas, D.F. (2005) A eukaryotic gene family related to retroelement integrases. *Trends Genet.* 21, 133–137
- 2 Wells, J.M. *et al.* (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
- 3 Wuitschick, J.D. *et al.* (2002) A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. *Nucleic Acids Res.* 30, 2524–2537
- 4 Wells, D.J. (1999) Tdd-4, a DNA transposon of *Dictyostelium* that encodes proteins similar to LTR retroelement integrases. *Nucleic Acids Res.* 27, 2408–2415
- 5 Craig, N.L. *et al.* (2002) *Mobile DNA II*, American Society for Microbiology Press