Commentary

Treasures in the attic: Rolling circle transposons discovered in eukaryotic genomes

Cédric Feschotte and Susan R. Wessler*

Departments of Botany and Genetics, University of Georgia, Athens, GA 30602

Since the advent of methodologies to analyze the content of whole genomes (e.g., renaturation kinetics and $C_o t$ analysis), it has been known that a large fraction of eukaryotic genomes is highly repetitive (1, 2). Recent computer-assisted analysis of several sequenced eukaryotic genomes, including *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and humans, has demonstrated that most repetitive DNA is composed of or derived from transposable elements (TEs). In the human genome, for example, TEs are the single most abundant component, accounting for over 40% of the total DNA (3). Although this amount of TEs is viewed as a hindrance to those engaged in the determination and assembly of DNA sequence, the availability of both complete and partial eukaryotic genome sequences is providing TE biologists with a bonanza of raw material that is being used to understand how genomes evolve.

Before the report in PNAS by Kapitonov and Jurka (4), all eukaryotic TEs were thought to use one of two mechanisms for transposition. Class 1, or retrotransposons, transpose via an RNA intermediate in reactions catalyzed by element-encoded proteins, including reverse transcriptase. In contrast, the transposon itself is the intermediate for class 2 elements where an element-encoded transposase catalyzes reactions, resulting in TE excision from one site and reinsertion elsewhere in the genome (the so-called cut-and-paste mechanism). In addition to these two mechanisms, some prokaryotic TEs (called IS or insertion sequences), move by another mechanism called rolling circle (RC) transposition (5, 6). This process is similar to the RC replication of some plasmids, single-stranded (ss) bacteriophage, and plant geminiviruses. In a recent issue of PNAS, Kapitonov and Jurka (4) report that RC transposons also occur in eukaryotes where, surprisingly, they comprise about 2% of the genomes of *A. thaliana* and *C. elegans*.

How could a group of TEs that account for such a large fraction of the genomes of these well-studied organisms remain until now essentially unknown? One answer to this question is that RC transposons have distinct structural features that are not easily detected by computer-assisted searches of DNA sequence databases. *Helitron* families of elements (as the eukaryotic RC transposons are called) do not generate target site duplications on insertion, as do all other eukaryotic TEs. These short duplications are derived from staggered endonucleolytic cleavage of the target DNA by element-encoded transposase or integrase. Instead, *Helitrons* target the dinucleotide AT, and insertion does not lead to the duplication of this sequence. Similarly, RC transposons do not have terminal inverted repeats, as do all other class 2 elements. Rather, *Helitrons* begin with a 5′ TC and end with a 3′ CTRR (Fig. 1*a*). Although there is a 16- to 20-nt palindrome just upstream of the 3′ CTRR, conservation of palindrome structure but not sequence would apparently preclude the use of a consensus sequence in the identification of *Helitrons* by computer-assisted searches. By analogy to RC mechanisms in prokaryotes, the distinct structural hallmarks of *Helitrons* are hypothesized to be essential for RC-mediated transposition (Fig. 1).

Helitrons may also have escaped classification for so long because the vast majority of family members are nonautonomous, defective elements that resemble internal deletion derivatives of their cognate autonomous element. It is important to note that up to 10 homogeneous subfamilies of nonautonomous *Helitrons*, with members ranging from 0.5 to 3 kb, were previously identified in the *Arabidopsis* genome as abundant repeats. These elements were first designated *AthE1* (7) and *AtREP* (8) and, later, *Basho* (9). However, in the absence of any obvious structural features of either class 1 or class 2 elements, these repeat families remained mysterious and unclassified. It was only when the complete genome sequence of *Arabidopsis* became available that Kapitonov and Jurka (4) were able to identify the much less abundant but very large (5.5 to 15 kb) *Helitrons* that have coding capacity for products related to RC replication proteins.

Although rare in prokaryotes, nonautonomous elements are common and abundant members of most eukaryotic transposon families. They are usually internally deleted derivatives of autonomous members and lack coding capacity for the transposase. Because most DNA transposon families contain distinct groups of nonautonomous elements that are conserved in both sequence and length, it is likely that most subfamilies arose from a single or a few deleted copies that were subsequently amplified with enzymes encoded in trans by an autonomous element. This seems to be the case for the RC-transposing *Helitrons*, because homogeneous groups of defective elements sharing their termini with autonomous copies are abundant in the *A. thaliana*, *Oryza sativa* (rice), and *C. elegans* genomes. Although nonautonomous RC transposons have not been reported in prokaryotes, engineered nonautonomous copies of the *Escherichia coli* RC element IS*91* transposed at high frequency when supplied with transposase in trans (5, 6).

What is still mysterious is how the RC mechanism generates nonautonomous elements. For other eukaryotic class 2 elements, it has been shown that such defective copies can arise by incomplete double strand gap repair after excision of an autonomous element (10–12). It is unlikely that a similar mechanism can account for the origin of nonautonomous *Helitrons* because they presumably do not excise as double-stranded molecules and thus do not create a double strand gap at the donor site. Nevertheless, recombination and slippage during the copying of the transposed single strand at the donor site may account for the origin of internally deleted *Helitrons* (see Fig. 1*b*). Alternatively, nonautonomous *Helitrons* may form *de novo* from host sequences given the minimal cis requirements that appear to be necessary for RC-mediated transposition.

Other open questions concern the function and origin of the putative genes encoded by the larger *Helitrons.* The preliminary analysis of Kapitonov and Jurka (4) suggests that *Helitrons* from *A. thaliana*, *O.*

See companion article on page 8714 in issue 15 of volume 98.

*To whom reprint requests should be addressed. E-mail: sue@dogwood.botany.uga.edu.

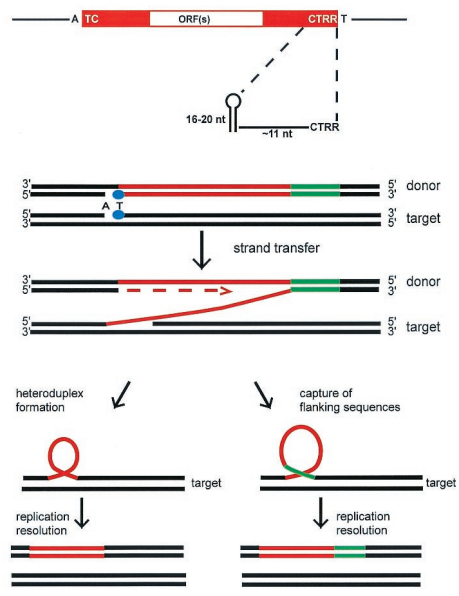

Fig. 1. Structure of *Helitron* elements and the rolling circle transposition mechanism. (*a*) A generic *Helitron* showing sequences and structural features that may be cis requirements for transposition (see text for details). *Helitrons* from *C. elegans* contain a single gene whereas *Helitrons* from *A. thaliana* and *O. sativa* contain two or three. (*b*) A hypothetical mechanism for *Helitron* transposition and gene acquisition based on the proposed rolling circle mechanism for bacterial transposons (e.g., IS*91*; refs. 5 and 18). The element (in red) could be either autonomous or nonautonomous. Two transposase molecules are shown (blue ellipses) cleaving at the donor and target sites and binding to the resulting 5′ ends. Replication at the cleaved donor site initiates at the free 3′ OH and proceeds to displace one strand of *Helitron*. If the palindrome and 3′ end of the element are recognized correctly, as is shown on the *Left*, cleavage occurs after the CTRR sequence and the one *Helitron* strand is transferred to the donor site where DNA replication resolves the heteroduplex. The illustration on the *Right* depicts one way by which DNA flanking the 3′ end of the element (in green) could be transferred along with the element to the donor site. This may be how *Helitrons* have acquired additional coding sequences.

sativa, and *C. elegans* have coding capacity for a large product of ≈1500 aa that contains an ≈500-aa domain similar to eukaryotic, prokaryotic, and viral 5′ to 3′ DNA helicases. These putative products of *Helitron* also share motifs with the replicator initiator proteins of RC plasmids and certain ssDNA viruses. More surprisingly, the plant *Helitrons* harbor additional genes that are related to RPA70, the largest subunit of replication protein A. RPA70 is a cellular ssDNA-binding protein that is conserved in plants, animals, and fungi. The gene richness of plant *Helitrons* is in sharp contrast with other class 2 transposons, including bacterial RC insertion sequences, which usually encode only one protein, a transposase. Whereas it has been shown *in vitro* that the transposase alone is sufficient to mediate the cut-and-paste mechanism (13–15), it is known that host-encoded factors are also required *in vivo* for most transposition reactions (16–18). Similarly, prokaryotic RC transposition has been shown to require host-encoded helicases and ssDNA-binding proteins (18). The identification of motifs for some of these functions among the *Helitron*-encoded products suggests a scenario whereby prokaryotic and eukaryotic RC elements arose from a common ancestral element, but that eukaryotic *Helitrons* have evolved further through the capture of additional functions from their host. A hypothetical mechanism for the acquisition of host genes by RC elements is depicted in Fig. 1*b* and is based on results showing that transposition of bacterial IS*91* and presumably of *Helitrons* has minimal cis requirements. That is, only the 5′ end of IS*91* is required to initiate transposition (5), whereas a cryptic downstream palindrome could furnish a new terminator if the normal terminator was bypassed. Whatever the mechanism, transduction events must occur with sufficient frequency to permit the eventual capture of useful genes or exons. In this regard, it is tempting to view *Helitrons* as "exon shuffling machines."

Although *Helitrons* are the first RC transposons identified in eukaryotic genomes, an RC mechanism is known to be responsible for the replication of geminiviruses, a group of ssDNA viruses that infect many plant species (19). Some of these viruses encode a *Rep* protein with both helicase and ssDNA-binding activities that can interact with the cellular machinery of DNA replication (20, 21). As suggested by Kapitonov and Jurka (4), it is possible that *Helitrons* represent the missing evolutionary link between prokaryotic RC elements and geminiviruses. Alternatively, *Helitrons* may have arisen from geminiviruses that were integrated into the genome of an early eukaryotic ancestor. On the surface, this scenario seems unlikely because integration into the host genome is not part of the geminivirus life cycle; that is, replication occurs extrachromosomally. However, it is noteworthy that multiple copies of geminivirus DNA have been found integrated into the chromosomes of tobacco (22). In the context of this commentary, this is probably not a surprising finding. Like integrated geminiviruses, RC transposons can now be added to a growing list of entities known to reside in eukaryotic genomes. More and more, genomes are beginning to resemble the family attic where the relics and mementos of several lifetimes are stored and await discovery.

1. Britten, R. J., Graham, D. E. & Neufeld, B. R. (1974) *Methods Enzymol.* **29,** 323–405.
2. Goldberg, R. B. (1978) *Biochem. Genet.* **16,** 45–68.
3. International Human Genome Sequencing Consortium (2001) *Nature (London)* **409,** 860–921.
4. Kapitonov, V. V. & Jurka, J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8714–8719. (First Published July 10, 2001; 10.1073/pnas.151269298)
5. Mendiola, M. V., Bernales, I. & de la Cruz, F. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1922–1926.
6. del Pilar Garcillan-Barcia, M., Bernales, I., Mendiola, M. V. & de la Cruz, F. (2001) *Mol. Microbiol.* **39,** 494–501.
7. Surzycki, S. A. & Belknap, W. R. (1999) *J. Mol. Evol.* **48,** 684–691.
8. Kapitonov, V. V. & Jurka, J. (1999) *Genetica* **107,** 27–37.
9. Le, Q. H., Wright, S., Yu, Z. & Bureau, T. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7376–7381.
10. Engels, W. R., Johnson-Schlitz, D. M., Eggleston, W. B. & Sved, J. (1990) *Cell* **62,** 515–525.
11. Plasterk, R. H. (1991) *EMBO J.* **10,** 1919–1925.
12. Rubin, E. & Levy, A. A. (1997) *Mol. Cell. Biol.* **17,** 6294–6302.
13. Kaufman, P. D. & Rio, D. C. (1992) *Cell* **69,** 27–39.
14. Vos, J. C., De Baere, I. & Plasterk, R. H. (1996) *Genes Dev.* **10,** 755–761.
15. Lampe, D. J., Churchill, M. E. & Robertson, H. M. (1996) *EMBO J.* **15,** 5470–5479.
16. Mizuuchi, K. (1992) *Annu. Rev. Biochem.* **61,** 1011–1051.
17. Beall, E. L. & Rio, D. C. (1996) *Genes Dev.* **10,** 921–933.
18. Mahillon, J. & Chandler, M. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 725–774.
19. Stenger, D. C., Revington, G. N., Stevenson, M. C. & Bisaro, D. M. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8029–8033.
20. Koonin, E. V. & Ilyina, T. V. (1992) *J. Gen. Virol.* **73,** 2763–2766.
21. Gutierrez, C. (2000) *EMBO J.* **19,** 792–799.
22. Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 759–764.