Review

The diversity of LTR retrotransposons

Ericka R Havecker, Xiang Gao and Daniel F Voytas

Address: Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA.

Correspondence: Daniel F Voytas. E-mail: Voytas@iastate.edu

Published: 18 May 2004

Genome ***Biology*** 2004, **5:**225

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/6/225

Abstract

Eukaryotic genomes are full of long terminal repeat (LTR) retrotransposons. Although most LTR retrotransposons have common structural features and encode similar genes, there is nonetheless considerable diversity in their genomic organization, reflecting the different strategies they use to proliferate within the genomes of their hosts.

Transposons are mobile genetic elements that can multiply in the genome using a variety of mechanisms. Retrotransposons replicate through reverse transcription of their RNA and integration of the resulting cDNA into another locus. This mechanism of replication is shared with retroviruses, with the difference that retrotransposons do not form infectious particles that leave the cell to infect other cells. The long terminal repeat (LTR) retrotransposons, one of the main groups of retroelements (which include both LTR and non-LTR retrotransposons as well as retroviruses), are among the most abundant constituents of eukaryotic genomes. The LTRs are the direct sequence repeats that flank the internal coding region, which - in all autonomous (functional) LTR retrotransposons - includes genes encoding both structural and enzymatic proteins. The *gag* gene encodes structural proteins that form the virus-like particle (VLP), inside which reverse transcription takes place. The *pol* gene encodes several enzymatic functions, including a protease that cleaves the Pol polyprotein, a reverse transcriptase (RT) that copies the retrotransposon's RNA into cDNA, and an integrase that integrates the cDNA into the genome.

Much of what we know about the mechanisms of LTR retrotransposition (Figure 1) comes from work on yeast retrotransposons [1,2], but it is generally assumed that the mechanism is very similar among LTR retrotransposons from divergent hosts. First, a retrotransposon's RNA is transcribed by the cellularly encoded RNA polymerase II from a promoter located within the 5′ LTR. The RNA is then translated in the cytoplasm to give the proteins that form the VLP and carry out the reverse transcription and integration steps. Typically, two RNA molecules are packaged into one virus-like particle, and the RNA is subsequently made into a full-length DNA copy through a reverse transcription reaction that is first primed from a tRNA that pairs to a sequence near the 5′ LTR (the primer-binding site). The resulting partial cDNA (called 'strong stop' DNA) is transferred from the 5′ LTR to the 3′ LTR, where reverse transcription proceeds. A second priming event initiates at a polypurine tract near the 3′ LTR. The cDNA primed from the polypurine tract undergoes an additional strand transfer, ultimately giving rise to a double-stranded cDNA molecule. Finally, the cDNA is integrated back into the host DNA, adding another copy of the retrotransposon to the genome.

LTR retrotransposon diversity

As genome-sequence data has accumulated for a large number of eukaryotes, it has become clear that the genomes of most organisms contain LTR retrotransposons from multiple distinct lineages. Although all are flanked by LTRs and encode *gag* and *pol* genes, the lineages diverge considerably in their DNA sequences and genomic organization. The International Committee on Taxonomy of Viruses has attempted to provide a taxonomic framework for understanding the relationships among the vast numbers of retrotransposons

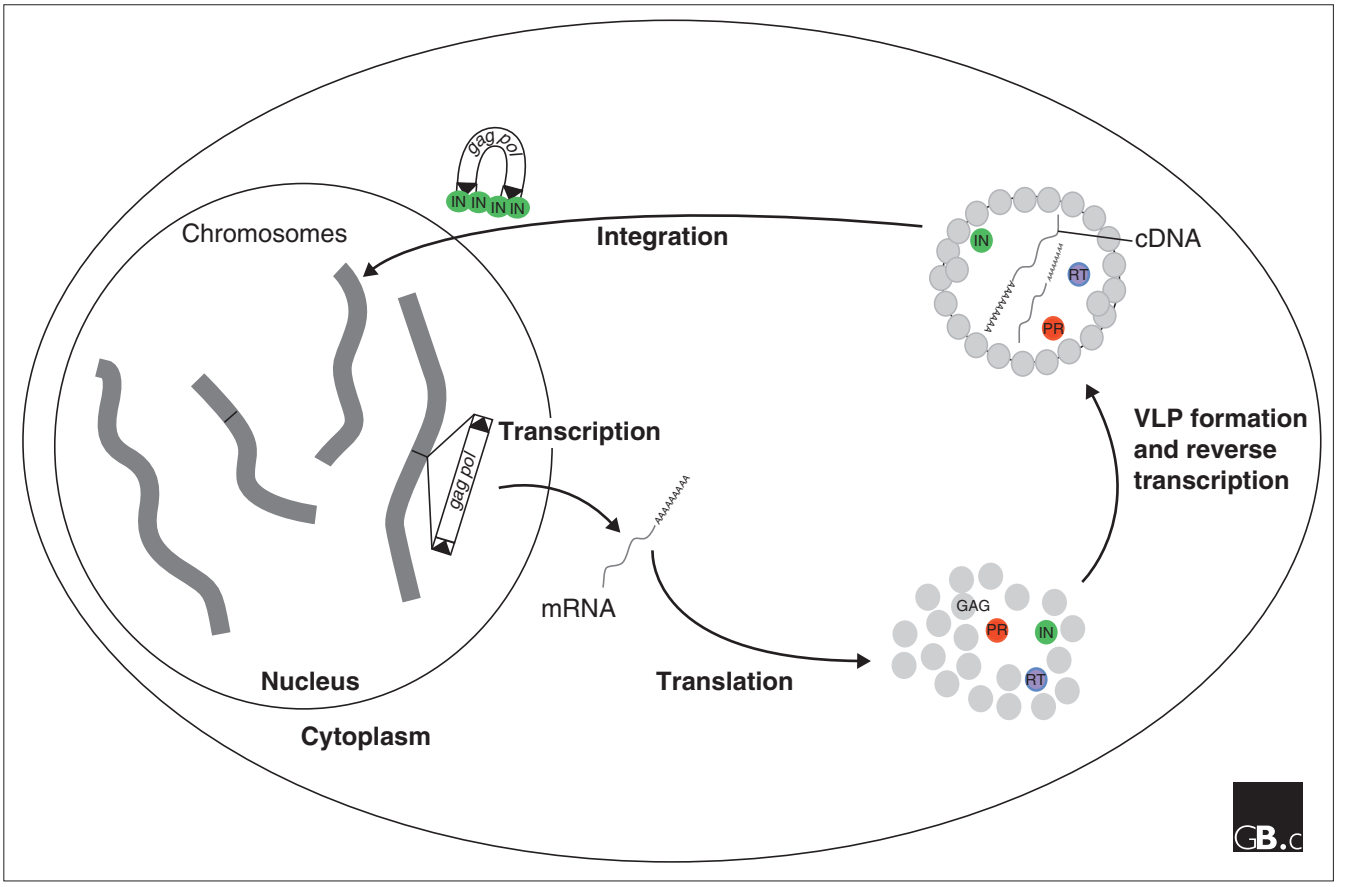

Figure 1
The life cycle of LTR retrotransposons. IN, integrase; PR, protease; RT, reverse transcriptase; VLP, virus-like particle. Black triangles represent the LTRs.

that have come to light through genome-sequence analysis [3,4] (Figure 2); this framework is based on relationships among the amino-acid sequences of the RT protein, the most highly conserved of the retrotransposon proteins. Two retrotransposon families - the *Pseudoviridae* and the *Metaviridae* - have been described in detail; both are found in most eukaryotes. The two families are also distinguished by the order of the coding regions within their *pol* genes (see Figure 3). Discovery of the *Gmr1* retrotransposon from Atlantic cod and related elements has shown that some members of the *Pseudoviridae* (on the basis of RT sequence) have a gene order characteristic of *Metaviridae* [5].

As with any taxonomic framework, the LTR retrotransposon classification system undergoes frequent revision as diverse elements are identified. This is particularly true for the genera that make up the two main families. Three genera have been proposed for the *Pseudoviridae* (Figure 2): pseudoviruses, hemiviruses and sireviruses (whose names do not necessarily indicate that they are viruses; Figure 2). The sireviruses derive from plant hosts and make up a distinct lineage according to their RT amino-acid sequences; the pseudoviruses and hemiviruses are distinguished by the primer used for reverse transcription (a full tRNA or a half tRNA, respectively). Note that this classification does not correspond directly with the phylogenetic relationships of the retrotransposons, so that the pseudoviruses make up three distinct lineages (Figure 2). The *Metaviridae* also comprises three genera - the metaviruses, the errantiviruses and the semotiviruses - which can be discriminated by phylogenetic analysis of RT amino-acid sequences. A distinct lineage of elements, the DIRS group (named after the founding member from *Dictyostelium discoideum*), has yet to be placed within the taxonomic framework. In addition to having characteristic RT sequences, the DIRS elements have some unusual features: they lack a protease and have a tyrosine recombinase instead of an integrase [6,7].

Organization of the *gag* and the *pol* genes

Whereas RT amino-acid sequences and the order of domains within *pol* are sufficiently conserved to be used to classify the LTR retrotransposons, the ways in which *gag* and *pol* are organized and expressed vary considerably. As multiple proteins are encoded on one mRNA, the *gag* and *pol* genes in some LTR retrotransposons are separated by a frameshift or

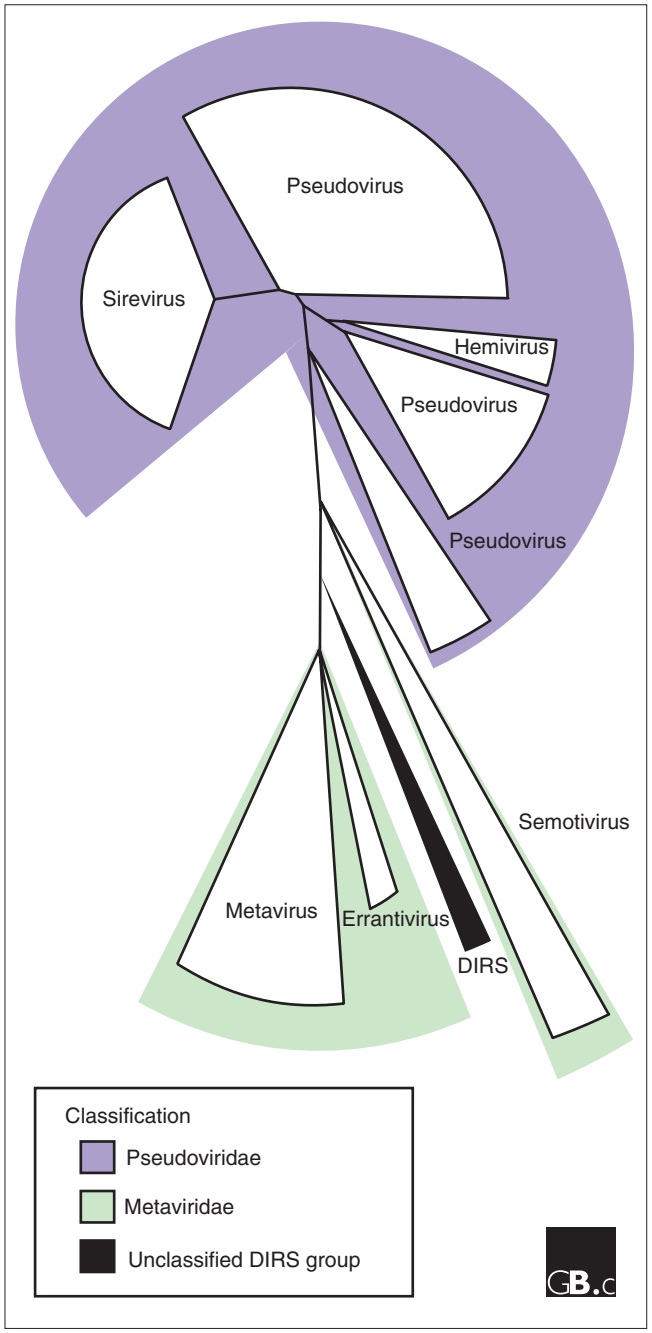

Figure 2
A schematic tree and classification of LTR retrotransposons. The sectors represent the diverse elements that make up each distinct lineage. The DIRS lineage is named for the founding member from *Dictyostelium discoideum*. Adapted from [3,4].

a stop codon, and occasionally these breaks in the reading frame are ignored by the translational machinery. Much more Gag than Pol is needed for productive VLP formation and consequently for replication of the retrotransposon; the use of either a stop codon that is occasionally ignored or ribosomal frameshifting (strategies called recoding) are used to regulate the ratio of the two proteins. We [8] have analyzed the genome sequences of *Caenorhabditis elegans*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Candida albicans* and *Arabidopsis thaliana* to predict the strategies used to express their *gag* and the *pol* genes. By analyzing the genomic structure and the nucleotide sequences surrounding the *gag-pol* junction, the type of recoding used for translation of the Pol protein could be inferred [8]. The results indicated that the mechanism used to express Pol is related to the host from which the retrotransposon originates. For example, about 50% of the retrotransposons identified in the study had a single open reading frame (ORF) fusing Gag and Pol, and this organization was the one found most often in plant elements. A single Gag-Pol ORF does not undergo recoding *per se* but is subjected to other mechanisms, such as differential protein degradation, to ensure a high ratio of Gag to Pol. Retrotransposons in the *Metaviridae* from the animal kingdom preferentially used -1 frameshifting to regulate Pol protein production. In contrast, a +1 frameshift was more rarely observed but was distributed equally among kingdoms and among *Pseudoviridae* and *Metaviridae*. Finally, stop-codon suppression was found in a total of only two possible cases.

Additional open reading frames in LTR retrotransposons

Although retrotransposon *gag* and *pol* genes are believed to be necessary and sufficient for transposition, a number of retrotransposon families with aberrant genomic organizations have now been identified (Figure 3). One frequent structural change is the addition of coding information.

Retrotransposons with '*env*-like' genes

One of the main differences between retrotransposons (with a wholly intracellular life-cycle) and their infectious retrovirus cousins is the presence of an *envelope* (*env*) gene in the latter, which allows a virus particle to infect another cell. A number of retroelements have an extra ORF in the same position as the *env* gene found in retrovirus genomes (Figure 3). The best characterized examples of *env*-containing retroelements are the *Drosophila* errantiviruses, including *gypsy* and *ZAM* [9,10]. The life-cycle of these elements has been examined in detail, and *gypsy* has been shown to be infectious [11,12].

The presence of an *env* gene within a retroelement is not limited to the errantiviruses; genomic studies have revealed that *env*-like ORFs are widespread among retrotransposons in both the *Pseudoviridae* (sireviruses) and *Metaviridae* (errantiviruses, metaviruses and semotiviruses) [13,14]. Elements contaning an *env*-like ORF in each of these lineages also originate from diverse host species. The retroelement most recently shown to have an *env*-like ORF, *Boudicca*, is a metavirus from a human blood fluke [15]. Other examples of metaviruses include the *Athila* elements, which represent a

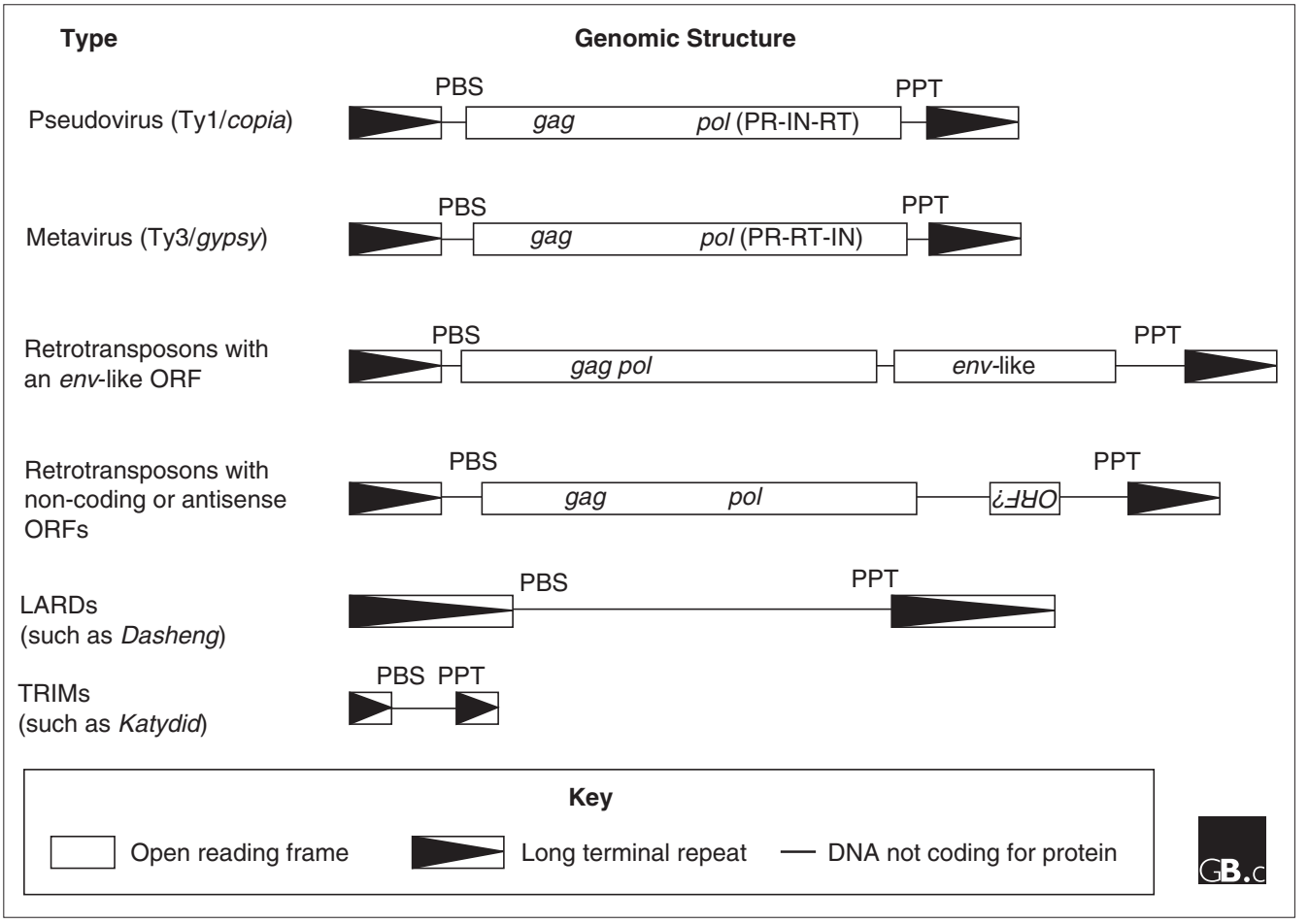

Figure 3
The genomic organization of different types of LTR retrotransposon. Abbreviations: IN, integrase; LARDs, large retrotransposon derivatives; ORF, open reading frame; PBS, primer-binding site; PPT, polypurine tract; PR, protease; RT, reverse transcriptase; TRIMs, terminal-repeat retrotransposons in miniature. The upside-down text indicates that the ORF is transcribed in the antisense direction. See text for descriptions of each type of element.

large proportion of the retroelements in *Arabidopsis* [16]. In a related element in barley, *Bagy*-2, the *env*-like transcript is spliced, similarly to the *env* transcripts of retroviruses [17]. Members of the sirevirus group make up half of the approximately 400 *Pseudoviridae* sequences present in GenBank, and of these, about one third have an *env*-like ORF (X.G. and D.V., unpublished observation). Semotiviruses (also called BEL retrotransposons) with *env*-like ORFs have also been described in nematode genomes as well as in pufferfish and *Drosophila* [18,19].

Do Env-like proteins enable these diverse retroelements to become infectious? In a few cases, the *env*-like genes have been shown to be significantly similar in sequence to genes of different viruses, suggesting that they were acquired by retrotransposons through transduction of a cellular gene [13]. Except for some errantiviruses, where the Env-like protein has been implicated in infection, the function of the Env-like proteins remains unclear. The amino-acid sequences of these proteins are highly divergent, making it difficult to assess whether or not they have a common function. That said, many Env-like proteins have predicted transmembrane domains (like retroviral Env proteins), although this is not a universal feature. It is possible that retroviral activity has evolved several times in the history of retrotransposons, or that these genes may confer novel function(s), such as movement between tissues of an organism (as suggested for the *gypsy* elements) or movement within cells (such as between the cytoplasm and the nucleus). Alternatively, the Env-like proteins could serve as chaperone proteins to facilitate replication. Functional studies are required to discern the biological roles of these interesting genes.

Other additional ORFs

Other novel coding regions have also been identified within various retrotransposons, but it is unclear how broadly these coding sequences are conserved. For example, *RIRE2* of rice - a metavirus - has a small ORF of unknown function

upstream of its *gag* gene [20]. Some plant retrotransposons carry ORF(s) that are antisense to the genomic RNA transcript (Figure 3), including the metaviruses *RIRE2* of rice and *Grande1* of maize [21,22]. The functions of the antisense ORFs are also unknown. In a few cases, retrotransposons have acquired sequences that probably do not have any role in the life cycle of the elements. The *Bs1* retrotransposon of maize, for example, has transduced a cellular gene sequence - in this case a part of a gene encoding an ATPase [23,24].

LTR retrotransposons lacking ORFs

An intriguing story is emerging about the presence of non-autonomous LTR retrotransposons in many eukaryotic genomes. Non-autonomous elements do not encode the proteins necessary for transposition; instead, they are mobilized *in trans* by proteins provided from functional (autonomous) elements. This mechanism is well documented for DNA transposons [25], and recent genome-mining studies have revealed many types of non-autonomous retrotransposons, suggesting that the process also occurs among retrotransposons. Typically, these elements lack all coding capacity but have retained LTRs, a primer-binding site and a polypurine tract (Figure 3). These are the minimal features required for replication, because the LTRs contain the promoter needed to produce a template RNA, and the primer-binding site and the polypurine tract are needed to prime reverse transcription. The success of some non-autonomous elements is staggering; for example, the non-autonomous *Dasheng* and *Zeon-1* elements are each represented by around 1,000 copies in the maize genome [26,27].

For most non-autonomous retrotransposons, it is unclear which autonomous element is involved in mobilization. Striking similarities between the non-autonomous *Dasheng* element and the autonomous *RIRE2* element, however, make it very probable that *RIRE2* provides the proteins needed to move *Dasheng* [28]. The evidence for this, mostly provided by the emerging rice genome sequence, includes a high degree of sequence similarity within and adjacent to the LTRs (suggesting that the promoters and/or sequences necessary for reverse transcription are the same), a similar distribution of *RIRE2* and *Dasheng* along the rice chromosomes (suggesting that they may be integrated by the same enzyme), the presence of chimeric *Dasheng*/*RIRE2* elements (suggesting that RNAs from both elements are packaged within a single virus-like particle), and the presence of young *Dasheng* and *RIRE2* elements (suggesting that these elements could be co-expressed).

The non-autonomous *Dasheng* elements are large, ranging in size from 5.5 kilobases (kb) to 8.5 kb [28]. Large non-autonomous elements like *Dasheng* have now been named 'large retrotransposon derivatives' (LARDS) [29]. The LARDs identified in barley and other members of the Triticeae have LTRs of 4.5 kb and an internal domain of 3.5 kb. The internal domain of the LARDs contains conserved non-coding DNA that may provide important secondary structure to the mRNA, although it is not known how these non-coding sequence features function in the life cycle of the LARDs. On the basis of sequence identity, it seems that barley LARDs may be mobilized by a retrotransposon related to the metaviruses *Erika-1* of the wheat *Triticum monococcum* and *RIRE3* of rice.

Finally, a second class of non-autonomous LTR retrotransposons has been identified in plants, called 'terminal-repeat retrotransposons in miniature' (TRIMs; Figure 3). They were originally identified in a potato *urease* gene intron and subsequently found in the *Arabidopsis* genome, where the founding element was named *Katydid* [30]. TRIMs also lack an internal coding domain but, in contrast to the LARD type of non-autonomous retrotransposon, TRIMs are very small - less than 540 bp overall. There are TRIMs in both monocotyledonous and dicotyledonous plants, but no autonomous partner has been found or proposed. The location of TRIMs within promoters and introns indicates that these elements have been important in restructuring plant genomes.

Non-coding information in LTR retrotransposons

Variation in retrotransposon genomic organization is not limited to the presence or absence of coding information. Some retrotransposons contain a large amount of conserved non-coding sequence. The barley LARD element with 3.5 kb of non-coding DNA (mentioned above) is one example; another is a group of plant metaviruses that carry several kilobases of non-coding DNA between *pol* and the 3′ LTR. Among these are the maize *Cinful* [31] and *Grande1* [22] elements, *RIRE2* from rice [21] and *Tat1* from *Arabidopsis* [32]. For *Grande1* and *RIRE2*, antisense ORFs have been described, but they do not account for the entire segment of non-coding DNA [21,22]. In addition, many retrotransposons, including the *Grande1* and *Cinful* elements, have a series of short tandem repeats very close to the 3′ end of the *pol* gene, or at a putative *pol-env* junction. This may suggest a potential function for the tandem repeats: they may facilitate recombination and acquisition of new coding information through gene transduction [31]. In support of this hypothesis, repeated non-coding information seems to be found between the *env*-like ORF and the 3′ LTR in both the *SIRE1* [33] and *Athila* retrotransposons [16]. In the retrotransposons with *env*-like ORFs, the repeats show similarity to polypurine tracts, suggesting that they might instead have a role in reverse transcription.

The sequenced eukaryotic genomes have provided a new appreciation of the diversity among LTR retrotransposons. As sequence data accumulate, additional novel elements are likely to be revealed. The challenge in the future will be to understand how diversity in retrotransposon genome

organization and coding sequences reflects differences in retrotransposition mechanisms and strategies employed by these elements to colonize their host genomes.

References

1. Voytas DF, Boeke JD: **Ty1 and Ty5 of *Saccharomyces cerevisiae*.** In *Mobile DNA II.* Edited by Craig NL, Craigie R, Gellert M, Lambowitz AL. Washington, DC: ASM Press; 2002:631-662.
2. Sandmeyer SB, Aye M, Menees T: **Ty3, a position-specific, *gypsy*-like element in *Saccharomyces cerevisiae*.** In *Mobile DNA II.* Edited by Craig NL, Craigie R, Gellert M, Lambowitz AL. Washington, DC: ASM Press; 2002:663-683.
3. Boeke JD, Eickbush T, Sandmeyer SB, Voytas DF: ***Pseudoviridae*.** In *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses.* Edited by Fauquet CM. New York: Academic Press; 2004: in press.
4. Boeke JD, Eickbush T, Sandmeyer SB, Voytas DF: ***Metaviridae*.** In *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses.* Edited by Fauquet CM. New York: Academic Press; 2004: in press.
5. Goodwin TJ, Poulter RT: **A group of deuterostome Ty3/*gypsy*-like retrotransposons with Ty1/*copia*-like *pol*-domain orders.** *Mol Genet Genomics* 2002, **267:**481-491.
6. Goodwin TJ, Poulter RT: **The DIRS1 group of retrotransposons.** *Mol Biol Evol* 2001, **18:**2067-2082.
7. Goodwin TJ, Poulter RT: **A new group of tyrosine recombinase-encoding retrotransposons.** *Mol Biol Evol* 2004, **21:**746-759.
8. Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF**: Translational recoding signals between *gag* and *pol* in diverse LTR retrotransposons.** *RNA* 2003, **9:**1422-1430.
9. Pelisson A, Mejlumian L, Terzian C, Bucheton A: ***Drosophila* germline invasion by the endogenous retrovirus *gypsy*: involvement of the viral *env* gene.** *Insect Biochem Mol Biol* 2002, **32:**1249-1256.
10. Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B, Vaury C: **Life cycle of an endogenous retrovirus, ZAM, in *Drosophila melanogaster*.** *J Virol* 2000, **74:**10658-10669.
11. Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG: **An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus.** *Genes Dev* 1994, **8:**2046-2057.
12. Kim A, Terzian C, Santamaria P, Pelisson A, Prud'homme N, Bucheton A: **Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 1994, **91:**1285-1289.
13. Malik HS, Henikoff S, Eickbush TH: **Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses.** *Genome Res* 2000, **10:**1307-1318.
14. Eickbush TH, Malik HS: **Origins and evolution of retrotransposons.** In: *Mobile DNA II.* Edited by Craig NL, Craigie R, Gellert M, Lambowitz AL. Washington, DC: ASM Press; 2002:1111-1144.
15. Copeland CS, Brindley PJ, Heyers O, Michael SF, Johnston DA, Williams DL, Ivens AC, Kalinna BH: ***Boudicca*, a retrovirus-like long terminal repeat retrotransposon from the genome of the human blood fluke *Schistosoma mansoni*.** *J Virol* 2003, **77:**6153-6166.
16. Wright DA, Voytas DF: ***Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses.** *Genome Res* 2002, **12:**122-131.
17. Vicient CM, Kalendar R, Schulman AH: **Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants.** *Genome Res* 2001, **11:**2041-2049.
18. Bowen NJ, McDonald JF: **Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements.** *Genome Res* 1999, **9:**924-935.
19. Frame IG, Cutfield JF, Poulter RTM: **New BEL-like LTR retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.** *Gene* 2001, **263:**219-230.
20. Kumekawa N, Ohtsubo H, Horiuchi T, Ohtsubo E: **Identification and characterization of novel retrotransposons of the *gypsy* type in rice.** *Mol Gen Genet* 1999, **260:**593-602.
21. Ohtsubo H, Kumekawa N, Ohtsubo E: ***RIRE2*, a novel gypsy-type retrotransposon from rice.** *Genes Genet Syst* 1999, **74:**83-91.
22. Martinez-Izquierdo JA, Garcia-Martinez J, Vicient CM: **What makes *Grande1* retrotransposon different?** *Genetica* 1997, **100:**15-28.
23. Bureau TE, White SE, Wessler SR: **Transduction of a cellular gene by a plant retroelement.** *Cell* 1994, **77:**479-480.
24. Jin YK, Bennetzen JL: **Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize.** *Plant Cell* 1994, **6:**1177-1186.
25. Feschotte C, Jiang N, Wessler SR: **Plant transposable elements: where genetics meets genomics.** *Nat Rev Genet* 2002, **3:**329-341.
26. Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J, Wing RA, McCouch SR, Wessler SR: ***Dasheng*: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice.** *Genetics* 2002, **161:**1293-1305.
27. Hu W, Das OP, Messing J: ***Zeon-1*, a member of a new maize retrotransposon family.** *Mol Gen Genet* 1995, **248:**471-480.
28. Jiang N, Jordan IK, Wessler SR: ***Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome.** *Plant Physiol* 2002, **130:**1697-1705.
29. Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH: **Large retrotransposon derivatives: abundant, conserved, but nonautonomous retroelements of barley and related genomes.** *Genetics* 2004, **166:**1437-1450.
30. Witte CP, Le QH, Bureau T, Kumar A: **Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes.** *Proc Natl Acad Sci USA* 2001, **98:**13778-13783.
31. Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL: **Structure and evolution of the *cinful* retrotransposon family of maize.** *Genome* 2003, **46:**745-752.
32. Wright DA, Voytas DF: **Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana* Ty3/*gypsy* retrotransposons that encode envelope-like proteins.** *Genetics* 1998, **149:**703-715.
33. Laten HM, Havecker ER, Farmer LM, Voytas DF: ***SIRE1*, an endogenous retrovirus family from *Glycine max*, is highly homogeneous and evolutionarily young.** *Mol Biol Evol* 2003, **20:**1222-1230.