PERSPECTIVES: GENETICS

L1 Retrotransposons Shape the Mammalian Genome

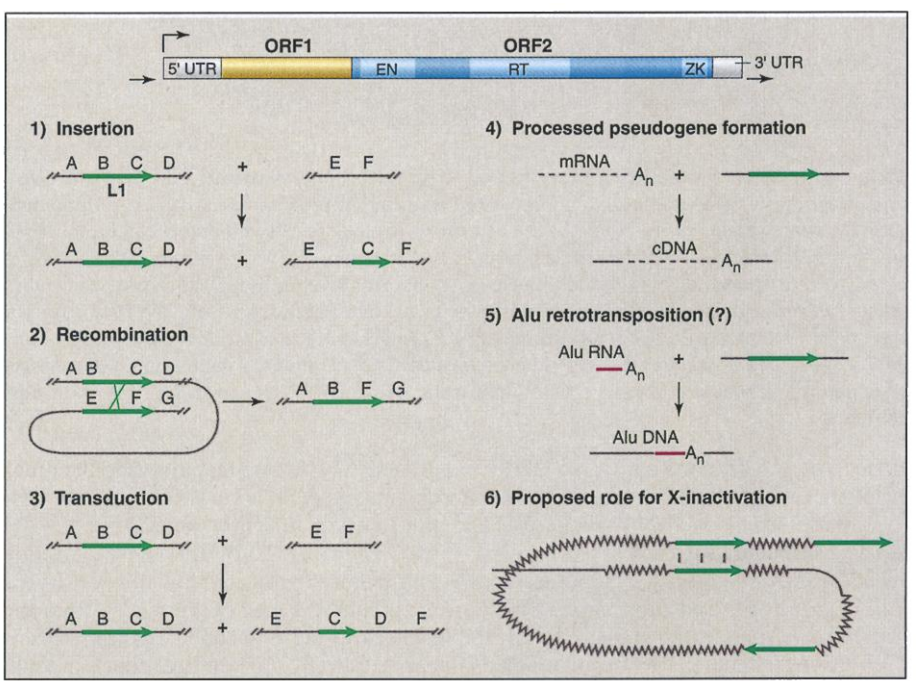
Haig H. Kazazian Jr.

Genomic mobile elements called retrotransposons make up about 40% of the mammalian genome (*1*). During retrotransposition, these small pieces of DNA are duplicated by a "copy and paste" mechanism—they are transcribed into RNA, reverse-transcribed into DNA, and the complementary DNA is then inserted back into the genome at a new site. This mechanism differs from the "cut and paste" process used by most transposable elements (that do not have an RNA intermediate in their mobility) found in prokaryotes and some eukaryotes. Although retrotransposons have been viewed as selfish DNAs that provide no benefit to their host cell, we now know that over evolutionary time they have increased the diversity of the genome through a variety of mechanisms, providing it with considerable "added value."

There are two classes of retrotransposons, those with long terminal repeats (LTRs) and those without LTRs (non-LTRs). The major non-LTR retrotransposon of mammals is the LINE-1 or L1 element. Roughly 500,000 truncated L1s and 3000 to 5000 full-length L1s are dispersed throughout the human genome, accounting for 15 to 17% of its mass (*1*). Full-length L1s have a poorly characterized internal promoter at their 5′ end and two open reading frames: ORF1, which encodes a nucleic acid binding protein, and ORF2, which encodes a protein with endonuclease, reverse transcriptase, and zinc knuckle (a zinc finger-like motif) domains (*2*) (see the figure). A short 3′ untranslated region ends in a polyadenylated (poly A) tail that is important for reverse transcription (*2*). The L1 element is usually flanked by a short duplication of genomic DNA, called a target site duplication. A cell culture assay that detects retrotransposition provides an estimate of 50 active L1s in human cells (*2*); in contrast, laboratory mouse strains have more than 3000 (*3*). The size of the active L1 population may fluctuate over evolutionary time. Throughout evolution it is likely that L1 retrotransposition has donated functional domains to other proteins, expanded the diversity of the genome, and increased the genome's size through mobilization of non-L1 sequences (see the figure).

L1 elements have been with us for a very long time. These sequences are present in all mammals, and mobile elements that share structural features with L1 are found in the frog *Xenopus laevis*, bony fish, the slime mold *Dictyostelium*, and in maize and *Arabidopsis* (*4*). Indeed, all non-LTR retrotransposons, including L1s, are evolutionarily related to one another. This entire class of elements is at least 600 million years old and perhaps dates to the origins of eukaryotes (*4*). Accordingly, L1-encoded protein domains are among the most ancient of all known functional domains. For example, the L1 endonuclease shares catalytic site residues with exonuclease III, an endonuclease found in the bacterium *Escherichia coli* (*5*). Moreover, the catalytic subunit of telomerase, an enzyme that adds DNA sequences to the ends of chromosomes, is a reverse transcriptase (RT) that functions in a manner analogous to the L1 RT. Telomerase RT shares considerable sequence similarities with L1 RT, and whether telomerase RT evolved from the RT of a non-LTR retrotransposon or vice versa is hotly debated (*6, 7*). I favor the idea that eukaryotic cells recruited retrotransposon RT to acquire telomerase activity for themselves. The evolutionary age of non-LTR retrotransposons is consistent with the very early eukaryotic origin of telomerase (*4*). A rational phylogenetic tree of eukaryotic RTs rooted by prokaryotic mobile elements also implies that telomerase RT was derived from retrotransposon RT (*7*). Beyond the possibility that early forms of L1 were responsible for the evolution of telomerase RT, these elements have played a major part in shaping the mammalian genome, not only through their own expansion but also through inducing the mobilization of non-L1 sequences (see the figure).

Retrotransposition of L1 elements is still a feature of some cells (for example, germ cells that form sperm and egg) but



The shape of things to come. (Top) A full-length L1 element with two open reading frames (ORFs) and 5′ and 3′ untranslated regions (UTRs). ORF2 has endonuclease (EN), reverse transcriptase (RT), and zinc knuckle (ZK) domains. **(1–6)** Various effects of L1 elements are shown (bold arrows are L1s, with the arrowhead at the 3′ end). **(1)** Insertion of truncated L1 (C) into a new site (EF); **(2)** intrachromosomal homologous recombination between L1s; **(3)** transduction of sequence D during retrotransposition of truncated L1 (C) into a new site (EF); **(4)** L1 proteins aid trans generation of processed pseudogenes; **(5)** L1 proteins aid in the retrotransposition of Alu elements; **(6)** possible role of L1s in spreading the X chromosome inactivation signal. The heterochromatinization signal (wavy line) may be transmitted from one L1 element to another through a contact process or through interaction with *Xist* RNA.

The author is in the Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. E-mail: kazazian@mail.med.upenn.edu

rarely causes genetic disease in humans. Fourteen individual disease-producing retrotransposition events are now known. Together with insertions of 17 Alu elements (short, high-repeat DNA sequences that do not encode proteins) and two other insertions, these 33 known retrotransposition events account for about 1 in 600 human mutations (*8*). Retrotransposition events have produced isolated cases of hemophilia A and B, β-thalassemia, and muscular dystrophy. However, it is estimated that retrotransposon insertions occur at a frequency of at least one event in every 50 to 100 human germ cells. In contrast, new mutations due to L1 retrotransposition events are roughly 30 times more frequent in laboratory strains of mice than in humans (*2*). In addition to insertional mutations, the large number of homologous L1 sequences in the genome leads to mutation through mispairing and unequal crossing-over between elements. There are three documented cases where this phenomenon has produced human disease. Furthermore, unequal crossing-over between L1 elements can influence genomic structure—for example, the duplication of the γ-globin genes present in most primates is due to an unequal crossing-over event between L1s in New World monkeys (*9*).

Retrotransposition of L1s has also remodeled the genome through the transduction of 3′ flanking sequences that have been carried along with L1 elements. The poly A signal of L1 RNA is weak, leading to failure of L1 RNA cleavage at its poly A site. In the event of this failure, cleavage then occurs after the next (downstream) poly A signal. The weak poly A signal of L1 is advantageous for gene expression because it allows many L1s to reside in introns without disrupting the transcription process through premature cleavage of mRNAs. The phenomenon of 3′ transduction was first recognized in a patient with muscular dystrophy whose dystrophin gene was disrupted by an L1 insertion that contained an additional 526 base pairs of single-copy sequence 3′ to the L1 (*2*). Transduction was later demonstrated experimentally in cultured cells (*10*). Recently, database searches have found that 3′ transduction occurs in 15 to 23% of L1 retrotransposition events, and that these transduced sequences account for 0.5 to 1% of the human genome (*11, 12*). Although transduction of functional sequences has not yet been documented, such events have the potential to shuffle, for example, exons, promoters, and enhancers, thus creating new genes or altering the function of old genes in the process (*11*). Moreover, because L1 insertions are often severely truncated at their 5′ end (that is, they are less than 1 kb in length), database searches of genomic DNA should uncover retrotransposition events consisting of 3′ transduced sequences without any L1 sequences.

It is paradoxical that L1 elements mobilize non-L1 sequences. There is considerable evidence that the protein products of a particular L1 preferentially facilitate the retrotransposition of that same L1, a so-called cis preference (*2*). Whereas cis preference is likely to have restricted the retrotransposition of many RNAs with 3′-poly A tails, such as defective L1 transcripts and cellular mRNAs, occasional trans mobilization of non-L1 RNAs has had an important impact on shaping the mammalian genome.

Processed pseudogenes are nonfunctional intronless copies of genes. They are derived from mRNAs that have been reverse-transcribed and reinserted into the genome, a process similar to the duplication of L1 sequences. On human chromosome 22, processed pseudogenes account for 0.5% of genomic DNA (*13*). In addition, some of the many processed pseudogenes contribute new activities to the cell, such as providing new exons for preexisting genes. Processed pseudogenes have been generated by active L1 elements in tissue culture (*14*), which suggests that L1 proteins rarely acting in trans are the driving force behind processed pseudogene formation.

Alu elements may be mobilized by the trans action of L1 proteins as well. Roughly 1 million retrotransposed Alu elements make up about 10 to 12% of the human genome. The 300–base pair Alu elements do not encode any protein but are transcribed by RNA polymerase III into RNA that ends in a poly A tail. L1-encoded proteins are likely candidates for Alu mobilization because Alu elements are flanked by target site duplications that bear a close resemblance to the target site duplications of L1 elements, and DNA sequences at the sites of Alu insertions are similar to those found at L1 insertion sites (*15*). However, Alu mobilization by L1s has not been demonstrated experimentally. Furthermore, it is difficult to explain the enormous Alu expansion in light of the apparent cis preference of L1 proteins. Perhaps Alu RNA becomes concentrated on large ribosomal subunits because it is bound by two signal recognition particle proteins, SRP9 and SRP14, that interact with ribosomes (*16*). A high concentration of Alu RNA near ribosomes may occasionally usurp L1 proteins, leading to Alu retrotransposition. In addition to genome expansion through retrotransposition, Alu elements have shaped the genome through mispairing and unequal crossing-over, leading to deletions and duplications.

Although the physiologic function, if any, of L1 proteins remains unknown, the large number of homologous L1 sequences in our chromosomes may be very important for the genome. Lyon has suggested that L1 elements on the X chromosome act as "booster stations" for a heterochromatinization signal (which leads to condensation of chromatin and inactivation of genes) transmitted by *Xist* RNA (*17*). *Xist* RNA is likely to play a key role in X chromosome inactivation because it is expressed from and specifically interacts with the inactive X chromosome. An alternative proposal suggests that heterochromatinization of the inactive X chromosome spreads from one L1 to another L1 through homologous pairing (*18*). So far, supporting evidence for both hypotheses is circumstantial. Such evidence includes observation of a very high concentration of L1 elements on the mammalian X chromosome with a significant L1 clustering around the X-inactivation center; the discovery of a reduced L1 content in X chromosome segments containing genes that escape X inactivation; and the finding of a positive correlation between the concentration of autosomal L1 elements and the extent of heterochromatinization of autosomal genes at sites of X autosome translocation (*17, 19*).

It is clear that L1 elements are the master mammalian retrotransposons. Although L1s may be selfish, they are clearly not junk, for they have played a major part in our evolution and the evolution of our genomes.

References and Notes

1. A. F. Smit, *Curr. Opin. Genet. Dev.* **9**, 657 (1999).
2. H. H. Kazazian Jr. and J. V. Moran, *Nature Genet.* **19**, 19 (1998).
3. R. J. DeBerardinis *et al.*, *Nature Genet.* **20**, 288 (1998).
4. H. S. Malik *et al.*, *Mol. Biol. Evol.* **16**, 793 (1999).
5. C. D. Mol *et al.*, *Nature* **374**, 381 (1995).
6. T. M. Nakamura *et al.*, *Science* **277**, 955 (1997).
7. T. H. Eickbush, *Science* **277**, 911 (1997).
8. University of Pennsylvania, Department of Genetics, *Retrotransposon Insertions into the Human Genome* (www.med.upenn.edu/genetics/labs/retrotrans_table.html).
9. D. H. Fitch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 7396 (1991).
10. J. V. Moran *et al.*, *Science* **283**, 1530 (1999).
11. J. L. Goodier *et al.*, *Hum. Mol. Genet.* **9**, 653 (2000).
12. O. K. Pickeral *et al.*, *Genome Res.* **10**, 411 (2000).
13. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
14. C. Esnault *et al.*, *Nature Genet.* **24**, 363 (2000).
15. J. Jurka, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1872 (1997).
16. J. D. Boeke, *Nature Genet.* **16**, 6 (1997).
17. M. F. Lyon, *Cytogenet. Cell Genet.* **80**, 133 (1998).
18. Y. Marahrens, *Genes Dev.* **13**, 2624 (1999).
19. J. A. Bailey *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6634 (2000).

20. I thank E. Luning Prak, E. Ostertag, and J. Goodier for helpful suggestions.