Mobile Elements: Drivers of Genome Evolution

Haig H. Kazazian Jr.*

Mobile elements within genomes have driven genome evolution in diverse ways. Particularly in plants and mammals, retrotransposons have accumulated to constitute a large fraction of the genome and have shaped both genes and the entire genome. Although the host can often control their numbers, massive expansions of retrotransposons have been tolerated during evolution. Now mobile elements are becoming useful tools for learning more about genome evolution and gene function.

Mobile, or transposable, elements are prevalent in the genomes of all plants and animals. Indeed, in mammals they and their recognizable remnants account for nearly half of the genome (*1*, *2*), and in some plants they constitute up to 90% of the genome (*3*). If, as many believe, the origins of life are in an "RNA world" followed by reverse transcription into DNA, then mobile elements could have been very early participants in genome formation (*4*). Indeed, mobile elements and genes appear to have forged a mutually beneficial relationship. How did this relationship come about? It is clear how mobile elements benefit from genes, because without genes they cannot survive from one generation to the next. But how have genes benefited from the genome shaping of mobile elements?

Important insights into genome evolution have emerged from the mining of multiple genome sequences. Here, I concentrate on how mobile elements have affected the evolution of genes and their function, particularly of humans and other mammals.

Mobile elements are DNA sequences that have the ability to integrate into the genome at a new site within their cell of origin (*5*). These elements include (i) DNA transposons, (ii) autonomous retrotransposons, and (iii) nonautonomous retrotransposons (Fig. 1). The mechanism by which many of these elements move is well known, but for others, such as mammalian retrotransposons, there is still much to learn.

DNA Transposons

DNA transposons are prevalent in bacteria (where they are called IS, or insertion sequences), but are also found in the genomes of many metazoa, including insects, worms, and humans. These elements are generally excised from one genomic site and integrated into another by a "cut and paste" mechanism. Because sequence specificity of integration is limited to a small number of nucleotides—e.g., TA dinucleotides for Tc1 of *Caenorhabditis elegans*—insertions can occur at a large number of genomic sites. However, daughter insertions for most, but not all, DNA transposons occur in proximity to the parental insertion. This is called "local hopping." Active transposons encode a transposase enyme between inverted-repeat termini. The transposase binds at or near the inverted repeats and to the target DNA. It then performs a DNA breakage reaction to remove the transposon from its "old" site and a joining reaction to insert the transposon into its "new" site. These reactions proceed with the hydrolysis of phosphodiester bonds between the transposon and flanking DNA to liberate 3′-OH residues that carry out the attack at the "new" site (*6*). Because the two strands of the "new" DNA are attacked at staggered sites, the inserted transposon is flanked by small gaps which, when filled in by host enzymes, leads to short duplications of sequence at the target sites. These are called target site duplications (TSDs), and their length is often characteristic for a particular transposon (*7*).

The reactions needed to move a piece of DNA use recombinase enzymes, of which there are two main classes. The first class is called conservative because the enzymes do not require high-energy cofactors, the total number of phosphodiester bonds remains unchanged, and no DNA degradation or resynthesis occurs. Examples of this recombinase type are the integrase protein of bacteriophage λ, Cre recombinase, and Flp recombinase. The second class is the transposases that catalyze a whole set of reactions necessary for DNA transposition. Examples are the transposases of Mu, P elements, and the Tc1/mariner family, and the integrases of long terminal repeat (LTR) retrotransposons and retroviruses. All of these enzymes share certain structural motifs such as a D,D35E sequence (aspartate, aspartate, 35 amino acid residues, then a glutamate) and a handlike three-dimensional structure (*6*, *8*).

Although these elements generally transpose to genomic sites less than 100 kb from their original site (e.g., the *Drosophila* P element), some are able to make distant "hops" (e.g., the fish Tc1/mariner element; see below).

LTR Retrotransposons

Retrotransposons are transcribed into RNA, and then reverse transcribed and reintegrated into the genome, thereby duplicating the element. The major classes of retrotransposons either contain long terminal repeats at both ends (LTR retrotransposons) or lack LTRs and possess a polyadenylate sequence at their 3′ termini (non-LTR retrotransposons).

LTR retrotransposons and retroviruses are quite similar in structure (Fig. 1). They both contain *gag* and *pol* genes that encode a viral particle coat (GAG) and a reverse transcriptase (RT), ribonuclease H (RH), and integrase (IN) to provide enzymatic activities for making cDNA from RNA and inserting it into the genome. They differ in that retroviruses encode an envelope protein that facilitates their movement from one cell to another, whereas LTR retrotransposons either lack or contain a remnant of an *env* gene and can only reinsert into the genome from which they came. Reverse transcription of retroviral RNA or LTR-retrotransposon RNA occurs within the viral or viral-like particle in the cytoplasm (*9*), and is a complicated, multistep process (Fig. 2). In contrast, reverse transcription of non-LTR retrotransposons occurs by a very different mechanism (see below).

Many LTR retrotransposons target their insertions to relatively specific genomic sites. For example, Ty3 elements of *Saccharomyces cerevisiae* target specifically to a few nucleotides from RNA polymerase III (Pol III) transcription initiation sites (*10*). Moreover, Pol III transcription factors, TFIIIB and TFIIIC, are essential for Ty3 integration. Ty1 finds a haven within 750 base pairs (bp) upstream of Pol III–transcribed genes (*11*), and Ty5 targets the heterochromatin of telomeres and the silent mating loci (*12*). Ty5 requires a specific protein partner, Sir4, for tethering its cDNA to telomeric DNA, and the interaction sites of Ty5 (six amino acids in the integrase domain) with Sir4 (a region near the C terminus) have been characterized (*12*). In contrast to the Ty elements of *S. cerevisiae*, Tf elements of *Schizosaccharomyces pombe* cluster 100 to

Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA.

*E-mail: kazazian@mail.med.upenn.edu

400 nucleotides upstream of Pol II–transcribed genes (*13*).

The retroviruses HIV (human immunodeficiency virus) and MLV (mouse leukemia virus) share many structural features with LTR retrotransposons. In general, HIV inserts into many sites throughout actively transcribed genes (*14*), whereas MLV integrates preferentially into the promoters of active genes (*15*). The preference of retroviruses for insertion sites in and around genes may explain the occurrence of leukemia-producing insertions into the promoter of the LMO-2 gene in 2 of 10 patients undergoing retroviral gene therapy for severe combined immunodeficiency (*16*).

Non-LTR Retrotransposons

Non-LTR retrotransposons are typified by LINE-1 (long interspersed nucleotide elements–1, or L1) elements of mammals. Full-length non-LTR retrotransposons are 4 to 6 kb in length and usually have two open reading frames (ORFs), one encoding a nucleic acid binding protein, and the other encoding an endonuclease and an RT (Fig. 1). Because these elements encode activities necessary for their retrotransposition, they are called autonomous even though they probably also require host proteins to complete retrotransposition.

Some non-LTR retrotransposons integrate at specific genomic sites. R1 and R2 of *Drosophila melanogaster* and *Bombyx mori* integrate at specific ribosomal RNA gene locations (*17*), whereas heT-A and TART elements help maintain the telomeres of *Drosophila melanogaster* chromosomes (*18*) and TRAS1 and SART1 integrate into telomeric repeats of *B. mori* (*19*). In contrast, mammalian L1 elements apparently integrate at a very large number of sites in the genome because their endonuclease prefers to cleave DNA at a short consensus sequence (5′-TTTT/A-3′, where / designates the cleavage site) (*20*, *21*).

Our knowledge of most of the steps leading to retrotransposition of non-LTR retrotransposons is sketchy except for the reverse transcription process. In contrast to reverse transcription of LTR retrotransposons and retroviruses, this process takes place on nuclear genomic DNA through target primed reverse transcription, or TPRT (Fig. 2) (*22*, *23*). The great majority of mammalian L1 insertions are 5′ truncated and much less than the full length of 6 kb. However, the mechanism of 5′ truncation is still unclear. In about 30% of mammalian L1 insertions, but not in *Drosophila* R1 or R2 insertions, the 5′ end of the insertion sequence is inverted. A likely explanation for this phenomenon is a variation on TPRT, called "twin priming" (Fig. 2 legend) (*24*).

Retroelements Distinct from Both LTR and Non-LTR Retrotransposons

Two infrequently observed families of retroelements distinct from both LTR retrotransposons and non-LTR retrotransposons have been described. One is the DIRS1-like family that lacks many characteristics of both LTR and non-LTR retrotransposons. Discovered in *Dictyostelium discoideum*, these elements have RT domains with homology to LTR retrotransposons, but they lack the aspartate protease and D,D35E integrase of LTR retrotransposons (*25*). They also lack typical LTRs, polyadenylate [poly(A)] tails, and target-site duplications. Their mechanism of integration is mysterious, but they may generate closed-circle DNA by reverse transcription, followed by integration using DNA recombination.

The second family is an unusual class of elements, exemplified by Penelope of *Drosophila virilis* and Athena of bdelloid rotifers, which contain characteristics of both non-LTR and LTR retrotransposons (*26*). Like non-LTR retrotransposons, they are frequently 5′ truncated and have variable-length TSDs. However, some have LTRs, either in a direct or inverted orientation. Importantly, their RT is disrupted by a short, classic intron that contains in-frame stop codons and frameshifts, and intronless elements have not been found. Moreover, their RT sequence is close-

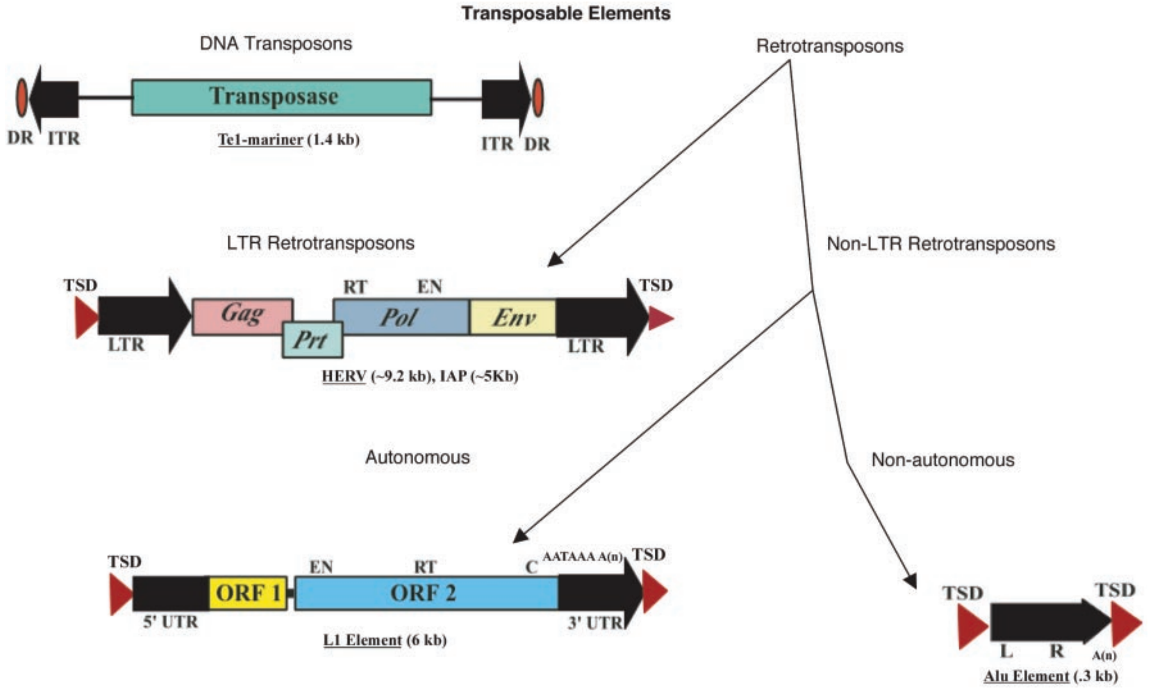

Fig. 1. Classes of mobile elements. DNA transposons, e.g., Tc-1/mariner, have inverted terminal inverted repeats (ITRs) and a single open reading frame (ORF) that encodes a transposase. They are flanked by short direct repeats (DRs). Retrotransposons are divided into autonomous and nonautonomous classes depending on whether they have ORFs that encode proteins required for retrotransposition. Common autonomous retrotransposons are (i) LTRs or (ii) non-LTRs (see text for a discussion of other retrotransposons that do not fall into either class). Examples of LTR retrotransposons are human endogenous retroviruses (HERV) (shown) and various Ty elements of *S. cerevisiae* (not shown). These elements have terminal LTRs and slightly overlapping ORFs for their group-specific antigen (*gag*), protease (*prt*), polymerase (*pol*), and envelope (*env*) genes. They produce target site duplications (TSDs) upon insertion. Also shown are the reverse transcriptase (RT) and endonuclease (EN) domains. Other LTR retrotransposons that are responsible for most mobile-element insertions in mice are the intracisternal A-particles (IAPs), early transposons (Etns), and mammalian LTR-retrotransposons (MaLRs). These elements are not present in humans, and essentially all are defective, so the source of their RT in trans remains unknown. L1 is an example of a non-LTR retrotransposon. L1s consist of a 5′-untranslated region (5′UTR) containing an internal promoter, two ORFs, a 3′UTR, and a poly(A) signal followed by a poly(A) tail (A_n). L1s are usually flanked by 7- to 20-bp target site duplications (TSDs). The RT, EN, and a conserved cysteine-rich domain (C) are shown. An Alu element is an example of a nonautonomous retrotransposon. Alus contain two similar monomers, the left (L) and the right (R), and end in a poly(A) tail. Approximate full-length element sizes are given in parentheses. [Modified from (*31*)]

ly related to that of telomerase. The presence of the RT strongly suggests that these elements are mobilized through an RNA intermediate, but the RT-disrupting intron means that they must have used an RT derived in trans from another genomic source.

Retrotransposons—Drivers of Genome Evolution

Genome evolution in eukaryotes has been driven by a number of processes, including the breakage and rejoining of different chromosomes (translocations), gene and segmental duplication, the shuffling of functional domains in exons, and gene conversion. Non-LTR retrotransposons have had a very long history over some 500 to 600 million years. They contain an RT that is similar to the RT of the mobile group II introns that occur in mitochondrial and chloroplast genomes of fungi and plants, and certain bacterial genomes (*27*, *28*). They also inhabit some yeast genomes, including that of *Candida albicans* (*29*). Their early evolutionary role is murky, but during recent times within mammals, they have been another important force in genome change.

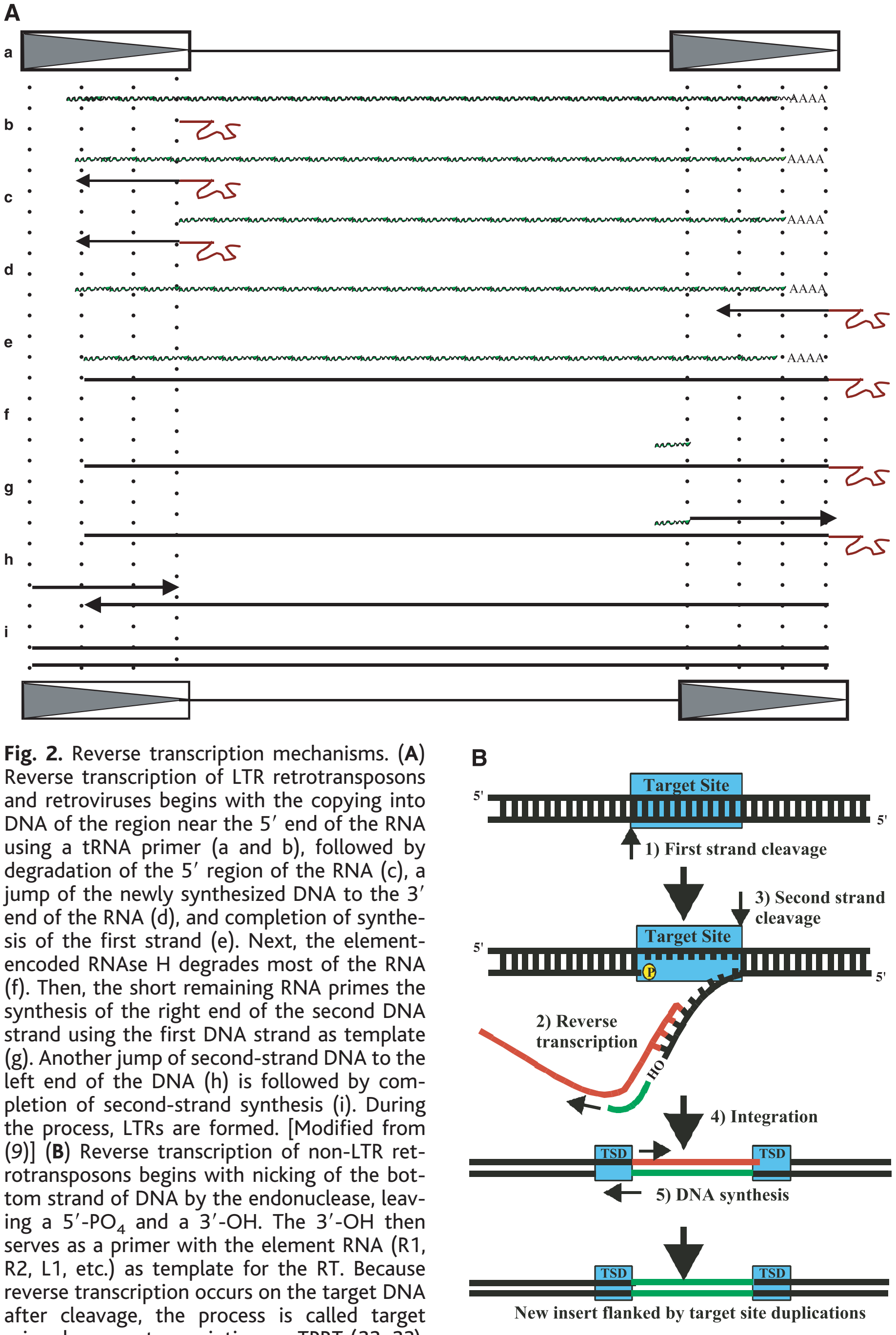

Fig. 2. Reverse transcription mechanisms. (**A**) Reverse transcription of LTR retrotransposons and retroviruses begins with the copying into DNA of the region near the 5′ end of the RNA using a tRNA primer (a and b), followed by degradation of the 5′ region of the RNA (c), a jump of the newly synthesized DNA to the 3′ end of the RNA (d), and completion of synthesis of the first strand (e). Next, the element-encoded RNAse H degrades most of the RNA (f). Then, the short remaining RNA primes the synthesis of the right end of the second DNA strand using the first DNA strand as template (g). Another jump of second-strand DNA to the left end of the DNA (h) is followed by completion of second-strand synthesis (i). During the process, LTRs are formed. [Modified from (*9*)] (**B**) Reverse transcription of non-LTR retrotransposons begins with nicking of the bottom strand of DNA by the endonuclease, leaving a 5′-PO_4 and a 3′-OH. The 3′-OH then serves as a primer with the element RNA (R1, R2, L1, etc.) as template for the RT. Because reverse transcription occurs on the target DNA after cleavage, the process is called target primed reverse transcription, or TPRT (*22*, *23*). [Modified from (*31*)] In a variation of TPRT, called "twin priming," inversions are formed (not shown). Here, it is proposed that the second strand of DNA is cleaved during reverse transcription of the first strand, and the 3′-OH of the second strand becomes a second primer for reverse transcription internally on L1 RNA. Resolution of this second cDNA produces the inversion (*24*).

Mammalian L1 elements affect the genome in many unusual ways, both destructive and constructive (Fig. 3). The destructive processes include insertion, and rearrangement due to homologous recombination. The average human diploid genome has 80 to 100 active L1s (*30*), and L1 insertions account for about 1 in 1200 human mutations, some of which cause disease (*31*). Moreover, at least 1 in every 50 humans has a new genomic L1 insertion that occurred in parental germ cells or in early embryonic development (*32*–*34*). In contrast, laboratory strains of mice have an estimated 3000 active L1 elements in their genomes (*34*), and L1 insertions are a much greater fraction of disease-producing mutations in the mouse than they are in humans (*31*). A canine L1 insertion disrupting the factor IX gene produces hemophilia B (*35*). Because active L1s have also been isolated from gorilla DNA (*36*), it seems likely that all mammals have active L1 elements that can be copied into new genomic locations and can occasionally produce disease.

In contrast to many other mobile elements, L1s have a marked cis preference, whereby their proteins greatly prefer to act on the RNA that encodes them (*37*). Nevertheless, they are still able on occasion to mobilize nonautonomous sequences in trans. Because the short interspersed nucleotide elements (SINEs) and LINEs of many species share homologous sequences at their 3′ end upstream of the poly(A), it is postulated that the RT encoded by these "stringent" LINEs interacts with the shared 3′-end sequence to mobilize the SINE in trans. Trans mobilization of an eel SINE by an eel LINE has been demonstrated in cultured human HeLa cells (*38*). Human Alu elements are another SINE that are probably mobilized by LINEs. These ~300-bp elements, derived from 7SL RNA, do not encode proteins, yet have expanded to 1.1 million copies, or 11%, of the human genome. Their B1 homologs make up almost 3% of the mouse genome. Alu insertions have accounted for over 20 cases of human genetic disease, and Alu retrotransposition events occur in at least 1 in every 30 individuals (*31*). Recently, trans mobilization of a transfected, marked Alu by an active human L1 was demonstrated in cultured HeLa cells (*39*). In addition, retrotransposition of a transfected Alu mediated by an endogenous L1 was demonstrated in cultured cells treated with an inhibitor of topoisomerase II (*40*). Moreover, a single mouse B1 insertion has recently been found, suggesting that present-day mouse L1s can also act occasionally in trans (*41*).

Two complementary reasons for the large number of Alus in the human genome in the face of present-day L1 cis preference have been suggested. They are the simultaneous occurrence at a particular evolutionary time of a highly trans-active L1 subfamily and transcription of Alu sequences susceptible to mobilization. Genome analysis suggests that a large burst of Alu insertions (and processed pseudogenes) occurred 40 million years ago when three presently inactive L1 subfamilies were prevalent and perhaps contained a large number of active members (*42*). At that time, Alu elements were special in their ability to gain access to the L1 retrotransposition machinery (*39*, *43*). Alu RNA binds the SRP9/14 subunit of the signal recognition particle, bringing it into proximity with ribosomes and nascent L1 proteins on L1 RNA. But Alu sequence evolution has resulted in a decline in SRP 9/14 binding to Alu RNA during primate evolution (*44*), suggesting that 40 million years ago Alu RNA had an enhanced ability to gain access to L1 proteins.

Processed pseudogenes and SVA elements are two other nonautonomous retrotransposons that are probably mobilized by human L1s because, like Alus, they end in poly(A), have L1-type TSDs, and insert at L1 endonuclease cleavage sites. A processed pseudogene arises by reverse transcription of a cellular mRNA followed by integration of the resulting cDNA into the genome. Roughly 5000 processed pseudogenes exist in the human genome, accounting for ~0.5% of its mass. Processed pseudogenes are not usually transcribed because they lack an external promoter. Human L1s probably drive low-level retrotransposition of processed pseudogenes in cultured cells (*36*, *45*). To date, no disease-causing insertions of processed pseudogenes have been found. SVA elements are nonautonomous, composite sequences containing a SINE derived from a human endogenous retrovirus (SINE-R), a variable number of tandem repeats (VNTR) segment, and a partial Alu sequence. Although there are only a few thousand of these elements in the human genome, SVA insertions have been found in three cases of human disease, and thus may be currently mobilized at a high frequency. One insertion has further hallmarks of

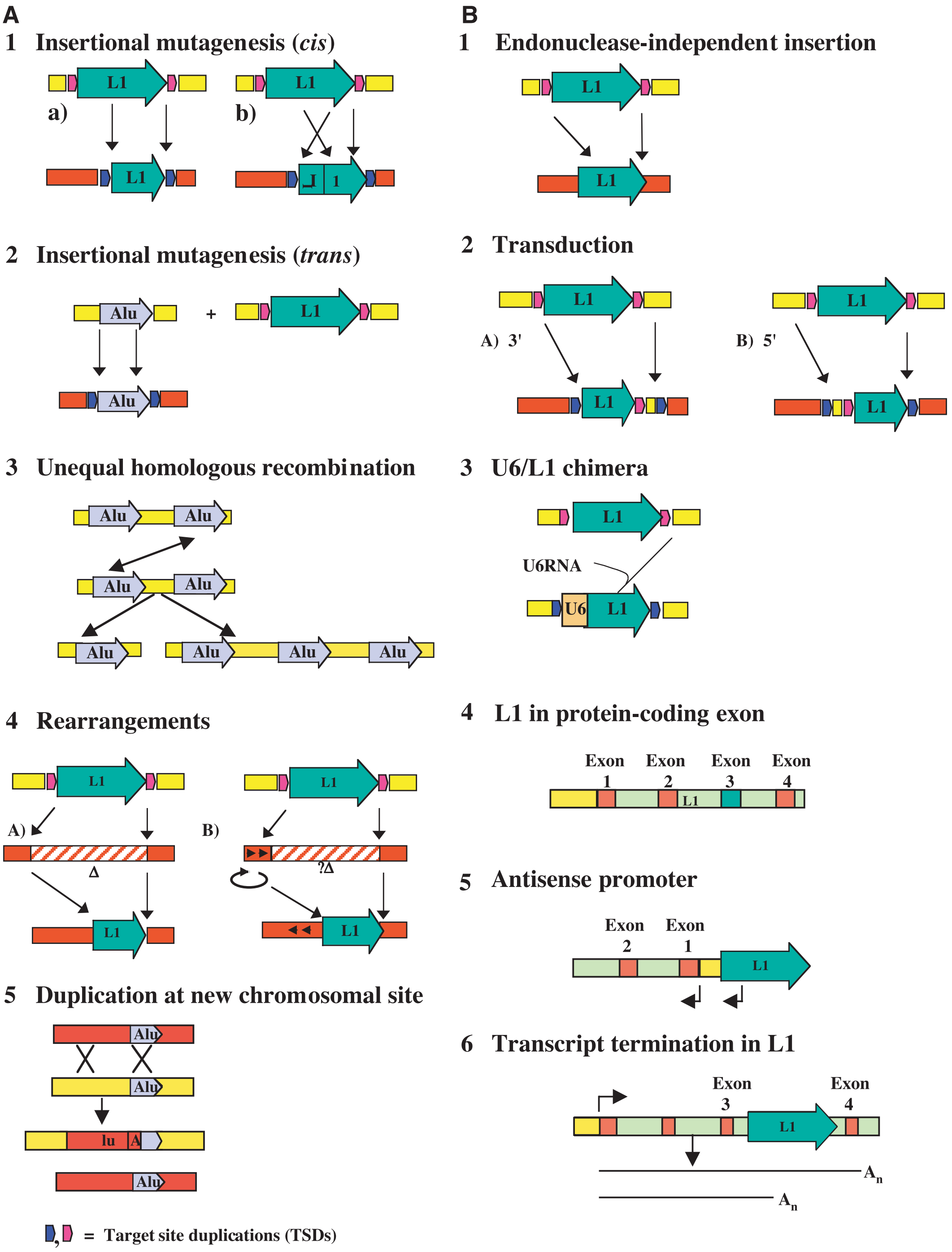
A
1 Insertional mutagenesis (*cis*)
L1
a)
L1
b)
L1
L1
2 Insertional mutagenesis (*trans*)
Alu
+
L1
Alu
3 Unequal homologous recombination
Alu
Alu
Alu
Alu
Alu
Alu
Alu
Alu
4 Rearrangements
L1
A)
Δ
L1
L1
B)
?Δ
L1
5 Duplication at new chromosomal site
Alu
Alu
lu
A
Alu
= Target site duplications (TSDs)
B
1 Endonuclease-independent insertion
L1
L1
2 Transduction
L1
A) 3'
L1
L1
B) 5'
L1
3 U6/L1 chimera
L1
U6RNA
U6
L1
4 L1 in protein-coding exon
Exon 1
Exon 2
Exon 3
Exon 4
L1
5 Antisense promoter
Exon 2
Exon 1
L1
6 Transcript termination in L1
Exon 3
Exon 4
L1
A$_n$
A$_n$

Fig. 3. Non-LTR retrotransposons are drivers of genome evolution. (**A**) Generally destructive mechanisms are (1) insertion of L1 elements, usually 5′ truncated or 5′ inverted; (2) trans-driven insertion of processed pseudogenes, Alus, and SVAs; (3) deletions and duplications due to unequal homologous recombination between Alus or L1s; (4) occasional deletions or inversions occurring upon insertion of L1s; and (5) segmental duplications leading to deletions and duplications. (Here a double crossing over facilitated by pairing at Alus moves a segment of DNA from one chromosome to another. Subsequent segregation places the two homologous segments in the same diploid genome.) (**B**) Generally constructive mechanisms are (1) repair of double-strand breaks by L1 insertion; (2) 3′ or 5′ transduction; (3) formation of chimeric retrogenes; (4) use of L1 or Alu sequence in coding regions of genes; (5) expression of genes 5′ to full-length L1s via an antisense promoter in L1; and (6) premature cleavage of gene transcripts at strong poly(A) signals in L1. Not shown are potential roles in the origin of eukaryotic telomerase and X-chromosome inactivation.

an L1-mediated event: namely, sequence derived from the 3′ flank of an element, called a 3′-transduced sequence (see below); and a 5′ inversion (*46*). Indeed, this event fulfills a prediction that, because of 3′ transduction followed by severe 5′ truncation, some L1-driven insertions could completely lack retrotransposon sequence (*47*). The insertion into an α-spectrin gene contains only 3′-transduced sequence that is partially inverted and completely lacks its full-length SVA parent.

L1s can also produce large DNA rearrangements upon insertion. Analyses of numerous L1 insertions in cultured cells have shown that about 10% are associated with large deletions of genomic DNA (*48*, *49*). One naturally occurring L1 insertion associated with a large deletion has been found in the mouse (*50*).

L1s and Alus provide material for DNA mispairing and unequal crossing over (homologous recombination), leading to deletion or duplication of sequences between the repeats. A number of these events have involved Alus, whereas only a few involving larger L1 elements have been described (*31*). The small number of mispairing and unequal crossing-over events between L1s is somewhat surprising, but may relate to the relatively low representation of L1s in regions of high gene density, in contrast to the much higher density of Alus presently in these regions. [Because Alu insertion is dependent on L1 machinery, Alus and L1s have similar insertion sites (*51*). Thus, the present distribution of these elements may reflect evolutionary selection against L1s in gene-rich regions.]

Table 1. Mobile element dynamics in model organisms. Tns, DNA transposons; Rtns, retrotransposons. Organisms are budding yeast, *S. cerevesiae*; mustard weed, *A. thaliana*; roundworm, *C. elegans*; fruit fly, *D. melanogaster*; mouse, *M. musculis*; human, *H. sapiens*.

Organism	Mobile element type (% of genome)			Active element(s)	Estimate of insertion freq. per generation	Estimate of removal freq.
	Tns	LTR Rtns	Non-LTR Rtns			
Budding yeast	0	3	0	LTR Rtn	10^{-5}–10^{-7}*	High (LTR recombination)
Mustard weed	5	5	0.5	Tn, LTR Rtn	?	?
Roundworm	12	0	<0.4	Tn	Very low	?(Low)
Fruit fly	0.3	2.7	0.9	Tn, LTR Rtn, non-LTR Rtn	10^{-1}–10^{-2}†	High (deletion and selection)‡
Mouse	0.9	10	27	LTR Rtn, non-LTR Rtn	$>10^{-1}$	Low
Human	3	8.5	35	Non-LTR Rtn	10^{-1}§	Low

*See (*63*). †Mobile element insertion rates for P and I element hybrid dysgenesis crosses are ~10°. In natural crosses, transposition and retrotransposition rates are 10^{-1} to 10^{-2} [for copia and Doc, see (*65*); for mariner, see (*66*)]. ‡See (*67*). §See (*31*).

Similarly, homologous recombination between Alus may have been involved in the genesis of segmental duplications, duplicated sequence blocks of 200 to 400 kb that account for up to 5% of the human genome. When these homologous sequence blocks are within 5 Mb of each other, they have an important role in human disease, producing large deletions, duplications, and inversions secondary to mispairing and unequal crossing over (*52*). A high proportion of Alu elements (29%) at the ends of segmental duplications suggests that many were generated by Alu mispairing followed by homologous recombination (*53*).

Offsetting these potentially destructive processes for the genome, L1s are constructive in numerous ways. First, they occasionally repair double-strand breaks in DNA by inserting into the genome via an endonuclease-independent pathway. Rare instances of this "bandage" phenomenon have been observed in vivo, but endonuclease-independent L1 insertions are common in cultured cells that are defective in DNA-repair proteins, e.g., XRCC4 (*54*).

Second, L1 retrotransposition can often move sequences 3′ to a parental L1 to a new genomic location. Because L1s contain a weak RNA cleavage and polyadenylation signal, their transcript is frequently not cleaved at the 3′ end of the L1 but instead is cleaved after a downstream poly(A) signal. By this mechanism, 10 to 20% of recent L1 retrotranspositions contain sequences derived from the 3′ flank of the parental L1, called 3′ transductions. These events have the potential to shuffle exons and regulatory sequences to new genomic sites (*47*). Occasionally 5′ transduction due to initiation of transcription from a promoter upstream of a full-length L1 also occurs.

Third, L1 retrotransposition can produce new chimeric retrogenes that are often expressed. These genes are probably generated through template switching of L1 RT from L1 RNA or Alu RNA to other small nuclear RNAs. In the human genome sequence, there are some 80 chimeric retrogenes whose 5′ regions originate from small nuclear RNAs, such as U6, U3, U5, and 5*S* RNA, and whose 3′ regions are the 3′ ends of L1 or Alu elements (*55*).

Fourth, retrotransposons have shaped mammalian genomes by providing their sequences for a number of protein-coding exons of genes. In the human genome, L1 or Alu sequences are present in nearly 200 confirmed and 2400 predicted protein-coding sequences (*56*). However, amino acids translated from these sequences still need to be demonstrated in the protein products of these genes.

Fifth, L1 retrotransposons can also affect gene expression. They contain an antisense promoter in the +400 to +600 region of their 5′ UTR, and a number of expressed genes located 5′ to full-length L1s have alternate transcription start sites in this L1 region (*57*). Moreover, because there are a number of strong poly(A) signals embedded in L1 sequence, L1 transcripts can also be cleaved prematurely (*58*). This means that an L1 positioned in the transcriptional sense orientation in an intron of a gene may cause a reduction in the gene's transcript level.

In addition, ancient mobile elements probably provided sequences for key host proteins and may have a role in other important biological processes. (i) A DNA transposon is the likely source of RAG1 and RAG2, the recombinase-activating proteins that carry out V(D)J recombination of immunoglobulin genes (*59*). (ii) An ancient retrotransposon may have provided an important enzymatic activity, telomerase, for the eukaryotic cell. DNA ends of chromosomes, telomeres, are maintained by telomerase, an RT that acts via TPRT and is closely related structurally to the RT of non-LTR retrotransposons (*60*, *61*). As we learn more about the vast array of non-LTR retrotransposons, it appears likely that eukaryotic telomerase had its origin from a retrotransposon RT (*26*). (iii) Although the evidence is only circumstantial, L1 elements may serve as "booster stations" for the spread of gene inactivation transmitted by *Xist* RNA in X-chromosome inactivation (*62*).

Genome Size and Mobile Element Clades

S. cerevisiae contains only a handful of retrotransposon types, or clades, and each clade contains less than 100 elements. Retrotransposons make up a small fraction of the yeast genome, probably because their rate of retrotransposition is rather low, about 10^{-5} to 10^{-7} per generation, and their rate of removal by recombination between LTRs is high (*63*) (Table 1). On the other hand, although the genomes of other organisms, such as *Drosophila* and various fish, contain a large number of different clades of both LTR and non-LTR retrotransposons, relatively little genome space is devoted to retrotransposons (4% of the *Drosophila* genome). Although

Drosophila elements, such as P and I, insert at frequencies of ~one per meiosis in hybrid dysgenesis crosses (*64*), and other elements such as copia, *doc*, and mariner move at relatively high frequencies of 10^{-1} to 10^{-3} per generation (*65*, *66*), both selection and deletion of elements after insertion probably account for the small number of each element type in the fly genome (*67*). Similarly, pufferfish have six clades of non-LTR retrotransposons and eight clades of LTR retrotransposons, but a total of only about 5000 retrotransposons (*68*).

In contrast, humans and mice have a very small number of non-LTR retrotransposon clades (~six), but a very large number of total non-LTR retrotransposons (~1,500,000) (*1*, *2*). Although the combined rate of retrotransposition for the autonomous (L1s) and nonautonomous (Alus, processed pseudogenes, SVAs) retrotransposons is probably $>10^{-1}$ per generation, the clearance rate due to deletion must be very much lower than that in the *Drosophila* and pufferfish genomes (*67*). Primarily because of these two factors, the human genome is 20 times as large as the *Drosophila* genome and 8 times as large as the pufferfish genome.

Controlling Mobile Elements

Although transposable elements are continuously entering new genomic sites, phenotype-altering mutations caused by their insertions are much less frequent than are point mutations in most organisms, with the exception of fruit flies, corn, and wheat. Indeed, transposable elements that alter phenotype were discovered only after many years of genetic analysis. Although many genomes contain a large number of active elements, they remain reasonably stable, perhaps because <10% of the genome in organisms with highly active mobile elements, such as mice and humans, consists of protein-coding and regulatory sequences (*1*, *2*). Similarly, only a small fraction of the maize genome consists of genes and regulatory sequences (*3*). Thus, with notable exceptions like *Drosophila*, transposable-element mobility is low in small genomes, where genes constitute a large fraction. In large genomes, with more active elements, only a small fraction of the genome is susceptible to deleterious insertions. Yet, in both of these scenarios, the host places further controls on mobility.

At least two control mechanisms are known: (i) cosuppression usually mediated by small interfering RNA (siRNA) and (ii) methylation. During cosuppression, both the expression of an introduced transgene and its endogenous homologs are suppressed. Both transcriptional and posttranscriptional cosuppression of Ty1 retrotransposition in *S. cerevisiae* have been demonstrated, although the mechanisms remain unknown (*69*, *70*). Cosuppression of *Drosophila* I factor, a non-LTR retrotransposon—probably by an siRNA mechanism—has also been observed (*71*). Perhaps the best-characterized regulation of a mobile element is that of siRNA action on the Tc1 transposon of *C. elegans* (*72*). Tc-1 transposition occurs only in somatic cells and is completely suppressed in germ cells. The mechanism underlying normal suppression begins with readthrough transcription of the transposon from an upstream *C. elegans* gene. Double-strand RNA (dsRNA) of the terminal inverted repeats (TIRs) forms as a result of "snap back" of one TIR onto the other. The 54-nucleotide (nt) TIR dsRNA is then cleaved to 20 to 27 nt by the RNAse III–like enzyme DCR-1 (dicer) to produce the siRNA, leading to destruction of Tc1 RNA by the standard RNA interference mechanism. Mutants of suppression lack Tc1 siRNA and allow germline transposition to occur.

Methylation of mobile elements is another control device used in nature (*73*). Mouse intracisternal A particles (IAPs) are LTR-containing, retroviral-like retrotransposons that frequently cause disease by insertion into genes (*31*). A direct correlation exists between demethylation of mouse IAPs and an increase in their expression (*74*). In addition, other mammalian retrotransposons are hypomethylated in germ cells and in very early development when they are able to retrotranspose, and hypermethylated in somatic cells where their expression is not detectable and they cannot be mobilized. However, the role of methylation in controlling retrotransposition is still unclear. Repetitive DNA, including multiple copies of an LTR retrotransposon, is largely unmethylated, whereas genes are mostly methylated in an invertebrate (*75*). Therefore, study of the rate of retrotransposition of a marked retrotransposon introduced into the genomes of both normal and methylation-defective mice would be useful.

Present and Future Uses of Mobile Elements

For many years, P transposable elements of *Drosophila* have been a powerful tool for insertional mutagenesis, providing a method to link phenotype with genomic sequence (*76*). Recently, bacterial transposons have also been successfully used as insertional mutagens to study the function of ~50% of the annotated genes in *S. cerevisiae* (*77*). To aid DNA sequencing, bacterial transposons have been inserted randomly into DNA from various sources, including fragmented bacterial artificial chromosomes and cDNAs. The mutagenized fragments are then separated, and sequencing reactions are performed using primers complementary to transposon end sequences (*78*).

Young L1s and Alus are polymorphic as to presence in the human genome, meaning that an L1 at a particular locus may be present at that site in <100% of human genomes. These polymorphic elements can then be used to track the migration of human populations, or if the elements are present in some species and not others, they can be used to determine the evolutionary history of those species (*79*). Moreover, because L1 alleles at a locus can also vary in their capability to retrotranspose (*80*), the potential for individual variation in retrotransposition capability is great.

Mobile elements will soon be useful in determining the function of many mammalian genes after gene knockout by insertional mutagenesis. A consensus sequence of the fish Tc1/mariner-type DNA transposon, called Sleeping Beauty (SB), has been constructed. The transposase of this rejuvenated element is 20 to 40 times as active as natural transposases of the Tc1/mariner family. When the transposon is inserted into the genome of mice that already contain the SB transposase, it is mobilized in the subsequent generation from its genomic location to another genomic site at a rate of one to two insertions per offspring (*81*, *82*). However, as expected, insertions are heavily concentrated close to the original transposon site; about 50% are within 3 Mb and 80% are on the same chromosome as the original transposon (*82*). On the other hand, L1 elements offer the potential for generating retrotranspositions at random sites throughout the genome. Retrotransposition from human L1 transgenes has been obtained in mice (*32*), and present insertion frequencies are 1 in every 15 to 20 offspring. With further improvements, this system may have substantial practical value for making random gene knockouts to determine gene function.

The SB transposon has also proven useful as a gene-delivery vector to liver cells in animal systems. In long-term studies in mice, factor IX deficiency and tyrosinase deficiency have been corrected with SB transposon vectors (*83*, *84*).

Summary

Over millions of years of evolution, mobile elements have achieved a balance between detrimental effects on the individual and long-term beneficial effects on a species through genome modification. Indeed, we may soon learn that the shaping of the genome by mobile elements has played an important role in events leading to speciation. Whether these repeated sequences are now "junk DNA" is a complex issue. Some may have had an important function long ago, but have lost that role today. Others may never have had a function, yet the cluttering of our genomes with nonfunctional DNA was a small price to pay for the genome malleability they provided.

References and Notes

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
3. P. SanMiguel *et al*., *Science* **274**, 765 (1996).
4. J. Brosius, H. Tiedge, *Virus Genes* **11**, 163 (1995).
5. N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds., *Mobile DNA II* (American Society for Microbiology, Washington, DC, 2002).
6. K. Mizuuchi, T. Baker, in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds. (American Society for Microbiology, Washington, DC, 2002), pp. 12–23.
7. N. L. Craig, in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds. (American Society for Microbiology, Washington, DC, 2002), pp. 3–11.
8. M. J. Curcio, K. M. Derbyshire, *Nature Rev. Mol. Cell Biol.* **4**, 865 (2003).
9. D. F. Voytas, J. D. Boeke, in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds. (American Society for Microbiology, Washington, DC, 2002), pp. 631–662.
10. D. L. Chalker, S. B. Sandmeyer, *Genes Dev.* **6**, 117 (1992).
11. S. E. Devine, J. D. Boeke, *Genes Dev.* **10**, 620 (1996).
12. Y. Zhu, J. Dai, P. G. Fuerst, D. F. Voytas, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5891 (2003).
13. N. J. Bowen, I. K. Jordan, J. A. Epstein, V. Wood, H. L. Levin, *Genome Res.* **13**, 1984 (2003).
14. A. R. Schroder *et al*., *Cell* **110**, 521 (2002).
15. X. Wu, Y. Li, B. Crise, S. M. Burgess, *Science* **300**, 1749 (2003).
16. S. Hacein-Bey-Abina *et al*., *Science* **302**, 415 (2003).
17. J. L. Jakubczak, Y. Xiong, T. H. Eickbush, *J. Mol. Biol.* **212**, 37 (1990).
18. M. L. Pardue, P. G. Debaryshe, *Annu. Rev. Genet.* **37**, 485 (2003).
19. H. Takahaski, S. Okazaki, H. Fujiwara, *Nucleic Acids Res.* **25**, 1578 (1997).
20. Q. Feng, J. V. Moran, H. H. Kazazian Jr., J. D. Boeke, *Cell* **87**, 905 (1996).
21. J. Jurka, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1872 (1997).
22. D. D. Luan, M. H., Korman, J.L. Jakubczak, T. H. Eickbush *Cell* **72**, 595 (1993).
23. G. J. Cost, Q. Feng, A. Jacquier, J. D. Boeke, *EMBO J.* **21**, 5899 (2002).
24. E. M. Ostertag, H. H. Kazazian Jr., *Genome Res.* **11**, 2059 (2001).
25. J. Cappello, K. Handelsman, H. F. Lodish, *Cell* **43**, 105 (1985).
26. I. R. Arkhipova, K. I. Pyatkov, M. Meselson, M. B. Evgen'ev, *Nature Genet.* **33**, 123 (2003).
27. H. S. Malik, W. D. Burke, T. H. Eickbush, *Mol. Biol. Evol.* **16**, 793 (1999).
28. M. Belfort, V Derbyshire, M. M. Parker, B. Cousineau, A. M. Lambowitz, in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds. (American Society for Microbiology, Washington, DC, 2002), p. 761.
29. T. J. D. Goodwin, J. E. Ormandy, R. T. M. Poulter, *Curr. Genet.* **39**, 83 (2001).
30. B. Brouha *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5280 (2003).
31. E. M. Ostertag, H. H. Kazazian Jr., *Annu. Rev. Genet.* **35**, 501 (2001).
32. E. M. Ostertag *et al*., *Nature Genet.* **32**, 655 (2002).
33. E. T. Luning Prak, A. W. Dodson, E. A. Farkash, H. H. Kazazian Jr., *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1832 (2003).
34. J. L. Goodier, E. M. Ostertag, K. Du, H. H. Kazazian Jr., *Genome Res.* **11**, 1677 (2001).
35. M. B. Brooks, W. K. Gu, J. L. Barnas, J. Ray, K. Ray, *Mamm. Genome* **14**, 788 (2003).
36. G. D. Swergold, personal communication.
37. W. Wei *et al*., *Mol. Cell. Biol.* **21,** (2001).
38. M. Kajikawa, N. Okada, *Cell* **111**, 433 (2002).
39. M. Dewannieux, C. Esnault, T. Heidmann, *Nature Genet.* **35**, 15 (2003).
40. C. R. Hagan, R. F. Scheffield, C. M. Rudin, *Nature Genet.* **35**, 219 (2003).
41. J. M. Bomar *et al*., *Nature Genet.* **15**, 270 (2003).
42. K. Ohshima *et al*., *Genome Biol.* **4**, R74 (2003).
43. J. D. Boeke, *Nature Genet.* **16**, 6 (1997).
44. H. Fan, J. L. Goodier, J. R. Chamberlain, D. R. Engelke, R. J. Maraia, *Mol. Cell. Biol.* **18**, 3201 (1998).
45. C. Esnault, J. Maestre, T. Heidmann, *Nature Genet.* **24**, 363 (2000).
46. E. M. Ostertag, J. L. Goodier, Y. Zhang, H. H. Kazazian Jr., *Am. J. Hum. Genet.* **73**, 1444 (2003).
47. J. V. Moran, R. J. DeBerardinis, H. H. Kazazian Jr., *Science* **283**, 1530 (1999).
48. N. Gilbert, S. Lutz-Prigge, J. V. Moran, *Cell* **110**, 315 (2002).
49. D. E. Symer *et al*., *Cell* **110**, 327 (2002).
50. S. M. Garvey, C. Rajan, A. P. Lerner, W. N. Frankel, G. A. Cox, *Genomics* **79**, 146 (2002).
51. I. Ovchinnikov, A. B. Troxel, G. D. Swergold, *Genome Res.* **11**, 2050 (2001).
52. B. S. Emanuel, T. H. Shaikh, *Nature. Rev. Genet.* **2**, 791 (20012)
53. J. A. Bailey, G. Liu, E. E. Eichler, *Am. J. Hum. Genet.* **73**, 823 (2003).
54. T. A. Morrish *et al*., *Nature Genet.* **31**, 159 (2002).
55. A. Buzdin *et al*., *Nucleic Acids Res.* **31**, 4385 (2003).
56. W.-H. Li, Z. Gu, H. Wang, A. Nekrutenko, *Nature* **409**, 847 (2001).
57. P. Nigumann, K. Redik, K. Matlik, M. Speek, *Genomics* **79**, 628 (2002).
58. V. Perepelitsa-Belancio, P. Deininger, *Nature Genet.* **35**, 363 (2003).
59. A. Agrawal, Q. M. Eastman, D. G. Schatz, *Nature* **394**, 744 (1998).
60. J. Lingner *et al*., *Science* **276**, 561 (1997).
61. M. Meyerson *et al*., *Cell* **90**, 785 (1997).
62. M. F. Lyon, *Cytogenet. Cell Genet.* **80**, 133 (1998).
63. M. J. Curcio, D. J. Garfinkel, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 936 (1991).
64. A. Bucheton, I. Busseau, D. Teninges, in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds. (American Society for Microbiology, Washington, DC, 2002), p. 796.
65. E. G. Pasyukova, S. V. Nuzhdin, D. A. Filatov, *Genet. Res.* **72**, 1 (1998).
66. D. Garza, M. Medhora, A. Koga, D. L. Hartl, *Genetics* **128**, 303 (1991).
67. T. H. Eickbush, A. V. Furano, *Curr. Opin. Genet. Dev.* **12**, 669 (2002).
68. J.-N. Volff, L. Bouneau, C. Ozouf-Costas, C. Fischer, *Trends Genet.* **19**, 674 (2003).
69. Y. W. Jiang, *Genes Dev.* **16**, 467 (2002).
70. D. J. Garfinkel, K. Nyswaner, J. Wang, J.-Y. Cho, *Genetics* **165**, 83 (2003).
71. S. Jensen, M.-P. Gassama, T. Heidmann, *Nature Genet.* **21**, 209 (1999).
72. T. Sijen, R. H. A. Plasterk, *Nature* **426**, 310 (2003).
73. T. H. Bestor, *Trends Genet.* **19**, 185 (2003).
74. C. P. Walsh, J. R. Chaillet, T. H. Bestor, *Nature Genet.* **20**, 116 (1998).
75. M. W. Simmen *et al*., *Science* **283**, 1164 (1999).
76. A. C. Spradling *et al*., *Genetics* **153**, 135)1999)
77. Y. Schevchenko *et al*., *Nucleic Acids Res.* **30**, 2469 (2002).
78. A. Kumar, M. Snyder, *Nature Rev. Genet.* **2**, 302 (2001).
79. M. A. Batzer, P. L. Deininger, *Nature Rev. Genet.* **3**, 370 (2002).
80. S. M. Lutz, B. J. Vincent, H. H. Kazazian Jr., M. A. Batzer, J. V. Moran, *Am. J. Hum. Genet.* **73**, 1431 (2003).
81. C. M. Carlson *et al*., *Genetics* **165**, 243 (2003).
82. K. Horie *et al*., *Mol. Cell. Biol.* **23**, 9189 (2003).
83. S. R. Yant *et al*., *Nature. Genet.* **25**, 35 (2000).
84. E. Montini *et al*., *Mol. Ther.* **6**, 759 (2002).

85. I acknowledge J. Goodier, E. Luning Prak, E. Ostertag, D. Babushok, and J. Moran for helpful comments on the manuscript, and the NIH for grant support.