

# 8

## Eukaryotic Transposable Elements: Teaching Old Genomes New Tricks

---

Susan R. Wessler

*Without transposable elements we would not be here and the living world would probably look very different from the one we know.<sup>1</sup>*

### Overview

When transposable elements were discovered in maize by Barbara McClintock over 50 years ago they were regarded as a curiosity—now they are known to be to the most abundant component of probably all eukaryotic genomes. As such, they make up the vast majority of the output of genome sequencing projects. The availability of so much new information has fueled a revolution in their analysis and studies of their interaction with the host. In addition to discovering transposable elements, McClintock also uncovered three ways that the elements can alter genetic information: by restructuring the genome through element-mediated chromosomal rearrangements; by inserting into and around genes and, in the process, generating new alleles; and by imposing their epigenetic marks on flanking chromosomal DNA. In the context of this book, what is implicit about transposable elements is that their presence and extraordinary abundance in genomes promotes a myriad of genome-altering events. By presenting recent case studies that illustrate each of the three modes of action, this chapter brings the reader up to date on the molecular consequences of transposable element activity on host gene expression and genome evolution.

### Introduction

Transposable elements (TEs) are fragments of DNA that can insert into new chromosomal locations, and often make duplicate copies of themselves in the process. With the advent of large-scale DNA sequencing, it has become apparent that, far from being a rare component of some genomes, TEs are the single largest component of the genetic material of most eukaryotes. They account for at least

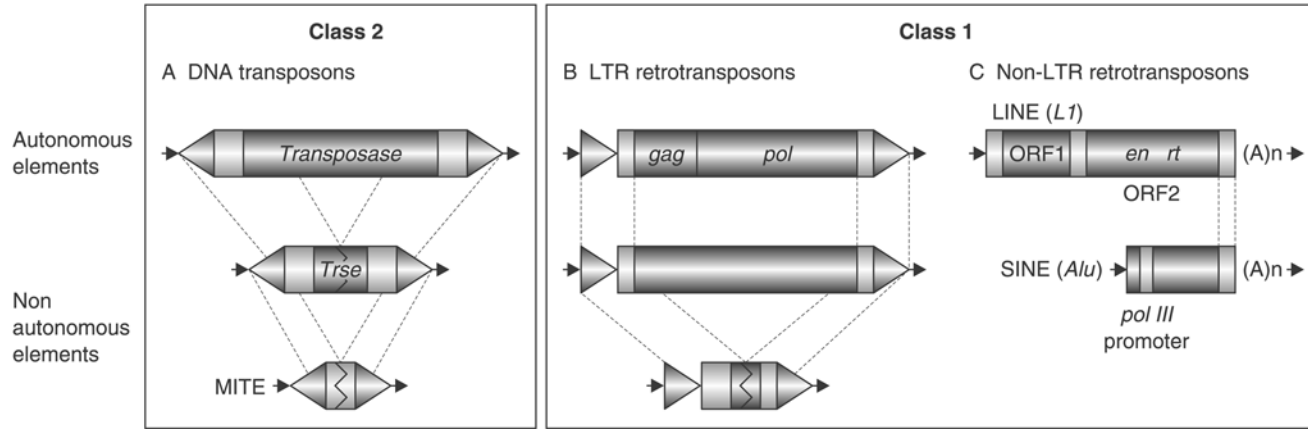
45% of the human genome and 50–90% of some plant genomes (reviewed in references 2 to 4).

TEs were discovered in maize by Barbara McClintock more than a half century ago as the genetic agents that are responsible for the sectors of pigmentation on otherwise colorless mutant kernels.<sup>5</sup> Each sector of colored tissue arises from the mitotic products of a single cell where a TE, which had inserted into and inactivated a gene whose expression is necessary for kernel pigmentation, has excised. Subsequent analysis of mutant alleles from *Drosophila melanogaster*, yeast (*Saccharomyces cerevisiae*), *Caenorhabditis elegans*, and other model eukaryotic organisms furnished the raw material from which molecular biologists could isolate active TEs (TEs have been extensively reviewed in reference 6). Active elements like these, however, constitute only a tiny fraction of the TE complement of the genomes of these model organisms and of most other eukaryotes. Instead, the genomes of higher eukaryotes are filled with thousands, even millions, of seemingly inactive TEs. However, as will be discussed, both active and inactive TEs can impact the evolution of genome structure and the regulation of gene expression.

## TE Classes and Mechanisms of Transposition

Eukaryotic TEs are divided into two classes, according to whether their transposition intermediate is RNA (class 1) or DNA (class 2) (figure 8.1). For all class 1 elements, the element-encoded transcript (mRNA) forms the transposition intermediate. In contrast, with class 2 elements, the element itself moves from one site to another in the genome. Each group of TEs contains autonomous and nonautonomous elements. Autonomous elements have open reading frames (ORFs) that encode the products required for transposition. In contrast, nonautonomous elements do not encode transposition proteins but are able to transpose because they retain the *cis*-sequences necessary for transposition. Integration of almost all TEs results in the duplication of a short genomic sequence (called a target site duplication, or TSD) at the site of insertion.

Eukaryotic DNA (class 2) transposons usually have a simple structure with a short terminal inverted repeat (TIR) (around 10–40 bp, but can be up to about 200 bp) and a single gene encoding the transposase. Transposase binds in a sequence-specific manner to the ends of its encoding element (called an autonomous element) and to the ends of nonautonomous family members. Once bound, transposase initiates a cut-and-paste reaction whereby the element is excised from the donor site (generating an “empty site”) and inserted into a new site in the genome. There are several possible fates for the empty donor site that can lead to different outcomes for the host. Repair of the double-strand break at the empty site can be precise (leaving no trace of the element or TSD) or imprecise (leaving a so-called “transposon footprint” of a few to several base pairs or deleting adjacent host DNA). Increase in element copy number occurs when the transposon sequence is restored to the empty donor site templated by the DNA sequence of the sister chromatid. This mechanism also can replace autonomous elements at the empty site with nonautonomous elements



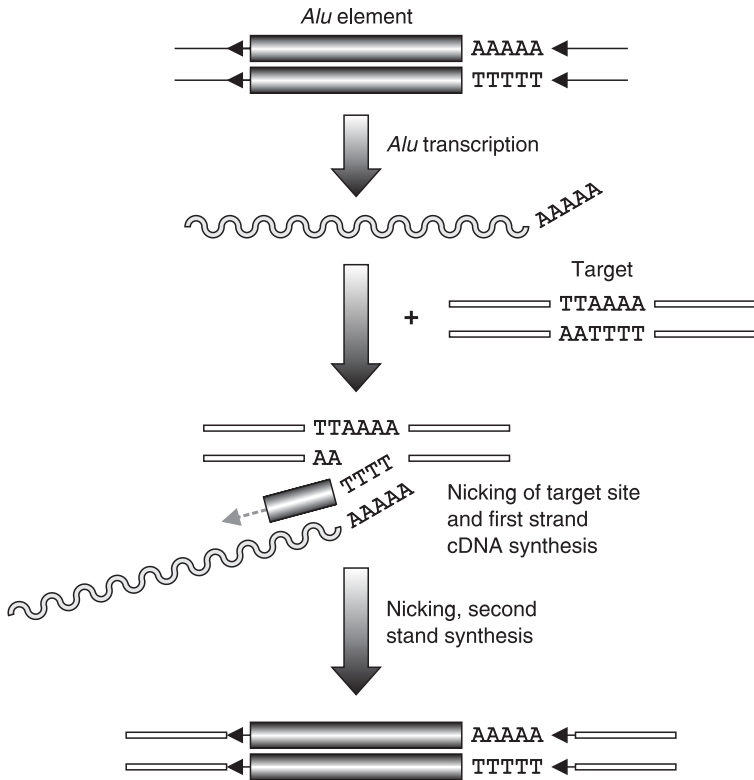
**Figure 8.1. Structural features and classification of eukaryotic transposable elements.** Elements are divided into two classes, according to whether their transposition intermediate is RNA (class1) (B and C) or DNA (class 2) (A). Class 1 elements are further divided into two groups on the basis of transposition mechanism and structure: (B) LTR retrotransposons, and (C) non-LTR retrotransposons. Each class contains autonomous and nonautonomous elements. Autonomous elements encode proteins required for transposition (*gag*, capsid-like protein; *pol*, reverse transcriptase; *ORF1*, a *gag*-like protein; *en*, endonuclease; *rt*, reverse transcriptase). Nonautonomous elements do not encode these proteins but retain the *cis*-sequences necessary for transposition. Target site duplications are the black arrowheads flanking each element, and the inverted repeats at the termini of class 2 elements (A) and the direct repeats at the ends of LTR retrotransposons (B) are represented by large gray arrowheads.

if, for example, a deletion occurs during template-mediated repair. The elements studied by McClintock, including the *Ac/Ds* and *Spm/dSpm* families, are DNA transposons capable of insertion and excision.

Class 1 retroelements can be divided into two groups on the basis of transposition mechanism and structure. LTR retrotransposons have long terminal repeats (LTRs) in direct orientation that can range in size from around 100 bp to several kb. Autonomous elements contain at least two genes, called *gag* and *pol*. The *gag* gene encodes a capsid-like protein and the *pol* gene encodes a polyprotein that is responsible for protease, reverse transcriptase (RT), RNase H, and integrase activities. LTR retrotransposons resemble retroviruses in both their structure and mechanism of transposition (called retrotransposition). An element-encoded transcript that initiates from a promoter in the 5' LTR and terminates in the 3' LTR is transported to the cytoplasm. There it serves as both mRNA and template for double-strand cDNA that is transported into the nucleus where it can then integrate into the genome, leading to massive increases in copy number (thousands, even hundreds of thousands). The host can mitigate this increase in genome size by mediating homologous recombination between the identical or near-identical LTRs of full-length elements, generating a much shorter solo LTR (where "solo" LTR refers to only one copy of the sequence that normally is repeated at the termini of the element). The efficiency of solo LTR formation depends on several factors, including the efficiency of host recombination mechanisms and the length of the LTR. LTR retrotransposons compose the largest fraction of most plant genomes, where they appear to be the major determinant of the tremendous variation in genome size.

Non-LTR retrotransposons are divided into the autonomous long interspersed elements (LINEs) and the nonautonomous short interspersed elements (SINEs). LINEs encode two ORFs, which are transcribed as a bicistronic mRNA composed of ORF1 (an RNA binding protein) and ORF2 (endonuclease and RT activities). Both LINEs and SINEs terminate by a simple sequence repeat, usually poly(A). LINE transcripts initiate at a promoter within the 5' end of the element and terminate at or often downstream of the simple repeat sequence. SINEs are characterized by an internal RNA *pol III* promoter (in contrast, all protein coding genes, including those in LINEs, have *pol II* promoters) near the 5' end. SINEs are a heterogeneous group of elements that range in length from 90 to 300 bp and are derived either from a variety of tRNA genes or from 7SL RNA. There is increasing evidence that SINEs rely on LINEs for the machinery necessary for their amplification.

LINEs amplify by an interesting mechanism (called target primed reverse transcription, TPRT) that appears to have played a critical role in the evolution of eukaryotes, especially primates.<sup>7</sup> As with LTR retrotransposons, an element-encoded transcript is transported to the cytoplasm where it serves as mRNA. However, unlike LTR element encoded transcripts, which serve as template for reverse transcription in the cytoplasm, non-LTR transcripts re-enter the nucleus where chromosomal DNA that has been nicked by the element-encoded endonuclease primes reverse transcription of the transcript into DNA (figure 8.2). SINEs have been spectacularly successful at utilizing LINE machinery to propagate; that is, a SINE transcript instead of a LINE transcript is inserted at the chromosomal nick. In addition, processed pseudogenes are thought to arise through TPRT utilizing



**Figure 8.2.** Transposition (also called “retroposition”) of non-LTR retrotransposons by target primed reverse transcription. An *Alu* element is shown; the same mechanism can also explain the increase in copy number of LINEs and other SINEs. An *Alu* RNA transcript (wavy line) anneals to a nicked site in the genome (the target) and the 3' OH of the T residue at the nick is used to prime first strand synthesis of a cDNA copy of the *Alu* transcript by reverse transcriptase (gray rectangle and dotted arrow; probably encoded by an *L1* LINE). Presumably, nicking of the other DNA strand must precede second strand synthesis. Black arrowheads represent the target site duplication, both at the original site (thick lines) and at the target site (thin open boxes). (Adapted from Batzer M. A. and Deininger P. L. *Alu* repeats and human genomic diversity. *Nat. Gen. Rev.*, 2002; 3: 370–379, box 1.)

LINE machinery to generate and insert cDNA copies of cellular mRNAs into the genome. A plausible model has also been proposed for the origin of introns through TPRT. As will be discussed below, non-LTR retrotransposons are the dominant element type in mammalian genomes, where they appear to account for most of the species-specific differences.

## Changing Views of the Impact of TEs on Evolution

Why do TEs predominate in most genomes and how much have they influenced the evolution of life? Ever since the discovery of TEs in maize, speculation has centered

on their possible role in genome evolution. McClintock called TEs “controlling elements” because her observations of mutant phenotypes led her to propose that TEs normally controlled maize development.<sup>8</sup> This idea was rejected and she later proposed that TEs were part of a global stress response that could potentially restructure genomes and promote survival (“genome shock”).<sup>9</sup> In time, TEs were recognized as ancient components of all genomes (both prokaryotic and eukaryotic). Their ubiquity and mutagenic potential led some—especially neo-Darwinian selectionists—to propose that they originated and thrived because they were important tools of evolution and were essential (integral) genome components (discussed in reference 10).

The field was transformed in 1980 with the publication of two influential papers heralding the view that TEs were selfish or junk DNA and that their evolutionary success could be explained solely by their ability to replicate themselves.<sup>11,12</sup>

As stated by Orgel and Crick:<sup>12</sup>

When a given DNA or class of DNA of unproven phenotypic function can be shown to have encoded a strategy (such as transposition) which ensures its genomic survival then no other explanation of its existence is necessary. The search for other explanations may prove if not intellectually sterile, ultimately futile.

These views had a chilling influence on the field of transposon biology, leading many investigators to change their research focus from the impact of TEs on their host to the characterization of TEs and transposition mechanisms.

### *From Genetics to Genomics*

The selfish DNA theory held until it became clear that TEs usually made up the largest fraction of the genomes of multicellular eukaryotes. Instead of there being one or two TEs near a gene, some human genes were found to contain up to 100! This revelation led to an entirely new set of questions about how organisms and their TEs coexist. What emerged was a new synthesis of prior ideas and current data, whereby TEs and their hosts are seen as being in an arms race—with the TEs trying to increase their copy number and the host attempting to protect its genetic information from mutation. This arms race leads to the development of genetic novelty that can be co-opted by the host. This view has been nicely summarized by Labrador and Corces:<sup>1</sup>

As replicative sequences, TEs are kept in check by their environment—which is the genome. Natural selection is thus responsible for the existing diversity of TEs and for the many different ways they employ to interact with their hosts. In the same manner TEs benefit from their hosts and evolve, improving their replicative efficiency. And, because evolution is an opportunistic process, the host benefits also from the genetic variability offered by TEs. As a result of the selective process, TEs have become a natural component of modern genomes, and their endurance is due not only to their ability to replicate themselves inside the cell but also to the fact that the eukaryotic genome found in these elements an excellent tool that is constantly used to generate evolutionary novelties and to maintain its own integrity.

The remainder of this chapter will discuss features of TEs that have been co-opted by their hosts by first defining the major mechanisms of genomic restructuring mediated by TEs and then providing examples, or case studies, that highlight different mechanisms.

### Three Major Ways TEs Restructure Genomes

McClintock's characterization of the genetic behavior of TEs revealed three distinct ways that TEs could restructure the host genome.

#### *TE-mediated Chromosome Breakage and Rejoining*

TEs were first observed in maize as specific sites of chromosome breakage (called *dissociation* or *Ds*) that could initiate the breakage–fusion–bridge cycle.<sup>13</sup> The name of this cycle derives from three events: DNA breakage at the chromatid stage; fusion of the broken ends to produce a dicentric chromosome; and formation of a bridge when the two centromeres of the dicentric chromosome are pulled to opposite poles during mitosis. We now know that this phenomenon is due to the ability of DNA transposons to mediate nonhomologous recombination events when transposase-generated single- and double-strand breaks are repaired (reviewed in reference 14). In addition, there is growing recognition that homologous recombination between the thousands, tens of thousands, or even hundreds of thousands of related TEs dispersed throughout eukaryotic genomes has had a major impact on genome structure and gene content. (See the case studies sections below.)

#### *TEs as Insertional Mutagens*

The ability of TEs to knock out or alter gene function via insertion has been recognized since McClintock's analysis of spotted corn kernels over 50 years ago. However, while TEs as insertional mutagens were originally thought to be rare, genome sequencing projects have revealed that the vast majority of normal plant and mammalian genes harbor several TE insertions. This is largely because the majority of eukaryotic transposons are small nonautonomous elements whose insertion into genes can alter rather than knock out gene function. Who would have guessed that over 200,000 of the 1 million *Alus* (SINEs) in the human genome are in (human) genes? As will be discussed below, TEs are prominent components of a large fraction of the regulatory regions of the genes of higher eukaryotes. (See the case studies sections below.)

#### *TEs and Epigenetic Regulation*

McClintock was also the first to note that the activity of some TEs cycled between active and inactive states, a phenomenon she later called “change in phase.”<sup>15</sup> Change in phase occurred during a single plant generation or from one generation to the next. In addition, many investigators determined that endogenous inactive

TEs could be reactivated by a variety of stresses (e.g., following the breakage–fusion–bridge cycle, chemical mutagenesis, and radiation). These observations led eventually to the recognition that TEs are the targets of inactivation by the host via epigenetic mechanisms that, as will be discussed in a later section, interfere with TE activity by preventing the production or accumulation of TE-encoded RNA. TEs that are inactivated by epigenetic mechanisms are said to be “silenced.” As illustrated below, a host’s attempts to protect its genetic information from insertional mutagenesis by silencing its TEs may result in epigenetic alterations that affect the activity of the genes it is trying to protect (see the case studies sections below).

### Huge Variations in TE Content May Impact Evolution

An unexpected finding from the analysis of genome sequences is that TE content varies from species to species in two important ways: by the classes of TEs present and their fractional representation in the genome, and by the level of TE activity. The yeast *Saccharomyces cerevisiae* has only LTR retrotransposons (called *Ty* elements), and the vast majority are solo LTRs, generated by the yeast’s very efficient homologous recombination machinery. In mammalian genomes, class 1 non-LTR retrotransposons predominate, with class 2 DNA transposons making up less than 5% of the TE fraction (reviewed in reference 16). A remarkable 25–30% of the human genome is derived from just two families of non-LTR elements: *L1* (*LINE-1*), with more than 500,000 copies (~17%); and the much smaller *Alu* (a SINE), with approximately 1–1.4 million copies (~10%). The genomes of flowering plants, including both monocots (e.g., grasses such as rice and maize) and dicots (e.g., *Arabidopsis* and tomato), have a rich collection of both class 1 and class 2 elements, with LTR retrotransposons comprising the largest fraction of most characterized genomes (reviewed in reference 17). *C. elegans* and *Drosophila melanogaster* also have both class 1 and class 2 elements, but class 2 elements predominate in the former and class 1 in the latter.

Genome-wide activity of TEs also varies from species to species. Given the rich genetic analysis of TE-mediated mutations in maize, *Drosophila* and *C. elegans*, it is not surprising that these genomes were found to contain many young, active TE families. Flowering plants, especially members of the grass clade (e.g., rice, maize, barley, and wheat), are in an epoch of TE-mediated genome diversification, with the participation of many families of active class 1 and class 2 elements. In contrast, while extant mammalian genomes have many fewer active TE lineages, TE activity varies dramatically between species. For example, although *L1* elements make up approximately the same fraction of the human and mouse genomes (~20%, ~500,000 copies), only about 80 to 100 *L1*s are active in humans while approximately 3000 are active in the mouse. The functional consequence of this difference is reflected in the fact that new mutations due to insertion of retroelements are rare in humans (about 1 in 500 human mutations) but represent approximately 10% of all mutations in the mouse. Analysis of the age of *Alu* elements in

the human genome (as estimated from the amount of sequence identity) indicates that the insertion rate may have been 100-fold higher (one new insertion per primate birth) earlier in primate evolution.

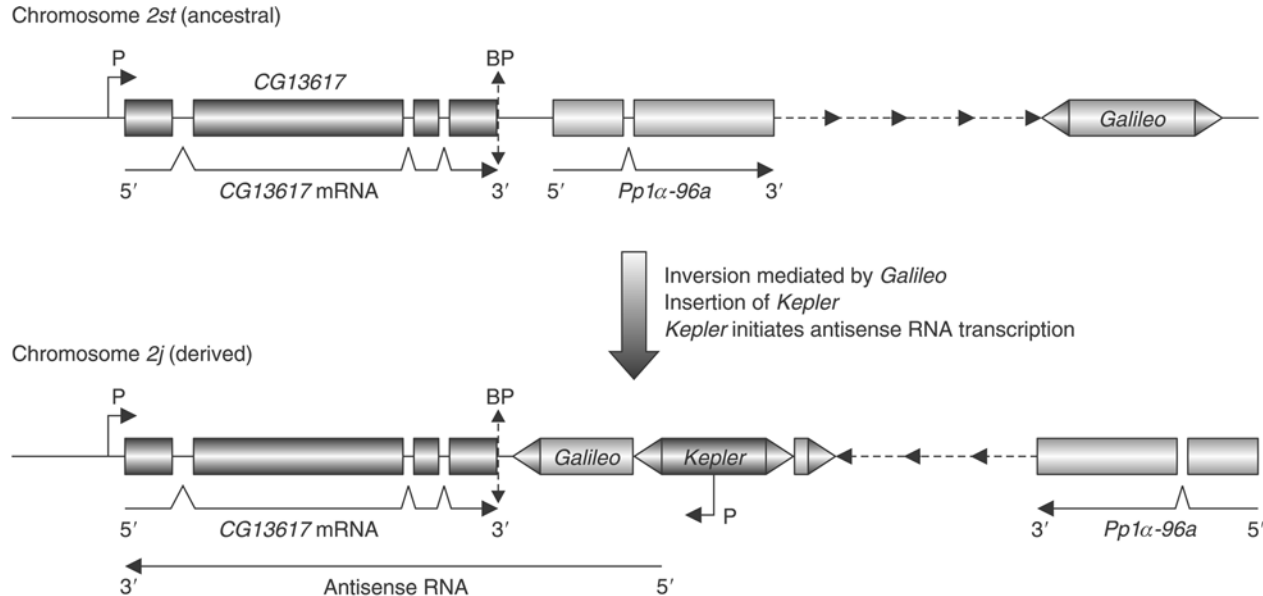
One of the objectives in presenting the following case studies is to illustrate how variations in TE content from species to species may influence both the quality and quantity of genomic variation and, in this way, impact the trajectory of genome evolution. To this end, the case studies are organized into one of the three categories of TE-mediated genomic change already discussed: recombination, insertion mutagenesis, and epigenetic.

## Case Studies of TE-mediated Recombination

### *TE-mediated Inversions in Drosophila buzzatii*

The *Drosophila* genus is known for a remarkable level of chromosomal inversions in natural populations, both within and between species. The best-characterized example at the molecular level is a large inversion containing about one third of the euchromatin found on chromosome 2 of *D. buzzatii*, a species found throughout South America.<sup>18</sup> Two chromosomal forms, the ancestral, *2 standard* (*2st*), and the derived, *2j*, are both found in populations at high frequency, possibly due to fitness advantages offered by each form; the *2j* derivative conditions a larger body while the *2st* form develops more quickly (figure 8.3). Thus, this is an example of a balanced polymorphism. The origin of these adaptive changes has been investigated by determining first how the inversion arose and then how it impacts the host phenotype. With regard to origin, the *2j* chromosomal inversion was caused by ectopic (homologous) recombination between two *Galileo* elements found in opposite orientation on *2st*. Perhaps significantly, *Galileo* is a member of the *Foldback* group, which is known to promote chromosomal rearrangements during transposition. The *2j* chromosome is itself highly polymorphic due to mutations that occurred subsequent to the original inversion. Many of these secondary mutations are insertions of TEs. The maintenance of these variants of *2j* is probably due to reductions in recombination associated with chromosomal inversions.

Having defined the chromosomal lesions, the next step was to understand the connection between the inversion and the host phenotype. Two models have been proposed to explain why some inversions are maintained at high frequency in natural populations. The “coadaptation” hypothesis posits that inversions maintain favorable allele combinations in heterozygotes because, as mentioned above, of a reduction in recombination in the inverted region. In contrast, the “position effect” hypothesis proposes changes in the expression or function of genes located near or within inversion breakpoints. In support of the position effect model, Puig et al.<sup>19</sup> identified an ORF (CG13617) whose 3′ end is located just 12 bp from the inversion breakpoint and whose expression is reduced in embryos homozygous for the *2j* chromosome. Surprisingly, the silencing of CG13617 appears to be caused, like the original inversion, by a TE. Specifically, a transcript originating in another *Foldback*-like TE (called *Kepler*) that inserted into the *Galileo* element (after the original inversion) is



**Figure 8.3.** Transposable elements mediate both a chromosomal inversion and altered gene expression in natural populations of *Drosophila buzzatii*. Part of ancestral chromosome 2 *standard* (*2st*) is shown at the top and derived chromosome 2*j* at the bottom. The *2j* inversion was caused by ectopic recombination between two *Galileo* transposable elements found in opposite orientation on *2st* (only one element and one breakpoint [BP] are shown). The breakpoint occurred between the 3' end of a gene of unknown function called *CG13617* and the 5' end of a gene called *Pp1α-96a*. Subsequent insertion of another element, *Kepler*, into the *Galileo* element near one breakpoint introduced a promoter that initiated an antisense transcript believed to be responsible for a reduction in *CG13617* mRNA in embryos of *D. buzzatii* with the *2j* chromosome.

transcribed across the breakpoint into CG13617, generating an antisense mRNA that appears responsible for gene silencing. The authors suggest the involvement of TEs at three stages—the inversion origin, the subsequent insertions into the inverted DNA, and the position effect—may contribute to the evolutionary success of inversions.

### ***Ty-mediated Adaptive Rearrangements in Yeast***

In multicellular eukaryotes such as *Drosophila*, it is much more challenging for the investigator to go beyond analyzing extant strains like those described in the previous section. It would be ideal if one could grow populations of eukaryotes under stress conditions, select survivors that adapt best, and determine whether and how the genome has been rearranged. Fortunately, such experiments have been done in *Saccharomyces cerevisiae* (yeast), where isogenic strains were grown in a chemostat under conditions of glucose-limitation for 100 to 500 generations and the survivors (so-called “evolved” clones) analyzed using DNA microarrays for changes in gene copy number and expression.<sup>20</sup>

Of eight evolved clones examined, the vast majority of the chromosomal rearrangements (deletions and translocations) had a TE-related sequence at the breakpoints. The genome of *S. cerevisiae* contains only one elements class, LTR retrotransposons (called *Ty* elements), and the vast majority of the approximately 330 copies in the genome are present as solo LTRs, not full-length elements. As with the *D. buzzatii* inversion, most of the yeast rearrangements appear to be due to ectopic recombination between *Ty* sequences. While the study does not offer direct evidence that the rearrangements promote fitness, the fact that multiple strains share the same breakpoint despite their independent origins points to the rearrangement as the basis for adaptation. For example, three strains shared a breakpoint at a *Ty* sequence near *CITI*, which encodes citrate synthase, an important enzyme in the TCA cycle. The authors speculate that the rearrangement may activate *Ty* sequences that lead to the derepression of *CITI* in the presence of glucose. As a key regulator of the TCA cycle, *CITI* activation may promote the derepression of other genes in the TCA cycle and result in the adaptive phenotype. The authors further speculate that the yeast genome may be populated with *Ty* element sequences that have become fixed because they offer selective advantages, in this case the promotion of adaptive chromosomal rearrangements in response to glucose limitation.

### ***Alu-mediated Chromosomal Rearrangements in Humans***

The increasing availability of genomic sequence, especially from primates and other mammals, provides an unparalleled opportunity to unravel the contributions of TEs to genome variation, hereditary disease, and speciation in complex organisms. Humans offer several advantages in this regard. There is not only a complete genome sequence but also increasing amounts of sequence from multiple individuals, including healthy people and those with genetic disorders. Almost 50% of the genome is derived from TEs, mostly *L1* and *Alu* elements. Finally, an extraordinary array of community resources (e.g., expression and metabolic profiles) is available

to elucidate the functional significance of TE-mediated genomic alterations (both within human populations and with close relatives, such as the chimpanzee).

As mentioned in an earlier section, although human genomic DNA is largely derived from TEs, fewer than 100 *LI* and *Alu* elements may be currently active (capable of transposition) (reviewed in reference 16). With so little activity, it was reasonable to assume, as many did, that the impact of TEs on genome evolution might be inconsequential. Recent studies make it clear that this assessment was dead wrong. As it turns out, TEs do not have to be active to be responsible for genomic rearrangements. What makes the human genome particularly susceptible to TE-mediated rearrangements is that with only two element families reaching huge copy numbers, the genome is packed with homologous sequences that are potential sites of unequal recombination and other mechanisms that promote rearrangements. Moreover, the enrichment of *Alu* in GC rich DNA (gene-rich regions) means that rearrangements are particularly likely to affect genes, with consequences for both disease and evolution.

### Low Copy Repeats in the Human Genome

One of the most surprising findings from the sequence of the human genome is the discovery that as much as approximately 5–6% is comprised of low copy repeats (LCRs) (also called “segmental duplications”).<sup>21</sup> LCRs are 10–250 kb in length and have greater than 95% sequence identity with each other, suggesting that they evolved over the past 35 million years. Most LCRs are dispersed, not tandemly arranged.

Evidence that LCRs are especially dynamic regions of the human is provided by their frequent association with hereditary diseases.<sup>22</sup> It appears that once generated, the LCR is more likely to undergo additional rearrangements due, in part, to nonallelic homologous recombination. Compared with other sequenced animal genomes, the human genome is enriched for longer LCRs (>10 kb) that preferentially contain genic sequences. Diseases caused by alterations in LCR sequences include Williams–Beuren, Prader–Willi, Angelman, and cat-eye syndromes.

What is the mechanism(s) underlying LCR formation? A role for *Alu* has long been suspected and has been documented for several deletions and rearrangements associated with genetic diseases (reviewed in reference 23). The abundance of *Alus*, especially in genic regions, makes it a prime candidate for mediating LCR formation via transposition and for subsequent LCR instability via unequal homologous recombination between *Alu* elements. In addition, the apparent formation of most LCRs since 35 million years ago (mya) coincides with the timing of bursts in *Alu* activity, which resulted in the enormous number of insertions of two *Alu* subfamilies, *AluS* (25–45 mya) and *AluY* (35 mya to present).

Data for a role of *Alu* in the origin of LCRs was provided by Bailey et al.,<sup>24</sup> who performed a comparative analysis of thousands of LCRs extracted from the human genome sequence. They found that *Alu* sequences appeared more frequently in LCRs than would be expected by chance and, more importantly, members of the young *Alu* subfamilies (*AluS* and *AluY*) were much more frequently associated with the LCR junctions than expected. They concluded:

We propose that the primate-specific burst of *Alu* retroposon activity (which occurred 35–40 mya) sensitized the ancestral human genome for *Alu-Alu*-mediated recombination events, which, in turn, initiated the expansion of gene-rich segmental duplications and their subsequent role in nonallelic homologous recombination.

The significance of gene-rich segmental duplications in the evolution of primates and of rodents is suggested by recent comparative analysis of gene content in humans, rats, and mice. While all three organisms have the same set of approximately 30,000 protein-coding genes, major differences in gene content arise from species-specific expansion and divergence of gene family members, especially those of possible adaptive significance, such as genes involved in olfaction and pathogen defense.<sup>25</sup>

### **Case Studies of TE Insertions: Diversifying Genes and Gene Expression**

Prior to the advent of genome sequencing, there were many reports of alterations in gene structure or function due to particular TE insertions. Several instances were reported, for example, of TEs in promoters influencing transcription initiation, or TEs in introns influencing pre-mRNA splicing patterns. Reports like these were dismissed by some as being anecdotal and rare and, as such, providing only marginal support for a significant role for TEs in evolution. The reader is referred to some reviews where these prior studies have been summarized and discussed.<sup>26–28</sup> The following section summarizes a few recent studies where large quantities of genomic sequence data have been used to argue that TEs may be influencing the expression or altering the function of hundreds, perhaps thousands, of host genes.

#### ***TE Insertions into Regulatory Regions of Human Genes: Guilt by Association***

Arguments have been made for the importance of TEs in human evolution based solely on the number of insertions in regulatory regions, many having occurred after divergence from our last common ancestor. The sheer number of insertions is staggering: over 20% of human genes have TEs (mostly *Alus*) in their 5' and 3' noncoding sequences, and approximately 25% of the entries in the human promoter database contain a TE-derived sequence.<sup>29</sup> At this time, very little is known about the actual role, if any, of TE sequences in the regulation of individual human genes. (See chapter 4 for evidence that some TEs may promote the origination of simple sequence repeats, which in turn may facilitate a mode of mutation that can quantitatively and reversibly adjust gene activity.) Furthermore, because most of the insertions occurred over 5 mya, it is likely that their impact on gene expression will never be known due to the accumulation of other mutations since the insertion event. The following case study describes the very recent insertion of hundreds of TEs into rice genes, thus permitting an examination of the impact of insertion on gene and genome evolution.

### **Miniature Inverted-repeat Transposable Elements Can Rapidly Diversify Rice Genes**

Miniature inverted-repeat transposable elements (MITEs) are a special class of nonautonomous DNA elements that are found in genomes at very high copy number, where they are preferentially in gene rich regions (reviewed in reference 30). What appears to make them “special” is how they originate and amplify. The majority of characterized nonautonomous class 2 (DNA) elements are more than 1 kb in length and can amplify to moderate copy number (usually fewer than 50 copies in a genome) after they arise by deletion from an autonomous (transposase-encoding) element. In contrast, MITEs are short (usually less than 500 bp) and appear to amplify from one or a few elements to over 1000 elements in a very short period of time (perhaps only a few hundred years). While class 1 *LI* and *Alu* elements are the most common TEs in the introns and regulatory regions of human genes, class 2 MITEs are the most common TEs in plant genes (they also are abundant in certain animal genomes, including insects and fish). Unlike *Alus*, where all one million plus elements have a common origin, there are many distinct MITE families of independent origin in a single genome, each with hundreds or thousands of related elements.

Plant genes have, on average, very short introns (about 200 bp, although there are plant introns longer than 3 kb) compared with their mammalian counterparts (about 2.5 kb on average). There is some evidence that plants cannot efficiently splice long introns; a requirement for short introns may be one reason why short elements such as MITEs (and, less frequently, SINES) predominate in plant genes. That is, there is a good chance that a MITE insertion into a plant gene will not disrupt gene expression. This apparently is the case as there are hundreds of normal genes already in databases with MITE sequences in their introns, 5' and 3' untranslated regions, and in their promoters. There also are numerous examples, especially in the grasses (e.g., rice and maize), of alleles that differ, in part, due to the presence or absence of MITEs. Unfortunately, in most of these cases the MITE insertion occurred a long time ago (perhaps over a million years) and it is virtually impossible to distinguish the impact of insertion on gene expression (if any) from the effect of other sequence changes that have accumulated in each allele.

To assess the impact of MITE insertions on genome evolution, it first was necessary to identify MITEs that are still transposing. Such a family recently was found in rice.<sup>31–33</sup> The rice genome is being sequenced because rice is the most important source of calories for humans and, fortuitously, because rice has the smallest genome among the cereals (~430 Mb, maize ~2500 Mb, barley ~5000 Mb). Whole genome draft sequences are available for *japonica* and *indica*, two of the three subspecies of rice that have been independently domesticated from wild relatives.<sup>34,35</sup> A 429 bp MITE called *mPing* is active in both *japonica* and *indica* rice. The difference in the estimated copy number of *mPing* elements in a *japonica* (Nipponbare) and an *indica* (93-11) genome (70 versus 14) suggested the recent amplification of this MITE family, perhaps during domestication. Furthermore, analysis of several *japonica* cultivars found that the temperate *japonicas* contained the highest number of *mPing* elements (over 1000 elements in a few cultivars!) whereas the tropical *japonicas* contain the least (many have only a single element).

This dramatic difference in *mPing* copy number between the two subgroups of *japonica* is significant because the temperate and tropical cultivars are thought to have diverged from a common ancestor since domestication (5000 to 7000 years ago). The two varietal groups are adapted to radically different temperature and water regimes: the tropical cultivars flourish in tropical and subtropical environments whereas the temperate cultivars represent an evolutionary extreme, having been selected for productivity in cool, temperate zones with very short growing seasons. Thus, in a situation reminiscent of McClintock's genome shock theory,<sup>9</sup> stress activation of *mPing elements* during the domestication of temperate *japonicas*, followed by their preferential insertion into genic regions, might have diversified these cultivars and hastened their domestication by creating new allelic combinations that might be favored by human selection.

As with most of the case studies in this chapter, the impact of the bursts of *mPing* insertions on genome evolution is unclear at this time. What is clear is that the thousands of new insertions, presumably into gene rich regions of the genome, will be the focus of detailed analyses to determine which, if any, contributed to adaptation and/or domestication.

### **L1 Insertions May Down-regulate Human Gene Transcription**

This review provides case studies that support the view that genetic novelty is an important outcome of the competition between TEs and their hosts. Given the recent availability of the human genome sequence, many findings regarding the impact of TEs are, like the examples above, preliminary in nature but still intriguing because of the potential to impact a very large number of genes. The possible impact of *L1* elements on human gene transcription provides a fitting example. Recall that *L1* elements are autonomous retrotransposons that encode two ORFs. The rarity of *L1*-encoded RNA and protein *in vivo* has been a longstanding puzzle given that 20% of the genome is derived from *L1* sequences. Clearly, the host tightly regulates *L1* expression, but how? Results from two studies indicate that *L1* RNA accumulation is inhibited in two ways: by premature polyadenylation within *L1* sequences, or by a block in the movement of RNA polymerase while *L1* sequences are being transcribed.<sup>36,37</sup>

*L1* sequences are found in around 79% of human genes; the vast majority is in introns. Prior to these studies it was reasonable to assume that most *L1* sequences were spliced with the surrounding intron from human gene transcripts and, as such, were phenotypically neutral. In light of the new findings, Han et al.<sup>36</sup> suggest an alternative model whereby *L1* sequences regulate gene expression on a global scale by serving as a "molecular rheostat" of transcription levels. According to this model, RNA polymerase may slow down or even terminate while transcribing *L1* sequences in the introns of certain genes due to features of the sequence that have yet to be determined. In support of this hypothesis, their computational analysis of genomic sequence and gene expression data revealed that genes with more *L1* insertions accumulate, on average, less mRNA than genes with fewer *L1* insertions.

### ***Alu Exonization: Diversifying Human Genes for Good and Bad***

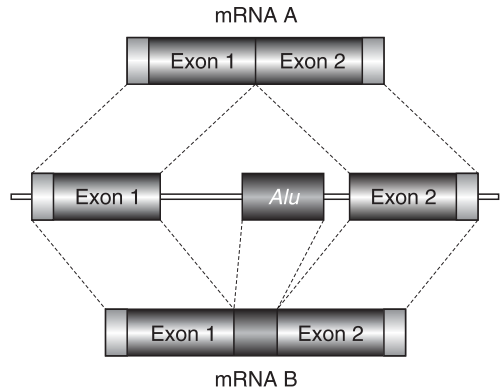
While humans, like other mammals, are estimated to have only around 30,000 protein-coding genes, the diversity of the proteome is much greater than this number because a large fraction of human genes (~40–60%) produce more than a single type of mRNA through alternative splicing (reviewed in reference 38 and chapter 15) (see also discussion of RNA editing in chapter 14). Alternative splicing in mammals is attracting increasing attention because species-specific splicing has been found to generate transcripts that encode distinct proteins from what appears to be the same gene. Furthermore, species-specific alternative splicing often is due to the presence of an *Alu* insertion in a human gene versus, for example, its mouse homolog.

First, some definitions are necessary. Exons that occur in all the transcripts from a single gene are called “constitutive” exons, while those that are not in all transcripts are called “alternative” exons. Among alternative exons are “major form” exons, which occur in a majority of the transcripts from a gene, and “minor form” exons, whose occurrence is less frequent and often rare. Comparative analysis of gene expression in mice and men show conservation of 67–98% of constitutive and major form exons but only 15–28% of minor form exons. One reason for this difference is that the hundreds of thousands of *Alu* elements that are in human but not mouse introns (mice introns have their own families of TEs that are not in human introns) have a significant impact on human splicing patterns.

Recent whole-genome analysis of human transcripts has revealed that some *Alu* elements in internal introns of protein-coding genes have become exons; a process that has been called “*Alu* exonization” (figure 8.4). Specifically, over 5% of alternatively spliced exons in humans are derived from *Alu* sequences.<sup>39,40</sup> Virtually all of the *Alu* exons are minor form exons that are transcribed from alternatively spliced genes. In this way, *Alu* sequences added to protein coding regions generate new protein isoforms that can be tested by evolution while the normal protein product is still being synthesized.

This sounded like an ideal situation to evolve new proteins until it was noticed that a few human diseases are caused by *Alu* exons that have become constitutive, that is, the *Alu*-containing transcript is the only transcript produced. In one case a patient with ornithine amino transferase deficiency had sustained a single base pair mutation that activated a cryptic 5′ splice site in an *Alu* element located in an intron of the ornithine aminotransferase gene. The mutant gene produced a constitutive *Alu* exon that encoded a truncated protein due to an in-frame stop codon in the *Alu* sequence. Another human disease, Alport syndrome, was found to result from a single base pair mutation that activated a cryptic 3′ splice site in an intron of a gene (*COL4a3*), which encodes collagen type IV, *a3*. This mutation transformed an *Alu* element that was never exonized into a constitutive *Alu* exon. It has been estimated that a staggering number of *Alu* elements, about 80,000, located in the introns of protein-coding genes may be, like these examples, a single base change away from becoming constitutive exons. While only a subset of these mutations might produce a disease phenotype, given the number of opportunities, it is likely that

**Figure 8.4. Exonization of *Alu* sequences.** A gene with an *Alu* element inserted into an intron can produce a major transcript (mRNA A) and a minor alternatively spliced transcript (mRNA B) where *Alu* sequences have been “exonized.” (Adapted from Makalowski W. Not junk after all (Perspectives) *Science*, 2003; 300: 1246–47.)



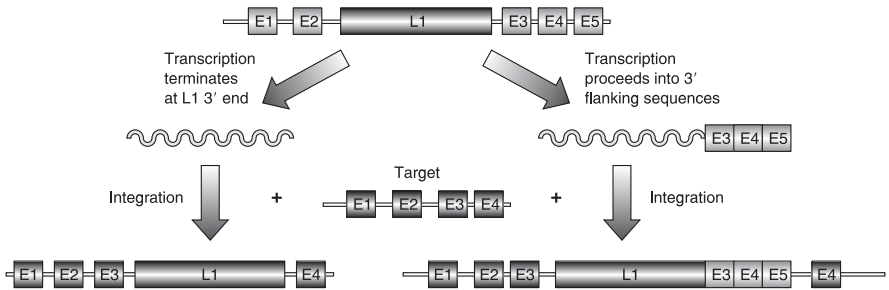
*Alu* exonization will ultimately be shown to be not only a source of diversity in human populations but also a significant causative agent of human genetic diseases.

### ***TE-mediated Exon Shuffling***

Unlike any other genomic component, TEs routinely move from one genomic location to another. Transposition may provide opportunities to rearrange host sequences through the acquisition of host DNA from one locus and its transposition, along with the TE, to another site in the genome. As such, TE activity could provide a mechanism for exon-shuffling, a 25-year-old hypothesis positing that new genes are assembled by stringing together fragments of existing genes.<sup>41</sup> The acquisition of host genes during the transposition of retroviruses is an important part of viral evolution. While LTR retrotransposons, which are structurally and mechanistically related to retroviruses, also have been reported to acquire fragments of host genes, these examples appear to reflect extremely rare events. In contrast, the two case studies below describe mechanisms of transposition where the acquisition of host sequence is a frequent outcome and, as such, may provide long sought after mechanisms for exon shuffling.

### **L1-mediated 3' Transduction**

As illustrated in figure 8.2, human LINE *L1* transposes through an RNA intermediate by a mechanism called target primed reverse transcription (TPRT). Recall that during TPRT, element-encoded transcripts are inserted into nicked sites in the chromosome, where they are copied into double-stranded DNAs (see figure 8.2). Interestingly, *L1* elements contain weak polyadenylation signals near their 3' end. During transcription of active elements, this signal is often bypassed in favor of stronger transcription stop sequences in flanking host sequences (figure 8.5). If these readthrough transcripts are reverse transcribed and reinserted back into the genome, a copy of the 3' host DNA also will be inserted into the new chromosomal



**Figure 8.5.** *Transduction of host sequences by the human L1 retrotransposon.* An L1 element is shown in the second intron of a hypothetical gene whose exons are light gray boxes (E1–E5). Integration of a new element copy via target primed reverse transcription (see figure 8.2) into a target gene (dark exons E1–E4) without 3' transduction of host sequences (on the left) or with 3' transduction of downstream spliced exons (on the right) is shown. In this example, E3–E5 are inserted into the coding region of a hypothetical target gene. (Adapted from Boeke J. D. and Pickeral O. K. Retroshuffling the genomic deck. *Nature*, 1999; 398: 108–109, figure 2.)

locus in a process that has been called L1-mediated 3' transduction.<sup>42</sup> This mechanism can potentially move any non-L1 sequence that happens to be next to an active *L1* element, including parts of regulatory regions, exons, or introns, to another chromosomal locus, where it could alter the regulation or function of a gene at or near the insertion site.

The impact of this mechanism on human genome evolution has begun to be addressed by determining the frequency of 3' transduction events in the complete human genome sequence. Initial studies focused on so-called “young” *L1*s, which represent the most recent insertion events. The rationale for this strategy is that the majority of 3'-transduced DNA is expected to be of no use to the host and will rapidly accumulate mutations that, over time, will make them unrecognizable. Analysis of young *L1*s established that the frequency of 3' transduction was a remarkably high 15–20% of all insertions, which led the authors to estimate that about 1% of human genomic DNA may have arisen in this way.<sup>43,44</sup>

### Pack-MULEs: Exon Shuffling Mediated by a DNA Transposon

DNA transposons are known to be important vectors in the transfer of genes between bacterial cells. As described in chapter 7, TE- and bacteriophage-mediated horizontal gene transfer now is recognized to be a significant factor in bacterial evolution. However, until recently there have been only a few tantalizing reports of eukaryotic DNA transposons containing fragments of host genes. A few of these reports involve *Mutator* elements, a family of DNA transposons that were first isolated from maize mutant alleles. One nonautonomous *Mutator* element, called

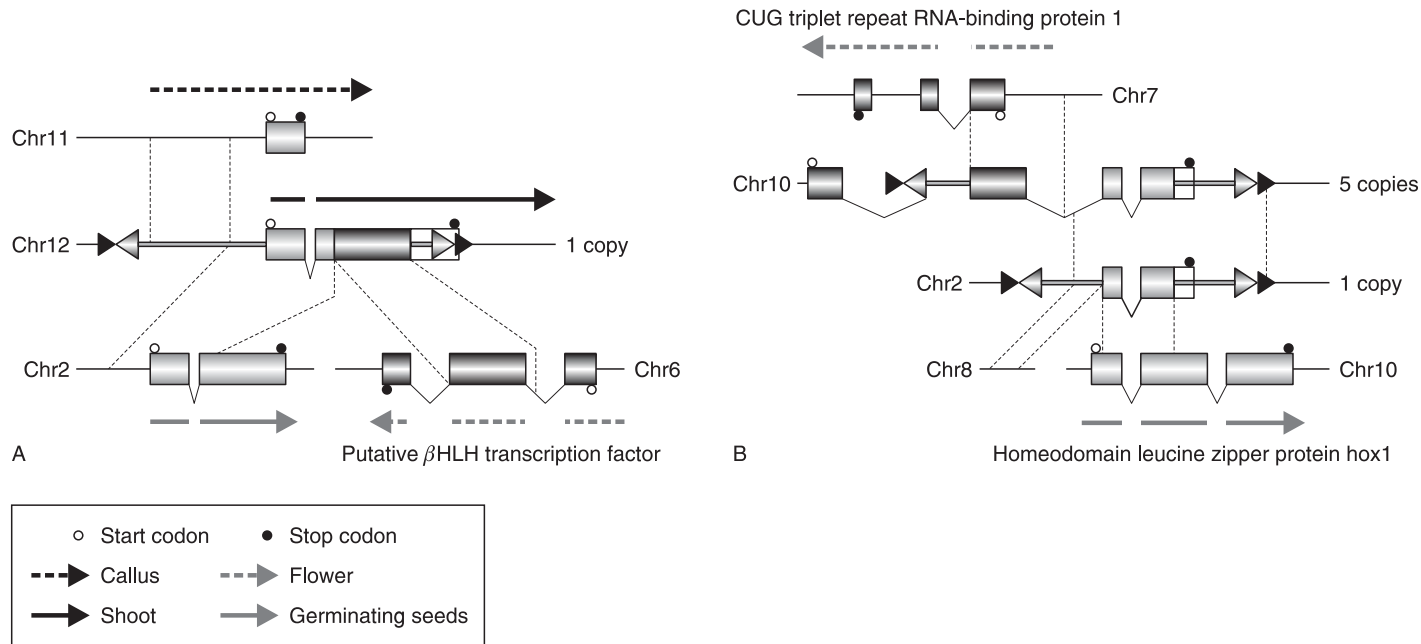
*Mu1*, was subsequently found to harbor a small fragment of a maize gene of unknown function.<sup>45</sup> That *Mu1* elements bearing this gene fragment have attained copy numbers of around 50 indicate that the captured DNA does not (at least in this case) interfere with transposition. The significance of this observation will soon become apparent.

The availability of complete genome sequences from two other plants, *Arabidopsis* and rice, led to the identification of *Mutator*-like transposable elements (known as MULEs) in their genomes and a handful of MULEs were reported to contain gene fragments. The name Pack-MULEs was given to these chimeric elements. MULEs are now recognized as a component of many eukaryotic genomes and are especially prevalent in higher plants. The potential impact of gene capture by MULEs on genome evolution was assessed by using a computational screen to identify Pack-MULEs in the rice genome.<sup>46</sup> Over 3000 Pack-MULEs containing fragments from over 1000 cellular genes were identified. About 5% of the Pack-MULEs are represented in cDNA collections, providing evidence that they are expressed. Comparison of the cellular genes and their Pack-MULE-borne counterparts indicates that fragments of genomic DNA have been captured, rearranged, and amplified over millions of years. About one fifth of the identified Pack-MULEs contain fragments acquired from multiple genomic loci, thus demonstrating their potential to promote exons shuffling through the duplication, rearrangement, and fusion of diverse genomic sequences (figure 8.6).

### Case Studies: Epigenetic Consequences of TE Insertions

As discussed above, McClintock discovered that maize TEs could be reversibly inactivated (epigenetically silenced). Inactivation could be for only a few cell divisions or it could be heritable for several generations.<sup>15</sup> Molecular analyses of members from three different maize TE families (*Ac*, *Spm*, and *Mutator*) later established that active elements differed from inactive elements in that the latter were not transcribed and their ends (containing the transposase gene promoter and the element TIRs) were hypermethylated at the C<sup>5</sup> position of cytosine residues (reviewed in reference 47). Interestingly, like inactive TEs, DNA methylation also can be epigenetically inherited because methylation patterns are copied during replication.

It took over a decade for the mechanistic connections between DNA methylation, element transcription, and the influence of TE activity on flanking gene expression to finally emerge. The early molecular results, along with other related phenomena, sat on the shelf for almost a decade as scientists unraveled the mechanisms underlying epigenetic regulation. This is now a very hot area of research, with hundreds, perhaps thousands, of publications each year. Rather than summarize the current state of knowledge, three case studies are presented in which a combined genetic and genomic approach has begun to open the black box of the epigenetic control of TEs by their host and the consequences for host gene regulation.



**Figure 8.6. Structure and genomic origin of chimeric Pack-MULEs.** (A) A Pack-MULE containing gene fragments from three genomic loci including one intron. (B) Possible step-wise formation of a chimeric Pack-MULE: a Pack-MULE on chromosome 10 with sequences acquired from three loci (on chromosomes 7, 8, and 10) and an apparent intermediate element (on chromosome 2) with the gene fragments from chromosomes 8 and 10. Pack-MULE terminal inverted repeats are shown as gray arrowheads and target site duplications are shown as black arrowheads. Homologous regions are associated with dashed lines. Long dotted arrows indicate sequences matching cDNAs from the designated tissues. Exons are depicted as dark gray boxes and the introns as the lines connecting exons. Light gray boxes represent exons (or part of an exon) where the origin of the sequence is not clear. The gene name is given for putative genes and hypothetical proteins; all other genes encode unknown proteins. (Adapted from reference 46, figure 3.)

### ***RNAi Regulates Transposition in the C. elegans Germline***

The *C. elegans*<sup>48</sup> genome harbors several families of class 2 TEs belonging to the *Tc1/mariner* superfamily. Distinct element families in *C. elegans* are designated Tc1, Tc2, Tc3, and so on, and Tc1, which has around 31 copies, is the most abundant family in the genome. In most isolates of *C. elegans*, transposition of Tc1 and the other Tc families is repressed in the germline but not in the soma. However, in the Bergerac strain, transposition of Tc elements occurs both in the soma and the germline and Tc1 insertions are the major cause of spontaneous mutation. That transposition could occur in the germline led the laboratory of Ron Plasterk to undertake both forward and reverse genetic screens to identify the gene(s) responsible for repressing germline transposition. While many loci identified by these screens are of unknown function, and others are only indirectly involved in repression, the screen identified a group of gene products that are now known to be part of the epigenetic regulatory process called RNA interference (RNAi).

In eukaryotes from yeast to plants to animals, RNAi has been shown to be a mechanism of posttranscriptional gene silencing (PTGS) that is triggered by double-strand RNA (dsRNA), which can be a by-product of genome perturbation. For example, insertion of a TE or foreign DNA (viruses or transgenes) next to a host gene may promote transcription into the gene on the antisense strand. The dsRNA formed when sense and antisense transcripts anneal is a substrate for an RNase III-like enzyme (called dicer or DCR-1) that cleaves dsRNAs into short interfering RNAs (siRNAs) of about 21–24 nt. Interaction of siRNAs with the RNA-induced silencing complex (RISC) leads to the degradation of RNAs that are specified by the siRNAs; in this example, the transcript of the host gene next to the TE insertion site. Recall from a previous section that antisense transcription promoted by a TE at an inversion breakpoint in *D. buzzatii* was implicated in the downregulation of a host gene adjacent to the breakpoint. We will return to these siRNAs in the next case study (additional discussion of siRNAs can be found in chapters 13 and 14).

The repression of Tc1 transposition in the germline implies that this process is very efficient and that dsRNA derived from Tc elements must be readily available to maintain repression. But what is this dsRNA and how does it originate? Unlike most class 2 elements, the Tc1 transposase gene does not appear to have its own promoter. Instead, transposase expression may rely on transcripts initiated in flanking host sequences that read through the entire element. The dsRNA trigger for RNAi is thought to arise from intramolecular pairing between their TIR sequences. Interestingly, while the internal sequences of other Tc families in *C. elegans* are distinct, their TIRs are very similar. Thus, siRNAs produced by the synthesis and processing of readthrough transcripts from a single Tc element may be sufficient to silence all Tc elements in the genome.

### ***TEs and Heterochromatin Formation in Arabidopsis thaliana***

Recall that inactive maize TEs are hypermethylated as are the TEs of other organisms, including other plants and mammals. However, not all organisms methylate

their DNA; among these exceptions are the model organisms *C. elegans*, *Drosophila*, and yeast. If *C. elegans* can repress its TEs without methylating them, why do most other organisms methylate their TEs? While there is still no definitive answer to this question, the search for an answer in plants has helped reveal connections between methylation, histone modification, and RNAi.

Because silencing of both TE and transgenes in plants had been associated with DNA methylation, methylation became the focus of many genetic screens in the model plant *Arabidopsis thaliana*. In one screen, mutagenized plants assayed for reduced methylation of repeat sequences (each plant was tested by Southern blot!) led to the identification of a gene, called *decrease in DNA methylation 1 (DDM1)*.<sup>49</sup> In addition to reduced DNA methylation, the mutant *DDM1* background gave rise to both developmental and unstable mutations that were not linked to the *ddm1* locus. One mutant, called *crab*, was found to be due to the insertion of a transposable element that was normally transcriptionally silent and transpositionally inactive in *A. thaliana*. While this study demonstrated that an *A. thaliana* transposon that must have been epigenetically silenced was released in the *DDM1* mutant background, the association between DNA methylation and TE inactivity remained unclear when it was determined that *DDM1* encoded a putative chromatin remodeling protein, not, as might be expected from its mutant phenotype, a methyltransferase. Recall that higher organisms have transcriptionally active chromatin and transcriptionally inactive or silent chromatin. The latter can often be observed cytologically as regions of heterochromatin, located around centromeres, telomeres, and, in some organisms, interstitially (where it is called “knobs”). While active and inactive chromatin often are defined biochemically by measuring acetylation and deacetylation of histones, respectively, actively transcribed chromatin can be identified by the methylation of lysine 4 in the amino terminus of histone 3 (H3mK4) while inactive chromatin is marked by methylation at lysine 9 (H3mK9). Recent studies indicate that mutations in *DDM1* decrease the methylation of H3mK9 in all loci tested.

To review up to this point, the research had shown that the TEs of *C. elegans* are not methylated but are rendered inactive through the production of siRNAs that targets TE transcripts for degradation. In *A. thaliana* (and in other plants and mammals), TEs are methylated and their chromatin contains H3mK9. In *A. thaliana* with mutant *DDM1*, TEs are both activated (H3mK9 decreases) and demethylated (H3mK4 increases). These results suggested that the product of the *DDM1* gene and a DNA methyltransferase are in a complex that can recognize TEs and silence them by methylating their DNA and associated histones. However one big question remained; what does this complex recognize when it distinguishes between TEs and genes?

This question was addressed by comparing a duplicated region in the *A. thaliana* genome where one duplicate is euchromatic and contains 33 genes and the other is heterochromatic (a chromosomal knob) and contains 8 of these 33 genes and also 73 TEs that inserted after duplication. In an experimental tour de force, the lab of Rob Martienssen compared the two regions by microarray analysis (in 1 kb segments) with respect to transcription, and histone and DNA methylation in both wild-type and mutant *DDM1* backgrounds.<sup>50</sup> They concluded that under the control of *DDM1*, TEs and repeats were responsible for heterochromatin formation because, as they put it: “In a *ddm1* background, TEs adopted gene-like chromatin

properties and the majority were expressed.” Therefore, DDM1 distinguishes TEs and related repeats from host genes. Given that many TEs, like host genes, encode proteins and both have similar GC content, the basis for discrimination was suspected to be sequence-based, which brings us back full circle to RNAi. In support of a role for an RNAi based mechanism for TE recognition, they detected siRNAs from the TEs that were inactivated in wild-type and became activated in *DDM1* mutants. Thus, as in *C. elegans*, *A. thaliana* TEs appear to be transcribed to generate siRNAs, which guides DNA and histone methylation a process that involves DDM1 and a DNA methyltransferase, leading to heterochromatinization and inactivation of the TEs.

Both in this instance and in others, TEs seem to be responsible in large part for the regions of heterochromatin in the genomes of most eukaryotes. For this reason, it has been suggested that TEs may be essential to centromere and telomere function as they too are sites of heterochromatin. This has been discussed in recent reviews<sup>51-53</sup> and will not be reviewed here. In keeping with the theme of this chapter, the last case study demonstrates that a mechanism to maintain centromeres and silence TEs can turn on the host to inadvertently silence its own genes.

### ***LTR Retrotransposons Can Silence Adjacent Host Genes: from Polyploidy to Cancer***

Organisms use a variety of interrelated epigenetic mechanisms, such as RNAi and DNA methylation, to inactivate the TEs in their genome. This strategy is absolutely essential in the very large plant and animal genomes where the majority of TEs are class 1 retroelements (LTR and non-LTR retrotransposons, figure 8.1) that require transcription for mobility. However, rendering TEs immobile is only one reason to repress their transcription. Equally important is to prevent so-called “readout” transcription into host genes. As discussed earlier in this chapter, LTR retrotransposons are the most abundant component of many plant genomes, accounting for an astounding 70% of the maize genome, for example. In addition, about 8% of the human genome is made up of a group of elements called human endogenous retroviruses (HERVs), which are derived from ancient infections of exogenous retroviruses. Like LTR retrotransposons, HERVs are flanked by LTRs (see figure 8.1 for the structure of LTR retrotransposons). Recall that a promoter in the 5′LTR initiates transcription into the element, producing an RNA that is reverse transcribed into a cDNA that can integrate elsewhere in the genome. However, because the two LTRs are identical at the time of insertion, the same promoter in the 3′LTR (or in a solo LTR) may initiate readout transcripts into flanking host sequences. The danger to the host is that their genes can be silenced if they are adjacent to LTRs that promote antisense readout transcripts (like the one promoted by the Kepler element into a flanking *D. buzzatii* gene in figure 8.3).

For this reason, the host must be able to protect its genes from the hundreds of thousands or millions of LTR-borne promoters that can be scattered throughout its genome. Fortunately, the host defends itself against both mobility and readout transcription by methylating the 5′ and 3′ LTRs and rendering their promoters

inactive. The retrotransposons in the genomes of plants and animals are said to be hypermethylated relative to the genes. However, more and more chinks in the host armor are being revealed as new studies report examples of what McClintock called “genome shock,”<sup>9</sup> where host DNA becomes hypomethylated and LTRs are transcriptionally reactivated.

Reactivation of LTRs can be potentially good or bad for the host. One of the major shocks to the integrity of a genome occurs during the formation of polyploids when the genome doubles. While polyploids were once thought to be rare, whole-genome sequencing has revealed evidence for one or multiple polyploidization events in the history of all plants and many animals (reviewed in reference 54). To understand the earliest events in polyploid formation, researchers have turned to domesticated plants such as wheat and cotton where synthetic polyploids can be created in the field and the genome-wide impact on TEs can be analyzed in the laboratory. In one study, genome-wide analysis of newly formed wheat polyploids revealed that widespread changes in DNA methylation in the LTRs of the *Wis 2-1A* retrotransposon activated readout transcription into adjacent wheat genes and, in some cases, led to gene silencing.<sup>55</sup> Because newly formed polyploids have a duplicate set of genes, gene silencing is probably not a serious threat and may in fact facilitate the successful merger of two genomes into one.

In contrast, the reactivation of HERV elements in humans may exacerbate an already bad situation. Hypomethylation of DNA is a consequence of many cancers and other human diseases (reviewed in reference 56). Recent studies have demonstrated that the methylation status of several HERV LTRs is altered and that some become transcriptionally reactivated in some human tumors. While the consequences of readout transcription on the cancerous state is currently under investigation, the wealth of human genomic resources promises to make this a lively area of future research.

## Conclusions

The thread that connects most of the case studies presented in this chapter is that TE-mediated genomic diversification is a by-product of the arms race between TEs and their hosts and that the novelty generated by this arms race has facilitated the evolutionary success of the host. When seen in this light, the term “coevolution” may be a more accurate term than “arms race” in describing the interaction between TEs and their hosts.

Scientists rely on negative controls to validate their experimental results. To bolster the argument that TEs are a significant component of the evolutionary success of eukaryotes, a comparison could be made between eukaryotes with and without TEs. Unfortunately, such a comparison is not possible at this time because all characterized sexually reproducing eukaryotes have TEs. But nature may have provided the next best thing, a eukaryote that has evolved a defense mechanism that appears to be completely successful in preventing the amplification of TEs. The eukaryote is the model organism *Neurospora crassa* and the mechanism of defense is called

repeat-induced point mutation (RIP). RIP first detects duplications and then mutates them by changing up to 30% of their G-C base pairs into A-T pairs (reviewed in reference 57). In fact, not a single intact TE was detected in the draft sequence.<sup>58</sup> But, the apparent consequence is that *N. crassa* has few highly similar duplicate genes. Of around 10,000 predicted protein-coding genes, there are only six pairs (12 genes) with greater than 80% nucleotide identity. Thus the route to evolution of new function through duplication and gradual divergence of genes is blocked by RIP. In addition, other TE-mediated mechanisms illustrated in the case studies, such as allele diversification via TE insertion and TE-mediated exon-shuffling, presumably cannot occur either. However, *N. crassa* does exist and has managed to survive and flourish for a very long time; clearly it has evolved mechanisms to evolve without TEs. What those mechanisms are will surely be the subject of future studies.

In addition, further studies will almost certainly reveal new ways that TEs diversify genomes. The availability of increasing amounts of eukaryotic genome sequence has permitted our first glimpse of the relationship between host genes and TEs. What is most striking about these initial results is that each genome has a different story to tell. As discussed above, there is tremendous variation in the TE content of the characterized genomes with respect to overall TE composition and level of activity. This variation reflects the distinct evolutionary trajectory experienced by each species. In this regard, we have only begun to understand how the coevolution of host genes and TEs impacts the mode and tempo of evolution. This situation is reminiscent of the statement at the end of each episode of *Naked City*, a TV show that I watched as a child: "There are eight million stories in the Naked City. This has been one of them."

## Chapter 8

1. Labrador, M. and Corces, V. Interactions between transposable elements and the host genome, in N. L. Craig, Craigie, R., Gellert, M., Lambowitz, A.M. (eds.), *Mobile DNA II*, pp. 1008–23, ASM Press, Washington DC (2002).
2. Feschotte, C., Jiang, N., and Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**: 329–41 (2002).
3. Kazazian Jr., H. H. Mobile elements: drivers of genome evolution. *Science* **303**: 1626–32 (2004).
4. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49–63 (2002).
5. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**: 344–55 (1950).
6. Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M. *Mobile DNA II*, ASM Press, Washington DC (2002).
7. Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**: 595–605 (1993).
8. McClintock, B. Chromosome organization and genic expression. *Cold Spg Hbr Symp. Quant. Biol.* **16**: 13–47 (1951).
9. McClintock, B. The significances of responses of the genome to challenge. *Science* **226**: 792–801 (1984).
10. Bowen, N. J. and Jordan, I. K. Transposable elements and the evolution of eukaryotic complexity. *Curr. Issues Mol. Biol.* **4**: 65–76 (2002).
11. Doolittle, W. F. and Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–3 (1980).
12. Orgel, L. E. and Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**: 604–7 (1980).
13. McClintock, B. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26**: 234–82 (1941).
14. Gray, Y. H. It takes two transposons to tango. *Trends Genet.* **16**: 461–8 (2000).
15. McClintock, B. The *Suppressor-mutator* system of control of gene action in maize. *Carnegie Institution of Washington Yearbook* **57**: 415–29 (1958).
16. Deininger, P. L. and Batzer, M. A. Mammalian retroelements. *Gen. Res.* **12**: 1455–65 (2002).
17. Kumar, A. and Bennetzen, J. L. Plant retrotransposons. *Ann. Rev. Genet.* **33**: 479–532 (1999).
18. Caceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**: 415–18 (1999).
19. Puig, M., Caceres, M., and Ruiz, A. Silencing of a gene adjacent to the breakpoint of a *Drosophila* inversion by a transposon-induced antisense RNA. *Proc. Natl. Acad. Sci. U. S. A.* **101**: 9013–18 (2004).
20. Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., and Botstein, D. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **99**: 16144–9 (2002).

21. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. Recent segmental duplication in the human genome. *Science* **297**: 1003–7 (2002).
22. Babcock, M., Pavlicek, A., Spiteri, E., Kashork, D. D., Ioshikhes, I., Shaffer, L. G., Jurka, J., and Morrow, B. E. Shuffling of genes within low-copy repeats on 22q11 (*LCR22*) by Alu-mediated recombination events during evolution. *Gen. Res.* **13**: 2519–32 (2003).
23. Deininger, P. L. and Batzer, M. A. *Alu* repeats and human genetic disease. *Mol. Genet. Metab.* **67**: 183–93 (1999).
24. Bailey, J. A., Liu, G., and Eichler, E. E. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**: 823–34 (2003).
25. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E.J., Scherer, S., et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521 (2004).
26. McDonald, J. F. Transposable elements: possible catalysts of organismic evolution. *Trends Ecol. Evol.* **10**: 123–6 (1995).
27. Weil, C. F. and Wessler, S. R. The effects of plant transposable element insertion on transcription initiation and RNA processing. *Ann. Rev. Plant Phys. Mol. Biol.* **41**: 527–52 (1990).
28. Kidwell, M. G. and Lisch, D. Transposable elements as sources of genomic variation, in N. L. Craig, Craigie, R., Gellert, M., Lambowitz, A. M. (eds.), *Mobile DNA II*, pp. 59–90, ASM Press, Washington DC (2002).
29. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., and Scherer, S. Transposable elements in mammals promotes regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**(10): 530–6 (2003).
30. Feschotte, C., Zhang, X., and Wessler, S. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons, in N. L. Craig, Craigie, R., Gellert, M., Lambowitz, A. M. (eds.), *Mobile DNA II*, pp. 1147–58, ASM Press, Washington DC (2002).
31. Jiang, N., Bao, Z., Zhang, X., McCouch, S. R., Eddy, S. R., and Wessler, S. R. An active DNA transposon in rice. *Nature* **421**: 163–7 (2003).
32. Kikuchi, K., Terauchi, K., Wada, M., Hirano, H. Y. The plant MITE *mPing* is mobilized in anther culture. *Nature* **421**: 167–70 (2003).
33. Nakazaki, T., Okumoto, Y., Horibata, A., Yamahira, S., Teraishi, M., Nishida, H., Inoue, H., Tanisaka, T. Mobilization of a transposon in the rice genome. *Nature* **421**: 170–2 (2003).
34. Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100 (2002).
35. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92 (2002).
36. Hans, J. S., Szak, S. T., and Boeke, J.D. Transcriptional disruption by the *LI* retrotransposon and implications for mammalian transcriptomes. *Nature* **429**: 268–74 (2004).
37. Perepelitsa-Belancio, B. and Deininger, P. L. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genet.* **35**: 363–6 (2003).
38. Kreauling, J. and Graveley, B. R. The origin and implications of Aluternative splicing. *Trends Genet.* **20**: 1–4 (2004).
39. Nekrutenko, A. and Li, W. H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* **17**: 619–21 (2001).

40. Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. The birth of an alternatively spliced exon: 3' splice-site selection in *Alu* exons. *Science* **300**: 1288–91 (2003).
41. Gilbert, W. Why genes in pieces? *Nature* **271**: 501 (1978).
42. Moran, J. V., DeBerardinis, R. J., and Kazazian Jr., H. H. Exon shuffling by *LI* retrotransposition. *Science* **283**: 1530–4 (1999).
43. Goodier, J. L., Ostertag, E. M., and Kazazian Jr., H. H. Transduction of 3'-flanking sequences is common in *LI* retrotransposition. *Hum. Mol. Genet.* **9**: 653–7 (2000).
44. Pickeral, O. K., Makalowski, W., Boguski, M. S., and Boeke, J. D. Frequent human genomic DNA transduction driven by *LINE-1* retrotransposition. *Gen. Res.* **10**: 411–5 (2000).
45. Talbert, L. E. and Chandler, V. L. Characterization of a highly conserved sequence related to *mutator* transposable elements in maize. *Mol. Biol. Evol.* **5**: 519–29 (1988).
46. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., and Wessler, S. R. Pack-Mule transposable elements mediate gene evolution in plants. *Nature* **431**: 569–73 (2004).
47. Fedoroff, N. V. and Chandler, V. Inactivation of maize transposable elements, in J. Paskowski (ed.), *Homologous Recombination and Gene Silencing in Plants*, pp. 349–85, Kluwer Academic Publishers (1994).
48. Vastenhouw, N. L. and Plasterk, R. H. RNAi protects *Caenorhabditis elegans* germline transcription against transposition. *Trends Genet.* **20**: 314–19 (2004).
49. Vongs, A., Kakutani, T., Martienssen, R. A., and Richards, E. J., *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**: 1926–8 (1993).
50. Lippman, Z. G., Black, A.-V., Vaughn, M. W., Dedhia, N., McCombie, R. W. et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–6 (2004).
51. Jeddelloh, J. A., Stokes, T. L., and Richards, E. J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genet.* **22**: 94–7 (1999).
52. Volpe, T., Schramke, V., Hamilton, G. L., White, S. A., Teng, G., Martienssen, R. A., and Allshire, R. C. RNA interference is required for normal centromere function in fission yeast. *Chrom. Res.* **11**: 137–46 (2003).
53. Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I., and Martienssen, R. A. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**: 1833–7 (2002).
54. Adams, K. L. and Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**: 135–41 (2005).
55. Kashkush, K., Feldman, M., and Levy, A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genet.* **33**: 102–6 (2003).
56. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**: 597–610 (2005).
57. Galagan, J. and Selker, E. RIP: the evolutionary cost of genome defense. *Trends Genet.* **20**: 417–23 (2004).
58. Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D. et al. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**: 859–68 (2003).