CHAPTER 3

Transposable Elements

MARGARET G. KIDWELL

Transposable elements (TEs) are discrete DNA sequences that move from one location to another within the genome. They are found in nearly all species that have been studied and constitute a large fraction of some genomes, including that of *Homo sapiens*. TEs are potent broad-spectrum mutator elements that are responsible for generating variation in the host genome and have a role as key players in the ecology of the genome. This chapter presents an overview that includes coverage of TE structures, regulation, distribution, and dynamics. A wealth of examples provides many illustrations of the diversity of TE types and behaviors as well as the rich variety of interactions between TEs and their host genomes. It is evident from this that knowledge of these elements is essential for a full understanding of genome evolution.

A BRIEF HISTORY OF THE STUDY OF TRANSPOSABLE ELEMENTS

Transposable elements comprise a group of distinct DNA segments with the capacity to move, or transpose, between many nonhomologous (unrelated) sites

The Evolution of the Genome, edited by TR Gregory

in the genome. The properties of these elements provide them with the capacity to mutate the DNA of the host organisms in which they reside in many different ways. Their biology is still the subject of active investigation, although their general existence has now been known for more than half a century. The following sections provide a brief history of the discovery and early study of these important genomic components.

The Discovery of Transposable Elements

The pioneering studies of Barbara McClintock in the mid-20th century led to her discovery of TEs in maize (*Zea mays*). In 1949 she showed that the change of unstable recessive alleles to the dominant form in this plant was due to the movement of a short segment of a chromosome. She first named these segments "controlling elements" because of her emphasis on their role in gene expression. Controlling elements were later described by Fincham and Sastry (1974) as

> TEs of apparently sporadic occurrence, which make themselves visible through their abnormal control of the activities of standard genes. Most simply, a controlling element may inhibit activity of a gene through becoming integrated, in, or close to, that gene. From time to time, either in germinal or somatic tissue, it may be excised from this site, and, as a result, the activity of the gene is often more or less restored, while the element may become reintegrated elsewhere in the genome where it may affect the activity of another gene.

Maize is a highly complex eukaryote whose genome is not easily amenable to analysis. It is therefore rather remarkable that this, rather than a simpler organism, was the species in which the discovery of transposition was made. In fact, it was not until a decade after McClintock's initial work that the insertion sequences of the bacterium *Escherichia coli* were discovered and immediate analogies were drawn and acknowledged between the maize and *E. coli* elements. The ease of study of bacteria at the molecular level allowed much faster progress to be made in the study of bacterial TEs than in maize and other eukaryotic species.

Today the familiar concept of transposition (i.e., the ability of mobile genetic elements to replicate themselves, or be replicated, and move around the genome) is often taken for granted. However, as recently as three decades ago, it was difficult for many traditional geneticists to change their view of genes and chromosomes as static entities, often likened to "beads on a string." The discovery of transposition was important in providing a new concept of genomes as fluid and dynamic entities. It represented a true paradigm shift whose importance in the history of science has probably not yet had time to be fully appreciated.

The explosion of knowledge about TEs that has occurred during the last 30 years is reflected in the size and scope of several sequential collections of papers describing the "state of the art" of TE biology. The first two volumes consisted of

the publication of the proceedings of meetings held at the Cold Spring Harbor Laboratory in 1976 (Bukhari *et al.,* 1977) and 1980 (Cold Spring Harbor Symposia on Quantitative Biology, vol. 45: "Movable Genetic Elements," 1981). Three later volumes are collections of invited papers updating research on TEs in an increasingly large range of species (Shapiro, 1983; Berg and Howe, 1989; Craig *et al.,* 2002).

Early TE Studies in Bacteria

Probably because of their small genome size, bacteria were some of the earliest organisms to be intensively investigated with regard to TEs. Two general types of elements were originally recognized. The *IS* (simple insertion) sequences were first defined as elements generally shorter than 2.5 kilobases (kb) and lacking genes unrelated to insertion function (Campbell *et al.,* 1977). These were distinguished from the second type, the *Tn* elements (referred to at that time as transposons), which were more complex in structure and generally larger than 2.5 kb. It was recognized that *Tn* elements often contain *IS* elements and behave like *IS* elements, but carry additional genes unrelated to insertion function.

Bacterial *IS* sequences were discovered during early investigations of the molecular genetics of gene expression in *E. coli* and bacteriophages. They were first isolated as highly polar, somewhat unstable mutations in the galactose and lactose operons of *E. coli,* and in the early genes of some bacteriophages. Many of these mutations were shown to be insertions of the same few segments of DNA in different positions and orientations. When it was realized that these segments of DNA were natural residents of the *E. coli* genome, their similarity to the TEs in maize discovered by McClintock (1952) was recognized.

The more complex *Tn* elements are members of a class of related transposons that are medically important because they confer antibiotic resistance to many pathogenic bacteria. For example, members of the *Tn3* family are usually found on plasmids from antibiotic-resistant bacteria, but they may transpose to bacteriophages and to the chromosome of *E. coli* and many other bacteria. Infectious antibiotic resistance (encoded on plasmids) usually results from production of an enzyme that inactivates the antibiotic. Space constraints prevent adequate coverage of the huge body of information on bacterial TEs in the present chapter, but fortunately several in depth reviews are available elsewhere (e.g., references in Berg and Howe, 1989; Craig *et al.,* 2002).

Early TE Studies in Fungi

The first fungal TEs to be studied were the *Ty* elements in the bakers' yeast *Saccharomyces cerevisiae* (Cameron *et al.,* 1979). Demonstration of the mobility

of *Ty* elements in the yeast genome came from the analysis of mutations at a number of genetic loci and the subsequent demonstration of insertions in those loci in some yeast strains, but not in others. Although interest in using TEs for gene tagging resulted in the cloning and analysis of an element called *Tad* in the bread mold *Neurospora crassa* (Kinsey and Helber, 1989), very little information about the TE complement of other fungal genomes was available until the advent of inexpensive DNA sequencing during the last decade. With the recent sequencing of the *N. crassa* genome (Galagan *et al.,* 2003), an interesting property has come to light concerning the TE complement of this species. Specifically, no active TEs were found, only inactive remnants. This observation is in fact consistent with the earlier discovery of a mechanism known as repeat-induced-point mutation (RIP), by which TEs and other relatively large repetitive sequences are inactivated by mutation (see later section for more details).

Early TE Studies in Plants

Following Barbara McClintock's original discovery, maize dominated the study of plant TEs, even though other suitable models were apparent. For example, variegation in flower color in the snapdragon *Antirrhinum majus* was described by Darwin and others in the 19th century. Although the genetic instability in *A. majus* was known to share many similarities with that in maize, it was not until the 1980s, when molecular techniques were becoming widely available, that the movement of TEs was identified as the cause (Saedler *et al.,* 1984). In particular, the *nivea (niv)* locus was cloned from *A. majus* genomic DNA, which allowed the isolation of an unstable allele that conferred a variegated flower color phenotype. This allele carried a 15 kb TE named *Tam* that turned out to belong to the same family as the *hobo* element discovered in *Drosophila melanogaster* and the *Ac* element in maize. This family was subsequently named the *hAT* family after these first three elements to be discovered.

Early TE Studies in Animals

Prior to 1980, *Drosophila* predominated in the study of animal TEs. More than 20 years ago, a number of different TE families had been described from *D. melanogaster* (Finnegan *et al.,* 1978). Prominent among these were seven different families of *copia*-like elements, a family of *foldback (FB)* elements (Rubin, 1983), and the *P* element family (Bingham *et al.,* 1982; Rubin *et al.,* 1982). The *copia*-like elements were found to be strikingly similar in structure to the *Ty* elements in yeast (Fink *et al.,* 1981) and to the integrated proviruses of RNA tumor viruses. It was not until the early 1980s that the identification of TEs in other

animal species such as humans (Singer, 1982) and the nematode worm (*Caenorhabditis elegans*) (Emmons *et al.,* 1983) was first made.

A number of factors contributed to this emphasis on species of *Drosophila*. Notably, the model organism *D. melanogaster* had been among the most intensively studied species at the genetic level throughout the 1970s and provided a number of clues to the activity of TEs. Among these were various unstable mutations that revert to a wild-type phenotype at unusually high rates, and generate deletions and chromosomal rearrangements that had one endpoint at the site of the mutation (reviewed by Green, 1977). Moreover, the giant polytene chromosomes of *D. melanogaster* salivary glands provide the added advantage that the genomic location of any isolated piece of DNA can be determined by the method of *in situ* hybridization. Thus TE insertion sites can be located even though they may not produce any visible phenotypic effect.

The molecular characterization of the *P* element family in 1982 was preceded by more than a decade of work on the phenotypic manifestations and phenomenology of this element family in natural populations of *D. melanogaster*. Following the discovery of an exception to the general rule of absence of recombination in *D. melanogaster* males (Hiraizumi, 1971) and other related observations (e.g., Kidwell and Kidwell, 1975), the phenomenon of hybrid dysgenesis was first described (Kidwell *et al.,* 1977). When male flies from natural populations of *D. melanogaster* were crossed with female flies from laboratory strains of the same species, the F_1 progeny produced high frequencies of mutation, chromosomal aberrations, and other genetic abnormalities. It was later discovered that these abnormalities were caused by the activation of *P* elements in dysgenic crosses between males from strains that carried *P* elements (*P* strains) and females from strains in which *P* strains were absent (*M* strains). The *P* element was not present in natural *D. melanogaster* populations until after the middle of the 20th century, and most common laboratory stocks were derived from these *M* strains. Following horizontal transfer from another species, the *P* element completely invaded natural populations of *D. melanogaster* during the last half-century.

A Recent Explosion of New Information from DNA Sequencing

Studies of the basic biology of TEs have received a major impetus with the development of inexpensive methods of DNA sequencing. This has led to an explosion of comparative data on a multitude of TE families identified in an ever-increasing range of species, including many nonmodel organisms. The quickening pace of genome sequencing, as well as the advent of sophisticated methods of computational analysis of sequenced genomes, is further fueling this trend. However, it is often not fully appreciated that because of the repetitive nature of these

sequences, they are frequently among the last to receive attention and the quality of data available is often relatively poor in comparison to that of gene-rich regions of the genome. For example, at the time of this writing—approximately three years after the publication of the first draft of the sequence of the *D. melanogaster* genome—about 80% of the centromeric heterochromatin had yet to be sequenced (Kapitonov and Jurka, 2003a).

WHO CARES ABOUT TRANSPOSABLE ELEMENTS?

It is informative to consider the range of perspectives represented by those researchers who are interested, either positively or negatively, in TEs. Four main groups of investigators have been identified (Holmes, 2002): (1) Geneticists have been particularly interested in the application of TEs as tools useful in research and phylogenetic analysis during the last two decades. These applications include the use of TEs as vectors for interspecies transformation (i.e., the transfer of naked DNA from one species into another) and as efficient markers for gene tagging and phylogenetic studies. Details of these applications are discussed more fully in a later section; (2) Genome annotators are interested in TEs and other repetitive DNA but in a more negative way. For them, these repetitive sequences often represent a major nuisance and increase the cost of genome sequencing. Again, this raises a continuing problem for those interested in the repetitive portions of the genome because these regions tend to be the last to be completely sequenced; (3) Structural molecular biologists have an interest in TEs because of their homologies with virus replication machinery, transcription factors, and binding proteins; (4) TEs have the potential to be of particular interest to evolutionary biologists because of the interactions with their hosts. For the most part, the focus of this chapter emphasizes the interests and approaches of this latter group.

HOW ARE TEs CLASSIFIED?

One of the most fundamental aspects of biological study is classification, and in this regard TEs are no exception. TEs are classified in two ways: according to their degree of functional self-sufficiency and according to their mechanism of transposition.

Autonomous and Nonautonomous Elements

TEs are described as being autonomous or nonautonomous based on whether or not they encode their own genes for transposition. Autonomous elements, such as long interspersed nuclear elements (LINEs) in humans, are defined as those

elements that essentially encode all the sequences that enable them to move. Nonautonomous elements are structurally deficient in one respect or another, and depend to at least some extent on other elements in the genome in order to move. Many nonautonomous elements are derived from autonomous elements through deletions of part of their structures that leave critical cis-acting sequences (i.e., ones on the same molecule) in place. Alternatively, some nonautonomous elements, such as short interspersed nuclear elements (SINEs) in humans, have evolved independently of, but in parallel with, autonomous counterparts. Sometimes the distinction between autonomous and nonautonomous elements is not clear, such as for the human endogenous retroviruses (HERVs). Also, some autonomous elements are not completely independent in that they require the cellular machinery of their hosts in order to transpose.

Classification Based on Mode of Transposition

Most TEs can be assigned to one of two broad classes according to their mechanism of transposition (Finnegan, 1989). Class I elements are generally referred to as "retroelements," and Class II elements are often called "DNA transposons" (or just "transposons"). Figure 3.1 shows the basic structural features of the two TE classes along with examples, and Table 3.1 provides some examples of TE families and host organisms.

Class I Transposable Elements

Class I elements are members of a large group of so-called retroelements that utilize reverse transcription of an RNA template to make additional copies of themselves, a process known as retrotransposition (Fig. 3.2). The original element is maintained *in situ,* where it is transcribed. Its RNA transcript is then reverse transcribed into DNA that integrates into a new location in the genome. This class includes the "retrotransposons" that are characterized by flanking long terminal repeats (LTRs) and the "retroposons" (also called "non-LTR retrotransposons") that lack terminal repeats, as well as the SINEs (see Fig. 3.1 for a comparison of the main structural features of these three subclasses). Class I mobile elements also include the endogenous retroviruses that are closely related to the retrotransposons and the mobile introns. These various types of elements are described in more detail in the following sections.

LTR Retrotransposons and Endogenous Retroviruses

The LTR retrotransposons were the first retroelements to be discovered in eukaryotes and are similar in structure and coding capacity to retroviruses.

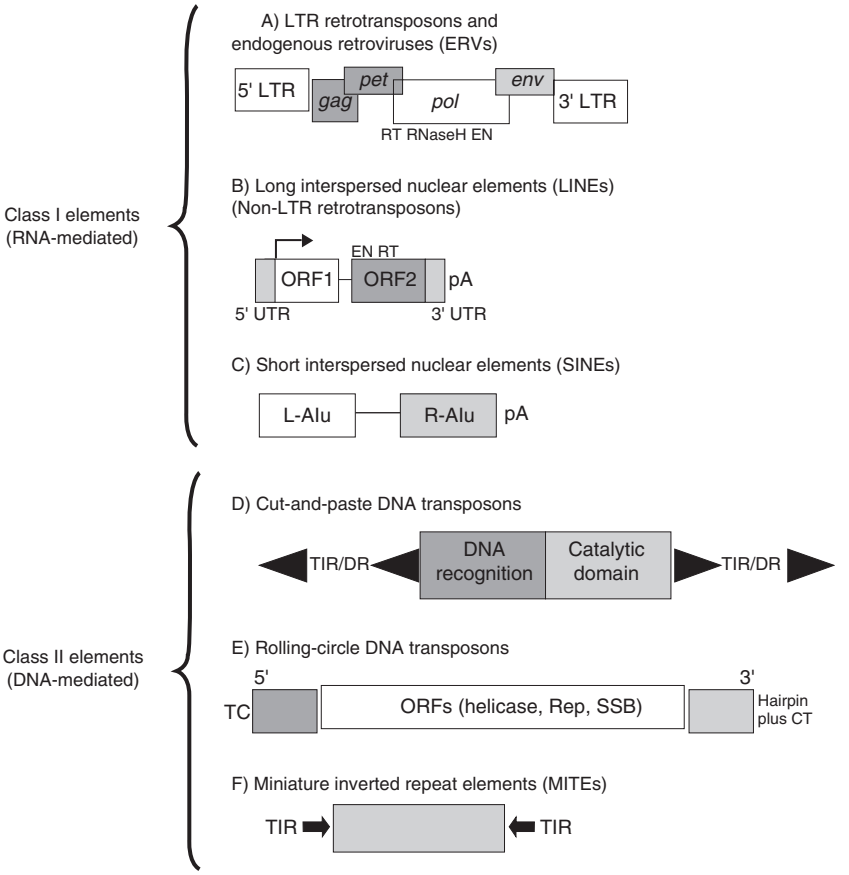

FIGURE 3.1 Generalized structures of the main types of transposable elements. Class I elements are those that transpose via an RNA intermediate, whereas the transposition of Class II elements is DNA-mediated. **(A)** Long terminal repeat (LTR) retrotransposons and endogenous retroviruses (ERVs) consist of partly overlapping coding regions for group-specific antigen (*gag*), protease (*prt*), and polymerase (*pol*) genes. These elements are flanked on both ends by LTRs with promoter capability. The *pol* gene contains domains for reverse transcriptase (RT), RNaseH, and integrase (IN) or endonuclease (EN). The distinction between LTR retrotransposons and ERVs is somewhat blurred by the fact that the envelope (*env*) genes typical of retroviruses have been both acquired by some LTR elements and rendered nonfunctional in some ERVs. This illustration is of the human endogenous retrovirus (HERV), which maintains an intact *env* gene. **(B)** Long interspersed nuclear elements (LINEs), or non-LTR retrotransposons, consist of a 5′ untranslated region (UTR) that has promoter activity, two open reading frames (ORF1 and ORF2) separated by an intergenic spacer, and a 3′ UTR with a poly-A tail (pA). LINEs are capable of autonomous transposition, since their ORF2 contains genes for EN and RT enzymes. ORF1 encodes a protein that binds nucleic acids. **(C)** The *Alu* element, the most common short interspersed nuclear element (SINE) in the human genome, consists of two GC-rich fragments, the left (L-*Alu*) and right (R-*Alu*) monomers which are connected by an A-rich linker and end in a poly-A tail. These elements are transcribed from an internal RNA polymerase III promoter, but do not code for any enzymes and are therefore incapable of transposing autonomously. Rather, SINEs appear to rely on LINEs for their transposition. LINEs and SINEs are also known collectively as "retroposons." **(D)** DNA transposons contain a large ORF that includes a DNA recognition and binding domain and a catalytic domain. These elements are flanked by short terminal inverted repeats (TIR) and may also include direct repeats (DR), and are capable of autonomous transposition.

Continued

TABLE 3.1 Examples of some common TE superfamilies, families, and host genomes

TE class	Superfamily	Family	Host example
I	*Gypsy*-like	*Gypsy*	*Drosophila virilis*
		297	*Drosophila melanogaster*
	Ty1-copia	*Ty1*	*Saccharomyces cerevisiae*
		Copia	*Drosophila melanogaster*
	Ty3	*Ty3*	*Saccharomyces cerevisiae*
		Micropia	*Drosophila melanogaster*
	LINEs	*L1*	*Homo sapiens*
		R1/R2	*Bombyx mori*
	SINEs	*Alu*	*Homo sapiens*
II	*Helitrons*	*Helitrons*	*Arabidopsis thaliana*
	Mariner-Tc1	*Mos1*	*Drosophila mauritiana*
		Tc1	*Caenorhabditis elegans*
	hAT	*Hermes*	*Musca domestica*
		Hobo	*Drosophila melanogaster*
	P	*P*	*Drosophila willistoni*
	MuDR	*MuDR*	*Zea mays*
	CACTA	*Tam1*	*Antirrhinum majus*
		En/Spm	*Zea mays*
Uncertain	*Foldback*	*Galileo*	*Drosophila buzzatii*

During transposition, a double-stranded DNA intermediate is synthesized by these elements from their RNA template by a mechanism similar to that used by DNA transposons (Craig *et al.,* 2002). The LTRs that are the hallmark of these retroelements play a key role in all aspects of their life cycle. Similar to retroviruses, these elements encode open reading frames (ORFs) called *gag* (specific group antigen) and *pol* (polymerase). The *pol* ORF is subdivided into different domains, including those of reverse transcriptase (RT), integrase (IN) or endonuclease (EN), and RNase H (RH) (Prak and Kazazian 2000; Eickbush and

FIGURE 3.1 *Cont'd.* **(E)** Rolling-circle DNA transposons replicate by a mechanism similar to the rolling-circle transposition mechanism in prokaryotes. They encode several ORFs, including helicase, replication initiator protein (Rep), and single-stranded DNA binding protein (SSB), and are flanked by a conserved TC dinucleotide on the 5′ end and a conserved hairpin and CT dinucleotide on the 3′ end. **(F)** Miniature inverted repeat elements (MITEs) have no coding potential and are flanked by TIR. Figure by T.R. Gregory, based on information provided by Prak and Kazazian (2000), reproduced by permission (© Nature Publishing Group), and Z. Tu (personal communication), reproduced by permission of the author.

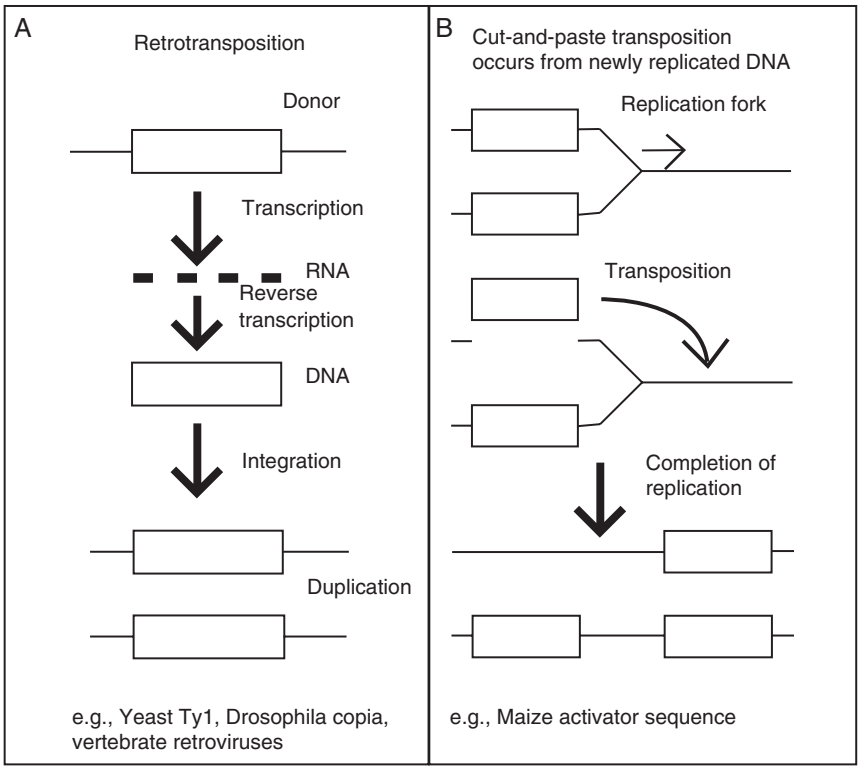

FIGURE 3.2 Two ways in which TEs can increase in copy number. **(A)** Retrotransposition, whereby a second copy of the TE is reverse transcribed to another location in the genome while the first copy remains in its original location. **(B)** Cut-and-paste transposition, whereby a DNA transposon is moved from a location on one newly replicated DNA segment into a region of the genome that has yet to be replicated, resulting in one daughter copy of the DNA that contains the TE only in its new location and one copy that includes the TE at both its original and novel locations. Adapted from Brookfield (1995), reproduced by permission (© Oxford University Press).

Malik, 2002) (Fig. 3.1). The arrangement, and even presence, of these different domains varies in different lineages. Recombination between domains is known to occur and thus different domains may exhibit different evolutionary histories. Based on the phylogeny of their reverse transcriptase, retrotransposons include four distinct lineages: *Ty1/copia, BEL, DIRS,* and *Ty3/gypsy* (Malik *et al*., 2000). Two of these lineages, *Ty1/copia* and *Ty3/gypsy,* are abundant in animals and plants and have been extensively characterized. The *BEL* and *DIRS* lineages are less abundant and have only recently been described.

Many species also harbor endogenous proviruses in their genomes (for more details, see Bushman, 2002). These proviruses may represent either recent insertions into the genome or ancient molecular fossils. Endogenous retroviruses (ERVs) originate

from the retroviral infection of germline cells, followed by fertilization involving the modified gamete. In turn, this gives rise to an offspring that differs from the parent by the presence of a proviral sequence insert. If the modified chromosome becomes fixed (homozygous) in the population then the presence of the provirus will become a permanent genetic property of the population unless changed by subsequent mutation. Human endogenous retroviruses (HERVs) have been particularly well characterized, and the integration of some of these is known to cause changes at the phenotypic level (see later section).

As noted in Figure 3.1, ERVs and LTR retrotransposons are very similar in structure, with the main difference sometimes given as the lack of an envelope (*env*) gene in the latter (Prak and Kazazian, 2000). However, the situation is rendered more complex than this by the fact that at least six lineages of LTR retrotransposons have independently acquired *env*-like ORFs (Eikbush and Malik, 2002). Furthermore, although the exception to the rule, some LTR retrotransposons (e.g., the *Gypsy* element in *D. melanogaster)* have potentially functional *env* genes, whereas many ERVs have nonfunctional *env* genes.

Retroposons (LINEs and SINEs)

Autonomous retroposons, also commonly called LINEs or LINE-like elements, are the second major subclass of TEs that utilizes reverse transcriptase (RT) during transposition. These retroelements are phylogenetically distinct from the retrotransposons. On the basis of RT phylogeny, they are most closely related to the Group II introns (see later section) of mitochondria and bacteria. Retroposons lack LTRs and transpose by simply reverse transcribing a complementary DNA (cDNA) copy of their RNA transcript directly onto the chromosomal target site. Although these elements are abundant and found in all eukaryotic lineages, it was some time before they were first recognized as a distinct self-propagating subclass of autonomous retroelements. Previously, they tended to be grouped with pseudogenes, SINEs, and other nonautonomous elements. Based on the phylogenetic relationship of their RT sequences, combined with the nature and arrangement of their protein domains, retroposons can be divided into five groups, *R2, L1, RTE, I,* and *Jockey,* each named after the first element to be discovered in that group (Eickbush and Malik, 2002).

Short interspersed nuclear elements (SINEs) are nonautonomous retroposons that exploit the enzymatic retrotransposition machinery of LINEs (Kajikawa and Okada, 2002). SINEs have a critical region that is homologous to tRNA, or 7SL RNA (a component of the signal recognition pathway), together with promoter sequences called A and B boxes (Nikaido *et al.,* 2003). SINEs are widely dispersed throughout many eukaryotic genomes and can be present in more than tens of thousands of copies per genome. For example, the SINE family *Alu* constitutes more than 10% of the human genome (International Human Genome Sequencing Consortium, 2001). The enormous number of SINE amplifications per organism makes them important evolutionary agents for shaping the diversity of genomes, and the

nature of their mode of insertion makes them useful for diagnosing common ancestry among host taxa (see later section). As such, they represent a powerful new tool for systematic biology that can be strategically integrated with other conventional phylogenetic characters, most notably morphology and DNA sequences.

Mobile Introns

Mobile introns are mobile elements that reside within genes, but the element sequences are spliced out of RNA transcripts after synthesis. Introns are divided into several distinct classes according to their sequence and structure, as well as their splicing mechanism (Belfort *et al.,* 2002). Many Group I introns have been identified in eukaryotes and bacteria, but none have been found in archaea. Group II introns occur in mitochondrial and chloroplast genomes of fungi and plants and in cyanobacteria, proteobacteria, and Gram-positive bacteria, but not in archaea or in the nuclear DNA of eukaryotes. Group II introns are thought to be the likely progenitors of eukaryotic spliceosomal introns. Both Group I and Group II introns have the capacity for "intron homing" (the process by which the intron invades the same site in a cognate intronless allele), but the two groups use a different homing mechanism. Many Group I and Group II introns as well as a few archaeal introns are characterized by the presence of an ORF within the intron that encodes a maturase that promotes mobility of the intron within the genome. Introns can mobilize by two mechanisms, intron homing and intron transposition. Intron homing involves the transposition of the intron between two alleles of the same gene, one of which starts out with a copy of the intron and the other of which does not. By contrast, intron transposition involves the invasion of a new genomic site or locus (Belfort *et al.,* 2002).

Class II Transposable Elements

DNA transposons are subdivided into two subclasses according to their mode of transposition. The majority of Class II elements previously described transpose by means of a classical cut-and-paste mode similar to that of the *Tn10* elements of bacteria. A second group transposes by means of a rolling circle (RC) mechanism reminiscent of the *IS91, IS801,* and *IS294* bacterial families of transposons (see Fig. 3.1 for a comparison of the main structural features of these two subclasses). The classification of a third group known as Miniature Inverted Repeat Transposable Elements (MITEs) has previously been unclear, but these elements probably represent nonautonomous Class II elements (see later section and Fig. 3.1).

DNA Transposons That Transpose by a Cut-and-Paste Mechanism

The structural hallmarks of Class II elements that transpose by a cut-and-paste mechanism (Fig. 3.2) include terminal inverted repeats (TIRs) of varying length

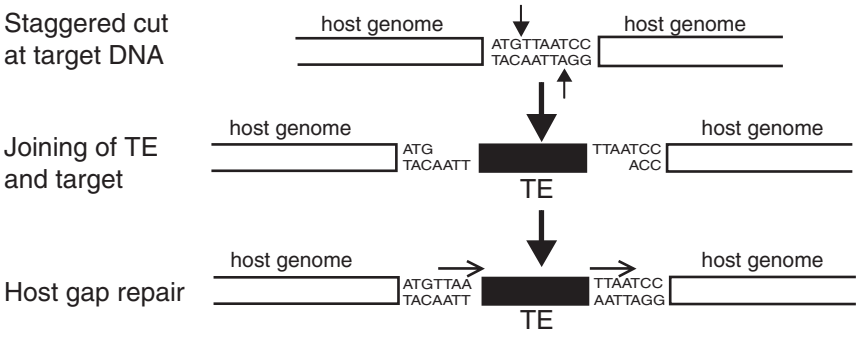

FIGURE 3.3 Generation of target site duplication (TSD) on insertion of a TE in host DNA. TE DNAs are indicated by solid black boxes. TSDs are produced by most TEs, with the exception of *Helitrons* and some long terminal repeat (LTR) retrotransposons. Adapted from Tu (2004), reproduced by permission (© Elsevier, Inc.).

and 2 to 10 base pair (bp) flanking direct repeats generated by target site duplications (TSDs) when the element inserts into a new site in the genome (Fig. 3.3). According to the cut-and-paste transposition model, element-encoded transposases perform both the cleavage and transfer reactions that are required to cut the transposon at both its termini and to insert it into a new position in the genome.

The majority of DNA transposons belong to families that are characterized by transposases of the DDE class named after the highly conserved Asp (D), Asp (D), and Glu (E) amino acid residues belonging to the catalytic core. They form a superfamily of Class II transposons that includes the *Tc1* and *mariner* element families that are widely distributed among animals and are prone to horizontal transfer across species. Class II elements that lack DDE signatures are subdivided into a number of superfamilies. Again, the *hAT* superfamily was named after the first members to be described: the *hobo* element in *D. melanogaster,* the *Activator* (*Ac*) element in maize, and the *Tam3* element of snapdragon. Additional superfamilies and families include the *P* elements found mainly in *Drosophila* species, the *piggyBac* family that was first identified in baculoviruses that had been associated with moth cell lines, and the *Mutator* and *En/Spm* families first described in maize.

Miniature Inverted Repeat Transposable Elements (MITEs)

It has been recognized for some time that many nonautonomous members of Class II element families are derived by internal deletion of autonomous elements. Although the position and size of these deletions varies widely within a given family, some internal sequence similarity with the full-sized elements of the same family is commonly observed. A group of TEs collectively referred to as Miniature Inverted Repeat Transposable Elements (MITEs) were first described about ten years ago and

were found to possess several properties reminiscent of nonautonomous Class II elements; notably, they are short (~100–500 bp in length) and have conserved terminal repeats. However, they have no coding potential and are frequently present in a copy number much higher than that of typical nonautonomous members of Class II families.

Target site preference is a hallmark of these elements. For example, *Tourist,* one of the first MITE families described, has a target site preference for TAA, and that of another family, *Stowaway,* is TA. MITEs were first discovered in plants, but they have subsequently been identified in many species of animals. They were first mistakenly identified as SINEs, but it is now evident that they are indeed nonautonomous DNA elements that originated from a subset of existing DNA transposons (Feschotte *et al.,* 2002). However, it remains unclear whether all DNA transposon families can give rise to MITEs.

Foldback (FB) Elements

The *FB* elements constitute a heterogeneous family of TEs whose transposition mechanism is still not fully understood. Nevertheless, it does seem clear that their transposition mechanism is quite unrelated to that of retroelements. *FB* elements were first described in *D. melanogaster* and are notable for their large inverted repeat termini that vary in length from several hundred base pairs to several kilobases. The inverted terminal repeats of *FB* elements have an unusual structure in that they are composed primarily of tandem copies of simple sequence DNA. Their central portions are heterogeneous with usually little or no additional information present in these regions. There is good evidence for deletions, inversions, and reciprocal translocations having one or both breakpoints at sites of preexisting *FB* insertions. For example, the origin of two polymorphic chromosomal inversions in *Drosophila buzzatii* can most likely be attributed to an *FB* element named *Galileo* (Caceres *et al.,* 2001; Casals *et al.,* 2003). Copies of this TE have been identified at all four inversion break points of these two inversions, and these breakpoints have become genetically unstable regions and hotspots for the accumulation of TE insertions and other structural changes. Thus *FB* elements appear to represent highly potent agents for generating genome rearrangements.

Rolling Circle Transposons

Previously it was thought that all eukaryotic Class II elements use a DNA-mediated mode of "cut-and-paste" transposition. However, recently a new family, the *Helitrons,* was discovered that is propagated by a mechanism similar to rolling-circle (RC) transposition in prokaryotes (Kapitonov and Jurka, 2001). *Helitrons* tend to be large and, surprisingly, constitute as much as 2% of the genomes of *Arabidopsis thaliana* and *Caenorhabditis elegans* (Kapitonov and Jurka, 2001).

The Relationship Between Class I and II Elements

A bacterial origin, either as transposons or retroelements, is generally assumed for most eukaryotic TEs. The structure and function of eukaryotic DNA elements is in fact very similar to those of bacteria. Eickbush and Malik (2002) have proposed an evolutionary scenario that includes a mosaic origin for LTR retrotransposons (Fig. 3.4). In this scheme, it is proposed that all eukaryotic mobile elements are descended from bacterial elements. LTR retrotransposons are suggested to have evolved from the fusion of bacterial transposons and bacterial retroelements, probably mobile Group II introns (see Fig. 3.4 for more details). This evolutionary scenario provides a rational basis for a new nomenclature

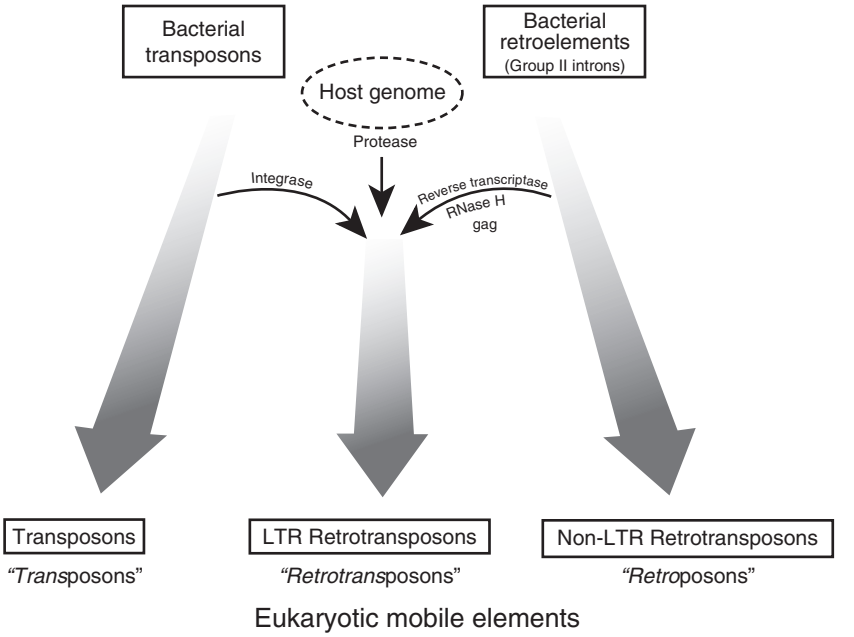

FIGURE 3.4 The probable bacterial origin of eukaryotic TEs. Transposons are DNA-mediated elements (Class II) and are probably derived from bacterial DNA transposons, whereas long terminal repeat (LTR) and non-LTR retrotransposons are RNA-mediated (Class I) (see Fig. 3.1). "Retroposons" refers to long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), which may be derived from bacterial Group II mobile introns. LTR retrotransposons are similar to retroviruses, and are believed to have a mosaic origin. Adapted from Eickbush and Malik (2002), reproduced by permission (© American Society for Microbiology Press).

(Eickbush and Malik, 2002) consisting of three main lineages: the "Transposons" (DNA-mediated elements), the "Retroposons" (non-LTR retrotransposons), and the "Retrotransposons" (LTR retrotransposons), the latter group having an origin that is a mosaic of the other two lineages. This nomenclature returns to the original usage of these terms by Howard Temin (1989).

HALLMARKS OF TE SEQUENCES

Most transposable elements share a number of properties that differentiate them from other types of DNA sequences, as discussed in the following sections.

Dispersed Multigene Families

There are two major groups of repeats in eukaryotic genomes: tandemly repeated satellites and repeats interspersed with genomic DNA. In contrast to satellite DNA, which tends to be highly repetitive and is usually confined to specific chromosomal regions, TEs belong to dispersed multigene families, which constitute a major component of the middle repetitive DNA that is abundant in many eukaryotic organisms. The degree of repetition (copy number) for any particular element family varies widely from a few to millions of copies, depending on the particular TE family and host species involved.

Target Site Duplications

At the DNA level, TEs are almost always recognized by small target site duplications (TSDs), which are induced at the point of insertion (Fig. 3.3). During transposition, the ends of the TE attack the target DNA at staggered positions such that the newly inserted element is flanked by short gaps. Host systems repair these gaps, resulting in the target sequence duplications that are characteristic of transposition. The length of this duplication is characteristic of each element, but typically many different target sites are used in the host DNA and thus different target sequences are duplicated.

Terminal Repeats

The termini of many TEs are characterized by the presence of repeats that are typically homologous among members of the same element family. In DNA-mediated transposition, inverted terminal repeats are binding sites for the transposase that

is encoded by complete, autonomous elements and whose role is to fuse the ends of the element with the target DNA. For example, *IS* elements in bacteria carry perfect or nearly perfect inverted repeats of about 10–40 bp. These terminal repeats are believed to serve as recognition sequences for the transposition enzymes (transposases) in their role of fusing the ends of the element with the target DNA. LTRs provide a signature for retrotransposons that is shared with retroviruses, to which they are closely related. The two direct repeat terminal sequences of any single element are identical at the time of insertion into the host DNA. Subsequently the termini of single elements diverge from one another over time and the amount of divergence can be used to estimate the age of the element in the host genome (e.g., SanMiguel *et al.,* 1998). In contrast to the LTR retrotransposons, the retroposons (such as LINEs and SINEs) are characterized by the absence of terminal repeats.

Coding Regions and Motifs

In addition to their terminal repeats, autonomous TEs contain various genes. For example, LTR retrotransposons encode genes in an order that is typical of their family (see Fig. 3.5). LINEs include two open reading frames (Fig. 3.1) that can be quite variable in sequence. SINEs are relatively short and are dependent on LINEs for transposition. Both LINEs and SINEs have internal promoters for transcription by RNA polymerase. Cut-and-paste transposons typically encode a transposase gene in one or more ORFs. Sometimes a truncated version of the transposase sequence functions as a repressor of transposition, as seen for the *P* element (Fig. 3.6).

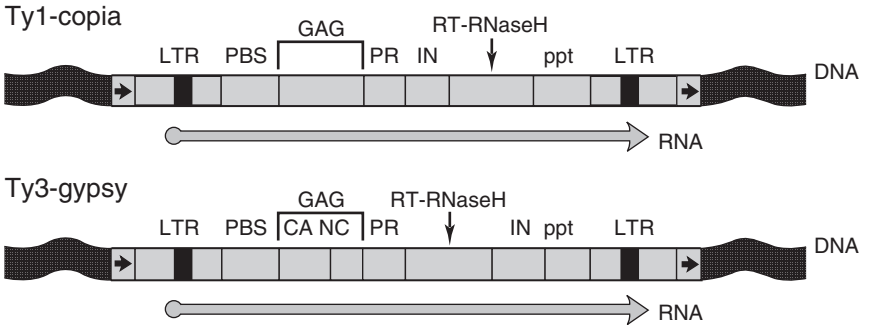

FIGURE 3.5 Structure of the two main families of long terminal repeat (LTR) retrotransposons, *Ty1-copia* and *Ty3-gypsy*. Some members of the *Ty3-gypsy* group also contain an envelope (*env*) gene between the integrase (IN) and polypurine tract (ppt) genes, and therefore replicate as retroviruses (see Fig. 3.1). Adapted from Kumar and Bennetzen (1999), reproduced by permission (© Annual Reviews).

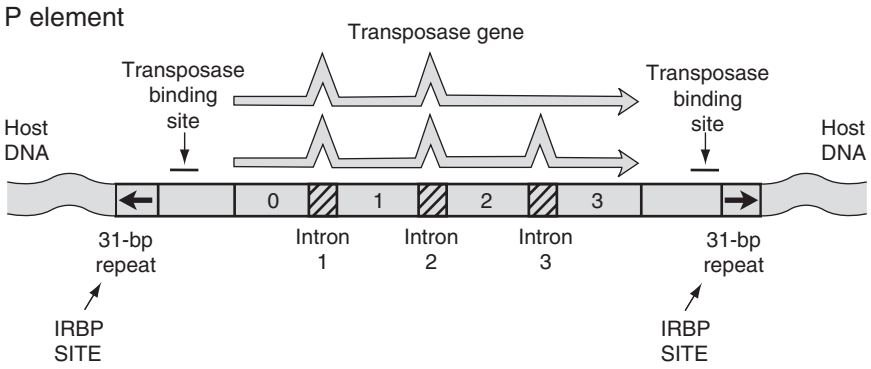

FIGURE 3.6 Structure of the *Drosophila P* element. Two transcripts are produced from the transposase gene, the triply spliced transposase message and the doubly spliced transposase inhibitor. IRBP = inverted repeat binding protein. Adapted from Clark *et al.* (1998), reproduced by permission (© Oxford University Press).

Fixed and Segregating Insertion Sites

Chromosomal sites in which TEs have inserted relatively recently will tend to be heterozygous and segregating in natural populations. In contrast, TEs and their derivatives that have been present for longer periods are likely to be homozygous and fixed in natural populations due to chance or selection. This is another way that the relative age of TEs can be assessed.

METHODS USED IN THE IDENTIFICATION AND STUDY OF TEs

Genetic Analysis of Naturally Occurring Unstable Mutations

Before the availability of modern molecular methods for identification and analysis, Barbara McClintock used purely genetic analyses to demonstrate the existence of elements in maize that can transpose to new chromosomal locations, alter the expression of nearby genes, and cause chromosome breakage, all in a developmentally regulated fashion. In *Drosophila,* hints of the presence of TEs were provided by some of their phenotypic manifestations, such as unstable mutations (Green, 1977) and hybrid dysgenesis (Kidwell *et al.,* 1977), long before they could be identified directly by molecular techniques. For example, the molecular nature of the *P* element responsible for *P-M* hybrid dysgenesis in *D. melanogaster* was only deducible with the

cloning and sequencing of the wild type allele at the *white* locus (Bingham *et al.,* 1981) nearly a decade after the first observations of hybrid dysgenesis at the molecular level. Seven independent mutations at the *white* locus that had been induced by hybrid dysgenesis were examined. It was possible to identify five mutations that were caused by DNA insertions of 0.5, 0.5, 0.6, 1.2, and 1.4 kb in the wild type allele. The DNA insertions in four of these mutations, although heterogeneous in size and pattern of restriction enzyme sites, were homologous in sequence.

Currently, the methods used in the identification and study of TEs are primarily those used in standard molecular genetic analyses. However, phylogenetic methods are also employed in the analysis of the evolutionary history of TEs and with the advent of large-scale genome sequencing, bioinformatics approaches are increasingly being used.

Methods of Molecular Analysis

The majority of data on variation in genomic copy number and location of members of a TE family have been obtained using three methods:

1. *Southern blot analysis*. This technique provides a method for detecting TE fragments that are complementary to a specific DNA or RNA sequence by probing Southern blots of restriction-digested genomic DNA. In the case of multigene families, such as TEs, Southern blot analysis can also provide estimates of the copy number of a specific TE family present in a genome. Determining whether the TE of interest is present and in what copy number requires probing the DNA on a nitrocellulose membrane with radio-labeled or biotinylated (biotin-labeled) probes. Probes are chosen with homology to the TE sequences of interest and should not bind to the host genomic DNA or the nitrocellulose in a nonspecific manner. If not fixed within a species, individual TEs are manifested as insertion–deletion polymorphisms. If fixed, they are seen as insertion–deletion variation between related species.
2. *In situ hybridization to polytene chromosomes*. *In situ* hybridization allows TEs and other specific DNA sequences to be localized to particular segments of chromosomes. Because of polytenization (multi-strandedness), *Drosophila* salivary gland chromosomes are relatively large and have a well-defined morphology, which allows the easy localization of individual members of specific families of TEs by *in situ* hybridization.

It has been shown that in several respects Southern blot analysis is inferior to *in situ* hybridization as a method for accurately describing the properties of any specific TE family in the *D. melanogaster* genome (Maside *et al.,* 2001). The Southern blotting technique had serious deficiencies in three ways: underestimating

TE abundance, revealing less than 30% of the new insertions detected by *in situ* hybridization, and spuriously identifying changes in the size of restriction fragments from any source as simultaneous insertion–excision events.

3. *Polymerase Chain Reaction (PCR) amplification using conserved regions.* The PCR technique involves annealing single stranded probe molecules and target DNA to form DNA duplexes. The PCR is a method for amplifying small amounts of DNA or RNA and therefore has obvious useful applications in the study of TEs. It can be used to isolate complete TEs, or fragments of TEs, and also for end labeling, cloning, sequencing, and mutating TE DNAs. Prior to whole genome sequencing, this method, more than any other, has enabled the detailed study of many families of TEs in myriad organisms.

Data from Genome Sequencing Projects

The sequencing of whole genomes should, in theory, be able to provide complete information about an organism's TE content. If the target site duplications that are the hallmarks of TE insertion are still intact, complete genome sequences can allow the identification of new TE families for which probes were not previously available. However, mutation or deletion of TSDs often prevents the identification of members of previously undescribed families, resulting in an underestimation of TE abundance. Furthermore, many early sequence drafts described as "complete" have included only the euchromatic portion of the genome. The heterochromatic fraction that is often particularly rich in TEs is often slow to be sequenced.

The sequencing of the *D. melanogaster* genome provides a good example of the underestimation of TE abundance. The initial draft sequence included only the euchromatic two thirds of the whole genome (~120 Mb of ~180 Mb) (Adams *et al.,* 2000). In this draft, the estimated proportion of TEs was only 3.10% (International Human Genome Sequencing Consortium, 2001). Subsequent study of the abundance and distribution of TEs in a representative part of the euchromatic genome was made by analyzing the sizes and locations of TEs of all known families in the genomic sequences of chromosomes 2R, X, and 4 (Bartolome *et al.,* 2002). This study came up with an even lower estimate of up to 2% TEs. More recently, a systematic computational analysis (Kapitonov and Jurka, 2003a) provided evidence that TEs in the *D. melanogaster* genome are three times more abundant than previously reported by Bartolome *et al*. (2002). Kapitonov and Jurka (2003a) identified about 80 new TE families in addition to the 50 TE families previously reported. TEs were estimated to account for 6% of euchromatin and 60% of heterochromatin, with an overall abundance of 22%. These figures may still be underestimates because even three years after publication of the original draft, the "complete" sequence is still incomplete.

Reconstruction of Ancestral TEs from Incomplete Contemporary Copies

The study of a new TE usually begins with the identification of other homologous family members, followed by sequence alignment, classification into subfamilies, and construction of consensus sequences. Ancestral TEs can be reconstructed from incomplete copies as demonstrated for the *Tc1*-like element *Sleeping Beauty* from fish (Ivics *et al.,* 1997). *Sleeping Beauty* is an active DNA-transposon system from vertebrates that is useful for genetic transformation and insertional mutagenesis. Molecular phylogenetic data were used to construct a synthetic version of *Sleeping Beauty,* which could be identical or equivalent to an ancient element that dispersed in fish genomes in part by horizontal transmission between species. A consensus sequence of a transposase gene of the salmonid subfamily of elements was engineered by eliminating the inactivating mutations.

Databases for Repetitive DNA Sequences

Repbase Update (RU) (www.girinst.org/Repbase_Update.html) is an electronic database assembled to organize the explosively growing number of repetitive sequences from different eukaryotic species. RU is being used in genome sequencing projects worldwide as a reference collection for masking and annotation of repetitive DNA (e.g., by RepeatMasker or CENSOR). It is particularly useful because it includes TE consensus sequences and their biological characterizations that are reported nowhere else.

Phylogenetic Analysis

The phylogenies of TEs based on DNA or amino acid sequences are usually reconstructed to study the evolution of a particular TE lineage over a long period. Frequently it is of interest to compare TE phylogenies with those of host genes from the same species. The main reason for comparing these two types of phylogenies is to determine whether the relationships between elements of a given family detected in different host species agree with the systematic classification of these species. Lack of congruence between the two types of phylogenies may be due to methodological problems, but it can also be due to TE horizontal transfer, which occurs relatively frequently as compared with that of host genes.

Retroelements appear to exhibit modular evolution—that is, different genes in the same element have different evolutionary histories. For example, McClure (1991) compared the phylogenies of the reverse transcriptase (RT) domain with those of the capsid protein and ribonuclease H (RH) among retroposons. The RT and

capsid protein trees were congruent with one another, but that from RH was clearly different. These differences could be accounted for by xenologous recombination (replacement of a resident gene by a homologous foreign gene), or by independent assortment (McClure, 1991).

APPLICATIONS OF TEs TO OTHER AREAS OF BIOLOGY

In addition to their basic biological interest, TEs have a number of uses for moving and marking genes in a variety of different genomes. Several examples of these types of applications are described in the following sections.

Transformation Systems Based on Transposable Elements

Genetic transformation involves the transfer and incorporation of foreign DNA into a host genome. In order for this transferred DNA to be transmitted to later generations, transformation of germline or other appropriate cells of the recipient species is essential. The transfer of DNA sequences into eukaryotic species usually requires a "vector" such as a bacterium, virus, or transposable element.

Considerable effort has been expended recently in the development of TE-based transformation systems, particularly in insects, and such systems are now being used in many areas of biology. For example, between 1998 and 2004, four new TE-based vector systems were successfully developed for stable germline transformation of nondrosophilid insects. These systems are based on the *Mos-1* (active *mariner*) element from *Drosophila mauritiana,* the *Hermes* element from *Musca domestica,* the *Minos* element from *Drosophila hydei,* and the *piggyBac* element from *Trichoplusia ni*. In addition to *Drosophila* species, successful transformation of mosquitoes, tephritid fruit flies, and other dipteran species has been achieved. In fact, insect transformation can now be considered routine (Handler, 2001).

Transposable Element Mutagenesis and Gene Tagging

TEs are commonly used for mutagenesis and to tag, or mark, host genes that contain an inserted copy. The basic idea is to activate and insert a labeled TE of known sequence, look for a phenotypic effect, and find the marked TE, thereby

identifying the gene responsible. Once the gene of interest has been isolated, it is amenable to further genetic analysis. One of the first examples of gene tagging was the use of a retrotransposon to isolate the *white* locus in *D. melanogaster* (Bingham *et al.,* 1981). Shortly following, dysgenesis-induced mutations at the *white* locus were used to identify the molecular basis of *P-M* hybrid dysgenesis (Bingham *et al.,* 1982). Later, the development and use of single genetically marked *P* elements simplified the identification and recovery of induced mutations (Cooley *et al.,* 1988). About the same time, the identification of single *P* elements carrying the transposase gene at a defined chromosomal location (Robertson *et al.,* 1988) made it possible to identify transposition events using genetic crosses rather than by embryo injection. This facilitated the removal of the transposase and subsequent stabilization of the new insertion. Single *P* element mutagenesis has made possible the identification of several thousand lethal *P* element insertions in distinct genes and has played an important role in the *Drosophila* Genome Project (Spradling *et al.,* 1995).

The development of efficient nonviral methodologies for genomewide insertional mutagenesis and gene tagging in mammalian cells is highly desirable for functional genomic analysis. A method of transposon-mediated mutagenesis (TRAMM), using naked DNA vectors, based on the *Drosophila hydei* TE *Minos,* has been developed with this goal in mind (Klinakis *et al.,* 2000). By simple transfections of plasmid *Minos* vectors, a high frequency of cell lines containing one or more stable chromosomal integrations was achieved. The *Minos*-derived vectors insert in different locations in the mammalian genome.

Transposable Elements as Markers in Evolutionary Studies

In order to be suitable markers for ecological and evolutionary studies, an element should be represented frequently in a genome, but should not have the potential for excision from its insertion site, nor be subject to frequent horizontal transfer. SINEs are tRNA-derived retroelements that most often have these properties. The irreversible, independent nature of their insertion frequently allows them to be used for diagnosing common ancestry among host taxa with extreme confidence. Also, many SINEs are specific to order, family, genus, and sometimes even species (Nikaido *et al.,* 2003).

It appears that new SINEs were created sporadically in a common ancestor of some lineages during evolution, a fact that makes them particularly useful for phylogenetic reconstruction (Nikaido *et al.,* 2003). For example, a family of SINEs called *AfroSINE*s is distributed exclusively among species of Afrotheria, a taxon that includes elephants, sea cows, and aardvarks. The use of *AfroSINE*s as markers in phylogenetic analysis confirmed the monophyletic relationships of the

Afrotheria that emerged as a result of physical isolation of the African continent from Gondwanaland (Nikaido *et al.,* 2003). In humans, the *Alu* family of SINEs has been used very successfully as markers in population genetic analysis. Again, this success depends on the facts that recently active SINEs can produce a high level of polymorphism and that their transposition does not involve excision.

The Use of Mobile Introns for Targeted Gene Manipulation

Mobile Group II intron RNAs insert directly into DNA target sites and are then reverse-transcribed into genomic DNA by the associated intron-encoded protein. The mechanism of homing employed by these introns has been used to allow reengineered mobile introns to be targeted to new sequences. Thus a highly efficient bacterial genetic assay was developed to determine detailed target site recognition rules and to select introns that insert into desired target sites (Guo *et al.,* 2000). It was shown that Group II introns can be retargeted to insert efficiently into virtually any target DNA and that the retargeted introns retain activity in human cells. The potential applications of this work on targeted Group II introns range from functional genomics to genetic engineering and gene therapy.

THE PREVALENCE OF TEs IN EUKARYOTIC GENOMES

Ancient Origins

It seems likely that some, but not all, eukaryotic TEs (or at least parts thereof) have an ancient bacterial origin. Although the origins of eukaryotic TEs are poorly understood, it appears that they have often evolved by the serial addition of domains. However, the reverse process, namely deletion or loss of domains, also appears to have occurred in some lineages (Capy *et al.,* 1997a). Both the integrase–transposase and the RT domains of Class I elements are likely to have evolved from those of bacteria, and have been reassembled in eukaryotes, leading to retrotransposons and then to retroviruses by the acquisition of an envelope gene (Capy *et al.,* 1997b).

A well-documented example of the ancient origin of TEs is provided by the *R1* and *R2* retroelements in insects. The *R1* and *R2* elements are two distantly related families of non-LTR retrotransposons, which insert at specific sites 74 bp apart in 28S ribosomal RNA genes. These elements have apparently been maintained by vertical transmission at least since the origin of the phylum Arthropoda, approximately 500 million years ago (Burke *et al.,* 1998).

Present-Day Prevalence

TEs have been identified in almost all eukaryotic species that have been examined. They are present in copy numbers ranging from just a few elements to tens or even hundreds of thousands per genome. In contrast to the high proportions found in large genomes, small genomes tend to have a low proportion of TEs. This wide range is illustrated for a sample of species in Table 3.2. In some cases TEs represent a major fraction of the genome, especially in some plants. For example, TEs have been estimated to constitute between 64 and 73% of the maize genome (Meyers *et al.,* 2001). The dispersed repetitive fraction of the human genome is currently estimated as ~46% (International Human Genome Sequencing Consortium, 2001), but may yet prove to be much higher when diverged and degraded TEs are fully included. Of course, when considering these high proportions, it is important to realize that the vast majority of TE-derived sequences in most genomes are inactive.

Examples of Common TEs in Familiar Organisms

Low Diversity of TEs in Bakers' Yeast

The genome of *Saccharomyces cerevisiae* was the first among eukaryotes to be sequenced, and thereby afforded the earliest glimpse of a full genomic TE complement. It has turned out that this species is exceptional in having only five families of TEs: *Ty1, Ty2, Ty3, Ty4,* and *Ty5,* all of them LTR retrotransposons. A total of 331 insertions were originally identified (Kim *et al.,* 1998) occupying 3.1% of this small genome (~12 Mb). Based on high sequence identities, the *Ty3* and *Ty4* families appear to have been more recent additions to this genome than the other three families. The genomic distribution of *Ty* elements was found to be highly nonrandom, with a high proportion of elements being inserted close to genes transcribed by RNA polymerase III such as tRNA genes.

Retrotransposons in Maize

Although the maize genome is not yet sequenced, analysis of a typical subregion indicated that TEs make up a major component representing numerous families of LTR retrotransposons (SanMiguel *et al.,* 1996). The largest families contain elements such as *Huck* and *Ji* (*copia*-like elements) and *Opie* (a *gypsy*-like element) whose copy numbers are estimated to range between 100,000 to 500,000 per genome (Meyers *et al.,* 2001). These retrotransposons are in general randomly distributed across the maize genome (Meyers *et al.,* 2001). All of the retrotransposons examined appear to have inserted within the last six million years, most in the last

TABLE 3.2 Breakdown of TE content of six species by class

	Yeast	Slime mold	Nematode worm	Mustard weed	Fruit fly	Human
Genome size (Mb)	12	34	100	157	180	3400
Retrotransposons						
No. families	5	6	1	70	22	104
% of genome	3.1	4.4	0.1	6.4		7.9
Retroposons						
No. families	0	7	12	10	5	6
% of genome	0	3.7	0.4	0.7		31.2
DNA elements						
No. families	0	7	12	80	4	63
% of genome	0	1.5	5.3	6.8		2.8
Total						
No. families	5	20	25	180	31	263
% of genome	3.1	9.6	6.5	14	10–22	44.8
References	Kim *et al.* (1998)	Glockner *et al.* (2001)	Duret *et al.* (2000); International Human Genome Sequencing Consortium (2001)	*Arabidopsis* Genome Initiative (2000); International Human Genome Sequencing Consortium (2001)	Vieira *et al.* (1999); Kapitonov and Jurka (2003a)	International Human Genome Sequencing Consortium (2001); Li *et al.* (2001)

three million years (SanMiguel *et al.,* 1998). Amazingly, it seems that retrotransposon insertions have increased the size of the maize genome from approximately 1200 Mb to 2400 Mb in the last three million years (SanMiguel *et al.,* 1998).

P Elements in *Drosophila*

P elements are a family of cut-and-paste TEs that were first described in *D. melanogaster* as the causal agent of *P-M* hybrid dysgenesis (Kidwell, 1994). There is good evidence that *P* elements were horizontally transferred from a distantly related species, *Drosophila willistoni,* and invaded the cosmopolitan species *D. melanogaster* during the last half of the 20th century. It is important to note that although they represent one of the most intensively studied elements in eukaryotes (such that much is known about their structure, transposition mechanisms, and evolution), *P* elements are only one of approximately 130 different TE families found in the *D. melanogaster* genome (Kapitonov and Jurka, 2003a).

The canonical *P* element found in *D. melanogaster* is ~ 3 kb long, and contains four open reading frames (ORFs) that together encode a transposase. In addition, a truncated peptide consisting of only the first three ORFs and part of the third intron encodes a repressor of transposition. Terminal 31-bp perfect inverted repeats and internal 11-bp inverted repeats are required for transposition. The copy number of *D. melanogaster P* elements varies from 0 to about 60 per haploid genome. Usually a minority of these copies are autonomous (transposase-competent) elements and the majority are internally deleted, nonautonomous elements that are generally smaller than autonomous *P* elements. The induction of internal deletions is associated with active transposition of *P* elements.

In addition to the canonical *P* element subfamily, many more diverged sequences have been identified and grouped into 15 subfamilies according to their level of sequence identity. Active *P* elements have only been found in one other subfamily.

LINEs and SINEs in the Human Genome

Almost half of the ~ 76% fraction of the human genome available for analysis has been estimated to consist of four main types of TEs (International Human Genome Sequencing Consortium, 2001; Li *et al.,* 2001). By total composition, LINEs are the most abundant type and make up about one fifth of the genome (protein-coding genes, by stark contrast, constitute only about 1.5%). A single LINE family, *L1,* accounts for ~16% of the human genome and is present in more than 800,000 copies. However, the majority of these are truncated or rearranged, with a mere 4000 or so being full-length and only 40 to 60 active (Prak and Kazazian, 2000).

Although not yet fully understood, LINE elements are thought to integrate into genomic DNA by a process called target primed reverse transcription (TPRT).

L1 insertions are frequently found within AT-rich DNA, or within other *L1* insertions. This distribution pattern may reflect a successful strategy to avoid elimination from the genome because coding sequences tend to be found most frequently in GC-rich regions. However, some LINE elements are found to be associated with human protein-coding genes (Nekrutenko and Li, 2001) and their associated regulatory regions (Jordan *et al.,* 2003).

*L1*s have shaped the human and other mammalian genomes in several major ways: (1) they have greatly expanded the genome both by their own retrotransposition and by providing the machinery necessary for the retrotransposition of other TEs such as SINEs; (2) they have shuffled host genome sequences by comobilization of flanking sequences, a process often referred to as 5′ transduction (discussed in more detail in a later section); (3) they have affected gene expression by a number of mechanisms. From an applied perspective, *L1* elements are useful as phylogenetic markers, in gene discovery, and in the delivery of therapeutic genes.

Numerically, the most abundant TEs in the human genome are SINEs, which are present in far more copies than LINEs even though they constitute a smaller total percentage (about 13%) of the sequence. The predominant family of SINEs, *Alu,* is present in more than a million copies in the human genome. In light of this staggering abundance, Doolittle (1997) has commented that human genomes "might be ironically viewed as vehicles for the replication of *Alu* sequences." Other types of TEs are less well represented but still contribute a significant portion of the total DNA. Thus LTR retrotransposons make up roughly 8%, whereas DNA transposons contribute only about 3%. In general, although their total contribution is large, human TEs are lacking in the diversity of families seen in some other genomes. Indeed, low diversity among TE families appears to be a feature common to many mammalian genomes, and may reflect competition between these families for replicative dominance, resulting in the survival of single rather than multiple lineages (Furano *et al.,* 2004).

Recent publication of the mouse genome sequence provides additional insights into the evolution of human TEs. Although both contain about 30,000 potential coding genes, the mouse genome (3250 Mb) is about 14% smaller than the human genome (3400 Mb) (Mouse Genome Sequencing Consortium, 2002). A lower proportion of the mouse genome (38.5%) appears to be TE-derived than that of humans (46%). However, this large difference is likely to be something of an artifact owing to the higher nucleotide substitution rate in mice than in humans, which makes it more difficult to recognize ancient repeat sequences in the mouse. Interestingly, marked differences exist in the present-day activity of TEs in the two genomes. The rate of transposition appears to have remained fairly constant in mice, whereas in humans, transposition activity apparently reached a peak about 40 million years ago and has since plummeted to its present low level. This difference in contemporary TE activity is reflected in the large difference

in TE-initiated mutations in mice (~10%) versus humans (<1%) (Prak and Kazazian, 2000).

THE DISTRIBUTION OF TEs WITHIN GENOMES

Many aspects of the transposition process are random and, in general, there are few (if any) genomic compartments in which TEs are never found. However, the observed genomic distribution of TEs is often highly nonrandom and distribution patterns appear to differ widely among different groups of organisms. Although reliable data are not yet available for many species, a few general patterns are starting to emerge.

It has previously been postulated that TEs can be loosely divided into two types that occupy two very different niches in the ecology of the genome (Kidwell and Lisch, 1997). One type preferentially inserts into regions distant from host gene sequences, such as heterochromatin or the regions between genes (e.g., the many retrotransposons found inserted between the genes on the third chromosome in maize). This type escapes inactivation (via methylation or heterochromatinization) in regions outside of single copy host genes that are relatively AT-rich and in which recombination is minimal. The second type lives more dangerously by inserting into, or near, single copy sequences that tend to be relatively GC-rich.

Nonrandomness of TE distributions can be accounted for by three main factors: purifying selection acting on inserts that are detrimental to host fitness, a negative correlation between recombination frequency and TE density, and TE target site specificity that may have evolved in response to differential survival owing to selection and recombination. These are discussed in more detail in the following sections.

Selection as a Mechanism for Reducing TE Copy Number

Population studies of the distribution of TEs on chromosomes have strongly suggested that copy number increase, due to transposition, is balanced by some form of natural selection (Charlesworth *et al.,* 1997). Negative purifying selection is expected to act against the deleterious effects of insertions, particularly those located in gene-coding regions. Selection is also expected to act against the gross chromosomal rearrangements caused by ectopic exchange between TE copies (unequal recombination). Whereas the action of both mechanisms in controlling copy number appears to be indisputable, there is continuing debate as to the relative importance of each (e.g., Biémont *et al.,* 1997; Charlesworth *et al.,* 1997).

TEs are powerful mutagenic agents and, like other mutagens, the changes they produce have a broad range of fitness values at the organismal level, with a high proportion being lethal, causing sterility, or being otherwise deleterious to a greater or lesser extent. As such, mutations in those regions of the genome that are most susceptible to disruption, the coding regions, will most likely be subject to negative purifying selection. Therefore, it is expected and observed that TEs are less dense in coding than in noncoding regions. Nevertheless, some TEs do survive for variable lengths of time in these coding regions, either because they have a neutral impact on host fitness or possibly because they confer some fitness benefit to the host. These latter elements exemplify the second type of TE strategy outlined earlier (see also Kidwell and Lisch, 1997).

It is possible that TEs that tend to insert in or near coding regions have evolved ways to take advantage of relatively accessible chromosomal architecture, a high concentration of transcription factors, host enhancer sequences, and horizontal transfer, to maximize replication advantage (Kidwell and Lisch, 1997). Likely examples are elements such as *Mu* in maize (which target single copy sequences) and *P* elements in *Drosophila*. In the latter example, at least 65% of insertions are located near enhancers (Spradling *et al.,* 1995). It can be argued that these elements trade the disadvantage of an increased risk of negative selection for the advantages of occupying genomic regions that are enriched for factors promoting efficient transcription and replication. However, teasing out the relative importance of the various factors involved in TE survival in any particular case is difficult.

The Role of Recombination in Determining TE Distributions

Because of their sequence homology, two TE copies inserted in nonhomologous positions in the same or different chromosomes may misalign and, if exchange occurs, this may result in ectopic recombination. The products of this recombination will include deletions and duplications that are likely to be deleterious to the host. Theory suggests that, as a consequence of deleterious ectopic meiotic exchange between TEs, selection can favor genomes with lower TE copy numbers. This predicts that TEs should be less deleterious, and hence more abundant, in chromosomal regions in which recombination is reduced. Indeed, a number of empirical studies have supported this theory. For example, a study of the distribution of nine families of TEs among a sample of autosomes isolated from a natural population of *D. melanogaster* (Charlesworth *et al.,* 1992) provided evidence to support the hypothesis that TE abundance is influenced by the deleterious fitness consequences of meiotic ectopic exchange between elements. However, Blumenstiel *et al.* (2002) concluded that the accumulation of TEs in heterochromatin and in euchromatic regions of low recombination is not a result of biased transposition

but of greater probabilities of fixation in these regions relative to regions of normal recombination.

TE Frequencies in Euchromatin and Heterochromatin

In general, the majority of genes are found in the euchromatic portions of genomes that exhibit relatively high rates of recombination. In contrast, the late replicating and highly condensed heterochromatic portions are usually depauperate with respect to genes, and have relatively low frequencies of recombination. It has been observed that for several relatively small genomes the euchromatic and heterochromatic compartments harbor widely differing frequencies of TEs. For example, TEs make up ~6% and 16% of the euchromatic portions of the *D. melanogaster* and *Anopheles gambiae* genomes, respectively, compared with greater than 60% of the heterochromatin in both species (Holt *et al.,* 2002; Kapitonov and Jurka, 2003b).

This concentration of TEs in heterochromatin is also seen in *Arabidopsis thaliana,* in which 95% of repetitive DNA by bulk is found in the left arm of Chromosome 2, which contains the centromeric region (Kapitonov and Jurka, 1999). Also, in the compact *Tetraodon nigroviridis* genome with a 10% complement of TEs, there is a marked compartmentalization of TEs in heterochromatin that is reminiscent of that in the small *Drosophila* and *Arabidopsis* genomes (Dasilva *et al.,* 2002). Comparison of TE densities in heterochromatin and euchromatin of the larger vertebrate genomes of human and mouse (~3300 Mb) with those of the smaller genomes are informative. In excess of 40% of these genomes is made up of TEs and other repeats. With the exception of the sex chromosomes, these are scattered relatively uniformly in both euchromatin and heterochromatin.

Inter- and Intrachromosomal Variation in TE Density

Although detailed distribution patterns are not yet available for very many species, it is possible to make a few tentative generalizations about the distribution of TEs among and within chromosomes. In agreement with theoretical predictions (Sniegowski and Charlesworth, 1994), it appears that in many cases the density of TEs is negatively correlated with recombination rate. TEs tend to accumulate in the pericentromeric and telomeric regions of chromosomes where recombination is reduced. For example, the abundance and distribution of TEs was studied in a representative part of the euchromatic genome of *D. melanogaster.* TEs were not distributed at random in the chromosomes and their abundance was

more strongly associated with local recombination rates than with gene density. The results are compatible with the ectopic exchange model, which predicts that selection on the deleterious products of ectopic recombination is a major factor constraining TE copy number (Charlesworth *et al.,* 1997).

Generally speaking, more TEs are found to be associated with sex chromosomes than with autosomes. This pattern is probably related to a greater concentration of heterochromatin in sex chromosomes than in autosomes. However, cause and effect are difficult to disentangle. On one hand, heterochromatin is characterized by low recombination rates. On the other hand, TEs may be agents in the formation and spread of heterochromatin. For example, Steinemann and Steinemann (1998) observed a massive accumulation of DNA insertions in the neo-Y chromosome of *Drosophila miranda*. They claim to present compelling evidence that the first step in Y chromosome degeneration is driven by the accumulation of TEs, especially retrotransposons. The switch from euchromatic to heterochromatic chromatin structure could be enabled by the enrichment of these elements along an evolving Y chromosome.

Overall differences in TE densities among different chromosomes and chromosome arms are a common feature of many sequenced genomes. In addition to differential recombination rates between sex chromosomes and autosomes, these may be related to genomic features specific to individual genomes. For example, TE densities in the different chromosome arms of *A. gambiae* are 59, 37, 46, 47, and 48 TEs per Mb for chromosomes X, 2R, 2L, 3R, and 3L, respectively. The relatively low overall repeat density for the right arm of the second chromosome (2R) may be related to the comparatively large number of paracentric inversions on this chromosome arm (Holt *et al.,* 2002). Significantly, a major effect of inversions is to reduce recombination rate.

TE Target Site Specificities

Some TEs appear to have evolved strategies to minimize the potentially devastating effects of their induced mutations on the fitness of their hosts. Such elements provide examples of the first type of TE strategy outlined previously, in which they preferentially insert into regions of the genome distant from host gene sequences. Some of these elements appear to target regions in which recombination is minimal (such as telomeres and centromeres) and where essential genes are scarce. In practice it is important, though often difficult, to determine whether nonrandom TE distributions result from the effects of selection or whether they can be accounted for by element target site specificity. *P* elements in *Drosophila* provide an example of how it is possible to experimentally discriminate between these two possibilities. The distribution of *de novo P* insertion mutations was examined by

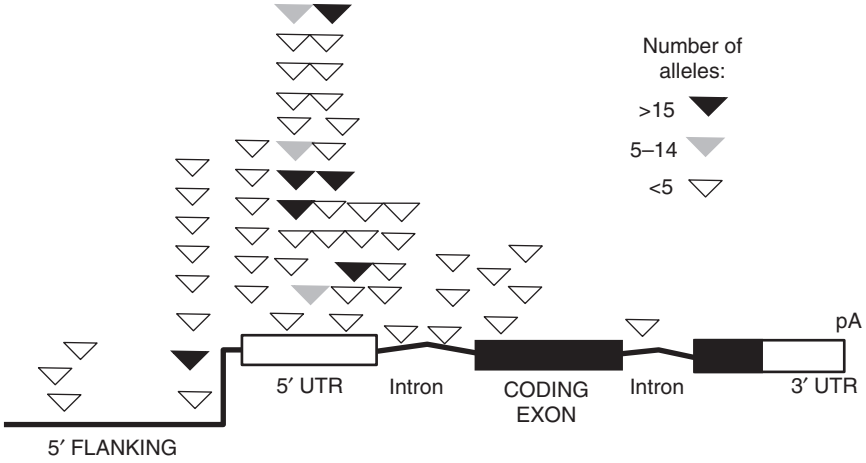

FIGURE 3.7 Illustration of preferential *P* element insertion near the 5′ end of *Drosophila melanogaster* gene transcription units. From Spradling *et al.* (1995), reproduced by permission (© National Academy of Sciences USA).

in situ hybridization before selection had had an opportunity to act on them (Spradling *et al.,* 1995). Whereas the insertion of members of most TE families was essentially random, *P* elements exhibited a preference for a particular subset of genes as well as noncoding regulatory regions within genes (Fig. 3.7). The implications of such target site preferences for the evolution of host gene regulation are discussed later.

THE DYNAMICS OF TE EVOLUTION

Long-Term Evolution and TE Life Cycles

It appears that many TE families have an ancient origin and exhibit common patterns of long-term evolution within their host lineages. Theoretical studies predicted that the ultimate fate of most TEs, over perhaps millions of years, is mutation resulting in loss of transposition activity, rapid divergence, and ultimately the loss of an identity separate from that of the host sequences (Kaplan *et al.,* 1985). In elements subject to horizontal transfer, this scenario is modified to include the possibility of a succession of life cycles in different host lineages with the potential for long-term evolutionary survival.

Mechanisms of Spread and Loss

The main mechanisms responsible for amplification and spread of TEs within and among genomes and populations are transposition, horizontal transfer, and sexual reproduction. The two mechanisms most responsible for loss of TE sequences within a genome are selection against deleterious insertions and selection against the products of ectopic recombination. However, loss can also occur through excision (deletion). There is no general agreement about which mechanism is most important for the removal of TEs. Some argue that ectopic recombination has the greatest influence (e.g., Charlesworth *et al.,* 1997). Others argue that purifying selection on insertions that reduce host fitness is the most important factor (e.g., Biémont *et al.,* 1997).

Transposition

Transposition involves the movement of DNA molecules from one chromosomal location to another in the same cell. This is achieved by means of one of three types of mechanism, depending on the family to which the TE belongs. Transposition may be replicative, nonreplicative (sometimes called conservative), or involve an RNA intermediate. The chemistry of the breakage and reunion reaction is identical for all three types (Plasterk, 1995). Even if the jump is not replicative, it is sufficient for genomic copy number increase that the overall frequency of jumping is higher from replicated to nonreplicated DNA (Plasterk, 1995). Autonomous TEs typically encode the enzymes that are involved in their own transposition, but host factors are often also involved. Transposition, together with out-crossing in sexually reproducing organisms, is responsible for the spread of TEs from one organism to another within a population and species. If unchecked over time, the copy number of a particular TE family may increase to very large numbers, as seen in some plants.

Selection

TEs can be subject to both positive and negative natural selection, which may act at one of two different levels: the molecular (sequence) level and the host organism level. Positive selection at the DNA sequence level results from the ability of TEs to replicate faster than those of the host genome. Such "intragenomic selection" forms the basis of the selfish (or parasitic) DNA hypothesis (Doolittle and Sapienza, 1980). This type of selection is common when TEs invade new genomes, but is probably quite infrequent at other stages of their life cycle. TEs may also be subject to negative selection at the molecular level (because wild type elements coding for a functional transposase tend to transpose more frequently than mutated elements), but several studies have concluded that such selection acts mainly at the time of horizontal transfer (Witherspoon, 1999).

At the host organism level, negative selection commonly results from either TE-induced insertional mutations that are deleterious to hosts or ectopic recombination between homologous TE sequences located in nonhomologous regions of the genome. Obviously, inactive TEs are neutral with regard to this type of selection. Although positive selection of TEs is possible at the organismal level, opinion has been divided about the frequency of the beneficial effects of TEs on host genomes. Some believe that positive selection is so rare that it has virtually no impact on host evolution. A second perspective that is becoming increasingly expressed is that TEs have beneficial effects on their host genomes more frequently than previously acknowledged, and sometimes in unexpected ways (McDonald, 1998; Brosius, 1999a; Kidwell and Lisch, 2001).

As discussed in Chapters 1 and 2, the total amount of DNA contained within a genome can have important fitness-related consequences. Parameters such as cell size, cell division rate, body size, metabolic rate, and development can all be affected by variation in DNA content. Given that a substantial fraction of most eukaryotic genomes is comprised of TEs (see later section), there is good reason to expect selection on genome size to pose limits on TE abundance in certain species. The potentially profound implications of such a process for evolutionary theory are discussed in Chapter 11.

Excision and Deletion

Many, but not all, TEs can excise or remove themselves from their site of insertion in the host genome either precisely or imprecisely. Precise excision leaves a "footprint" in the form of a target site duplication of the original site of insertion in the host genome. Imprecise excision can take many forms, resulting in various mutations in the host genome, including residual partial insertions, deletions, and duplications.

On the scale of small (<400 bp) neutral insertions and deletions, there is a bias toward DNA loss, which some authors have proposed would tend to remove inactive ("dead-on-arrival") TEs over long timescales (Petrov, 2001). Although this is probably true in some organisms with small genomes (e.g., *Drosophila*), the data on which this theory is based are limited, and caution is needed in their interpretation (Gregory, 2003, 2004) (see Chapter 1).

Ectopic Recombination

Multigene families are susceptible to nonhomologous, or ectopic, recombination when homologous regions of different members of the family misalign and genetic exchange takes place either intra- or interchromosomally. Such events can lead to duplications, deficiencies, and new linkage relationships, often with consequent host fitness reduction and constraints on increase in copy number.

A good example of the importance of ectopic recombination is provided by the *BARE-1* retrotransposon in barley (e.g., Kalendar *et al.,* 2000). In addition to full-length copies, the *BARE-1* element is also represented by numerous solo LTRs, which are the relics of intraelement recombination between the LTRs of complete elements, and the consequent loss of their internal domains. The excess of LTRs relative to full-length elements suggests that recombination is additive among elements in the barley genome. Also, the greater the number of solo LTRs relative to full-length *BARE-1* elements in a given part of the genome, the lower the observed density of these elements. This is consistent with the expectation that high recombination rates result in loss of intact elements and an increase in the frequency of solo LTRs.

Horizontal Transfer

In addition to being transmitted vertically from parent to offspring, as with the normal inheritance of host genes, TEs can occasionally be transmitted horizontally (or laterally) from one species to another. A discrepancy between the phylogeny of a mobile element and that of a host gene is one of the most common ways that horizontal transfer is detected. However, it is important to note that

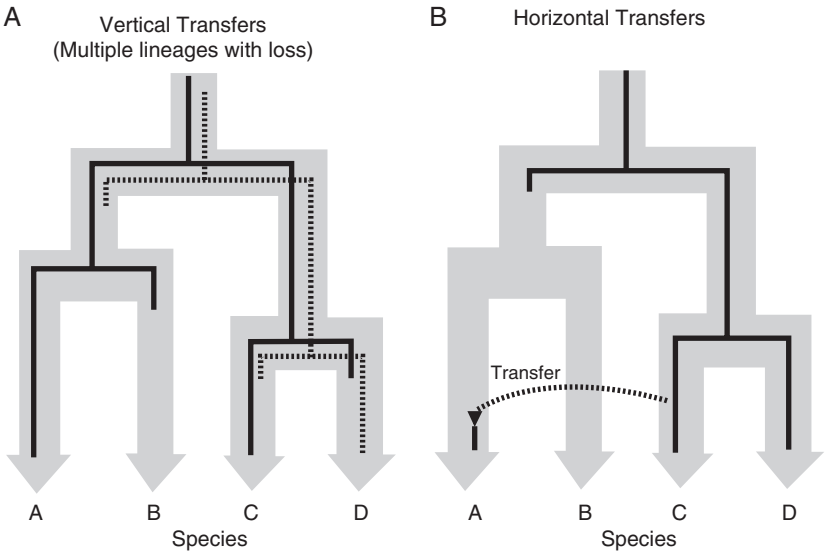

FIGURE 3.8 Alternative ways in which the incongruence of host and TE phylogenies can be explained. **(A)** Vertical transfer, in which the host gene and TE are lost differentially in the various lineages. **(B)** Horizontal transfer, in which a host gene or (more likely) a TE is transferred from one unrelated species to another. From Eickbush and Malik (2002), reproduced by permission (© American Society for Microbiology Press).

such discrepancies may have alternative explanations (Fig. 3.8). There is evidence that TEs transfer horizontally more frequently than nonmobile genes (Kidwell, 1993) and that Class II elements are more prone to this behavior than Class I elements. The Class II elements *mariner* and *P* provide good examples (Kidwell, 1994; Lohe *et al.,* 1995), but a major puzzle remains regarding the mechanism by which horizontal transfer is achieved.

REGULATION OF TE ACTIVITY

The unbridled activity of TEs has the potential to cause chaos in the host genome with a consequent reduction of fitness. Not surprisingly, a variety of different mechanisms have evolved to keep transposition in check. These mechanisms can be divided into two groups according to whether they are mediated by the elements themselves, or by the host organism. The regulation of *P* element activity in *Drosophila* is one of the most thoroughly studied in this regard and will serve to provide examples of many of the different kinds of regulation described in the following sections.

ELEMENT-MEDIATED REGULATION

The ability of TEs to regulate themselves is a property not shared with physical and chemical mutagenic agents. Presumably, these self-regulatory mechanisms have evolved to modulate the extent of damage to the host genome upon which the TEs depend.

Regulation by Maternal Inheritance

In the *P* element system in *D. melanogaster,* a *P* element–encoded 66-kDa truncated transposase, formed from mRNA transcripts from which the ORF 2-3 intron is spliced out, acts as one type of repressor of transposition. This was first observed as differences in the frequency of hybrid dysgenesis in reciprocal crosses between *P* (*P*+) and *M* (*P*–) strains of this species. *P* strain females produced a maternally inherited repressor that inhibited transposition in F_1 progeny, a condition termed the *P* cytotype. In contrast, *M* females that lacked any *P* elements in their genomes produced no *P* repressor and their progeny were unprotected from the effects of transposition.

Regulation by Multimer Poisoning

Still considering the *P* element system, certain deletion derivatives such as the *KP* element have been shown to encode a product other than the 66-kDa protein that also represses transposition, but is biparentally inherited. These elements

may be involved in forming inactive multimers with the transposase, or with a host protein required for transposition.

Regulation by Overproduction Inhibition (OPI)

The *mariner* element is regulated by a mechanism called OPI in which an excess of the transposase reduces the activity of the element as measured by an excision assay. For example, increasing the number of copies of a *Mos1* construct decreased the rate of germline excision by 13% with one copy of the *Mos1* construct and by 37% with two copies of the construct (Lohe and Hartl, 1996).

Regulation by Transposase Titration

Defective TEs that retain their transposase binding sites may play a role in the regulation of transposition through titration of the active transposase. Such titration effects have been suggested to regulate the *P* element in *D. melanogaster* (Simmons and Bucholz, 1985) and the *mariner*-like elements in other *Drosophila* species (Hartl *et al.,* 1997). This mechanism might explain why certain deletions have replaced almost all functional *mariner*-like elements in a number of species (Hartl *et al.,* 1997).

Antisense RNA Regulation

Several TE families are known to produce antisense transcripts that can regulate transposition by means of RNA–RNA interactions. For example, maize contains two antisense transcripts that correspond to the two major transcripts of the *MuDR* TE (Hershberger *et al.,* 1995). Also, evidence for heat shock–induced antisense regulation of the *P* element has been found in *D. melanogaster* (Simmons *et al.,* 1996).

Host-Mediated Regulation of TE Activity

The broad range in copy number of a single TE family over a number of species attests to the importance of host genomic environment in element copy number control. As an example, the copy number of *Ty-copia* group retrotransposons shows extreme variation, particularly among plant species; *A. thaliana* has only 100–200 copies of *Ty-copia* elements, whereas the bean *Vicia faba* has about one million copies (Flavell *et al.,* 1997). The primary mechanisms employed by host genomes to control transposition are outlined in the following sections.

Host Factors

Specific host factors are needed for the excision and transposition of *P* elements. Two good examples are provided by the *P* element Inverted Repeat Binding Protein

(IRBP) and *P* element Somatic Inhibitor (PSI). IRBP is related to the 70-kDa subunit of the human KU autoimmune antigen and is encoded by a mutagen-sensitive gene, *mus309*. IRBP interacts with 16 bp of the terminal 31 bp inverted repeats and may be involved with the repair of double strand DNA breaks.

DNA Methylation

Cytostine methylation is common, but not ubiquitous, among eukaryotes. In mammals and the bread yeast, about 2–3% of cytosines are methylated. Methylation plays an important role in the regulation of repetitive sequences. In both mammals and plants, the vast majority of methylated DNA is within TEs (SanMiguel *et al.,* 1996; Yoder *et al.,* 1997), but whereas most mammalian exons are methylated, plant exons are not. Thus targeting of methylation specifically to transposons appears to be restricted to plants (Rabinowicz *et al.,* 2003). Maize elements, for instance, undergo reversible methylation associated with changes in their activity (Banks *et al.,* 1988). In *Aspergillus,* a highly efficient system operates to methylate repetitive sequences.

The process known as methylation induced premeiotically (MIP) generates heavy methylation of repetitive DNA and is hypothesized to have evolved to silence TEs (Selker, 1997). It is known from studies of several eukaryotes, such as *Arabidopsis* and mice, that otherwise quiescent TEs become transpositionally active in methylation-deficient backgrounds. Following the sequencing of the *N. crassa* genome (Galagan *et al.,* 2003), it was observed that the methylated part of the genome consists almost entirely of relics of TEs (Selker *et al.,* 2003). This observation strongly suggests that, at least in *N. crassa,* DNA methylation has evolved as a partial defense against the invasion of TEs.

Repeat-Induced Gene Silencing

Repetitive DNA induces diverse silencing mechanisms, including cosuppression in nematodes, repeat-induced-point mutation (RIP) in *Neurospora,* position-effect variegation in *Drosophila,* and RNA-mediated interference (RNAi) in a number of eukaryotic species.

Repeat-Induced-Point Mutation (RIP)

A fascinating mechanism known as "repeat-induced-point mutation," exists in the bread mold *N. crassa,* by which TEs and other relatively large repetitive sequences are inactivated by mutation. The recent sequencing of the genome of this species (Galagan *et al.,* 2003) revealed that the total repeat fraction of the genome is limited to only 10% and, consistent with the hypothesis that RIP has evolved as a defense against the

invasion of TEs, not a single intact TE was identified. This situation differs from that of other fungi such as yeast, which have not evolved the RIP mechanism.

RNA-Mediated Interference (RNAi)

RNAi involves the targeting of complementary mRNAs for degradation by double-stranded RNAs (dsRNAs). Proteins of the Dicer family cleave dsRNAs to generate small interfering RNAs, which target the RNA-induced silencing complex (Dernburg and Karpen, 2002). The Dicer proteins play a number of essential roles, including as a host defense mechanism against TEs and viruses. The RNAi mechanism of TE silencing has been particularly well characterized in *C. elegans,* but has also been found in certain protists, yeasts, insects, and plants (for review, see Vastenhouw and Plasterk, 2004).

Disruption of TE Regulation by Environmental Stresses

There is some evidence that various environmental factors affect the transcription and transposition of some TEs, particularly in plants. For example, activation of the *Tnt1* family in tobacco, tomato, and *Arabidopsis* is affected by pathogen attacks, biotic elicitors such as fungal extracts, and abiotic stresses such as wounding. TE-induced mutations have been recorded to occur in transpositional bursts (Gerasimova *et al.,* 1990) whose causes are not well understood, but which are likely related to inbreeding and other forms of genomic or environmental stress, possibly akin to the genomic stress referred to by McClintock (1984) in her Nobel Prize lecture.

Observations of stress-induced high transcription rates of SINEs have been reported in insects. For example, the level of transcripts of *Bm1*, a SINE found in the silk worm *Bombyx mori,* increases in response to either heat shock, inhibiting protein synthesis by cycloheximide, or viral infection (Kimura *et al.,* 1999). Post-transcriptional events may partially account for stress-induced increases in *Bm1* RNA abundance. In light of these results, it has been proposed that SINE RNAs may serve a role in the cell stress response that predates the divergence of insects and mammals (Kimura *et al.,* 1999), implying that SINEs may have been coopted as a class of cell stress genes.

Temperature is another external factor that may influence TE activation. For example, several cases of disruption of *hsp70* regulatory regions by TE insertions have been reported (Lerman *et al.,* 2003), which underlies natural variation in expression of the stress-inducible molecular chaperone Hsp70 in *D. melanogaster*. It is hypothesized that the distinctive promoter architecture of *hsp* genes may make them vulnerable to TE insertions. Thus transposition may create

quantitative genetic variation in gene expression within populations, on which natural selection can act.

A Continuum of TE-Host Interactions from Parasitism to Mutualism

TEs are often summarily dismissed in the literature as simply being selfish or even junk. It has been argued more recently that a more accurate and enlightened approach is to consider them and their hosts in the coevolutionary terms of host–parasite relationships (Kidwell and Lisch, 2000). Such an approach is considerably more flexible, and envisions a continuum from total parasitism (or selfishness) at one extreme, through a middle ground of neutrality, to "molecular domestication" (or mutualism) at the other extreme. Indeed, the relationship between an element and its host may vary along this continuum over time. For example, when first invading a naïve genome, an element such as the *P* element in *Drosophila* may multiply rapidly in the host genome and represent the epitome of a selfish element. After perhaps millions of years, in rare cases molecular domestication occurs, representing a mutualistic relationship in which the element and the host genome are interdependent on one another. This very process has been demonstrated for *P* elements in several *Drosophila* species (Miller *et al.,* 1999). Other examples of important interactions between TEs and their hosts are described in the following section.

TEs AS MUTAGENS AND SOURCES OF GENOMIC VARIATION

Coding Sequences and the Evolution of Novel Host Genes

Because of the deleterious effects of insertion into protein-coding genes, it has not usually been considered likely that TEs play an important role in the evolution of coding regions. Although a number of examples of associations between TEs and coding sequences have been reported, comprehensive data on the genomic frequency of such associations are only just beginning to emerge. For example, Nekrutenko and Li (2001) examined the sequences of 13,799 human genes and found that 533 (~4%) of them contained TEs or fragments thereof. Among these 533 genes, ~40% were *Alu* elements (SINEs), ~27% were *L1* elements (LINEs), ~24% were LTR retrotransposons, and 9% were DNA transposons. Extrapolation of this result to the ~30,000 genes in the entire human genome suggests that ~1200 human genes contain TEs or fragments of TEs (Nekrutenko and Li, 2001).

It is becoming apparent that novel host genes may be mosaic in origin. The stationary *P* element–related gene clusters of *D. guanche, D. madeirensis,* and *D. subobscura* provide an interesting example of mosaic molecular domestication (Miller *et al.,* 1997). Each cluster unit consists of a cis-regulating section composed of different insertion sequences followed by the first three exons of a *P* element that encodes a 66 kDa repressor-like protein. In contrast to this normal repressor function, these stationary *P* element repeats appear to have evolved the function of transcription factors. Remarkably, the *D. guanche* P-protein produces an enhancer-like effect, rather than repressing canonical *P* element activity in transgenic *D. melanogaster* (W. Miller, personal communication). The insertion sequence, which gave rise to the *de novo* A-type promoter of this *P*-gene cluster, has been identified as belonging to a MITE-like TE family (designated *SGM*) that is related to poorly characterized bacterial *IS* elements of other *obscura* group species (Miller *et al.,* 2000).

Of particular interest is the mechanism by which TEs contribute to new coding sequences. As described previously, approximately 4% of coding sequences in the human genome are associated with TE-derived sequences (Nekrutenko and Li, 2001). Two possibilities exist for how TEs are integrated in coding regions: either the TE can insert directly into a protein-coding region or it can first be inserted into a noncoding region (e.g., an intron) and subsequently be recruited as a new exon (Fig. 3.9). Interestingly, the study by Nekrutenko and Li (2001) indicated that only about 10% of TEs have inserted directly into coding regions, whereas almost 90%

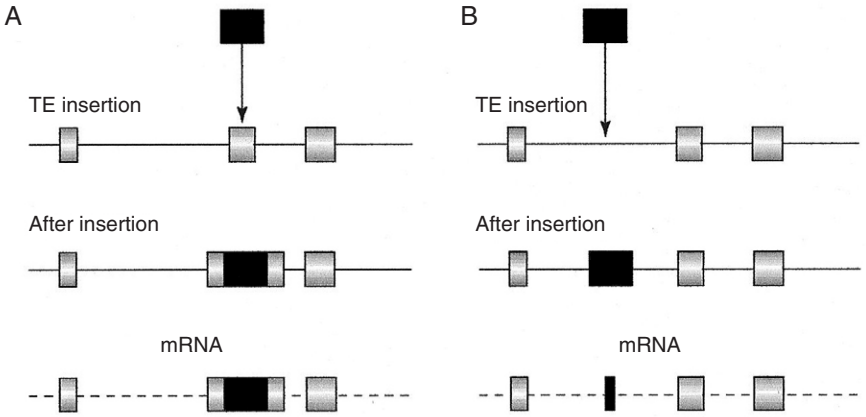

FIGURE 3.9 Two possible ways in which TE insertion into a protein-coding region can alter host genes. **(A)** A TE (black box) is inserted directly into a protein-coding exon (gray boxes), and therefore would automatically become part of the resulting mRNA if the gene is transcribed. However, note that in the majority of cases, the effects of such direct insertion are probably deleterious because TEs often contain multiple stop codons and would destroy the target exon. **(B)** A TE (black box) is inserted into an intron (black lines), and later a portion of the TE is recruited as a novel exon. Adapted from Nekrutenko and Li (2001), reproduced by permission (© Elsevier Inc.).

were recruited from nearby noncoding regions. This high rate of recruitment is possible because TEs carry potential splice sites (Nekrutenko and Li, 2001). This result explains the relatively high rate of insertion despite the expectation that the majority of new insertion in coding regions will be selected against. It will be interesting to see whether this pattern is also found in other species.

Introns

One way that TEs can increase their probability of survival in the genome is to camouflage themselves as introns in order to mitigate the negative selection pressure that accompanies insertion into host genes. Several lines of evidence suggest that the splicing of TEs may be a generalized mechanism to remove insertions in pre-mRNA. The ability to be spliced appears to be widespread, occurring in various species and both classes of TEs (Purugganan and Wessler, 1992). Several examples have been described in maize, including the *wx-m9* allele. Partial restoration of activity of this allele was shown to be due to the splicing of the *Ds* TE from the gene transcript (Wessler *et al.,* 1987).

Alternative Splicing

Alternative splicing provides an important mechanism for generating the observed proteomic diversity that is derived from a relatively small number of protein-coding genes in vertebrates. Alternative splicing is mediated by alternative exons that are included in only a fraction of mRNAs synthesized from a given gene. This is in contrast to constitutive exons that are included in all mRNAs (Kreahling and Graveley, 2004). For example, at least 5% of human alternative exons are derived from a process called exonization by which *Alu* elements are inserted into mature mRNAs via a splicing-mediated process (Lev-Maor *et al.,* 2003; Kreahling and Graveley, 2004). Although the disease-producing potential of constitutively spliced *Alu* exons is considered important in medical genetics, the evolutionary importance of the more common alternative splicing of these exons should not be overlooked. When alternative splicing occurs, the normal version of the encoded protein may still be synthesized at a level sufficient to maintain its function, enabling an *Alu* exon to diversify and possibly produce a protein with improved functionality (Kreahling and Graveley, 2004) under changing environmental conditions.

Gene Regulatory Sequences

The potential evolutionary importance of host genome regulatory sequences (as opposed to the protein-coding sequences themselves) has been recognized

for some time. TEs can affect gene regulation in several ways. A new insertion can either disrupt an existing host regulatory element, or the TE may contribute its own regulatory sequences to a host gene into which it has inserted at an appropriate location.

Inserted sequences may also provide the potential for future evolution of regulatory sequences. For example, LTR retrotransposons carry a promoter within each terminal repeat that provides the potential for contributing to new patterns of host gene expression. The preference for insertion into genic regions also makes certain elements good candidates for cooption as regulatory elements. MITEs seem to be particularly likely to assume such a role, as shown in examples from the yellow fever mosquito *Aedes aegypti* (Tu, 1997), *Arabidopsis* (Feschotte and Mouches, 2000), and rice (Bureau *et al.,* 1996). Relatively small elements like SINEs are most likely to be coopted, but larger elements may sometimes act in a similar way.

Another important example has recently been provided by analysis of the human genome. Jordan *et al.* (2003) analyzed more than 2000 human promoter sequences, each about 500 bp in length (promoters can be defined as the sequence regions that are located directly 5′ of transcription initiation sites and that regulate their 3′ adjacent genes), and identified TE-derived sequences in almost 25% of these. In fact, TE-derived sequences comprised 8% of the total nucleotides found in all promoters. SINEs were represented in promoters at a higher frequency than in the genome as whole, but LINEs were relatively less frequent.

Many TEs can act as movable carriers of regulatory elements, such as promoters or enhancers, and may integrate into or near genes. They can thus contribute to the functional diversification of genes by supplying cis-regulatory domains and can have the potential to alter tissue-specific expression patterns. A second possibility is that, following their initial insertion into or near genic regions, and after lying immobilized and dormant, perhaps for some time, TEs can mutate to be able to regulate nearby genes. The most common regulatory functions served by coopted TEs include those of transcriptional enhancers, reducers and modulators, and polyadenylation signals (Brosius, 1999b). A number of examples of vertebrate regulatory elements and parts of coding regions that have been generated by retroelements illustrate some of the diversity of possible new host functions (Brosius, 1999b) (see http://exppc01.uni-muenster.de/expath/alltables.htm for a list).

TELOMERES

Although most organisms examined to date use telomerase to prevent the loss of sequences at the ends of chromosomes following DNA replication, insects in the order Diptera use two non-LTR elements, *HetA* and *TART,* for this purpose. These retroelements transpose specifically to the ends of *D. melanogaster* chromosomes and are present in tandem arrays at the ends of these chromosomes (reviewed by

Pardue and DeBaryshe, 2003). It is the serial addition of these elements to the ends of dipteran chromosomes that maintains their length following DNA replication. *HetA* and *TART* are completely dedicated to their role in maintaining the ends of chromosomes. Interestingly, the mechanism that these TEs use for extending *Drosophila* chromosomes is essentially the same as that used by telomerase (Pardue *et al.,* 1997). This suggests that the telomeric TEs may have evolved from telomerase. However, an alternative hypothesis, that *Drosophila* might have lost its telomerase and had the role taken over by a "domesticated" TE, cannot be ruled out.

In addition to the retrotransposons *HetA* and *TART* that transpose specifically to telomeres in flies, several nondipteran retrotransposons are commonly found to be associated with telomeres, although they may also transpose to other sites (Pardue and DeBaryshe, 2003). In the yeast *S. cerevisiae, Y′* retrotransposons are found only in subterminal regions of chromosomes while *Ty5* is found preferentially at telomeres and silent mating loci. The silkworm *Bombyx mori* has at least two non-LTR retroposons that insert specifically in the TTAGG telomerase repeat arrays. The parasitic protozoan *Giardia lamblia* reproduces asexually and has only two active retroelement families, called *GilM* and *GilT,* in its genome. These elements are found in tandem head-to-head arrays in the subtelomeric regions of chromosomes, but do not form the terminal sequences.

Centromeres

Centromeres are largely heterochromatic regions of chromosomes devoted to segregation of sister chromatids during cell division. Until very recently the inability of current technology to determine contiguous sequence for highly repetitive regions meant that centromeres largely fell within multimegabase gaps, analogous to black holes from which no information escapes (Henikoff, 2002). Now the previously intractable centromeric boundary is starting to be bridged. For example, analysis of the border of the X chromosome centromere in humans reveals an intriguing gradient (Schueler *et al.,* 2001). Human centromeres consist primarily of alpha satellite DNA (made up of highly repetitive tandemly repeating units). In contrast to the interior centromere sequences, which appear to consist of relatively young alpha satellite repeats that are kept homogeneous by a process such as unequal crossing over, the centromere border reveals repeats that are successively older as the border with euchromatin is approached. Divergence of the repeats increases with age and proximity to the boundary and is associated with an increasing frequency of *L1* insertions (Schueler *et al.,* 2001). Indeed, the frequency of *L1* elements can be used for dating the repeat units in the boundary regions. Recombination among *L1* elements may also be a major mechanism for homogenization of the repeat units. It is argued that this sequence organization is likely to be quite general in complex eukaryotes (Henikoff, 2002), but many more

data will be needed to determine the general frequency with which centromeres are formed from or maintained by TEs.

TRANSDUCTION

Transduction is usually defined as the transfer of a small segment of DNA from donor to recipient bacteria via a bacteriophage (see Chapter 10). Recently, however, the term has apparently been used more generally, including the transfer of short DNA sequences from one site to another in the genome by the action of TEs. Specifically, some TEs can provide a vehicle for the mobilization of flanking nucleotide sequences accompanying aberrant transposition events. In addition to nongenic sequences, gene sequences such as exons or promoters can sometimes be transduced and inserted into other existing genes. This may provide a general mechanism for the evolution of new genes. For example, human *L1* retroposons can modify the genome by three separate mechanisms: (1) insertional mutagenesis during retrotransposition, (2) homologous recombination between different *L1* elements, and (3) the initiation of 3′ transduction events by the mobilization of unique DNA sequences downstream of *L1* elements as a result of aberrant transposition events (Fig. 3.10).

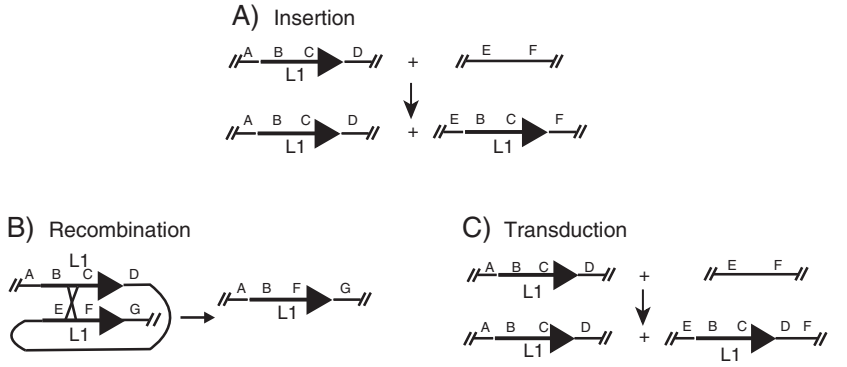

FIGURE 3.10 *L1* elements (a type of long interspersed nuclear element [LINE]) may modify the genome by three means: insertion, recombination, and transduction. Large arrows indicate *L1* elements, letters indicate different regions of the sequence that are affected by *L1* activity. **(A)** Insertional mutagenesis during retrotransposition. Before insertion, only one of the two genome segments (AD) contains the *L1* element and its genes (BC); the other segment (EF) lacks the element. After insertion, both regions contain the *L1* element. This insertion may disrupt coding regions and act as a powerful mutagen. **(B)** Intramolecular (as illustrated here) or intermolecular homologous recombination between different *L1* elements can lead to different combinations of genes in host coding regions (e.g., recombination generating the new arrangement ABFG from ABCD and EFG). **(C)** 3′ transduction occurs when the *L1* element transposes and takes a host sequence (in this case, region D) along with it. This process bears similarities to the transduction of bacterial genes by bacteriophages (see Chapter 10). Redrawn from Goodier *et al.* (2000), reproduced by permission (© Oxford University Press).

Genome Size

Although estimates of genome size are available for several thousand eukaryotes (see Chapters 1 and 2), estimates of the TE fraction are currently available for only a relatively few well-studied species (see Table 3.3 for some examples). At best, these provide only crude estimates of TE proportions in the noneuchromatic regions of genomes, especially because data regarding repetitive fractions of genomes are often slow to be published. In order to explore the extent that TEs contribute directly to genome size variation, the total DNA contributed by TEs was plotted against genome size for 12 species (Kidwell, 2002). Overall, an

TABLE 3.3 Genome sizes and TE proportions for ten species

Species	Genome size (Mb)	No. coding genes	% TE DNA*	References
Saccharomyces cerevisiae	12	6300	3	Kim *et al*. (1998)
Caenorhabditis elegans	100	19,500	6	Waterston and Sulston (1995); International Human Genome Sequencing Consortium (2001)
Arabidopsis thaliana	157	26,000	14	*Arabidopsis* Genome Initiative (2000)
Drosophila melanogaster	180	13,600	10–22	Vieira *et al*. (1999); Kapitonov and Jurka (2003a)
Anopheles gambiae	278	14,000	16 (E), 60 (H)[a]	Holt *et al*. (2002)
Takifugu rubripes	400	31,000	2	Aparicio *et al*. (2002)
Oryza sativa	490	32,000–62,000	16	Yu *et al*. (2002)
Zea mays	2500	50,000	64–75	SanMiguel *et al*. (1996); Meyers *et al*. (2001)
Homo sapiens	3400	30,000	44	International Human Genome Sequencing Consortium (2001)
Mus musculus	3250	30,000	40	Mouse Genome Sequencing Consortium (2002)

[a]E, euchromatin; H, heterochromatin.

approximately linear relationship between total TE DNA and genome size was observed. On the basis of this preliminary analysis, it was suggested that the contribution of TEs to genome size variation is greater, relative to other sources of variation, in larger (>500 Mb) than in smaller (<500 Mb) genomes. Thus, TEs may play a more important role in the increase in size of relatively large plant and animal genomes than in smaller ones. Because increasingly larger genomes provide proportionately more noncoding sites for the insertion of TEs with minimum host damage, the process once started will tend to feed-back on itself and produce larger and larger genomes unless opposed by genomewide selection and/or a mutational proclivity for DNA loss.

Lynch and Conery (2003) have pointed out the potential importance of nonadaptive processes in mediating increases in genome complexity, including increases in the abundance of spliceosomal introns and mobile elements. Under their model, negative selection against the deleterious effects of newly arising introns and insertions is expected to increase in intensity with increasing effective population size. Based on the negative relationship between population size and genome size in their dataset, they argued that TEs have a threshold genome size below which they are unable to become established, an intermediate range in which they are harbored by only a fraction of species, and an upper threshold (~100 Mb) above which all species are infected. On a broader scale, it should be noted that this particular analysis was based on a comparison of sequenced genomes, which were in almost all cases chosen specifically because they are small. It is not at all clear that genome size and population size would correlate strongly across broader samples, and this model therefore is unlikely to explain the extensive variation in genome size found in groups such as animals or plants (see Chapters 1 and 2).

Host Genome Structure

In contrast to mutator mechanisms that act to mediate only small DNA changes such as nucleotide substitutions, TEs can be responsible for both small and large structural changes in the genome, including deletions, inversions, duplications, and translocations. Indeed, TEs in *D. melanogaster* have been shown to have a major impact on genome structure via such processes (e.g., Montgomery *et al.,* 1991).

Two main mechanisms have been implicated in TE-induced karyotypic changes. The best-known mechanism is ectopic recombination, in which homologous recombination occurs between multiple copies of a TE present in a genome. The second mechanism for inducing genomic rearrangements is alternative transposition of Class II elements in bacteria, plants, and animals (Gray, 2000). Ectopic recombination is a meiotic process that is a significant source of the molecular rearrangements causing human disease (Reiter *et al.,* 1999).

Further, a study of mobilization rates of nine TEs in *D. melanogaster* indicated that most changes in restriction patterns were consistent with rearrangements, rather than with true transposition (Dominguez and Albornoz, 1996). A survey of ectopic recombination in the region flanking the *white* locus of *D. melanogaster* indicated that inter- and intrachromosomal recombinants were generated in about equal numbers.

Several species of *Drosophila* provide good examples of TE-induced genomic structural changes. During the evolution of this genus, the molecular organization of the major chromosomal elements has been repeatedly rearranged via the fixation of inversions, and the rate of chromosomal reshuffling appears to be higher than that of any other animal or plant taxon that has been similarly studied (Ranz *et al.,* 2001), with the notable exception of *C. elegans* (Coghlan and Wolfe, 2002). For example, detailed analysis of the breakpoints of large inversions in natural populations of *Drosophila buzzatii* (Caceres *et al.,* 2001; Casals *et al.,* 2003) demonstrated an unprecedented frequency and complexity of molecular rearrangements in the relatively short chromosomal regions surrounding the inversion breakpoints and the remarkable rapidity of these changes on an evolutionary timescale. The *foldback*-like element *Galileo* was found to be present at inversion breakpoints and is the most likely inducer of the inversions (Caceres *et al.,* 2001; Casals *et al.,* 2003).

The human genome appears to be relatively rich in segmental duplications that are produced by duplicative transposition between nonhomologous chromosomes. Some of these play an important role in human evolution and disease, and are prone to promote further rearrangements because of misalignment (Eichler, 2001). Examination of the junctions of a large number of human segmental duplications revealed that 27% terminated within an *Alu* sequence (Bailey *et al.,* 2003). Although several mechanisms are likely to be responsible for generating these duplications, the role of *Alu–Alu* recombination appears to be among the most important (Bailey *et al.,* 2003).

CONCLUDING REMARKS AND FUTURE PROSPECTS

Until quite recently, the majority of information on TEs was obtained from a restricted number of model organisms. Now that genome sequencing data are becoming increasingly available, it is of considerable interest to determine which TE characteristics commonly vary between species and which have properties in common across a broad range of taxa. Some of the questions to be addressed in this context are: What are the relative frequencies of different families and classes of TEs in different organisms? Why are different element classes differentially abundant in eukaryotic genomes? (See Table 3.2.)

A number of outstanding questions also relate to the relationship between TEs and their hosts. Among the most interesting are: How frequently are TEs found in coding and regulatory sequences of host genes? How do patterns of TE distribution vary within and between different TE families and different host species? What is the relative importance of TEs in determining genome size? Why are some genomes relatively streamlined, whereas others have expanded so greatly as a result of retroelement proliferation?

The propensity of TEs to undergo horizontal transfer is a fascinating aspect of their evolution, and a number of outstanding questions remain in this area, including: How frequent is TE horizontal transfer in nature? Why does TE horizontal transfer appear to be rarer in plants than in animals? Can TEs act as vehicles for ferrying host genes between species? And, most fundamentally, how are TEs transferred (vectored) between species? This last question is one for which very few answers are currently available. Possible vectors currently considered are viruses, bacteria, mites, parasitic wasps, and other parasites. It is possible that multiple vectors might be involved in any particular case, making this an especially challenging puzzle to unravel in the future.

Finally, there are questions of both theoretical and practical significance still to be addressed in the study of TEs. For example, it will be important from the standpoint of evolutionary theory to determine the relative roles of selection and chance in the evolution of TE families. From a pragmatic perspective, there is still a great deal to be learned about the relevance of TEs for genetic engineering in agriculture and medicine, and in creating risks for xenotransplantation (Bromham, 2002).

Although a number of aspects of the relationship between TEs and their hosts are still quite controversial, there does appear to be general agreement that these elements play an important role as mutagenic agents that provide new sources of genomic variation on which natural selection may act. An important aspect of their role as mutators in evolution is the broad spectrum of mutations produced by their activity. TE-induced genetic changes range from substitutions, deletions, and insertions of single nucleotides to modifications in the size and arrangement of entire genomes. TEs can produce small, silent changes that are detectable only at the DNA sequence level or may exert major effects on phenotypic traits. Indeed, the spectrum of TE-induced mutations is broader than that produced by any other mutator mechanism.

TEs produce their mutagenic effects not simply on initial insertion into host DNA, but also when they excise imprecisely, leaving either no identifying sequence or only small "footprints" of their previous presence. Of special evolutionary significance to their hosts may be TE-induced mutations that affect the regulatory sequences of the genome (Britten, 1997). Unfortunately, the identifiable properties of TE sequences present in the genomes of contemporary species of animals and plants may inadequately reflect the full range of elements that have been present in the past because of the rapid divergence of inactive elements.

Simple DNA base substitutions are well suited for the generation, diversification, and optimization of local protein space (Maeshiro and Kimura, 1998), but a hierarchy of natural mutational events is required for the rapid generation of protein diversity (Bogarad and Deem, 1999). Sequence shuffling has the potential to improve protein function significantly better than does point mutation alone. Because they are uniquely competent to reshuffle DNA sequences, TEs and viruses are important generators of the more complex types of mutations in the mutational hierarchy. The relative silence of plant TEs during normal development and their activation by stresses including wounding, pathogen attack, and cell culture provide additional interesting aspects of their mutational roles.

In summary, although many more studies are needed to flesh out the details, it is becoming increasingly accepted that TEs have the potential to exert a major influence on the evolution of their hosts. The characteristics that allowed TEs to be labeled "selfish DNA" might have allowed them to furnish genomes with the plasticity to evolve new mechanisms for generating genetic diversity. Thus it is probable that a balance between fidelity and exploration (Kidwell and Lisch, 2000) has evolved through the operation of natural selection and chance on the products of both recent and ancient interactions between TEs and their hosts. In a very real sense, TEs have played a prominent role in shaping the evolution of the genome.

REFERENCES

Adams MD, Celniker SE, Holt RA, *et al*. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.

Aparicio S, Chapman J, Stupka E, *et al*. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301–1310.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Bailey JA, Liu G, Eichler EE. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73: 823–834.

Banks JA, Masson P, Fedoroff N. 1988. Molecular mechanisms in the developmental regulation of the maize *Suppressor-mutator* transposable element. *Genes Dev* 2: 1364–1380.

Bartolome C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol* 19: 926–937.

Belfort M, Derbyshire V, Parker MM, *et al*. 2002. Mobile introns: pathways and proteins. In: Craig NL, Craigie R, Gellert M, Lambowitz AM eds, *Mobile DNA II*. Washington, DC: ASM Press, 761–783.

Berg DE, Howe MM. 1989. *Mobile DNA*. Washington, DC: American Society for Microbiology.

Biémont C, Tsitrone A, Vieira C, Hoogland C. 1997. Transposable element distribution in *Drosophila*. *Genetics* 147: 1997–1999.

Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of *P-M* hybrid dysgenesis: the role of the *P* element, a *P*-strain-specific transposon family. *Cell* 29: 995–1004.

Bingham PM, Levis R, Rubin GM. 1981. Cloning of DNA sequence from the *white* locus of *Drosophila melanogaster* by a novel and general method. *Cell* 25: 693–704.

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* 19: 2211–2225.

Bogarad LD, Deem MW. 1999. A hierarchical approach to protein molecular evolution. *Proc Natl Acad Sci USA* 96: 2591–2595.

Britten RJ. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* 205: 177–182.

Bromham L. 2002. The human zoo: endogenous retroviruses in the human genome. *Trends Ecol Evol* 17: 91–97.

Brookfield JFY. 1995. Transposable elements as selfish DNA. In: Sherratt DJ ed. *Mobile Genetic Elements*. Oxford: IRL Press, 130–153.

Brosius J. 1999a. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107: 209–238.

Brosius J. 1999b. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238: 115–134.

Bukhari AI, Shapiro JA, Adhya SL. 1977. *DNA Insertion Elements, Plasmids, and Episomes*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Bureau TE, Ronald PC, Wessler SR. 1996. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc Natl Acad Sci USA* 93: 8524–8529.

Burke WD, Malik HS, Lathe WC, Eickbush TH. 1998. Are retrotransposons long-term hitchhikers? *Nature* 392: 141–142.

Bushman F. 2002. *Lateral DNA Transfer: Mechanisms and Consequences*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Caceres M, Puig M, Ruiz A. 2001. Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. *Genome Res* 11: 1353–1364.

Cameron JR, Loh EY, Davis RW. 1979. Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* 16: 739–751.

Campbell A, Berg DE, Botstein D, *et al*. 1977. Nomenclature of transposable elements in prokaryotes. In: Bukhari AI, Shapiro JA, Adhya SL eds. *DNA Insertion Elements, Plasmids, and Episomes*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Capy P, Bazin C, Higuet D, Langin T. 1997a. *Dynamics and Evolution of Transposable Elements*. Austin, TX: Landes Bioscience.

Capy P, Langin T, Higuet D, *et al*. 1997b. Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* 100: 63–72.

Casals F, Caceres M, Ruiz A. 2003. The foldback-like transposon *galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. *Mol Biol Evol* 20: 674–685.

Charlesworth B, Langley CH, Sniegowski PD. 1997. Transposable element distributions in *Drosophila*. *Genetics* 147: 1993–1995.

Charlesworth B, Lapid A, Canada D. 1992. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II. Inferences on the nature of selection against elements. *Genet Res* 60: 115–130.

Clark JB, Kim P, Kidwell MG. 1998. Molecular evolution of *P* transposable elements in the genus *Drosophila*. III. The *melanogaster* species group. *Mol Biol Evol* 15: 746–755.

Coghlan A, Wolfe KH. 2002. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res* 12: 857–867.

Cooley L, Kelley R, Spradling A. 1988. Insertional mutagenesis of the *Drosophila* genome with single *P* elements. *Science* 239: 1121–1128.

Craig NL, Craigie R, Gellert M, Lambowitz AM. 2002. *Mobile DNA II*. Washington, DC: ASM Press.

Dasilva C, Hadji H, Ozouf-Costaz C, *et al*. 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. *Proc Natl Acad Sci USA* 99: 13636–13641.

Dernburg AF, Karpen GH. 2002. A chromosome RNAissance. *Cell* 111: 159–162.
Dominguez A, Albornoz J. 1996. Rates of movement of transposable elements in *Drosophila melanogaster*. *Mol Gen Genet* 251: 130–138.
Doolittle WF. 1997. Why we still need basic research. *Ann R Coll Phys Surg Can* 30: 76–80.
Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603.
Duret L, Marais G, Biemont C. 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* 156: 1661–1669.
Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* 17: 661–669.
Eickbush TH, Malik HS. 2002. Origin and evolution of retrotransposons. In: Craig NL, Cragie R, Gellert M, Lambowitz AM eds. *Mobile DNA II*. Washington, DC: ASM Press, 1111–1144.
Emmons SW, Yesner L, Ruan KS, Katzenberg D. 1983. Evidence for a transposon in *Caenorhabditis elegans*. *Cell* 32: 55–65.
Feschotte C, Mouches C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol Biol Evol* 17: 730–737.
Feschotte C, Zhang X, Wessler SR. 2002. Miniature inverted repeat transposable elements and their relationship to established DNA transposons. In: Craig NL, Cragie R, Gellert M, Lambowitz AM eds. *Mobile DNA II*. Washington, DC: ASM Press, 1147–1158.
Fincham JRS, Sastry GRK. 1974. Controlling elements in maize. *Annu Rev Genet* 8: 15–50.
Fink G, Farabaugh P, Roeder G, Chaleff D. 1981. Transposable elements (*Ty*) in yeast. *Cold Spring Harb Symp Quant Biol* 45 Pt 2: 575–580.
Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet* 5: 103–107.
Finnegan DJ, Rubin GM, Young MW, Hogness DS. 1978. Repeated gene families in *Drosophila melanogaster*. *Cold Spring Harb Symp Quant Biol* 42: 1053–1063.
Flavell AJ, Pearce SR, Heslop-Harrison P, Kumar A. 1997. The evolution of *Ty1-copia* group retrotransposons in eukaryote genomes. *Genetica* 100: 185–195.
Furano AV, Duvernell DD, Boissinot S. 2004. *L1 (LINE-1)* retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* 20: 9–14.
Galagan JE, Calvo SE, Borkovich KA, *et al*. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422: 859–868.
Gerasimova TI, Ladvishenko A, Mogila VA, *et al*. 1990. Transpositional bursts and chromosome rearrangements in unstable lines of *Drosophila*. *Genetika* 26: 399–411.
Glockner G, Szafranski K, Winckler T, *et al*. 2001. The complex repeats of *Dictyostelium discoideum*. *Genome Res* 11: 585–594.
Goodier JL, Ostertag EM, Kazazian HH. 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* 9: 653–657.
Gray YH. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* 16: 461–468.
Green MM. 1977. The case for DNA insertion mutations in *Drosophila*. In: Bukhari AI, Shapiro JA, Adhya SL eds. *DNA Insertions, Plasmids and Episomes*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 437–445.
Gregory TR. 2003. Is small indel bias a determinant of genome size? *Trends Genet* 19: 485–488.
Gregory TR. 2004. Insertion-deletion bias and the evolution of genome size. *Gene* 324: 15–34.
Guo H, Karberg M, Long M, *et al*. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* 289: 452–457.
Handler AM. 2001. A current perspective on insect gene transformation. *Insect Biochem Mol Biol* 31: 111–128.
Hartl DL, Lohe AR, Lozovskaya ER. 1997. Regulation of the transposable element *mariner*. *Genetica* 100: 177–184.

Henikoff S. 2002. Near the edge of a chromosome's black hole. *Trends Genet* 18: 165–167.
Hershberger RJ, Benito MI, Hardeman KJ, *et al*. 1995. Characterization of the major transcripts encoded by the regulatory *MuDR* transposable element of maize. *Genetics* 140: 1087–1098.
Hiraizumi Y. 1971. Spontaneous recombination in *Drosophila melanogaster* males. *Proc Natl Acad Sci USA* 68: 268–270.
Holmes I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. *Genome Res* 12: 1152–1155.
Holt RA, Subramanian GM, Halpern A, *et al*. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129–149.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
Ivics Z, Hackett PB, Plasterk RH, Izsvak, Z. 1997. Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell* 91: 501–510.
Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 19: 68–72.
Kajikawa M, Okada N. 2002. LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell* 111: 433–444.
Kalendar R, Tanskanen J, Immonen S, *et al*. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc Natl Acad Sci USA* 97: 6603–6607.
Kapitonov VV, Jurka J. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* 107: 27–37.
Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* 98: 8714–8719.
Kapitonov VV, Jurka J. 2003a. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci USA* 100: 6569–6574.
Kapitonov VV, Jurka J. 2003b. A novel class of SINE elements derived from 5S rRNA. *Mol Biol Evol* 20: 694–702.
Kaplan N, Darden T, Langley CH. 1985. Evolution and extinction of transposable elements in Mendelian populations. *Genetics* 109: 459–480.
Kidwell MG. 1993. Lateral transfer in natural populations of eukaryotes. *Annu Rev Genet* 27: 235–256.
Kidwell MG. 1994. The evolutionary history of the *P* family of transposable elements. *J Heredity* 85: 339–346.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115: 49–63.
Kidwell MG, Kidwell JF. 1975. Cytoplasm-chromosome interactions in *Drosophila melanogaster*. *Nature* 253: 755–756.
Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci USA* 94: 7704–7711.
Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. *Trends Ecol Evol* 15: 95–99.
Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55: 1–24.
Kidwell MG, Lisch DR. 2002. Transposable elements as sources of genomic variation. In: Craig NL, Cragie R, Gellert M, Lambowitz AM eds. *Mobile DNA II*. Washington, DC: ASM Press, 59–90.
Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics* 36: 813–833.
Kim JM, Vanguri S, Boeke JD, *et al*. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res* 8: 464–478.

Kimura RH, Choudary PV, Schmid CW. 1999. Silk worm *Bm1* SINE RNA increases following cellular insults. *Nucleic Acids Res* 27: 3380–3387.

Kinsey JA, Helber J. 1989. Isolation of a transposable element from *Neurospora crassa*. *Proc Natl Acad Sci USA* 86: 1929–1933.

Klinakis AG, Zagoraiou L, Vassilatis DK, Savakis C. 2000. Genome-wide insertional mutagenesis in human cells by the *Drosophila* mobile element *Minos*. *EMBO Rep* 1: 416–421.

Kreahling J, Graveley BR. 2004. The origins and implications of *Alu*ternative splicing. *Trends Genet* 20: 1–4.

Kumar A, Bennetzen JL. 1999. Plant retrotransposons. *Annu Rev Genet* 33: 479–532.

Lerman DN, Michalak P, Helin AB, *et al*. 2003. Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. *Mol Biol Evol* 20: 135–144.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* 300: 1288–1291.

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* 409: 847–849.

Lohe AR, Hartl DL. 1996. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Mol Biol Evol* 13: 549–555.

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. *Mol Biol Evol* 12: 62–72.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302: 1401–1404.

Maeshiro T, Kimura M. 1998. The role of robustness and changeability on the origin and evolution of genetic codes. *Proc Natl Acad Sci USA* 95: 5088–5093.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 10: 1307–1318.

Maside X, Bartolome C, Assimacopoulos S, Charlesworth B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: *in situ* hybridization vs Southern blotting data. *Genet Res* 78: 121–136.

McClintock B. 1952. Chromosome organization and gene expression. *Cold Spring Harb Symp Quant Biol* 16: 13–47.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* 226: 792–801.

McClure MA. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol Biol Evol* 8: 835–856.

McDonald JF. 1998. Transposable elements, gene silencing and macroevolution. *Trends Ecol Evol* 13: 94–95.

Meyers BC, Tingey SV, Morgante M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res* 11: 1660–1676.

Miller WJ, McDonald JF, Nouaud D, Anxolabehere D. 1999. Molecular domestication—more than a sporadic episode in evolution. *Genetica* 107: 197–207.

Miller WJ, McDonald JF, Pinsker W. 1997. Molecular domestication of mobile elements. *Genetica* 100: 261–270.

Miller WJ, Nagel A, Bachmann J, Bachmann L. 2000. Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. *Mol Biol Evol* 17: 1597–1609.

Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* 129: 1085–1098.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.

Nekrutenko A, Li WH. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 17: 619–621.

Nikaido M, Nishihara H, Hukumoto Y, Okada N. 2003. Ancient SINEs from African endemic mammals. *Mol Biol Evol* 20: 522–527.

Pardue ML, Danilevskaya ON, Traverse KL, Lowenhaupt K. 1997. Evolutionary links between telomeres and transposable elements. *Genetica* 100: 73–84.

Pardue ML, DeBaryshe PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* 37: 485–511.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* 17: 23–28.

Plasterk RA. 1995. Mechanisms of DNA transposition. In: Sherratt DJ ed. *Mobile Genetic Elements*. Oxford: IRL Press, 18–37.

Prak ET, Kazazian HH. 2000. Mobile elements and the human genome. *Nat Rev Genet* 1: 134–144.

Purugganan M, Wessler S. 1992. The splicing of transposable elements and its role in intron evolution. *Genetica* 86: 295–303.

Rabinowicz PD, Palmer LE, May BP, *et al*. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res* 13: 2658–2664.

Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res* 11: 230–239.

Reiter LT, Liehr T, Rautenstrauss B, *et al*. 1999. Localization of *mariner* DNA transposons in the human genome by PRINS. *Genome Res* 9: 839–843.

Robertson HM, Preston CR, Phillis RW, *et al*. 1988. A stable genomic source of *P* element transposase in *Drosophila melanogaster*. *Genetics* 118: 461–470.

Rubin GM. 1983. Dispersed repetitive DNAs in *Drosophila*. In: Shapiro JA ed. *Mobile Genetic Elements*. New York: Academic Press, 329–361.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of *P-M* hybrid dysgenesis: the nature of induced mutations. *Cell* 29: 987–994.

Saedler H, Bonas U, Gierl A, *et al*. 1984. Transposable elements in *Antirrhinum majus* and *Zea mays*. *Cold Spring Harb Symp Quant Biol* 49: 355–361.

SanMiguel P, Gaut BS, Tikhonov A, *et al*. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* 20: 43–45.

SanMiguel P, Tikhonov A, Jin YK, *et al*. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765–768.

Schueler MG, Higgins AW, Rudd MK, *et al*. 2001. Genomic and genetic definition of a functional human centromere. *Science* 294: 109–115.

Selker EU. 1997. Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet* 13: 296–301.

Selker EU, Tountas NA, Cross SH, *et al*. 2003. The methylated component of the *Neurospora crassa* genome. *Nature* 422: 893–897.

Shapiro JA. 1983. *Mobile Genetic Elements*. New York: Academic Press.

Simmons MJ, Bucholz LM. 1985. Transposase titration in *Drosophila melanogaster*: a model of cytotype in the *P-M* system of hybrid dysgenesis. *Proc Natl Acad Sci USA* 82: 8119–8123.

Simmons MJ, Raymond JD, Grimes CD, *et al*. 1996. Repression of hybrid dysgenesis in *Drosophila melanogaster* by heat-shock-inducible sense and antisense *P*-element constructs. *Genetics* 144: 1529–1544.

Singer MF. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28: 433–434.

Sniegowski PD, Charlesworth B. 1994. Transposable element numbers in cosmopolitan inversions from a natural population of *Drosophila melanogaster*. *Genetics* 137: 815–827.

Spradling AC, Stern DM, Kiss I, *et al*. 1995. Gene disruptions using *P* transposable elements: an integral component of the *Drosophila* genome project. *Proc Natl Acad Sci USA* 92: 10824–10830.

Steinemann M, Steinemann S. 1998. Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of *Drosophila miranda* a model for sex chromosome evolution. *Genetica* 103: 409–420.

Temin HM. 1989. Reverse transcriptases. Retrons in bacteria. *Nature* 339: 254–255.

Tu Z. 1997. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc Natl Acad Sci USA* 94: 7475–7480.

Tu Z. 2004. Insect transposable elements. In: Gilbert L, Latrous K, Gill S eds. *Comprehensive Molecular Insect Science*. Oxford: Elsevier.

Vastenhouw NL, Plasterk RHA. 2004. RNAi protects the *Caenorhabditus elegans* germline against transposition. *Trends Genet* 20: 314–319.

Vieira C, Lepetit D, Dumont S, Biemont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol Biol Evol* 16: 1251–1255.

Waterston R, Sulston J. 1995. The genome of *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* 92: 10836–10840.

Wessler S, Baran G, Varagona M. 1987. The maize transposable element *Ds* is spliced from RNA. *Science* 237: 916–918.

Witherspoon DJ. 1999. Selective constraints on *P*-element evolution. *Mol Biol Evol* 16: 472–478.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 13: 335–340.

Yu J, Hu S, Wang J, *et al*. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.