
Computational Analysis and Paleogenomics of Interspersed Repeats in Eukaryotes

2

Cédric Feschotte and Ellen J. Pritham

Abstract

Interspersed repeats occupy a significant fraction of the genetic material and represent the single major component of large eukaryotic genomes. They result from the persistent activity and gradual accumulation of transposable elements (TEs), sequences that are able to replicate in virtually all organisms and that have been successfully maintained throughout the evolution of life. Despite their selfish nature, the movement and amplification of TEs have had an enormous impact on the evolution of genes and the dynamics of genomes. Improvements to the efficiency of DNA sequencing coupled with decreases in its associated costs have fueled the sequencing of dozens of eukaryotic genomes. This has resulted in the rapid accumulation of large quantities of DNA sequences in the public databases. As such, the identification and annotation of repeats has become an integral facet of genome biology and has provoked a shift from the study of single TEs to huge populations of elements. Here we review the approaches and methods by which TEs are identified, classified and analyzed in complete eukaryotic genome sequences. We provide examples illustrating how these processes greatly facilitate genome annotation and illuminate the extent of the role of TEs in the evolution of genomes and species.

An introduction to transposable elements

In a series of seminal studies initiated more than four decades ago, Britten, Davidson and colleagues demonstrated that the nuclear genome of diverse eukaryotes contained a large fraction of repetitive DNA (Britten and Kohne, 1968; Davidson *et al.*, 1974; Waring and Britten, 1966). This conjecture was first formulated in response to the observation that the reassociation of mammalian DNA occurred far more rapidly and imprecisely than would be expected given the size of the genome (Britten and Kohne, 1968; Waring and Britten, 1966). It was later confirmed by the identification and molecular cloning of representative sequences from two abundant families of repetitive elements in the human genome, *Alu* and L1 (Deininger and Schmid, 1976; Rubin *et al.*, 1980; Singer, 1982). Further genetic and molecular work made it increasingly evident

that a large fraction of most metazoan genomes is composed of repetitive DNA. More recently, large-scale genome sequencing has ascertain the ubiquitous nature of repetitive DNA and its prevalence in most eukaryotic species (for review, Brookfield, 2005; Feschotte *et al.*, 2002a; Kidwell, 2002). The human genome is a striking example: repetitive DNA accounts for more than half of the nuclear DNA and the aforementioned L1 and *Alu* repeat families alone occupy nearly one third of the genomic space (IHGSC, 2001; Smit, 1999).

Repetitive DNA can be divided into two distinct categories based on its organization within the genome: there are tandem and interspersed repeats. Tandem repeats typically are amplified through slippage during replication, unequal crossing-over and gene conversion and are composed mainly of satellite DNA and of simple sequence repeats, also referred to as mini- and microsatellites (for review, Schlotterer, 2000). The bulk of tandem repeats are primarily represented by centromeric and pericentromeric satellite DNA whose evolutionary origin and dynamic remain poorly understood (Ugarkovic and Plohl, 2002). These highly repetitive regions are essentially excluded from large-scale genome sequencing efforts due to technological limitations in their cloning, sequencing and assembly. Tandem repeats may account for a significant fraction of a given genome (e.g. ~15% of the human genome) but interspersed repeats generally represent the major component of the repetitive DNA (IHGSC, 2001). Most interspersed repeats are derived from mobile genetic elements or transposable elements (TEs). These are fragments of DNA that can move around and insert into new chromosomal locations and are often duplicated in the process thereby increasing their copy number (for a comprehensive review of transposition mechanisms, see Craig *et al.*, 2002). Thus, TEs or their remnants often represent the single largest component of eukaryotic genomes, accounting for 10% of the tiny genome of the nematode *C. elegans*, ~45% of the human genome and ~80% of the maize genome (Fig. 2.1).

Traditionally eukaryotic TEs have been divided into two major classes according to whether their transposition intermediate is RNA (class 1) or DNA (class 2) (Capy *et al.*, 1998; Finnegan, 1989). For class 1 elements, it is the element-encoded transcript (mRNA), and not the element itself (as with class 2 elements), that forms the transposition intermediate. Class 1 elements can be further divided into two major subclasses with distinct mechanisms of transposition (Eickbush and Malik, 2002). The short and long interspersed elements (SINEs and LINEs) are (retro)transposed by reverse transcription of mRNA directly into the site of integration. In contrast the long terminal repeat (LTR) retroelements are reverse transcribed from RNA intermediates within viral-like particles and are integrated as double-stranded DNA molecules. This last step of the transposition cycle of LTR retroelements is mechanistically similar to the integration of class 2 elements or DNA transposons. Class 2 elements transpose directly via a DNA intermediate through the so-called cut-and-paste mechanism: the element is excised from the chromosome via a staggered double stranded cleavage and reinserted elsewhere in the genome (Craig *et al.*, 2002).

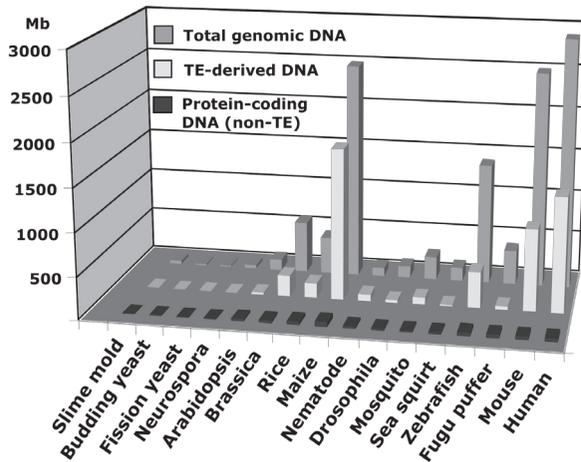


Figure 2.1 Genome size, amount of TE-derived DNA and amount of protein-coding DNA (non-TE proteins) in various eukaryotes with complete genome sequences available. The data was compiled either from the corresponding papers reporting draft genome sequences or from the following sources: *Brassica* (Zhang and Wessler, 2004) and X. Zhang, personal communication; rice (Jiang *et al.*, 2004) and N. Jiang, personal communication; maize (Meyers *et al.*, 2001); nematode, C.F., unpublished data. The species are: slime mold, *Dictyostelium discoideum*; budding yeast, *Saccharomyces cerevisiae*; fission yeast, *Schizosaccharomyces pombe*; *Neurospora crassa*; *Arabidopsis thaliana*; *Brassica oleracea*; rice, *Oryza sativa*; maize, *Zea mays*; nematode, *Caenorhabditis elegans*; *Drosophila melanogaster*; mosquito, *Anopheles gambiae*; sea squirt, *Ciona intestinalis*; zebrafish, *Danio rerio*; fugu pufferfish *Takifugu rubripes*; mouse, *Mus musculus*.

Besides the dramatic variation in total TE content among eukaryotic genomes (see Fig. 2.1), there is also tremendous variation in the relative representation of the different classes (class 1 vs. class 2) and subclasses (LTR vs. non-LTR) of TEs among species (Fig. 2.2). For example, the compact genome of the bakers' yeast contains only about 300 TEs and all are complete or fragmented copies of LTR-retrotransposons (Kim *et al.*, 1998). At the other end of the spectrum, the human genome harbors about 3 million copies of TEs and 3/4 can be classified as non-LTR retrotransposons (IHGSC, 2001). In contrast, class 2 DNA transposons are the most prevalent type of TEs in the rice and nematode genomes (Jiang *et al.*, 2004; Surzycki and Belknap, 2000). Other genomes, such as those of the plant *Arabidopsis thaliana* (AGI, 2000), the fungi *Neurospora crassa* (Galagan *et al.*, 2003) and the mosquito *Anopheles gambiae* (Holt *et al.*, 2002) contain intermediate quantities of TEs (~10–30% of the genomic DNA) with comparable proportions of DNA transposons, LTR and non-LTR retrotransposons (Fig. 2.2).

Dramatically different assortment of TEs can also occur among closely related species. This was recently illustrated through a comparative genomic analysis of TEs in four species of *Entamoeba* protozoans (Pritham *et al.*, 2005). The genomes of *Entamoeba histolytica* and *E. dispar*, two human parasites, contain

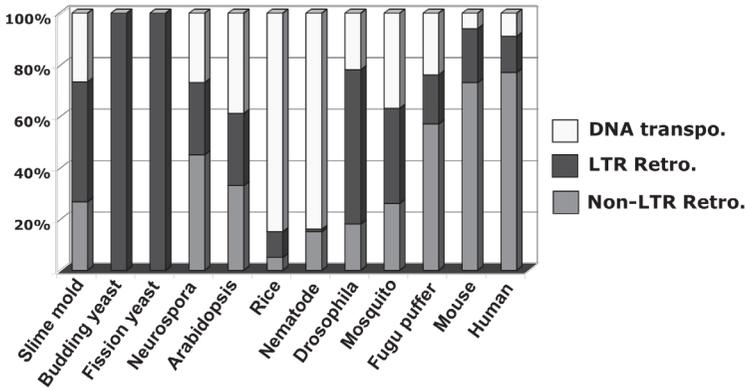


Figure 2.2 Variation of TE composition across eukaryotic genomes. For each species, the relative proportion of DNA transposons, non-LTR and LTR retrotransposons was calculated. The data were obtained as for [Figure 2.1](#).

many dozen non-LTR retrotransposons, some of which recently amplified, but only a handful of degenerate DNA transposons. The reptile parasite *E. invadens* and the free-living *E. moshkovskii* display the exact opposite assortment, with very few non-LTR retrotransposons but a plethora of recently amplified DNA transposons (Pritham *et al.*, 2005). The causes and mechanisms underlying the differential success of various TE types in different species remain a puzzling and vastly unanswered question (for discussion see also Brookfield, 2005; Feschotte, 2004; Kidwell, 2002; Zhang and Wessler, 2004).

An important and virtually universal feature of TEs, regardless of their classification, is that the large majority of the elements found in a genome at any given time are not capable of transposition (Feschotte *et al.*, 2002a; Smit, 1999). Indeed most of the elements are found as defective copies that cannot encode the proteins necessary for their movement. For some elements this is their “natural” form (non-autonomous elements which must rely on autonomous elements present elsewhere in the genome, e.g. SINEs) while for others, this is the result of sequence degradation or deletion at the time of insertion (“dead-on-arrival” elements, i.e. most LINE copies). Consequently, once integrated, most TEs will never transpose again and can be regarded as molecular fossils (Smit *et al.*, 1995).

Regardless of their origin and of the mechanisms responsible for their inactivation, it is widely accepted that fossilized TEs, as a whole, do not assume function to the host. Consequently, inactive TE copies are progressively eroded by mutations accumulating at a neutral rate until they become unrecognizable and are ultimately erased or removed from the genome. With a few exceptions (Doak *et al.*, 2003; Witherspoon *et al.*, 1997), this appears to be the general pattern of evolution of TEs irrespective of the host species (e.g. Petrov *et al.*, 2000; Robertson, 2002; MGSC, 2002). Given this fate, it is difficult to explain how

TEs have been maintained throughout billions of years of evolution and how they have managed to colonize the genomes of virtually every organism.

One explanation is that TEs are able to repeatedly and frequently invade genomes through nonsexual routes, via horizontal transfer (Capy *et al.*, 1998; Robertson, 2002). This mode of transmission is well documented for both class 1 and class 2 elements in several invertebrate (e.g. Jordan *et al.*, 1999; Robertson, 2002; Sanchez-Gracia *et al.*, 2005; Silva and Kidwell, 2000) and fungal species (Daboussi and Capy, 2003). Horizontal transmission has also been suggested to be responsible for the maintenance of active TEs in asexual organisms, such as the bdelloid rotifers (Arkhipova and Meselson, 2005). However neither for the bdelloid rotifers nor for many other eukaryotes, such as vertebrates, plants or protists, has horizontal transfer of TEs been unequivocally demonstrated.

As alternative mechanism to explain this paradox, one would have to evoke episodes of purifying selection acting on TEs and/or the rapid vertical diversification of the elements so that a reservoir of active elements can be perpetuated at low frequency in populations (Feschotte and Wessler, 2002; Lampe *et al.*, 2001; Le Rouzic and Capy, 2005). Also it cannot be ruled out that dead elements could somehow be reincarnated and provide the raw material to assemble new active founder copies. Supporting this provocative hypothesis is the well-established observation that many dead or nonautonomous elements are indeed able to proliferate for extensive evolutionary time despite their very ancient origin and coding incapacity (Han *et al.*, 2005; Johanning *et al.*, 2003). For both retrotransposons and DNA transposons, such amplification is accomplished via trans-activation by enzymes encoded by a limited number of distantly related autonomous elements coexisting within the same genome (Dewannieux *et al.*, 2003; Feschotte *et al.*, 2005; Kajikawa and Okada, 2002). This scenario may well explain the rise and recurrent explosion of MITEs and SINEs in several plant and animal genomes (Feschotte *et al.*, 2005; Feschotte *et al.*, 2003; Feschotte *et al.*, 2002b).

It is clear that the largest portion of most multicellular and even some single-celled eukaryotic genomes is home to an enormous TE graveyard. This decaying material is often referred to as “junk DNA” because of its lack of obvious function and usefulness for the organism. However, just like junk accumulating in an attic, TE-derived DNA might occasionally resurface and get reused in a novel and sometimes evolutionarily significant way. It is becoming increasingly clear that fossilized repeats represent a rich reservoir of raw sequence material (both coding and non-coding) for the emergence of *cis*-regulatory elements and for the assembly of new genes (e.g. Britten, 1996; Britten, 2004; Cordaux *et al.*, 2006; Kidwell and Lisch, 2001; Smit, 1999). Because of its repetitive nature, this material also provides an abundant substrate for large- and small-scale chromosomal rearrangements via illegitimate or unequal recombination events (Deininger *et al.*, 2003; Kazazian, 2004). As such, TEs are persistent and active players in the structural evolution and dynamics of eukaryotic genomes.

On practical grounds, TEs and other repeats present one of the most challenging obstacles to genome sequence assembly and annotation. On the other hand, the extreme abundance of TEs in some genomes and the observation that most are evolving under no selective constraint make them a useful fossil record to adjust molecular clocks and estimate neutral evolutionary rate in different species lineages (e.g. CSAC, 2005; MGSC, 2002). As such identifying and classifying interspersed repeats and understanding the mechanisms and forces driving TE amplification have become an essential facet of genome biology. In the following sections, we will first provide an overview of some of the existing computational methods to identify and classify repeats in large amount of genomic sequence data. We will then describe how repeat sequences can be used not only to extract crucial information regarding the evolutionary dynamics of the repeats themselves, but also those of the host genomes.

Identification and annotation of repeats

The problem of computational repeat identification and annotation in a given genome sequence can be broken down into three different tasks: (1) assembly of a repeat library representative of the species, (2) curation of the library (i.e. classification of the repeats into different classes and superfamilies) and (3) genome annotation (also called “repeat masking”). The assembly of the repeat library consists of defining all the repeats within the genome, demarcating the boundaries of each repeat, subdividing related repeats into families and reconstructing an accurate consensus sequence representative of each family. The curation of the repeat library involves inspecting the repeat family or its consensus for a variety of criteria to allow the assignment of the family to a particular class and/or superfamily. Genome repeat annotation includes the mapping (or masking) of all regions of the genome that correspond to a given repeat family on the basis of sequence similarity to one of the consensus sequences compiled in the library. In this section, we will give an overview of the principles and computational methods currently used to identify, classify and annotate the repeats in whole genome sequences.

Repeat identification

Two distinct approaches have been applied to the task of genomic repeat identification. The first relies on detecting similarities in genome sequences through comparisons to existing and previously described repeats and therefore is referred to here as a homology-based identification. In contrast, the second approach is the identification of repeats *de novo* based on their repetitive nature alone.

Homology-based identification

The identification of repeats is greatly facilitated through the use of homology-based techniques. Here “homology” is defined as the recognizable relationship between two sequences by genetic descent. In this process, the query genome sequence is compared to a database of nucleotide or protein sequences of repeats

and transposable elements previously described from other species. This is traditionally achieved using basic local alignment search tools (BLAST, see [Table 2.1](#)). For example, the program `blastn` is used to compare a nucleotide query to a nucleotide database (e.g. non-redundant database, nr at GenBank), while `blastx` allows comparison of a nucleotide query translated into the six reading frames to a protein database (e.g. Swiss-Prot/UniProt at EMBL). Although these popular databases contain most of the described repeats and their encoded proteins, homology-based searches can be made more sensitive and comprehensive by using a more specific database of “manually” curated repeat libraries. Such repeat databases are generally the result of years of meticulous mining of partial genome sequences and of the TE literature. The most comprehensive reference libraries are the Repbase libraries (Jurka *et al.*, 2005), which are compiled and maintained by the Genetic Information Research Institute (www.girinst.org). Currently, Repbase offers reference libraries for about two dozens of eukaryotic species. Other specialized databases are often dedicated to particular types of repeats or organisms ([Table 2.1](#)).

The major limitation of the homology-based approach is that only sequences with significant similarity to repeats previously described or deposited in Repbase will be identified. It could be a satisfying approach if the query sequence originates from a species that is very closely related to one for which a comprehensive library is available. Otherwise, homology-based searches will tend to reveal only those repeats that are relatively intact and contain protein-coding sequences similar to those typically found in known TEs (e.g. reverse transcriptase, transposase ... etc.). As our sequence coverage of the tree of life expands and the number of curated repeat databases increases, this approach will become increasingly efficient at identifying the protein-coding regions of most typical repeats.

De novo repeat identification

De novo strategies are designed to identify repeated DNA within a genome simply due to the repetitive nature sequences and not based on homology to repeats in a known library. Two basic approaches have been employed for *de novo* identification of repeats: query vs. query similarity searches and word counting/seed extension. The first relies on a self-comparison, for example an entire genome aligned to itself. This step generates a series of pairwise or local alignments (e.g. using BLAST or BLAST-like programs). Pairwise alignments are then converted into multiple alignments and clustering methods are used to group related sequences into families, based on a pre-set or user-defined threshold of similarity and alignment length (Bao and Eddy, 2002). For example, we usually refer to sequences as members of the same family when they share at least 85% nucleotide similarity over 75% of their length. Miropeats (Parsons, 1995), RepeatFinder (Volfovsky *et al.*, 2001) and RECON (Bao and Eddy, 2002) are examples of programs using this approach ([Table 2.1](#)).

One disadvantage of these programs is that they are computationally intensive, requiring vast memory capacity and anywhere from hours to days of processing, depending on the algorithm or heuristic used and on the size and complexity of the query sequence. Another frequent issue is the relative lack of sensitivity of the programs used for the initial self-comparison (e.g. BLAST) or subsequent problems related to the clustering methods, which often lead to imprecise definition of the ends of the repeats.

An alternative and increasingly popular approach involves word counting and seed extension. These methods bypass the need for whole genome alignments by building a set of repeat families starting with short sequence strings (seeds) representing repeats in the genome. These strings are progressively extended into a consensus sequence through dynamic comparisons of their multiple iterations in the query sequence. RepeatScout (Price *et al.*, 2005) and ReAS (Li *et al.*, 2005) are two recent programs that have been developed along those lines (Table 2.1).

In a direct comparison with the query vs. query RECON package, RepeatScout was reported to require less computational power and to be more sensitive than RECON, defining more accurately the biological boundaries of the repeats (Price *et al.*, 2005). Conveniently, several of these programs generate a consensus ancestral sequence for each repeat family identified in the query (RepeatScout, ReaS) and/or may provide the nucleotide positions of each repeat and the copy number per family (RECON). Some may also be able to distinguish tandem from interspersed repeat families. However, currently none of these programs permit a detailed biological classification of the repeats.

De novo repeat identification methods are powerful because they will generate the most comprehensive catalogue of repeats present within a query sequence without *a priori* knowledge of the characteristics and classification of the repeats. One major pitfall of these approaches is that they can only be applied to queries representing a considerable amount of sequence data (e.g. whole-genome shotgun sequences). A related problem is the identification of low-copy number repeat families (e.g. fewer than 10 copies). This is because of the increasing number of “false positive” (host-gene families and segmental duplications) as the copy number cut-off of the program is decreased. *De novo* repeat identification programs will also tend to miss or split composite repeats (i.e. repeats made of several independently repeated units) or families with highly variable structure.

Another persistent problem with *de novo* methods is a lack of accuracy in the definition of the biological boundaries (ends) of the repeats. One of the reasons for this is that local sequence alignments (e.g. those generated by BLAST in a query vs. query approach) may not correspond to the biological ends of the repeats because of degraded or truncated copies or because of segmental duplications covering more than one repeat. This causes a number of subsequent problems in clustering related sequences into families. Although some programs such as RECON have improved in their abilities to accurately detect

the biological termini of repeats (Bao and Eddy, 2002), this remains an inherently difficult problem for the downstream steps of genome annotation, and in particular for the automated classification of the repeats. In theory, short seed extension approaches such as those used by RepeatScout or ReaS should be less sensitive to this problem and generate more faithfully defined consensus sequences. Nevertheless, these programs still appear to suffer from the problem of repeat fragmentation, in particular when low coverage whole genome shotgun sequences are used as input (N. Ranganathan and C. F., unpublished observation).

Classification of the repeats

Recent advances in the development of *de novo* repeat identification tools have been fueled by the pressing need to automate the construction of repeat libraries to assist in the assembly of the multitude of genome sequencing projects. All of the programs described above, in particular when used in combination (Quesneville *et al.*, 2005), have the potential to greatly accelerate the construction of consensus repeat libraries and as such they should largely fulfill the needs for raw repeat masking and assisting in genome assembly. However, the output produced by programs such as RECON (Bao and Eddy, 2002) or RepeatScout (Price *et al.*, 2005) provide little if any biological information on the repeats themselves and therefore cannot be used to annotate and classify the repeats. At present, there is no program to completely automate this tedious yet biologically important phase of genome annotation.

Developing such software represents a very challenging issue in computational biology. Not only are there currently only a few experts whose knowledge of the elements can cover the extreme diversity of TEs, but also new groups of elements are continuously being discovered and existing groups being reclassified, merged, or subdivided. Consequently, the body of TE literature is enormous and rapidly expanding. As of September 2005 a single keyword search in Medline using the combination “(retro)transposable OR (retro)transposon” retrieves over 21 000 publications, including ~1500 review articles! Thus, synthesizing the diversity of TEs into a rational and comprehensive classification upon which one could develop an automatic system of repeat annotation is extremely difficult.

As mentioned earlier, TEs are divided into two classes depending on their transposition intermediate. Class 1 elements transpose via reverse-transcription of a transcribed RNA intermediate, while class 2 elements move directly through a DNA intermediate (Finnegan, 1989). The two classes are further divided into distinct subclasses and a plethora of clades and superfamilies using other structural features related to other aspects of their transposition mechanism. For example, non-LTR retrotransposons used a target-primed mechanism of transposition where reverse-transcription and integration are coupled, while LTR retrotransposons are reverse transcribed within a viral-like particle and the resulting cDNA integrated in the genome as a DNA intermediate.

Table 2.1 Useful databases and tools to analyze repeats

	Use	URL/download
<i>Databases</i>		
Repbase Update	Eukaryotic DNA repeats	http://www.girinst.org/repbase/index.html
IS finder	Insertion sequences (IS) from eubacteria and archaea	http://www-is.biotoul.fr/
ACLAME	Prokaryotic mobile genetic elements, including transposons, plasmids and phages	http://aclame.ulb.ac.be/
TIGR Plant Repeat DB	Collection of plant repeats organized by genera	http://www.tigr.org/tdb/e2k1/plant.repeats/
Retrobase	Depository for TE sequences identified by the Poulter lab	http://biocadmin.otago.ac.nz/retrobase/home.htm
HERVd	Human endogenous retrovirus database	http://herv.img.cas.cz/
ICTVdB	International committee on taxonomy of viruses	http://www.ncbi.nlm.nih.gov/ICTVdb/
<i>Programs/tools</i>		
BLAST and related	Nucleotide or protein sequence similarity searches	http://www.ncbi.nlm.nih.gov/BLAST/
WU-BLAST	Nucleotide or protein sequence similarity searches	http://BLAST.wustl.edu/
ClustalW	Multiple sequence alignment	http://www.cf.ac.uk/biosi/research/biosoft/Downloads/clustalw.html
PileUp	Multiple sequence alignment	http://helix.nih.gov/docs/gcg/pileup.html

Felsenstein's list	A profusion of links to various phylogeny programs	http://evolution.genetics.washington.edu/phylip/software.html
CENSOR	Repeatmasking	http://www.girinst.org/censor/index.php
RepeatMasker	Repeatmasking	http://www.repeatmasker.org/
RECON	<i>De novo</i> repeat identification (query vs. query)	http://www.genetics.wustl.edu/eddy/recon/
RepeatFinder	<i>De novo</i> repeat identification (query vs. query)	http://cbbcb.umd.edu/software/RepeatFinder/
REPuter	<i>De novo</i> repeat identification (query vs. query)	http://www.genomes.de/
Miropeats	<i>De novo</i> repeat identification (query vs. query)	http://www.genome.ou.edu/miropeats.html
RepeatScout	<i>De novo</i> repeat identification (seed extension)	http://repeatscout.bioprospects.org/
ReAS	<i>De novo</i> repeat identification (seed extension)	ReAS@genomics.org.cn
FindMITE	Structure-based identification of MITEs	http://jaketu.biochem.vt.edu/dl_software.htm
TRANSPO 1.0	Evolutionary analysis of MITEs and DNA transposons	http://aliggen.lsi.upc.es/
MAK	MITE analysis kit	http://perl.idmb.tamu.edu/mak.htm
Tu lab TE toolset	TE analysis toolset based on BLAST homology-based search	http://jaketu.biochem.vt.edu/download.html
LTR_STRUC	Structure-based identification of LTR retrotransposons	http://www.genetics.uga.edu/retrolab/data/LTR_Struct.html
TSDfinder	Identification of TE boundaries by locating TSD, primarily for L1 LINES	http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/

One of the problems with this scheme is that one needs to have sufficient knowledge of the transposition mechanism of the elements for accurate classification. However relatively few elements have been investigated functionally and therefore in many cases the classification is made on the basis of sequence similarity with members for which the transposition mechanism has been experimentally determined and on the presence of conserved signatures and motifs in coding sequences (Capy *et al.*, 1998). For example, the detection of coding sequences similar to reverse transcriptases in a given element strongly indicates that the element transpose through a RNA intermediate and thus should be classified as a class 1 element. Unfortunately, as mentioned above, a very large fraction of TEs are nonautonomous and/or degenerate copies that have only partial or no coding capacity and thus cannot be classified upon protein similarity searches. The classification of these elements (e.g. SINEs, MITEs) is more ambiguous as it generally relies on structural characteristics or weak sequence similarities limited to non-coding terminal regions important for transposition (Feschotte *et al.*, 2002b; Ohshima and Okada, 2005).

As an alternative to sequence similarity searches, a number of programs have been developed to mine particular group of TEs in genomic sequences based on their structural features and assist in their classification and characterization. Some of these programs are listed in [Table 2.1](#). For example, FindMITE is a C program designed to rapidly search a database for sequences that have the characteristics of MITEs (Tu, 2001). The program searches the database for inverted repeats flanked by user-defined direct repeats within a specified distance range. One advantage of FindMITE is that one does not necessarily need a large amount of input sequence data and long-sized contigs to identify representatives of high copy number MITE families in any given species. Indeed FindMITE program was used to identify eight different MITE families in the mosquito *Anopheles gambiae* based on a search of ~17 500 sequence tagged sites (STS) with an average size of only 829 bp (Tu, 2001). In fact, when a whole genome sequence is available, it is advisable to restrict the FindMITE search to a relatively small subsequence (depending on genome size and complexity) because the program is prone to generate a large number of false positives and requires a substantial amount of “visual” or semi-assisted downstream filtering. The analysis of candidate MITEs identified by FindMITE can be complemented and extended using TRANSPO (Santiago *et al.*, 2002) and MAK (Yang and Hall, 2003), two computational tools recently designed to assist in the evolutionary sequence analysis of MITE and other class 2 transposon families at the genome-wide level ([Table 2.1](#)).

Another alternative to traditional homology-based classification is the use of hidden Markov models (HMMs) as probabilistic models to detect and classify TEs in genomic sequences (Quesneville *et al.*, 2003). Interestingly, models based on nucleotide composition biases were able to reliably discriminate TEs from host genes and class 1 from class 2 elements in three distant species (*A. thaliana*, *C. elegans*, *D. melanogaster*). Furthermore, the HMM method could distinguish

the coding region of a TE from its non-coding region. These results suggest that this innovative approach could be very useful for the automated annotation of TE sequences. Recently, this method was combined with other tools to assemble a TE annotation pipeline that could reproduce faithfully a comprehensive annotation *de novo* of TEs in the *D. melanogaster* genome (Quesneville *et al.*, 2005).

Many repetitive and mobile elements may also be reticent to classification and annotation simply because they have not been characterized in sufficient detail or have not even been described as bona fide TEs. Some elements may simply lack defined common features or may not conform to any of the established rules of TE classification. This problem is illustrated by the recent discoveries of several atypical groups of TEs that remain difficult to classify despite their widespread distribution in eukaryotes. These include *Helitrons* (Kapitonov and Jurka, 2001), *Penelope*-like elements (Evgen'ev and Arkhipova, 2005), *Cryptons* (Goodwin *et al.*, 2003) and *Maverick* elements (Feschotte and Pritham, 2005). The transposition mechanisms of these elements are not fully understood so their classification remains uncertain.

The structure and coding capacity of these “exotic” elements suggest that some may represent novel TE classes or subclasses. For example, *Helitrons* appear to be related to a set of bacterial mobile elements and single-stranded DNA viruses that replicate through a rolling-circle mechanism (Kapitonov and Jurka, 2001). Thus the transposition of *Helitrons* likely involves displacement of a single-stranded DNA intermediate (Feschotte and Wessler, 2001). In contrast, the structural features and the conserved genes encoded by *Mavericks* indicate a possible evolutionary relationship with linear plasmids found in fungi and plants and with various double-stranded DNA viruses that infect animals, most clearly adenoviruses (Pritham *et al.*, 2006). The transposition cycle of *Mavericks* is unclear at this point, but there is indirect evidence that the elements are integrated as double-stranded DNA molecules, similarly to LTR retroelements and DNA transposons (Feschotte and Pritham, 2005; Pritham *et al.*, 2006). With the rapid growth of the databases and the accumulation of newly discovered types of elements, a revision of the current classification scheme has become necessary.

Masking of repeats in nucleotide sequences

Once a reference library of annotated repeats is assembled, it can be used to identify all sequences related to each repeat family in the genome sequence of interest. This process is also referred to as repeat masking because the nucleotide symbols identified as repetitive regions in the query sequence are replaced with “N” (or “X”) in the output sequence (Jurka, 1998). Subsequently, masked regions may be “ignored” or eliminated all together to facilitate the process of genome assembly, gene discovery and comparative genomic analyses. Repeat masking also allows related repeat sequences to be extracted automatically along with their positional information, which greatly assists in the study of their pattern of evolution, genomic distribution and so forth.

The efficiency and precision of genome masking strongly depends on the quality of the repeat library, both in terms of the accuracy of consensus sequences and of their faithful classification. Currently, the task of repeat annotation is overwhelmingly dominated by the use of RepeatMasker, a program designed nearly 10 years ago by Arian Smit and Phil Green. RepeatMasker remains unpublished to date, but an online server allows one to run the program on any query sequence. The latest version (3.1.0 at the time of this publication) of the program is also freely downloadable through the server (see [Table 2.1](#) for URL).

Sequence comparisons in RepeatMasker are typically performed by the program `cross_match` developed by Phil Green. This program uses an implementation of the Smith-Waterman-Gotoh algorithm, which is known for its high sensitivity (compared to BLAST, for example). RepeatMasker can also accommodate similarity searches powered by the WU-BLAST 2.0 package (Lopez *et al.*, 2003). Wu-BLAST offers less sensitivity than `cross_match`, but is less demanding in terms of computational power and processing time. The RepeatMasker output also includes a table describing the location (nucleotide positions), identity (repeat name and classification) as well as the length and nucleotide divergence of the aligned sequence to the consensus. This later feature is particularly useful to those interested in the biology of the repeats themselves, because it provides an instant estimation of the relative age of the repeat copy identified (see below). The online server of RepeatMasker compares the query sequence to the Repbase libraries, but a locally installed version can be used with any sequence library such as a repeat catalogue compiled by the user.

One person's junk is another person's treasure: interspersed repeats as markers of evolution

Despite the conundrum that repeats pose to genome assembly and annotation, the once popular view that all repeats are genomic junk has now become obsolete. Repeats provides a rich “fossil” record of ancient and recent genome history and represent an extraordinary source of information about molecular evolutionary processes in general. Many recent landmark publications illustrate how this “paleogenomic” record holds crucial clues about evolutionary events and forces acting to shape the eukaryotic genome (some examples are developed below). Because they are generally evolving under no selective constraint, interspersed repeats provide useful neutral markers for studying processes of mutation and selection. It is possible to recognize groups of TEs that arose at the same time and to follow their fates in different regions of the genome or in different species. Furthermore, most ancient TE insertions are very rarely excised (especially retrotransposons) and removed precisely and therefore they can be used as phylogenetic markers essentially free of homoplasy (Schmitz *et al.*, 2005; Shedlock and Okada, 2000). Note: homoplasy refers to a collection of phenomena (e.g. convergence) that leads to similarities in character states for reasons other than inheritance from a common ancestor.

Consensus sequences and ancestral repeat reconstruction

The consensus sequence is generated from a multiple alignment of a sample of interspersed copies assigned to the same TE family using a simple majority rule for each base position in the alignment (Fig. 2.3). The consensus may be further improved by manually correcting ambiguous or overlooked positions that could introduce stop codons and frameshifts in the ORF(s). Given that all TEs from a family arose from a single or few active TE copies (also known as source or master elements), the consensus can be considered a particularly accurate approximation of the sequence of the original master copy that gave rise to the family under consideration (Jurka and Milosavljevic, 1991; Smit *et al.*, 1995). This principle has been experimentally validated through the demonstration that the reconstructed consensus sequence of three different animal DNA transposase families restored their enzymatic activity *in vitro* and/or *in vivo* (Ivics *et al.*, 1997; Lampe *et al.*, 1996; Miskey *et al.*, 2003). Another early successful exploitation of the consensus approach was the resurrection of an extinct promoter for the human L1 retrotransposon based on a phylogenetically reconstructed consensus (Adey *et al.*, 1994).

An extension of the concept of consensus is that sequence comparisons of individual repeat copies to their consensus can be used to infer an approximate age of the repeat family, i.e. the time lapsed since transposition and insertion of the copies (Kapitonov and Jurka, 1996; Smit *et al.*, 1995). This is because at

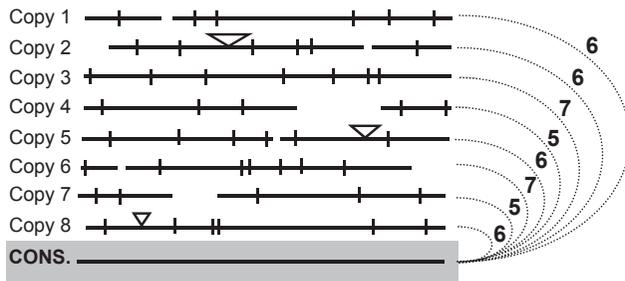


Figure 2.3 The concept of TE consensus sequence and its utility to infer the age of a TE family. In this schematic representation, eight related TE sequences are shown as thick horizontal lines. The consensus (CONS.) is deduced from a multiple alignment of TE sequences that vary at several positions. Most sequence variations occur at discrete positions, but some are also shared between copies. All shared and nonshared nucleotide changes may be used as phylogenetic characters to infer the relationship of the elements within the family (see Figure 2.4). Changes include single nucleotide substitutions (vertical ticks) and insertions (triangle above the sequence) and deletions (interrupted line) of variable size. The number of pairwise nucleotide substitutions to the consensus is shown for each copy on the right hand side of the figure. The average number of substitution to the consensus (here $n = 6$) can be used to infer the approximate age of the family if the neutral rate of nucleotide substitution per year is known for the host species. For example, if the neutral rate is 1 substitution per site per million year, the family depicted here would be ~6 million year old.

the time of insertion, each repeat was in theory identical to the ancestral active founder copy and therefore to the consensus. Following insertion, it is assumed that the different copies accumulate discrete nucleotide substitution at a neutral rate. Thus the average percent divergence of the copies to their consensus is positively correlated to the time elapsed since the bulk of the insertions occurred. The average percent divergence can be used to estimate the approximate age of the repeat family, provided that a neutral substitution rate per year has been determined for the lineage and timeframe under consideration.

This method was employed to determine the approximate age of each of the three million TE-derived repeats dispersed in the human genome (IHGSC, 2001). These elements fall into about 500 different families and subfamilies for which a consensus sequence was determined. For each family, the average sequence divergence to the consensus was calculated assuming equal frequency of all nucleotide substitutions but excluding CpG positions in the consensus since these tend to evolve several orders of magnitude faster than other dinucleotides. This data was used to estimate the age of each family based on three-way species comparisons (“relative rate test”) and fossil data to calibrate the molecular clock (IHGSC, 2001).

This analysis revealed several provocative findings. First, the majority (~55%) of the TEs recognizable in the human genome were inserted prior to the radiation of mammals, some 80–100 million year (myr) ago. Second, there was an exceptional burst of TE activity ~40 myr ago, largely attributable to a major wave of amplification of *Alu* SINEs. Since this peak, overall transpositional activity has rapidly tumbled in the hominid lineage. Indeed, only a handful of TE families (mostly members of the *Alu* and L1 groups) remain active in present human populations (Deininger *et al.*, 2003).

In contrast, a similar analysis shows that > 85% of the TEs present in the mouse genome were inserted specifically during the rodent radiation (MGSC, 2002). These insertions were therefore specific to the rodent lineage. This lineage-specific expansion of TEs now accounts for nearly one third (about 818 Mb) of the mouse genome. In addition, age distribution analysis of mouse TE families reveals that transposition activity has remained roughly constant over the last 100 myr and that dozens of diverse families are still active in this species (MGSC, 2002). This is in sharp contrast to the history and fate of TEs in the hominoid lineage, in agreement with the observation that the rate of spontaneous mutations caused by TE insertions is estimated to be ~60 times higher in mouse than in humans (Deininger *et al.*, 2003). The reasons for the radically different levels of TE activity in the human and mouse lineage remain unclear, but they could be in part explained by recent fluctuations in population size and different life history and genetics of the species (MGSC, 2002).

Phylogenetic analyses of transposable elements

The examples above illustrate how the history of large populations of TEs, which often represents the single largest component of eukaryotic genomes,

can be deciphered through a relatively simple comparison of the divergence of individual repeat sequences to their respective family consensus sequences. This approach can be complemented by phylogenetic analyses of separate families at the whole-genome level. Such an analysis allows one to infer possible relationships among members of the same family (e.g. existence and age of subfamilies) and to explore the amplification dynamics of the family (for review, Capy *et al.*, 1998; Feschotte *et al.*, 2002a).

In order to carry out phylogenetic analyses, it is first necessary to produce a multiple alignment of the TE sequences using programs such as ClustalX or PileUp (Table 2.1). The multiple sequence alignment is the raw material used for phylogenetic reconstruction, regardless of the method employed, so it is essential to obtain the most reliable alignment. Therefore it is often necessary to edit and refine “manually” (i.e. by eye) multiple alignments in order to obtain biologically significant phylogenetic trees. There are currently four major methods used to construct phylogenetic trees: (1) distance methods, such as Neighbor Joining, (2) maximum parsimony, (3) maximum likelihood, and (4) Bayesian analysis. It is beyond the scope of this review to describe the pros and the cons of each method (for more information see Felsenstein, 2004), but it is important to keep in mind that no single method is the best for all circumstances. Phylogenetic analyses can be carried out at the DNA level or at the protein level by aligning conceptual translations of protein coding DNA. This type of analysis has been used extensively to decipher the evolutionary history of TE families and to classify new elements based on clustering of elements with sequences representing known families or superfamilies in phylogenetic trees (for a list of references see Capy *et al.*, 1998; Feschotte *et al.*, 2002a). An example of the application is shown in Fig. 2.4.

In some instances, peculiar sequence or biological features of some TE groups can also be exploited to infer the evolutionary history of individual TE families or copies. For example, when a LTR retrotransposon inserts in the genome, its two LTRs are identical in sequence because both are copied from the same template during the reverse transcription of the element. As time passes, nucleotide changes accumulate in each LTR in a relatively random fashion. If the average (neutral) rate of nucleotide substitution per year is known for the host organism, then sequence divergence between the LTRs provides an estimate of when the insertion occurred. For example, this method has been applied to date the numerous insertions of LTR retrotransposons nested in the intergenic regions that surround the maize alcohol dehydrogenase 1 gene (SanMiguel *et al.*, 1998). This study reveals that several massive bursts of LTR retrotransposon activity were responsible for a doubling of the maize genome within the past 6 myr. A similar approach was used to determine the timing of various waves of retroviral integrations during primate evolution (for review, Sverdlov, 2000).

One of the major drawbacks of the methods and approaches outlined above to determine the age and evolutionary history of TE is that they strongly rely on sequence comparisons and alignments. Moreover, the age of the elements is

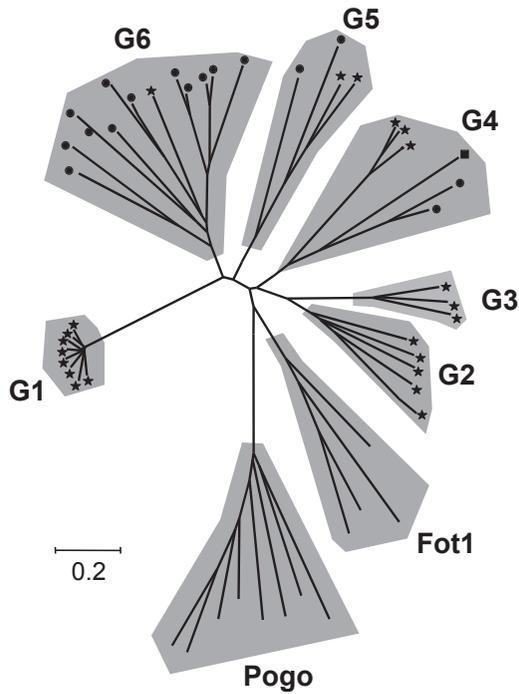


Figure 2.4 Phylogenetic analysis of TE sequences. This example shows an unrooted distance tree constructed using the neighbor-joining method using the MEGA3.1 software. The phylogeny is based on a multiple alignment of protein sequences from *Gemini* DNA transposons identified in the protozoans *Entamoeba moshkovskii* (represented by stars), *E. invadens* (dots) and *E. terrapinae* (square) and from transposases of the *pogo*/*Fot1* superfamily extracted from various animal, fungi and plant species. For a list of the sequences used for this alignment, see (Pritham *et al.*, 2005). For clarity, the origin and accession number of the sequences have been omitted. In distance tree, the branch length is proportional to the genetic distance between the sequences. The scale bar refers to 20% sequence divergence. The topology of the tree shown here emphasizes several aspects of TE dynamics. First, it shows the heterogeneity of *Gemini* sequences; they cluster into six distinct subfamilies (G1–6), while the *Fot1* and *pogo* transposases form separate groups. These groupings received strong statistical support in bootstrap analysis (not shown). Second, it illustrates that *Gemini* sequences are closer to the *Fot1* group than to the *pogo* group. Third, it reveals the tempo and succession of *Gemini* subfamily amplification. Three *Gemini* subfamilies contain sequences from only *E. moshkovskii* (G1–3) while the other three groups include a mix of sequences from two or three species. This, along with shorter branch lengths, indicates that sequences within groups G1 and G3 likely transposed in the *E. moshkovskii* lineage after its divergence from the other species, while the other sequences have a more ancient origin. Also, based on the branch lengths, elements within the G1 group amplified very recently in *E. moshkovskii* and are younger than elements of the G2 and G3 groups.

estimated based on the availability of a neutral molecular clock, with the overly simplistic assumptions that the clock ticks regularly over a given period of time and evenly across the genome and that most TE sequences are evolving under a neutral model of nucleotide substitution.

Concluding remarks

Improvements to the efficiency of DNA sequencing coupled with decreases in its associated costs have fueled the sequencing and annotation of hundreds of genomes. This has resulted in the rapid accumulation of huge quantities of sequences in the public databases. The comparison of closely related genomes has made clear the significance of the repetitive portion of the eukaryotic genome and revealed TEs as a major force shaping the structure of genes and genomes. As such, the identification and annotation of repeats has become an integral facet of genome biology and has recently benefited from the development of several powerful computational tools to automate these processes. Two major obstacles remain to further accelerate and improve repeat discovery and analysis. First, the programs used for *de novo* identification of repeats are performing poorly at the task of accurately defining the biological ends of some repeats. Second, there is an urgent need to develop software that can classify and annotate repeats identified *de novo*—a necessity for the automation of repeat library curation. This represents an exciting challenge for the years to come for both computational and transposon biologists because of the complexity and variety of TE structure and their intricate pattern of evolution. This would nonetheless be a rewarding task as the accurate annotation of interspersed repeats will allow for an improved assembly and annotation of genomes and will illuminate the extent of the role of TEs in the evolution of genomes and species.

Acknowledgments

Research in the Feschotte and Pritham laboratories is supported by start-up funds from the University of Texas at Arlington and by a Research Enhancement Program grant (to C.F.).

References

- Adey, N.B., Tollefsbol, T.O., Sparks, A.B., Edgell, M.H., and Hutchison, C.A. (1994). Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl. Acad. Sci. USA* 91, 1569–1573.
- Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.
- Arkipova, I., and Meselson, M. (2005). Deleterious transposable elements and the extinction of asexuals. *Bioessays* 27, 76–85.
- Bao, Z., and Eddy, S.R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276.
- Britten, R.J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. USA* 93, 9374–9377.
- Britten, R.J. (2004). Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc. Natl. Acad. Sci. USA* 101, 16825–16830.
- Britten, R.J., and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161, 529–540.
- Brookfield, J.F. (2005). The ecology of the genome—mobile DNA elements and their hosts. *Nat. Rev. Genet.* 6, 128–136.
- Capy, P., Bazin, C., Higuier, D., and Langin, T. (1998). *Dynamics and Evolution of Transposable Elements* (Austin, TX, Springer-Verlag).

- Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.
- Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. USA* 103, 8106–8110.
- Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. (2002). *Mobile DNA II* (Washington, DC, American Society for Microbiology Press).
- Daboussi, M.J., and Capy, P. (2003). Transposable elements in filamentous fungi. *Annu. Rev. Microbiol.* 57, 275–299.
- Daboussi, M.J., Daviere, J.M., Graziani, S., and Langin, T. (2002). Evolution of the *Fot1* transposons in the genus *Fusarium*: discontinuous distribution and epigenetic inactivation. *Mol. Biol. Evol.* 19, 510–520.
- Davidson, E.H., Graham, D.E., Neufeld, B.R., Chamberlin, M.E., Amenson, C.S., Hough, B.R., and Britten, R.J. (1974). Arrangement and characterization of repetitive sequence elements in animal DNAs. *Cold Spring Harb. Symp. Quant. Biol.* 38, 295–301.
- Deininger, P.L., Moran, J.V., Batzer, M.A., and Kazazian, H.H., Jr. (2003). Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13, 651–658.
- Deininger, P.L., and Schmid, C.W. (1976). An electron microscope study of the DNA sequence organization of the human genome. *J. Mol. Biol.* 106, 773–790.
- Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* 35, 41–48.
- Doak, T.G., Witherspoon, D.J., Jahn, C.L., and Herrick, G. (2003). Selection on the genes of *Euplotes crassus* *Tec1* and *Tec2* transposons: evolutionary appearance of a programmed frameshift in a *Tec2* gene encoding a tyrosine family site-specific recombinase. *Eukaryot. Cell* 2, 95–102.
- Eichler, E.E., and Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science* 301, 793–797.
- Eickbush, T.H., and Malik, H.S. (2002). Origins and evolution of retrotransposons. In: *Mobile DNA 2*, N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz, eds. (Washington, DC, ASM Press), pp. 1111–1144.
- Evgen'ev, M.B., and Arkhipova, I.R. (2005). *Penelope*-like elements—a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet. Genome Res.* 110, 510–521.
- Felsenstein, J. (2004). *Inferring Phylogenies* (Sunderland, MA, Sinauer Associates).
- Feschotte, C. (2004). *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Mol. Biol. Evol.* 21, 1769–1780.
- Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* 3, 329–341.
- Feschotte, C., Osterlund, M.T., Peeler, R., and Wessler, S.R. (2005). DNA-binding specificity of rice *mariner*-like transposases and interactions with *Stowaway* MITEs. *Nucleic Acids Res.* 33, 2153–2165.
- Feschotte, C., and Pritham, E.J. (2005). Non-mammalian *c*-integrase are encoded by giant transposable elements. *Trends Genet.* 21, 551–552.
- Feschotte, C., Swamy, L., and Wessler, S.R. (2003). Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. *Genetics* 163, 747–758.
- Feschotte, C., and Wessler, S.R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 98, 8923–8924.
- Feschotte, C., and Wessler, S.R. (2002). *Mariner*-like transposases are widespread and diverse in flowering plants. *Proc. Natl. Acad. Sci. USA* 99, 280–285.
- Feschotte, C., Zhang, X., and Wessler, S. (2002b). Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In: *Mobile DNA II*, N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz, eds. (Washington, DC, American Society for Microbiology Press), pp. 1147–1158.

- Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5, 103–107.
- Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., *et al.* (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422, 859–868.
- Goodwin, T.J., Butler, M.I., and Poulter, R.T. (2003). Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology* 149, 3099–3109.
- Han, K., Xing, J., Wang, H., Hedges, D.J., Garber, R.K., Cordaux, R., and Batzer, M.A. (2005). Under the genomic radar: the stealth model of Alu amplification. *Genome Res.* 15, 655–664.
- Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., *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, P.B., Plasterk, R.H., and Izsvak, Z. (1997). Molecular reconstruction of *Sleeping Beauty*, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* 91, 501–510.
- Jiang, N., Feschotte, C., Zhang, X.Y., and Wessler, S.R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* 7, 115–119.
- Johanning, K., Stevenson, C.A., Oyeniran, O.O., Gozal, Y.M., Roy-Engel, A.M., Jurka, J., and Deininger, P.L. (2003). Potential for retroposition by old Alu subfamilies. *J. Mol. Evol.* 56, 658–664.
- Jordan, I.K., Matyunina, L.V., and McDonald, J.F. (1999). Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. *Proc. Natl. Acad. Sci. USA* 96, 12621–12625.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 110, 462–467.
- Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human *Alu* genes. *J. Mol. Evol.* 32, 105–121.
- Kajikawa, M., and Okada, N. (2002). LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* 111, 433–444.
- Kapitonov, V., and Jurka, J. (1996). The age of Alu subfamilies. *J. Mol. Evol.* 42, 59–65.
- Kapitonov, V.V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA* 98, 8714–8719.
- Kazazian, H.H., Jr. (2004). Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632.
- Kidwell, M.G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115, 49–63.
- Kidwell, M.G., and Lisch, D.R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* 55, 1–24.
- Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* 8, 464–478.
- Lampe, D.J., Churchill, M.E., and Robertson, H.M. (1996). A purified mariner transposase is sufficient to mediate transposition *in vitro*. *EMBO J.* 15, 5470–5479.
- Lampe, D.J., Walden, K.K., and Robertson, H.M. (2001). Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. *Mol. Biol. Evol.* 18, 954–961.
- Le Rouzic, A., and Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169, 1033–1043.

- Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Yang, H., Yu, J., and Wong, G.K. (2005). ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.* *1*, e43.
- Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., and Gish, W. (2003). WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.* *31*, 3795–3798.
- Meyers, B.C., Tingey, S.V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* *11*, 1660–1676.
- Miskey, C., Izsvak, Z., Plasterk, R.H., and Ivics, Z. (2003). The Frog Prince: a reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. *Nucleic Acids Res.* *31*, 6873–6881.
- Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* *420*, 520–562.
- Ohshima, K., and Okada, N. (2005). SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res.* *110*, 475–490.
- Parsons, J.D. (1995). Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* *11*, 615–619.
- Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. (2000). Evidence for DNA loss as a determinant of genome size. *Science* *287*, 1060–1062.
- Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics* *21 Suppl 1*, i351–i358.
- Pritham, E.J., Feschotte, C., and Wessler, S.R. (2005). Unexpected diversity and differential success of DNA transposons in four species of *Entamoeba* protozoans. *Mol. Biol. Evol.* *22*, 1751–1763.
- Pritham EJ, Putliwala T and Feschotte C (2006). *Mavericks*, a new class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene in press*
- Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* *1*, 166–175.
- Quesneville, H., Nouaud, D., and Anxolabehere, D. (2003). Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. *J. Mol. Evol.* *57 Suppl 1*, S50–59.
- Robertson, H.M. (2002). Evolution of DNA transposons. In: *Mobile DNA II*, N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz, eds. (Washington, DC, American Society for Microbiology Press), pp. 1093–1110.
- Rubin, C.M., Houck, C.M., Deininger, P.L., Friedmann, T., and Schmid, C.W. (1980). Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. *Nature* *284*, 372–374.
- Sanchez-Gracia, A., Maside, X., and Charlesworth, B. (2005). High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet.* *21*, 200–203.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* *20*, 43–45.
- Santiago, N., Herraiz, C., Goni, J.R., Messeguer, X., and Casacuberta, J.M. (2002). Genome-wide analysis of the emigrant family of MITES of *Arabidopsis thaliana*. *Mol. Biol. Evol.* *19*, 2285–2293.
- Schlotterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* *109*, 365–371.
- Schmitz, J., Roos, C., and Zischler, H. (2005). Primate phylogeny: molecular evidence from retrotransposons. *Cytogenet. Genome Res.* *108*, 26–37.
- Shedlock, A.M., and Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *Bioessays* *22*, 148–160.
- Silva, J.C., and Kidwell, M.G. (2000). Horizontal transfer and selection in the evolution of P element. *Mol. Biol. Evol.* *17*, 1542–1557.
- Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* *9*, 657–663.
- Smit, A.F., Toth, G., Riggs, A.D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* *246*, 401–417.

- Surzycki, S.A., and Belknap, W.R. (2000). Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl. Acad. Sci. USA* 97, 245–249.
- Sverdlov, E.D. (2000). Retroviruses and primate evolution. *Bioessays* 22, 161–171.
- Tu, Z. (2001). Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA* 98, 1699–1704.
- Ugarkovic, D., and Plohl, M. (2002). Variation in satellite DNA profiles—causes and effects. *EMBO J.* 21, 5955–5959.
- Volfovsky, N., Haas, B.J., and Salzberg, S.L. (2001). A clustering method for repeat analysis in DNA sequences. *Genome Biol.* 2, RESEARCH0027.
- Waring, M., and Britten, R.J. (1966). Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* 154, 791–794.
- Wessler, S.R., Bureau, T.E., and White, S.E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* 5, 814–821.
- Witherspoon, D.J., Doak, T.G., Williams, K.R., Seegmiller, A., Seger, J., and Herrick, G. (1997). Selection on the protein-coding genes of the TBE1 family of transposable elements in the ciliates *Oxytricha fallax* and *O. trifallax*. *Mol. Biol. Evol.* 14, 696–706.
- Yang, G., and Hall, T.C. (2003). MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res.* 31, 3659–3665.
- Zhang, X., and Wessler, S.R. (2004). Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proc. Natl. Acad. Sci. USA* 101, 5589–5594.

