Genes Genet. Syst. (2019) **94**, p. 233–252

Structural and sequence diversity of eukaryotic transposable elements

Kenji K. Kojima[1,2*]

[1]*Genetic Information Research Institute, 465 Fairchild Drive, Suite 201, Mountain View, CA 94043, USA*
[2]*Department of Life Sciences, National Cheng Kung University, No. 1, Daxue Road, East District, Tainan 701, Taiwan*

(Received 10 May 2018, accepted 12 July 2018; J-STAGE Advance published date: 9 November 2018)

The majority of eukaryotic genomes contain a large fraction of repetitive sequences that primarily originate from transpositional bursts of transposable elements (TEs). Repbase serves as a database for eukaryotic repetitive sequences and has now become the largest collection of eukaryotic TEs. During the development of Repbase, many new superfamilies/lineages of TEs, which include *Helitron*, *Polinton*, *Ginger* and *SINEU*, were reported. The unique composition of protein domains and DNA motifs in TEs sometimes indicates novel mechanisms of transposition, replication, anti-suppression or proliferation. In this review, our current understanding regarding the diversity of eukaryotic TEs in sequence, protein domain composition and structural hallmarks is introduced and summarized, based on the classification system implemented in Repbase. Autonomous eukaryotic TEs can be divided into two groups: Class I TEs, also called retrotransposons, and Class II TEs, or DNA transposons. Long terminal repeat (LTR) retrotransposons, including endogenous retroviruses, non-LTR retrotransposons, tyrosine recombinase retrotransposons and *Penelope*-like elements, are well accepted groups of autonomous retrotransposons. They share reverse transcriptase for replication but are distinct in the catalytic components responsible for integration into the host genome. Similarly, at least three transposition machineries have been reported in eukaryotic DNA transposons: DDD/E transposase, tyrosine recombinase and HUH endonuclease combined with helicase. Among these, TEs with DDD/E transposase are dominant and are classified into 21 superfamilies in Repbase. Non-autonomous TEs are either simple derivatives generated by internal deletion, or are composed of several units that originated independently.

Key words: transposase, transposon, Repbase, retrotransposon, reverse transcriptase

INTRODUCTION

Transposable elements (TEs), also known as transposons, mobile DNA, or mobile elements, include a variety of DNA segments that can, in a process called transposition, move (or duplicate) from one location in the genome to another.

Repbase was first established as a database of human repeat sequences in 1992 (Jurka et al., 1992). Now, Repbase contains diverse eukaryotic repeat sequences that are categorized by organism and repeat type (Bao et al., 2015). In the development of Repbase, two things became clear. First, the majority of eukaryotic interspersed repeat sequences are originated from TEs, which are active now or were active in the past. The majority of *Medium reiterated repeats* families found in the human genome have been classified into various TE superfamilies (Kojima, 2018a). The origins of these interspersed repeats were not initially obvious. Eukaryotic repeat sequences not derived from TEs are microsatellites, satellite repeats arrayed in tandem, multicopy genes (such as ribosomal RNA genes), histone genes, and occasionally integrated viruses (Bao et al., 2015).

Second, the mechanisms and components responsible for transposition vary among TEs. Repbase contributed significantly to reveal the diversity of TEs. Many TE superfamilies were described by the team at the Genetic Information Research Institute (GIRI) who

Edited by Kenji Ichiyanagi
* Corresponding author. E-mail: kojima@girinst.org
DOI: http://doi.org/10.1266/ggs.18-00024

have maintained and expanded Repbase (Bao et al., 2015). The discovery of *Helitron* opened a new window in the world of TE studies because this superfamily encodes a unique protein set (Kapitonov and Jurka, 2001). Characterization of the superfamily of gigantic TEs, *Polinton*, allowed us to create a vague boundary between TEs and viruses (Kapitonov and Jurka, 2006; Krupovic et al., 2014a). Some recently characterized groups of TEs include *Ginger1*, *Ginger2* (Bao et al., 2010), *Dada* (Kojima and Jurka, 2013b) and *SINEU* (Kojima, 2015). In addition, studies outside Repbase cannot be neglected. Recent examples that were characterized by other teams are *Zisupton* (Böhne et al., 2012), *Spy* (Han et al., 2014) and *Teratorn* (Inoue et al., 2017).

After transposition, many types of TEs are flanked by short (1–20 bp) direct repeats called target site duplications (TSDs), which are derived from the target sequence (Kapitonov and Jurka, 2008). However, certain TE types, such as *Helitron*, some terminal inverted repeat (TIR)-bearing TEs, and *CR1* retrotransposons, do not produce TSDs. The length of a TSD is usually characteristic of the TE's group and its relatives, but may also vary across groups in a specific superfamily. TEs constitute the majority of repetitive sequences in most eukaryotic genomes. In fact, TEs can be viewed as intra-genomic parasites. Some viruses, such as retroviruses, behave like TEs. TEs also have diverse evolutionary impacts on their host genome.

The aim of this review is to introduce and summarize our present understanding of the diversity in eukaryotic TEs in sequence, protein domain composition, as well as structural hallmarks that include TSDs or terminal signatures (long terminal repeats, terminal inverted repeats, polyA tail, etc.). I focus on protein domain composition because (1) it is tightly related to the mechanism of transposition, and (2) it can be easily detected by bioinformatics analysis during the initial characterization of TEs.

TE CLASSIFICATION BASED ON BIOINFORMATICS

The concept that the highest rank of classification in TEs is linked to the mechanism of mobilization is well accepted. Historically, eukaryotic TEs are divided into two classes: Class I and Class II (Finnegan, 1989). Despite several objections by critiques (Piégu et al., 2015; Arensburger et al., 2016), this simple classification has worked very well to date. Class I includes retrotransposons, which transpose through an RNA intermediate. Because reverse transcriptase (RT) is the only enzyme that can efficiently catalyze reverse transcription, all autonomous retrotransposons encode RT. Class II includes DNA transposons, which do not use RNA as transposition intermediates. In other words, Class I includes all transposons that encode RT and their non-autonomous derivatives, while Class II includes all other autonomous transposons that lack RT and their non-autonomous derivatives.

Class I is subdivided into two large categories that are distinguished by the presence of long terminal repeats (LTRs): LTR retrotransposons and non-LTR retrotransposons. Recent studies have revealed additional groups of eukaryotic retrotransposons that are distinguishable from these two by the transposition mechanism and/or the phylogeny of RT. They are *DIRS* retrotransposons (or tyrosine recombinase-encoding retrotransposons) (Glöckner et al., 2001; Poulter and Goodwin, 2005) and *Penelope*-like retrotransposons (*Penelope*-like elements) (Arkhipova et al., 2003). It should be mentioned that even though *DIRS* is the abbreviation of *Dictyostelium* intermediate repeat sequence, retrotransposons related to *DIRS* have been found in diverse species and the term *DIRS* is now used as the name of a group whose members show similar protein domain composition. In this review, names representing a superfamily or group are not shown as abbreviations to avoid confusion about their distribution. These four groups are distinct in the origins of the catalytic components (endonuclease or recombinase) that are responsible for integration into the host genome. In the classification implemented in Repbase, *DIRS* retrotransposons are included in LTR retrotransposons and *Penelope*-like retrotransposons in non-LTR retrotransposons. Currently, this expedient classification was primarily introduced for practical reasons to avoid over-subclassification, and it does not mean that Repbase ignores the unique properties of *DIRS* and *Penelope*-like retrotransposons.

Due to the lack of any conserved protein domains among DNA transposons, the classification of DNA transposons is less widely accepted than that of retrotransposons. The machinery of transposition is the framework for classification of TEs. In general, the machinery is tightly linked to the composition of the protein domains encoded by TEs. When considering eukaryotic and prokaryotic TEs together, the transposases encoded by DNA transposons are classified into six types: DDD/E transposase, DEDD transposase, tyrosine recombinase (YR), serine recombinase (SR), HUH nuclease and Cas1 nuclease (Siguier et al., 2006; Chandler et al., 2013; Krupovic et al., 2014b). Among these, DEDD transposase, SR and Cas1 nuclease have not been found in any eukaryotic TEs. YR is encoded by *Crypton* (Goodwin et al., 2003; Kojima and Jurka, 2011a), while HUH nuclease is encoded by *Helitron* (Kapitonov and Jurka, 2001). All other groups of eukaryotic DNA transposons are thought to encode DDD/E transposase.

Table 1 is a brief comparison between the classification systems of Repbase (Bao et al., 2015), Wicker et al. (2007), and Arkhipova (2017). They are largely consistent, except for several minor conflicts. It is notewor-

Table 1. Classification of eukaryotic autonomous TEs based on the combination of protein domains

Polymerase	Nuclease/ recombinase	Superfamilies/clades in Repbase	Wicker (2007)	Arkhipova (2017)	Common name
RT	APE	*L1*, *Proto1*, *Tx1*, *Proto2*, *RTE*, *RTEX*, *RTETP*, *I*, *Nimb*, *Ingi*, *Vingi*, *Tad1*, *Loa*, *R1*, *Outcast*, *Jockey*, *CR1*, *L2*, *L2A*, *L2B*, *Kiri*, *Rex1*, *Crack*, *Daphne*, *Ambal*	RI_	RA	Non-LTR retrotransposon (LINE)
RT	APE + RLE	*Dualen*	n/d	n/d	*Dualen*
RT	RLE	*CRE*, *NeSL*, *R4*, *R2*, *HERO*	RIR	RP	Non-LTR retrotransposon
RT	GIY-YIG	*Penelope*	RPP	RG	*Penelope*-like element
RT	–	*Penelope*	n/d	RO	*Athena*, *Coprina*
RT	DDE	*Copia*, *BEL*, *Gypsy*, *endogenous retrovirus*	RL_	RD	LTR retrotransposon
RT	YR	*DIRS*	RY_	RY	*DIRS*
–	DDE	*Academ*, *Dada*, *EnSpm*, *Ginger1*, *Ginger2*, *Harbinger*, *hAT*, *IS3EU*, *ISL2EU*, *Kolobok*, *Mariner*, *Merlin*, *MuDR*, *Novosib*, *P*, *piggyBac*, *Sola*, *Transib*, *Zator*, *Zisupton*	DT_	DD	DNA transposon
–	YR	*Crypton*	DYC	DY	*Crypton*
–	HUH	*Helitron*	DHH	DH	*Helitron*
PolB	DDE	*Polinton*	DMM	BD	*Polinton*

'_' represents any character for subdivision.
Abbreviations: RT, reverse transcriptase; APE, apurinic-like endonuclease; RLE, restriction-like endonuclease; GIY-YIG, GIY-YIG endonuclease; DDE, DDD/E transposase; YR, tyrosine recombinase; HUH, HUH endonuclease; PolB, DNA polymerase B; n/d, not defined.

thy that Repbase attempts to avoid fixed higher-rank classification, mainly to avoid frequent revision of the classification. The most recently proposed system by Arkhipova is the simplest and the most adjustable for newly recognized groups of TEs. However, this classification is limited because it integrates the type of nuclease into the system and cannot designate TEs that have more than one nuclease/recombinase. *Dualen* encodes two endonucleases, apurinic-like endonuclease (APE) and restriction-like endonuclease (RLE) (Kojima and Fujiwara, 2005a), while *Helentron* is a group of *Helitron* families that has an APE in addition to the canonical HUH nuclease (Poulter et al., 2003). *Fanzor* is another group of unclassified TE families, which is seen in combination with diverse autonomous TEs (Bao and Jurka, 2013b).

It is difficult to classify TEs into more detailed groups. This is primarily due to the absence of reliable methods for predicting the mechanisms of transposition based solely on sequence information. Curcio and Derbyshire (2003) classified the mechanism of transposition of TEs that encoded DDD/E transposases into "copy-in", "cut-out and paste-in", and "copy-out and paste-in". The "cut-out and paste-in" group could be further divided into several mechanisms based on the structure of transposition intermediates. However, it is quite difficult, if not impossible, to determine the mechanism of transposition for newly recognized groups of TEs using only sequence information. A protein family that has a prim-pol domain and a helicase domain, called insertion sequence (IS)-excision enhancer (IEE), can change the transposition mechanism from "copy-in" to "cut-out and paste-in" in bacteria (Kusumoto et al., 2011). IEE coding sequences are located outside TEs. Therefore, the detailed mechanism of transposition cannot be determined by the TE sequence itself. To maintain a database without frequent reclassification, it is better to avoid integrating the mechanism of transposition into a higher-rank classification system.

CLASS I TRANSPOSONS (RETROTRANSPOSONS)

LTR retrotransposons LTR retrotransposons contain LTRs at both ends, and between these ends there are protein-coding regions. Proteins may contain several catalytic domains: protease, RT, ribonuclease H (RNase H) and integrase; there are also structural proteins called Gag and occasionally Env. LTR retrotransposons mobilize through reverse transcription of their own mRNA, catalyzed by RT. cDNA is generated as extrachromosomal DNA and is then integrated into the genome by integrase. Integrase of LTR retrotransposons shows similarity to the transposase of some DNA transposons, especially the *Ginger1* and *Ginger2* superfamilies, which indicates the composite origin of LTR retrotransposons (Bao et al., 2010). LTR retrotransposons are subdivided

into four superfamilies: *Copia*, *Gypsy*, *BEL* and endogenous retroviruses (ERVs) (Table 2). The International Committee on Taxonomy of Viruses (ICTV) classifies some LTR retrotransposons as virus families. These families include Pseudoviridae for *Copia*, Metaviridae for *Gypsy*, and Belpaoviridae for *BEL* (https://talk.ictvonline.org/). The most recent update (2017) determined the order "Ortervirales", which includes Retroviridae, Pseudoviridae, Metaviridae, Belpaoviridae and Caulimoviridae (https://talk.ictvonline.org/taxonomy/).

***Copia* (Pseudoviridae)** One feature that distinguishes *Copia* from other LTR retrotransposons is the position of the integrase domain, which is upstream of the RT domain. With some exceptions, *Gypsy*, *BEL* and retroviruses encode an integrase downstream of RT.

Repbase does not offer further classification for *Copia*, *Gypsy* and *BEL* in the rapid classification of new LTR retrotransposons. The taxonomy of ICTV contains three genera that are under the family Pseudoviridae: Hemivirus, Pseudovirus and Sirevirus. The representatives of the three ICTV genera are: *Ty1* from the budding yeast *Saccharomyces cerevisiae* for Pseudovirus; *SIRE* from the soybean *Glycine max* for Sirevirus; and *Copia* from the fruit fly *Drosophila melanogaster* for Hemivirus. *SIRE*-like elements have a third, *env*-like ORF downstream of the RNase H domain. The Gypsy Database (GyDB) divides *Copia* into two branches and further into 19 clades (Llorens et al., 2011).

***Gypsy* (Metaviridae)** The taxonomy of ICTV contains two genera under the family Metaviridae: Errantivirus and Metavirus. The representative family of Errantivirus is *Gypsy* from the fruit fly *D. melanogaster*. Metavirus corresponds to most of the *Gypsy* superfamily of LTR retrotransposons. *Ty3* from the budding yeast, *Tf1* from the fission yeast *Schizosaccharomyces pombe*, *Athila* from the thale cress *Arabidopsis thaliana* and *Sushi-ichi* from the pufferfish *Takifugu rubripes* are members of Metavirus. GyDB classified the *Gypsy* superfamily into two branches and further into 34 clades (Llorens et al., 2011).

Chromoviruses, which correspond to branch 1 in GyDB, usually bear a chromodomain (chromatin organization modifier domain) at the C-terminal end of their integrases. The term "Chromoviridae" (Marín and Llorens, 2000) was used to describe this branch within the *Ty3*/*Gypsy* phylogeny. The chromodomain is a domain of approximately 50 residues, and is generally involved in chromatin remodeling and regulation of gene expression (Koonin et al., 1995; Cavalli and Paro, 1998).

Non-chromoviral families of the *Gypsy* superfamily correspond to the GyDB branch 2 (Llorens et al., 2011) and include errantivirus, *Athila* and *Tat* from the thale cress, *Gmr1* from the Atlantic cod *Gadus morhua* and many others. Some of these families encode an additional protein, besides Gag and Pol. The *env* genes of errantiviruses have similarity to the *env* genes from baculovirus, a group of large double-stranded DNA viruses that infect insects (Malik et al., 2000). The *Athila* families also encode additional proteins that have a transmembrane domain and likely an *env* (Malik et al., 2000). *Gmr1* and its relatives have a unique domain structure. They encode an integrase downstream of protease and upstream of RT, like the *Copia* superfamily of LTR retrotransposons. Some *Tat* families encode an additional RNase H domain as well as their canonical RNase H domain, which is shared by all LTR retrotransposons. These additional RNase H domains are more similar to archaeal RNase H domains than to the RNase H domains of LTR retrotransposons and retroviruses (Ustyantsev et al., 2015). Importantly, the archaeal RNase H domain is not restricted to archaea, but is also found in bacteria and plants.

Table 2. Classification, distribution and the number of entries of LTR retrotransposons in Repbase

Superfamily	Total
Copia	10,595
Gypsy	6,694
BEL	1,855
ERV	
ERV1	1,967
ERV2	1,266
ERV3	657
ERV4	187
Lentivirus	4
Unclassified ERV	325
Unclassified LTR	719
DIRS	418

***BEL* (Belpaoviridae)** Belpaoviridae includes only one genus, Semotivirus, in the taxonomy of ICTV. Semotivirus corresponds to the *BEL* superfamily of LTR retrotransposons. GyDB divided *BEL* into three branches and further into five clades (*BEL*, *Tas*, *Suzu*, *Sinbad* and *Pao*) (Llorens et al., 2011). de la Chaux and Wagner (2011) added two more "superfamilies" (*Dan* and *Flow*), which are closely related to *Pao* and *Sinbad*. Some *BEL* families, such as *Roo* from the fruit fly, encode an additional protein that is similar to the errantiviral Env (Llorens et al., 2011). Some *Tas*-like families from *Caenorhabditis elegans*, including *Cer7* and *Cer13*, encode a protein that is similar to the Env proteins, which are encoded by Phleboviruses, a class of single-stranded RNA viruses (Malik et al., 2000).

ERVs Retroviruses are a specialized branch inside LTR retrotransposons. Retroviruses generally have an enve-

lope protein gene, *env*, in addition to other genes encoded in LTR retrotransposons. Env typically contains two domains: a transmembrane domain and a host receptor-binding domain. ERVs are retroviruses that omit the extracellular stage of their life cycle and replicate themselves in germ cells. Some retain the coding ability for Env, but most do not. The loss of *env* and the expansion of ERVs by intracellular retrotransposition are strongly correlated (Magiorkinis et al., 2012).

ERVs are traditionally classified based on the length of TSDs. ERVs with 4-bp TSDs are classified as *ERV1*, ERVs with 6-bp TSDs as *ERV2* and ERVs with 5-bp TSDs as *ERV3* (Kapitonov and Jurka, 2008). This scheme works well, even if there is no information regarding the internal portions of ERVs. It is natural, however, that the classification system for ERVs is combined with the classification for infectious retroviruses (Table 3). Based on the classification of infectious (exogenous) retroviruses, which are classified into eight genera, ERVs can be classified into more groups. *ERV1* corresponds to two retroviral genera, Gammaretrovirus and Epsilonretrovirus, and *ERV2* corresponds to Alpharetrovirus and Betaretrovirus. *ERV3* does not have a corresponding infectious retrovirus group.

Recent genome analyses revealed not only the traditional ERV lineages (*ERV1*, *ERV2* and *ERV3*), but also other groups of infectious retroviruses, which left traces on the genome. The identification of endogenous lentiviruses (Katzourakis et al., 2007) and endogenous foamy viruses (Katzourakis et al., 2009) allowed us to trace their evolutionary history to an origin much older than previously thought. The finding of endogenous foamy viruses revealed that *ERV3* is not the lineage of endogenous spumaviruses (foamy viruses), because endogenous foamy viruses show closer relationships to infectious spumaviruses (Katzourakis et al., 2009; Han and Worobey, 2012). Since Deltaretrovirus integrated into the genomic DNA was finally reported (Farkašová et al., 2017), it is now clear that all genera of retroviruses can be endogenized. *ERV4* has features that are similar to those of *ERV3*, but phylogenetic analysis suggests that the *ERV4* branch is independent from *ERV3* (Chong et al., 2014).

Table 3. Relationships between endogenous and infectious retroviruses

Endogenous retrovirus	Infectious retrovirus (genus)
ERV1	Gammaretrovirus, Epsilonretrovirus
ERV2	Alpharetrovirus, Betaretrovirus
ERV3	n/d
ERV4	n/d
Endogenous deltaretrovirus (EDV)	Deltaretrovirus
Endogenous lentivirus (ELV)	Lentivirus
Endogenous foamy virus (EFV)	Spumavirus

n/d, not defined.

YR retrotransposons YR retrotransposons are located as a branch inside that of LTR retrotransposons in the RT phylogeny. This indicates that YR retrotransposons were generated via recombination between a *Crypton*-like DNA transposon and an LTR retrotransposon, although the origin and the monophyly of this group have not yet been determined (Goodwin and Poulter, 2004; Kojima and Jurka, 2011a). Retrotransposons designated with the names *DIRS* (Glöckner et al., 2001), *PAT* (de Chastonay, 1992), *Ngaro* (Goodwin and Poulter, 2004), *VIPER* (Lorenzi et al., 2006) and *TATE* (Peacock et al., 2007) encode a YR. They share a coding ability for RT, RNase H and YR. Another domain is likely an analog of Gag that is encoded by LTR retrotransposons. *DIRS* and *PAT* encode an additional domain, methyltransferase, which is downstream of YR (Goodwin and Poulter, 2004). In Repbase, all YR retrotransposons are classified into one superfamily, *DIRS*.

Even though these YR retrotransposons appear to originate from LTR retrotransposons, they do not have LTRs. Instead, they have either split repeats (SRs) or inverted terminal repeats (ITRs). In the elements having split repeats, sequences homologous to the left and right termini are also present in the middle of the elements. SRs or ITRs are very likely the key modules in transposition, but the mechanism of transposition of YR retrotransposons has not been adequately identified. The proposed model assumes a circular intermediate (Goodwin and Poulter, 2004).

Pararetroviruses Based on the RT phylogeny, besides retrovirus, LTR retrotransposons and YR retrotransposons are related to two virus families: Hepadnavirus and Caulimovirus. Hepadnavirus and Caulimovirus are called pararetroviruses, although they cluster separately in the RT phylogeny. These two groups of viruses are sometimes present as repetitive sequences in the genome, but they appear to be accidental integrants in the genome, rather than true TEs.

The identification of Hepadnaviral fossils in avian and reptile genomes revealed a higher diversity of hepadnaviruses than is found in the current hepadnaviruses (Gilbert and Feschotte, 2010; Liu et al., 2012). Caulimoviruses (the family Caulimoviridae in the ICTV taxonomy) are classified into eight genera: Badnavirus, Caulimovirus, Cavemovirus, Petuvirus, Rosadnavirus, Solendovirus, Soymovirus and Tungrovirus. In addition,

a new group, "florendovirus", was proposed from the analysis of integrated Caulimovirus sequences (Geering et al., 2014). Florendoviruses are closest to Petuvirus in the RT phylogeny. Caulimoviruses are the most abundant endogenous viral elements, next to retroviruses; Repbase contains 157 Caulimovirus sequences.

Non-LTR retrotransposons Non-LTR retrotransposons lack LTRs and usually have poly(A) or simple repeats at their 3'-terminus. Non-LTR retrotransposons encode one of two types of endonucleases, RLE or APE. *Dualen* is an exception that encodes both RLE and APE (Kojima and Fujiwara, 2005a). Endonuclease nicks one strand of DNA and RT initiates reverse transcription using the exposed 3' end as a primer and the mRNA of non-LTR retrotransposons as a template (Luan et al., 1993). This mechanism is called target-primed reverse transcription (TPRT). TPRT is also used as a mechanism for the integration of group II self-splicing introns (Zimmerly et al., 1995), and probably of *Penelope*-like elements (Pyatkov et al., 2004). However, no intact group II intron is present in eukaryotic nuclear genomes.

Non-LTR retrotransposons are classified into many clades. The classification "clade" was first proposed by Malik et al. (1999), who introduced the term to cluster non-LTR retrotransposons that (1) share the same structural features, (2) are grouped together with ample phylogenetic support, and (3) date back to the Precambrian era. They originally introduced 11 clades (*CRE*, *R2*, *R4*, *L1*, *RTE*, *I*, *R1*, *LOA*, *Tad1*, *Jockey* and *CR1*). Three years after the proposal of clade, the term "group" was designated as a higher-order classification than the clade by Eickbush and Malik (2002), who classified non-LTR retrotransposons into five groups (*R2*, *L1*, *RTE*, *I* and *Jockey*). However, these groups are not always monophyletic; for instance, the *R2* group is paraphyletic.

Now, two decades later, the number of clades has increased significantly because of additional lineages or splits in original clades. More than 30 clades have been proposed, which complicates the classification of non-LTR retrotransposons. GIRI offers a simple classification tool designated RTclass1, which is based on the neighbor-joining tree and a reference set of non-LTR retrotransposons (Kapitonov et al., 2009). As of January, 2018, Repbase uses 32 clades (*CRE*, *NeSL*, *R4*, *R2*, *Hero*, *RandI/Dualen*, *L1*, *Proto1*, *Tx1*, *Proto2*, *RTE*, *RTEX*, *RTETP*, *I*, *Nimb*, *Ingi*, *Vingi*, *Tad1*, *Loa*, *R1*, *Outcast*, *Jockey*, *CR1*, *L2*, *L2A*, *L2B*, *Kiri*, *Rex1*, *Crack*, *Daphne*, *Ambal* and *Penelope*) in its classification (Bao et al., 2015), where, due to practical reasons, *Penelope* is included as a non-LTR retrotransposon clade. Except for *Penelope* and SINEs, non-LTR retrotransposons in Repbase are classified into eight groups (Table 4): *CRE*, *R2*, *Dualen*, *L1*, *RTE*, *I*, *CR1* and *Ambal*.

Table 4. Classification, and the number of entries of non-LTR retrotransposons in Repbase

Group	Clade	Total
CRE	*CRE*	43
R2	*R4*	46
	Hero	23
	NeSL	106
	R2	159
Dualen	*RandI/Dualen*	13
L1	*Proto1*	6
	L1	1,690
	Tx1	273
RTE	*RTETP*	1
	Proto2	47
	RTEX	138
	RTE	487
I	*Outcast*	23
	Ingi	17
	Vingi	141
	I	195
	Nimb	108
	Tad1	141
	Loa	74
	R1	237
	Jockey	243
CR1	*Rex1*	95
	CR1	803
	Kiri	91
	L2	285
	L2A	5
	L2B	27
	Crack	140
	Daphne	227
Ambal	*Ambal*	8
Penelope	*Penelope*	477
SINE	*SINE1/7SL*	95
	SINE2/tRNA	539
	SINE3/5S	30
	SINEU	17
	Unclassified SINE	112
	Unclassified non-LTR retrotransposon	179
	Total	7,341

***CRE* group** The *CRE* clade is the first branched lineage in non-LTR retrotransposons (Malik et al., 1999). The *CRE* clade in Repbase includes families in the original *CRE* clade (Malik et al., 1999) and the *Genie/Gil* lineage

(Burke et al., 2002). The *CRE* clade is the sister group of all other non-LTR retrotransposons. The first identified families in the *CRE* clade (*CRE1*, *CRE2*, *SLACS* and *CZAR*) are spliced-leader exon-specific retrotransposons (Aksoy et al., 1990; Gabriel et al., 1990). Another group, called *MoTeR*, from several fungi are specifically inserted into telomeric repeats (Starnes et al., 2012). However, it is now clear that many *CRE* families are not necessarily sequence-specific. The *CRE* clade can be considered an independent group (Putnam et al., 2007), or as a part of the *R2* group (Eickbush and Malik, 2002).

***R2* group** The *R2* group is one of the five original groups (Eickbush and Malik, 2002). The *R2* group includes the clades *R2*, *R4*, *NeSL* and *Hero*. The *R2* clade (or superclade) may be divided into four clades (*R2A*, *R2B*, *R2C* and *R2D*) based on the phylogeny and the structures of N-terminal zinc-finger motifs (Kojima and Fujiwara, 2005b). The *R2* group and *CRE* group share one feature, namely that an RLE is encoded downstream of the RT. The other structures are not conserved throughout the group, although members often have zinc-finger motif(s) at the N-terminus of their encoded protein. Some families encode a Ulp1-type protease upstream of the RT. Some have two open reading frames (ORFs), while others have only one. This group includes many target sequence-specific families that include *R2*, *R4*, *NeSL* and *Utopia* (Burke et al., 1995; Malik and Eickbush, 2000; Kojima and Fujiwara, 2005b; Kojima and Jurka, 2015).

***Dualen* group** The *Dualen* group includes only one clade, *Dualen* (from dual endonucleases), also called *RandI* (Kojima and Fujiwara, 2005a). The *Dualen* clade is the only clade that encodes both RLE and APE simultaneously, even though some *Dualen* families, such as *RandI-1_ACas*, lack RLE. *Dualen* is a family of gigantic retrotransposons that are longer than 10 kb and encode a single protein that is longer than 3,000 residues. Although their termini are not determined, some *Dualen* families such as *Dualen-5_CCu* and *Dualen-1_GCr* encode a protein longer than 5,000 residues (Lescot et al., 2016). The structure and phylogenetic position of *Dualen* indicate that it is a descendant of non-LTR retrotransposons that exchanged their endonucleases from RLE to APE.

***L1* group** The *L1* group is one of the five original groups (Eickbush and Malik, 2002). It originally included a single clade, *L1*, but now includes two additional clades (*Tx1*, *Proto1*). In this group, the *L1* clade appears paraphyletic. Canonical elements that belong to the *L1* group encode two proteins. The sequence of the ORF1 protein is highly diverged: the ORF1 protein of human *L1* has a leucine-zipper motif, while others have zinc-finger motifs. The second protein (ORF2) includes an APE, RT, and often a CCHC-type zinc finger motif. The *L1* group does not encode an RLE, but some lineages of *L1*, especially *L1* families from plants, encode an RNase H domain downstream of the RT domain.

The *L1* clade is represented by LINE1 (long interspersed element 1), found in various mammals. *L1* is the only active autonomous non-LTR retrotransposon family in the human genome and causes cancers and genetic diseases by transposition. The *Tx1* clade is derived from the *L1* clade. Most families that belong to the *Tx1* clade have target sequence specificity (Kojima and Fujiwara, 2004; Kojima, 2015). The *Proto1* clade was first proposed with elements from *Naegleria gruberi* (Kapitonov and Jurka, 2009). *Proto1* encodes two proteins, one of which includes three domains: APE, RT and RNase H.

***RTE* group** The *RTE* group is also one of the five original groups (Eickbush and Malik, 2002). It originally included a single clade, *RTE*, but now includes several more (*RTEX*, *RTETP* and *Proto2*). The *RTE* group has been found in animals, fungi, plants and algae. However, the distribution of clades, except for *RTE* and *RTEX*, are is quite restricted. The *RTETP* clade has been only found in diatoms.

The *RTE* clade is one of the original clades (Malik et al., 1999). *Bov-B* from the bovine *Bos taurus*, *Expander* from the pufferfish *T. rubripes*, and *SR2* from the bloodfluke *Schistosoma mansoni* belong to this clade. Elements belonging to the *RTE* clade are generally short and encode a protein with two functional domains: APE and RT. Some *RTE* elements are reported to be horizontally transferred (Kordis and Gubensek, 1999; Walsh et al., 2013).

In contrast to the *RTE* clade, canonical *RTEX* elements encode two proteins. The ORF1 protein sometimes includes an esterase domain and/or a PHD (plant homeodomain) domain. *ORTE* families from the yellow fever mosquito *Aedes aegypti* encode an OTU cysteine protease upstream of APE (Kojima and Jurka, 2011c).

***I* group** Originally the *I* group included five clades (*I*, *Ingi*, *R1*, *LOA* and *Tad1*) (Eickbush and Malik, 2002). The distinctive feature of this group is an RNase H that is downstream of RT, although many elements have lost the RNase H. It had been considered that the last common ancestor of the *I* group acquired an RNase H domain, but recent findings for RNase H domains from the *Dualen*, *L1* and *Proto1* families indicate that the acquisition of RNase H was an earlier event (Kojima and Fujiwara, 2005a; Kapitonov et al., 2009). The *Jockey* clade was considered a representative of the "*Jockey* group"; however, there is an accumulation of evidence that the *Jockey* clade is more closely related to the *I* group than the *CR1* clade (Kojima and Fujiwara, 2005a; Putnam et al., 2007). Thus, here, the *Jockey* clade is proposed to be included in the *I* group,

along with the other eight clades (*I*, *Ingi*, *Vingi*, *R1*, *LOA*, *Tad1*, *Nimb* and *Outcast*). The *I* group has also been found in animals, fungi and trypanosomatids.

The *I* clade is one of the original clades (Malik et al., 1999). The *Ingi* and the *Nimb* clades were originally part of the *I* clade. The present *I* clade is probably paraphyletic. *Loner* has been reported only in two species of mosquitoes, *Anopheles gambiae* and *Ae. aegypti* (Biedler and Tu, 2003). In the classification of Repbase, *Loner* is included in the *I* clade.

The *Ingi* clade was split from the *I* clade (Eickbush and Malik, 2002). *L1Tc*, from *Trypanosoma cruzi*, is often misrecognized as an *L1* family, but is actually a close relative of *Ingi* from *T. brucei*. One characteristic feature of *Ingi* elements (shared with *Vingi* and *RTE*) is that they frequently have non-autonomous derivatives both termini of which are similar to those of autonomous elements. The *Ingi* clade is paraphyletic, as the *Vingi* clade was split from the *Ingi* clade (Kojima et al., 2011). *Vingi* generally lacks an RNase H domain, in contrast to *Ingi* elements.

Monophyly of three clades, *R1*, *LOA* and *Tad1*, is well supported. Elements belonging to the *R1* clade frequently show target sequence specificity that is achieved by their APE (Kojima and Fujiwara, 2003). The first reported element belonging to the *LOA* clade, *LOA*, from *D. silvestris*, is a fusion with a *Gypsy*-like LTR retrotransposon (Felger and Hunt, 1992), but the structures of other elements are similar to other families in the *I* group. Elements belonging to the *Tad1* clade have been found only in fungi.

The *Jockey* clade is one of the original clades proposed (Malik et al., 1999). Its members do not have an RNase H domain in their ORF2 protein. Some elements belonging to the *Jockey* clade (*TART*, *TAHRE* and *HeT-A*) are specifically transposed onto the telomere (Abad et al., 2004).

***CR1* group** The "*Jockey* group" was originally proposed to include two clades: *Jockey* and *CR1* (Eickbush and Malik, 2002). The *Jockey* clade is now thought to be closer to the *I* group in the RT phylogeny (Kojima and Fujiwara, 2005a; Putnam et al., 2007). The "*CR1* group" (Putnam et al., 2007) includes the *CR1* clade and clades split from it (*L2*, *Rex1*, *L2A*, *L2B*, *Daphne*, *Crack*, *Kiri*). One common feature of the *CR1* group is the lack of an RNase H domain. The *CR1* group has been found exclusively in animals.

***Ambal* group** Elements belonging to the *Ambal* clade have been identified in two species of diatoms, *Fragilariopsis cylindrus* and *Thalassiosira pseudonana* (Kapitonov and Jurka, 2010). *Ambal* elements are longer than 10 kb and encode two proteins. The ORF2 protein contains APE, RT and RNase H. The domain composition of *Ambal* resembles those of the *L1* and *I* groups. Despite this, the phylogenetic position of *Ambal* elements in the RT phylogeny is close to that of *CRE*. *Ambal* may be a chimeric retrotransposon, or may be a remnant of an ancient retrotransposon. A proposal for the *Ambal* group is not yet in the literature, although *Ambal* elements are distinct from any other non-LTR retrotransposons in structure and phylogeny.

Group unknown The clades below are not classified into any group because they were positioned as an outgroup of certain group(s) in the phylogeny. The *Odin* clade includes families found only in the tunicate *Oikopleura dioica* (Volff et al., 2004). *Odin* is closer to the *I* and *CR1* groups than the *RTE* and *L1* groups. Unfortunately, no related retrotransposons have been found in other organisms, and, therefore, the *bona fide* position of this clade is still unclear. The APEs coded by *Odin* elements have DGH residues instead of canonical SDH residues in the catalytic core and the functionality of this endonuclease is unknown. The *REP* clade, proposed in an analysis of non-LTR retrotransposons from the ciliate *Tetrahymena thermophila*, is close to the *L1* clade in phylogeny (Fillingham et al., 2004). *Deceiver* and *Inkcap* are the other proposed clades whose phylogenetic positions remain unsolved. *Deceiver* branched earlier than the *RTE* clade, but later than the *L1* clade (Novikova et al., 2009). *Inkcap* branched earlier than the *CR1* and *I* groups, but later than the *RTE* clade (Novikova et al., 2009).

Compared with other groups of TEs, such as LTR retrotransposons and DNA transposons, non-LTR retrotransposons have been classified into too many subgroups (clades). Considering the high number of clades, describing a new clade is not useful and the last clade integrated in the Repbase classification system was *Kiri* (Kojima and Jurka, 2011b).

***Penelope*-like element** *Penelope* was first described in *D. virilis* (Evgen'ev et al., 1997). Because of its long terminal repeats, it was considered expediently as a member of the LTR retrotransposons, although its features differ from other LTR retrotransposons that are described above. The presence of GIY-YIG-type endonuclease downstream of the RT domain led to a new definition of *Penelope* and its relatives as a new group of retrotransposons (Lyozin et al., 2001; Volff et al., 2001). This GIY-YIG endonuclease works analogously to APE and RLE in non-LTR retrotransposons; *Penelope*-like elements likely transpose via the TPRT mechanism (Pyatkov et al., 2004).

Two lineages of *Penelope*-like elements, *Athena* and *Coprina*, lack a GIY-YIG-type endonuclease (Gladyshev and Arkhipova, 2007). They may represent an ancestral state that preceded the acquisition of an endonuclease. They are found at telomeres. Targeting

chromosome ends is known for the transposition of an endonuclease-deficient human *L1* non-LTR retrotransposon (Morrish et al., 2007). Analogously, *Athena* and *Coprina* are expected to transpose to the chromosome ends via the TPRT mechanism, in which the 3' end of chromosomal DNA is used as a primer.

The RT phylogeny clustered *Penelope*-like elements and telomerase RT (TERT) together (Gladyshev and Arkhipova, 2007). This, as well as the features of endonuclease-lacking *Penelope*-like elements, raises the possibility that *Penelope*-like elements are close relatives of the putative retroelements that gave rise to telomerases.

Some *Athena* elements have introns, but the biological meaning of these remains unknown (Arkhipova et al., 2003).

CLASS II TRANSPOSONS (DNA TRANSPOSONS)

DNA transposon superfamilies encoding DDD/E transposase/integrase The dominant group of DNA transposons is the TEs that encode DDD/E transposase as an enzyme for mobilization, both in eukaryotes and prokaryotes (Siguier et al., 2006; Bao et al., 2015). As of January 2018, Repbase contains 23 Class II TE superfamilies (Bao et al., 2015). Among them, 21 (*Mariner/Tc1*, *hAT*, *MuDR*, *EnSpm/CACTA*, *piggyBac*, *P*, *Merlin*, *Harbinger*, *Transib*, *Polinton*, *Kolobok*, *ISL2EU*, *Sola*, *Zator*, *Zisupton*, *Ginger1*, *Ginger2/TDD*, *Academ*, *Novosib*, *IS3EU* and *Dada*) are known to encode DDD/E transposase for catalysis during integration. This type of transposase shares the same catalytic core with integrases of the LTR retrotransposons. *Ginger1*, *Ginger2/TDD* and *Polinton* superfamilies have the highest sequence similarity with integrases of the LTR retrotransposons (Bao et al., 2010). Based on the core and other highly conserved residues, some superfamilies can join together (Yuan and Wessler, 2011): *Harbinger* and *ISL2EU*; *MuDR*, *Rehavkus*, *P*, *hAT* and *Kolobok*; and *EnSpm*, *Mirage*, *Chapaev* and *Transib*, based on signature motifs inside the DDD/E transposase. Importantly, because the sequences are extremely divergent excluding the catalytic residues, the presence of a conserved DDD/E core sequence does not guarantee their common origin; they may have independently evolved. DDD/E transposase/integrase is related to RNase H in its protein ternary structure and is classified in RNase H fold.

Prokaryotic DNA transposons have more variety than their eukaryotic counterparts, and are classified into many families based on ISfinder (Siguier et al., 2006, 2015). IS1, IS3, IS6, IS30, IS21, IS982, IS630, IS4, IS5, IS256, IS481, IS1380, ISL3 and Tn3 (and possibly also IS66) encode a DDD/E transposase, and IS110 encodes a DDED transposase whose structure is more similar to Holliday junction resolvase, RuvC, than to DDD/E transposases.

Similarities between eukaryotic and prokaryotic DDD/E transposases are sometimes observed: *Mariner* and *Zator* to IS630 (Doak et al., 1994; Bao et al., 2009), *MuDR* to IS256 (Eisen et al., 1994; Hua-Van and Capy, 2008), *Merlin* to IS1016 (Feschotte, 2004), *piggyBac* to IS4 and IS5 (Sarkar et al., 2003), *Harbinger* and *ISL2EU* to IS5 (Kapitonov and Jurka, 1999; Zhang et al., 2001), *Ginger1*, *Ginger2*, *Polinton* and LTR retrotransposons to IS481 (Bao et al., 2010), and *IS3EU* to IS3 (*IS3EU* families in Repbase; http://www.girinst.org/repbase/).

Majorek et al. (2014) compared the proteins structurally related to RNase H, including DDD/E transposases. In their phylogenetic analysis, these RNase H-like proteins were classified into 12 lineages, among which seven clades (A, B, C, D, II, III and IV) include DDD/E transposase. The clade A includes *Mariner*, *MuDR*, HIV-1 integrase (LTR retrotransposon, *Ginger1*, *Ginger2* and *Polinton*) as well as IS1016 (related to *Merlin*) from bacteria. Clade B includes *hAT* and *P*. Clade C includes *Harbinger*, *piggyBac*, IS4 and Tn5 from bacteria. Clade D includes COG3547 (IS116/IS110/IS902) from bacteria. Clade II includes RAG (*Transib*) and *Chapaev* (now merged with *EnSpm*). Clade III includes *EnSpm* and *Mirage* (now merged with *EnSpm*) and Tn3 from bacteria. Finally, Clade IV includes IS66 from bacteria. The clustering here is not always consistent with the relationships inferred by Yuan and Wessler (2011). The relationships between eukaryotic TE superfamilies supported by these analyses are that of *EnSpm* and *Mirage*, and that of *hAT* and *P*.

Table 5 reveals the number of entries in Repbase for each superfamily. *hAT* and *Mariner* are dominant DNA transposon superfamilies in humans and vertebrates. *MuDR*, *Harbinger*, *Helitron* and *EnSpm* are dominant, especially, in higher plants (angiosperms). Hereafter, each superfamily of DNA transposons found in Repbase is described briefly.

Mariner/Tc1 The eukaryotic superfamily *Mariner/Tc1* is related to the bacterial IS630 family (Doak et al., 1994). This group is also referred to as the *IS630-Tc1-Mariner (ITm)* family. Many subdivisions in the *Mariner/Tc1* superfamily have been proposed. These include *Tc1*, *Mariner*, *Pogo*, *MaT*, *ITmD37D*, *ITmD37E*, and so on (Shao and Tu, 2001; Claudianos et al., 2002; Coy and Tu, 2005; Tellier et al., 2015). These studies suggest that the distance between the second D and the third D/E is one distinguishable characteristic in each of these subgroups. *Sagan* has a long insertion between the second D and the last E residues, unlike other *Mariner/Tc1* families (Kojima and Jurka, 2011d). The centromeric protein CENP-B is a Pogo transposase that acquired a biological function (Tudor et al., 1992; Smit, 1996). *Mariner/Tc1* elements exclusively generate TSDs of TA dinucleotide.

Table 5. Classification, terminal features, TSD features and the number of entries of DNA transposons in Repbase

Group	Superfamily	Termini	TSD	Entries
IS630/Mariner	*Mariner/Tc1*	YR..YR	TA	2,539
	Zator	GG..CC	3	54
IS481/Ginger	*Ginger1*	TGT..ACA	4	39
	Ginger2/TDD	TGT..ACA	4–5	20
IS3/IS3EU	*IS3EU*	TAY..RTA	6	23
IS1016/Merlin	*Merlin*	GG..CC	8–9	75
IS256/DxxH	*hAT*	YA..TR	5–8	2,955
	MuDR	GR..YC	8–9	1,345
	P	CA..TG	7–8	189
	Kolobok	RR..YY	TTAA	286
	Dada	?	6–7	36
IS1380/piggyBac	*piggyBac*	YY..RR	TTAA	377
IS5/PHIS	*Harbinger*	RR..YY	3	1,097
	ISL2EU	RR..YY	2	88
CCHH	*EnSpm/CACTA*	CAC..GTG	2–4	715
	Transib	CAC..GTG	5	123
KDZP	*Zisupton*	?	8	18
Sola	*Sola*			
	Sola1	?	4	100
	Sola2	GRG..CYC	4	90
	Sola3	GAG..CTC	TTAA	28
	Unclassified *Sola*			1
?	*Academ*	YR..YR	3–4	90
?	*Novosib*	CA..TG	8	9
Crypton	*Crypton*			
	CryptonF		0	23
	CryptonA	TTA..	0	17
	CryptonI	?	0	9
	CryptonS	TATGG..	0	59
	CryptonV	?	0	46
	Unclassified *Crypton*			80
Helitron	*Helitron*	TC..CTRR	0	955
Polinton	*Polinton*	AG..CT	6	108
	Unclassified DNA transposon			2,357
	Total			13,960

Zator *Zator* is related to the bacterial TP36 family of transposases (Bao et al., 2009). Along with the *Mariner/Tc1* superfamily, *Zator* and TP36 are clustered with the bacterial IS630 family. Unlike the *Mariner/Tc1* superfamily, *Zator* generates 3-bp TSDs.

Ginger1 Transposases of *Ginger1* and *Ginger2*, and integrases of *Polinton*, LTR retrotransposons (*Copia*, *Gypsy*, *BEL*) and retroviruses, are related to each other (Bao et al., 2010). Their transposase/integrase is distantly related to the transposases of the bacterial IS3 and IS481 families. Most IS3 elements terminate with 5'-TG..CA-3' (Siguier et al., 2015). IS481 is much shorter than the IS3 family members, although their transposases are quite similar.

Ginger1 DNA transposons likely originated from a *Gypsy* LTR retrotransposon that was possibly related to the *Athila* and *Tat* families. The integrases of *Ginger1*

and *Gypsy* LTR retrotransposons share the YPYY motif, the four conserved residues upstream of the integrase core. Most of the *Ginger1* families contain a Ulp1 cysteine protease or OTU cysteine protease that is downstream of their transposase.

Ginger2/TDD Compared with *Ginger1*, *Ginger2* has a weaker relationship to *Gypsy* LTR retrotransposons (Bao et al., 2010). Although statistically not significant, the integrases encoded by *Ginger2* families are clustered together with those encoded by *Polinton* DNA transposons. It remains to be examined whether *Ginger2* is a remnant lineage of DNA transposons that contributed to the birth of LTR retrotransposons or *Polinton*s.

IS3EU *IS3EU* is a superfamily of DNA transposons that has only been published in Repbase (Bao et al., 2015). *IS3EU* encodes two proteins, one of which is a DDD/E transposase. These DDD/E transposases are most similar to those from the bacterial IS3 family. *IS3EU* has been identified in various animals and a species of fungi, *Puccinia graminis*.

Merlin *Merlin* is related to the prokaryotic IS1016 and IS1595 families (Feschotte, 2004). The IS1595 family does not always share the DDE residues (some members contain N instead of E). The IS1016 family has DDE residues, and is most similar to the eukaryotic *Merlin* elements. *Merlin* generates 8-bp or 9-bp TSDs.

MuDR The two-component system *MuDR*/*Mu* from maize comprises the first reported DNA transposon family that belongs to the superfamily currently recognized as *MuDR* or *MULE* (*Mutator*-like element) (Robertson, 1978). *MuDR* encodes two proteins, MURA and MURB; MURA is the transposase. *MuDR* families are primarily identified in plants, but have also been reported in animals, fungi and stramenopiles. TSDs are 8 bp or 9 bp.

The majority of *MuDR* families have relatively long TIRs at both ends. *Arnold* and *Vandal*, although they are the members of the *MuDR* superfamily, lack TIRs (Kapitonov and Jurka, 1999). *Vandal* encodes a third protein, which is reported to function in counteracting transcription suppression by the host (Fu et al., 2013).

Transposases of *MuDR* elements and prokaryotic IS256 elements share some features (Eisen et al., 1994; Hua-Van and Capy, 2008). *MuDR* can be clustered with *Kolobok*, *hAT*, *P*, and *Rehavkus* based on the presence of the C/D(2)H motif between the second D and the last E of the catalytic residues (Yuan and Wessler, 2011). *Rehavkus*, previously present in Repbase as a superfamily, is now integrated into *MuDR* in the Repbase classification.

The N-terminus of *MuDR* contains a DNA-binding domain. The zinc-finger motif seen in the *MuDRF* families is called the GCM1 domain (Cantu et al., 2011). The DNA-binding domains of other *MuDR* elements are called WRKY or FLYWCH (Babu et al., 2006).

hAT The *hAT* superfamily is one of the most abundant DNA transposon superfamilies. The name *hAT* originated from the initials of three well-studied *hAT* transposons: *hobo* from *D. melanogaster*, *Activator*/*Dissociation* (*Ac*/*Ds*) from maize, and *Tam3* from the snapdragon *Antirrhinum majus* (McClintock, 1950; Blackman et al., 1989; Hehl et al., 1991). *Hermes* and *Tol2* are two *hAT* families that are used for transgenesis and mutagenesis (O'Brochta et al., 1996; Kawakami and Shima, 1999). Although no *hAT* DNA transposons are active in the human genome, many ancient *hAT* transposons, *Charlie* and its non-autonomous derivatives, preserve their traces on the human genome (Kojima, 2018a).

In general, *hAT* families encode a single protein that includes a transposase domain. TIRs of *hAT* families are usually short, up to 50 bp. The majority of *hAT* families generate 8-bp TSDs. However, *hAT5* families generate 5-bp TSDs, *hAT6* families generate 6-bp TSDs and *hATw* generates 7-bp TSDs. *hATm* and *hATx* are distinct lineages inside the *hAT* superfamily of DNA transposons.

P The representative of the *P* superfamily, *P element*, was found in the genome of *D. melanogaster* (O'Hare and Rubin, 1983). The *P* superfamily is a relatively small group, despite the long research history: fewer than 200 families belonging to the *P* superfamily have been deposited in Repbase. The *P* superfamily is widely distributed among animals, plants, fungi and protozoans. The human genome retains a catalytically active transposase of an ancient *P* family member as THAP9 (Majumdar et al., 2013).

Kolobok *Kolobok* was reported by Kapitonov and Jurka (Kapitonov and Jurka, 2007c), and encodes two proteins. One is a protein in which a DDD/E transposase follows a THAP DNA-binding domain. The THAP domain is also found in some families in the *P* superfamily. The other protein has no motifs that are conserved with known domains. *Kolobok* generates TSDs of TTAA. The distinguishable characteristics of *Kolobok* and *piggyBac* are the sequences of their termini. *Kolobok* ends with 5'-RR..YY-3', while *piggyBac* ends with 5'-YY..RR-3'. *Kolobok* has been found in many animals, plants, stramenoplies, heterolobosea and parabasalia. *Kolobok1* is a subgroup of *Kolobok* (Jurka and Bao, 2008).

Dada *Dada* is the only superfamily of DNA transposons having strict target sequence specificity (Kojima and Jurka, 2013b). The most widely distributed lineage is *Dada-U6*, which is seen from various teleost fishes, water flea and the polychaete worm *Capitella teleta*, and it is specifically inserted into a site within the U6

small nuclear RNA genes. *Dada* lacks terminal inverted repeats and instead has a short sequence that is similar to the sequence with the same distance from the integration site.

piggyBac *piggyBac* was originally isolated from a baculovirus infecting a cell culture of the cabbage looper *Trichoplusia ni* (Fraser et al., 1983; Cary et al., 1989). The members of the *piggyBac* superfamily target a specific sequence, TTAA. The transposases in the *piggyBac* superfamily have three conserved D residues and show similarity to that encoded by the bacterial IS4 family (Sarkar et al., 2003).

piggyBacA is a distinct group related to *piggyBac* and generates ATAT TSDs instead of TTAA TSDs (Kapitonov and Jurka, 2014). *piggyBacX*, from the red seaweed *Chondrus crispus* and several species of the oomycete *Phytophthora*, also encodes a distinct transposase, which shows weak similarity to other *piggyBac* transposases (Bao and Jurka, 2014).

Recently, a family of gigantic (~180-kb) *piggyBac* transposons was characterized and designated as *Teratorn* (Inoue et al., 2017). The coded proteins of *Teratorn* revealed that *Teratorn* is a composite DNA transposon born as a fusion between a *piggyBac* DNA transposon and a herpesvirus that belongs to Alloherpesiviridae.

Harbinger The *Harbinger* superfamily, or the *PIF/Harbinger* superfamily, has two founder members, *Harbinger* and *PIF* (Jurka and Kapitonov, 2001). *Harbinger* was described from *A. thaliana* (Kapitonov and Jurka, 1999), while *P instability factor* (*PIF*) was characterized in maize (Zhang et al., 2001). *PIF* and related autonomous TEs are responsible for the mobilization of *Tourist*, which is one of the two predominant nonautonomous TE groups in plants. These TEs encode two proteins, ORF1 and transposase. The transposases show similarity to those encoded by IS5 and ISL2 in bacteria. The ORF1 protein usually contains a Myb-like DNA-binding domain and is required for transposition besides the transposase (Sinzelle et al., 2008). *HarbingerS* is a group of *Harbinger* families that encode three proteins: DDD/E transposase, SET histone methyltransferase and an unknown protein (Kojima and Jurka, 2014b).

In Repbase, two eukaryotic superfamilies, *Harbinger* and *ISL2EU*, are related to the bacterial IS5 family. Han and colleagues (Han et al., 2014, 2015) proposed several other lineages that were designated as *Spy*, *NuwaI*, *NuwaII* and *Pangu*, and referred to the whole group as PHIS. *NuwaI* and *NuwaII* show similar protein-coding capacity to *Harbinger*. They encode two proteins, transposase and Myb-like DNA-binding protein. *Pangu* also encodes two proteins, but the protein that is other than transposase does not contain any recognizable domain. Phylogenetic analysis indicated that *HarbingerS* families are a branch inside *Pangu*. *Harbinger*, *NuwaI*, *NuwaII* and *Pangu* generate 3-bp TSDs. *Spy* is reported to generate no TSDs.

ISL2EU *ISL2EU* shows strong similarity to ISL2 and related bacterial ISs (Kapitonov and Jurka, 2007b). Due to mis-annotation of bacterial ISL2 as IS4, some families in this group were named with the header *IS4EU*. Autonomous *ISL2EU* families such as *IS4EU-1_DR* and *ISL2EU-4_HM* encode two proteins: transposase and the YqaJ exonuclease. The transposase protein contains two domains, THAP DNA-binding domain and DDD/E transposase domain. *ISL2EU* generates 2-bp TSDs in contrast to other related TEs, which generate 3-bp TSDs.

EnSpm/CACTA Peterson characterized an autonomous TE insertion designated *Enhancer* (*En*) (Peterson, 1953). McClintock independently characterized a TE insertion and designated it as *Suppressor-Mutator* (*Spm*) (McClintock, 1954). These two TE insertions were sequenced and revealed to be almost identical, and we now refer to this family of TEs as *Enhancer/Suppressor-Mutator* (*En/Spm*). The *EnSpm* superfamily is also called the *CACTA* superfamily because many plant *EnSpm* family sequences begin with the pentanucleotide CACTA. *EnSpm* families usually encode two proteins and plant *EnSpm* families generate 3-bp TSDs.

The *Mirage* superfamily was proposed with new families found in the nematode *C. elegans* (Kapitonov and Jurka, 1999, direct submission to Repbase Update). *Mirage* families generate 2-bp TSDs. The *Chapaev* superfamily was proposed in 2007 (Kapitonov and Jurka, 2007a). *Chapaev* families generate 4-bp TSDs. The transposase proteins encoded by *Chapaev* contain a unique zinc-finger motif, the Chapa domain, at their N-terminus. The Chapa domain and its downstream RING finger domain show similarity to recombination activating gene 1 protein (RAG1).

Yuan and Wessler (2011) proposed the clustering of *EnSpm*, *Mirage*, *Chapaev* and *Transib*, based on the presence of C(2)C and H(3-4)H motifs between the second D and the last E catalytic residues. *Mirage* and *Chapaev*, which were previously present in the classification in Repbase, have been integrated into *EnSpm/CACTA* based on their similarity.

Transib The resemblance between V(D)J recombination and the transposition of DNA transposons was recognized just several years after V(D)J recombination was discovered (Sakano et al., 1979). The *Transib* superfamily encodes a transposase that is most similar to the RAG1 protein, which is responsible for V(D)J recombination (Kapitonov and Jurka, 2005). A protein similar to RAG2, another protein responsible for V(D)J recombination, was identified in a lineage of *Transib*, *TransibSU*

(Kapitonov and Koonin, 2015). A long-standing debate regarding the origin of V(D)J recombination was concluded upon the discovery of a *Transib* DNA transposon in the lancelet, designated *ProtoRAG* (Huang et al., 2016). *ProtoRAG* encodes two proteins that are similar to RAG1 and RAG2, and its termini resemble recombination signal sequences.

Zisupton The superfamily *Zisupton* was proposed by Bohne et al. (2012). Three related TE superfamilies (*Kyakuja*, *Dileera*, *Plavaka*) were also proposed (Iyer et al., 2014), although no consensus sequence for these three superfamilies has been reported. Fungal insertions of these groups of TEs are often associated with TET/JBP genes, which are responsible for the removal/modification of cytosine, or for the modification of thymine. Some *Zisupton* families in Repbase show similarity to *Kyakuja*, *Dileera* and *Plavaka*. This group of TEs is currently found only in chordates, fungi and red algae, but the presence of *Zisupton*-like proteins in other organisms indicates that they are distributed more widely. *Zisupton* families in fish encode a single protein containing one or two CCHH zinc fingers, a SWIM zinc finger, a DDD/E transposase, and SAP and Ulp1 protease domains.

Sola Three weakly related lineages of TEs, *Sola1*, *Sola2* and *Sola3*, constitute the *Sola* superfamily (Bao et al., 2009). The three *Sola* subgroups are quite different from one another, and Yuan and Wessler (2011) recognized them as superfamilies. *Sola* has no close relative in either bacteria or eukaryotes. *Sola3* shows target specificity against TTAA, and ends with GAG..CTC.

***Crypton*: a DNA transposon superfamily encoding tyrosine recombinase** DNA transposons that encode tyrosine recombinase (YR) are known from bacteria. *Tec* DNA transposons (*Tec1*, *Tec2* and *Tec3*) from ciliates are among the first eukaryotic DNA transposons that encode YR (Doak et al., 2003; Jacobs et al., 2003). *Crypton* was first reported in fungi (Goodwin et al., 2003), and is now known to be distributed among various eukaryotes that include fungi, animals and stramenopiles (Kojima and Jurka, 2011a). *Crypton* is proposed to be transposed via a circular DNA intermediate. The presence of YR suggests a relationship between *Crypton* and YR retrotransposons, like *DIRS*, but phylogenetic analysis does not support such a relationship.

Unlike DNA transposons encoding DDD/E transposase, *Crypton* does not have TIRs. Instead, at the termini of some *Cryptons* there are short direct repeats. Taking into account the mechanism of transposition of bacterial TEs encoding a YR, it is likely that one of these repeats is the terminus of the TE and the other is the target.

Crypton is subdivided into several groups (*CryptonA*, *CryptonF*, *CryptonI*, *CryptonS* and *CryptonV*), which may or may not share common ancestry in eukaryotes; they may have independently evolved from prokaryotic DNA transposons. In general, these *Crypton* groups have limited distribution. *CryptonF* is distributed among fungi and oomycetes; *CryptonF* in oomycetes is likely to have been horizontally transferred from fungi. *CryptonA* is distributed among animals such as medaka, sea urchins and sea anemones, and is the origin of several human genes (KCTD1, KIAA1958, ZMYM2, ZMYM3, ZMYM4 and QRICH1). *CryptonI* is distributed among insects that include mealworms, mosquitoes and triatomid bugs. *CryptonS* is distributed among stramenopiles (oomycetes and diatoms).

CryptonF encodes a protein that includes two domains, the YR and GCR1 DNA-binding domains. *CryptonA* encodes a protein that has only one known domain, YR. *CryptonS* encodes a protein that includes only one known domain, YR, but its C-terminal region (downstream of YR) is much longer than those of *CryptonA* and *CryptonI*. Many *CryptonS* families encode a second protein that includes a SET histone methyltransferase domain.

CryptonV is the latest characterized *Crypton* group (Kapitonov and Jurka, 2012). Some *CryptonV* families show target sequence specificity for microsatellites. The zebrafish genome harbors several autonomous and non-autonomous *CryptonV* families. There are many *Crypton*-type DNA transposons that are not yet characterized in detail. *CryptonH* is one such lineage and is found mainly in *Hydra magnipapillata* (Kojima and Jurka, 2014a). *LRS* repeats from zebrafish (Tracey, 2010) were revealed to be members of *CryptonH*. *CryptonC*, *CryptonR* and *CryptonX* are DNA transposons that encode a YR, and are found only in the Irish moss *Chondrus crispus* (Bao and Jurka, 2013a; Kojima and Jurka, 2013a).

***Helitron*: a DNA transposon superfamily encoding HUH nuclease** *Helitron* is a unique group of DNA transposons in eukaryotes (Kapitonov and Jurka, 2001). *Helitron* usually encodes one protein, which includes two enzymatic domains: one is helicase and the other is the rolling-circle replication initiator (Rep). Rep is also called a "Y2 transposase", because the conserved residues that are essential for transposition are two tyrosines. Upstream of these Y2 motifs is a HUH motif, in which the U is any bulky hydrophobic residue. This HUH motif, as well as the conserved tyrosines, are known in other groups of mobile genetic elements, such as the IS91 and IS605 families of bacterial DNA transposons.

The transposition mechanism of *Helitron* was experimentally characterized recently (Grabundzija et al., 2016). *Helitron* nicks and peels only one strand of its own DNA and integrates it at another site of the genome. Both single-stranded copies are healed by DNA repair machinery. *Helitron* can be subdivided into two

groups, *Helitron1* and *Helitron2*, although Repbase has not yet implemented this classification (Bao and Jurka, 2013b). *Helentron* is a group of *Helitron* families that encode an APE (Poulter et al., 2003). APEs of *Helentron*s are clearly close to those encoded by non-LTR retrotransposons that belong to the *CR1* group.

Like non-LTR retrotransposons, 3'-transduction is seen in *Helitron*s. During the transposition, *Helitron* proteins "peel through" the original 3'-terminus. As a result, the 3' downstream sequence from the original *Helitron* can be duplicated (Lai et al., 2005).

***Polinton*: an endogenous virus encoding DDD/E transposase/integrase** *Polinton*, also called *Maverick*, was reported as a long, complex DNA transposon superfamily (Kapitonov and Jurka, 2006; Pritham et al., 2007). The structure of *Polinton* indicates that it is a DNA transposon because it encodes a DDD/E transposase, has terminal inverted repeats, and generates 5-bp TSDs upon integration. *Polinton* is expected to transpose similarly to other DNA transposons, but it likely generates extrachromosomal DNA and replicates by itself using the encoded DNA polymerase B. Recently, *Polinton* was proposed to be a genome-integrated endogenous virus, and its viral form is designated Polintovirus (Krupovic et al., 2014a). This virus is analogous to bacteriophages and vertebrate endogenous retroviruses. *Tlr1* from ciliates is a DNA transposon family related to *Polinton* (Krupovic et al., 2016).

NON-AUTONOMOUS TEs

Almost all TE families potentially have non-autonomous derivatives. DNA transposons with DDD/E transposase usually have non-autonomous derivatives that only contain short fragments of both termini. Since DDD/E transposase recognizes only the terminal sequences and flanking nucleotides, such non-autonomous derivatives can successfully transpose and increase their copy number. These non-autonomous derivatives can be very short and sometimes it is hard to recognize the relatedness to their autonomous counterparts. Some non-autonomous DNA transposons are classified into designated groups, such as miniature inverted-repeat transposable elements (MITEs) (Bureau et al., 1996). *Tourist* and *Stowaway* are MITEs that depend for their mobilization on *Harbinger* and *Mariner*, respectively (Jurka and Kapitonov, 2001; Turcotte and Bureau, 2002). *Crypton* also has non-autonomous TE families (Kojima and Jurka, 2011a), composed of the short left and right terminal portions of the autonomous counterpart. *Polinton* elements also have non-autonomous derivatives, such as *Polinton-2N1_DR*.

Transposases usually recognize only the terminal sequences, raising the possibility that there are parasitic TE families that have unrelated sequence between two TE-derived terminal sequences. *Pack-MULE* is a term to describe non-autonomous *MuDR*-type DNA transposons that contain fragments of host genes, instead of the transposase genes (Jiang et al., 2004).

LTR retrotransposons and non-LTR retrotransposons also have non-autonomous derivatives. Extremely short non-autonomous LTR retrotransposons are called terminal-repeat retrotransposons in miniature (TRIMs), which are sometimes shorter than 500 bp (Witte et al., 2001). *Cassandra* is a unique group of non-autonomous LTR retrotransposons, because it has 5S rRNA-derived sequences inside its LTRs (Kalendar et al., 2008). Solo LTRs are not non-autonomous LTR retrotransposons, although they are frequently observed due to recombination after integration.

The frequent truncation of non-LTR retrotransposons, template switching during reverse transcription, and the short essential sequence for mobilization in the 3'-terminus lead to the evolution of composite non-autonomous TE families. One large group of non-autonomous non-LTR retrotransposons is the short interspersed elements (SINEs). SINEs are classified into four groups in Repbase based on the origin of their 5' part: *SINE1* for 7SL RNA (Ullu and Tschudi, 1984; Kriegs et al., 2007), *SINE2* for tRNA (Daniels and Deininger, 1985; Okada and Hamada, 1997), *SINE3* for 5S rRNA (Kapitonov and Jurka, 2003), and *SINEU* for U1 or U2 snRNA (Kojima, 2015). The former three groups contain internal promoters for RNA polymerase III for their transcription. The transcription of *SINEU* is not yet characterized, and if transcription by RNA polymerase III is a requirement of SINEs, *SINEU* may be excluded from the SINE category. *SINE28*, which has 28S rRNA-derived sequences, and SINEs with GC-rich sequences at the 5'-termini, have also been proposed (Longo et al., 2015; Suh et al., 2016).

Another way of classifying SINEs is based on the similarity of their central regions. *CORE-SINE* (Gilbert and Labuda, 1999), *V-SINE* (Ogiwara et al., 2002), *Deu-SINE* (or *Nin-SINE*) (Nishihara et al., 2006; Piskurek and Jackson, 2011), *Ceph-SINE* (Akasaki et al., 2010) and *Meta-SINE* (Nishihara et al., 2016) have been proposed, although Repbase does not use this classification because it contradicts the classification that is based on the origin of the 5' regions. Recently, similarity between *CORE-SINE* and *Ceph-SINE* was reported (Kojima, 2018b).

Besides SINEs, there are other groups of non-autonomous non-LTR retrotransposons. One is the bipartite non-autonomous non-LTR retrotransposons, which originated from the internal deletion of an autonomous non-LTR retrotransposon; *Vingi-1N1_EE* and the putative *Bov-A* family are examples (Ogiwara et al., 1999; Kojima et al., 2011). Considering their origin as a fusion of the 5'- and 3'-termini, this group corresponds to the canonical non-autonomous families of

DNA transposons. For unknown reasons, only a few clades of the non-LTR retrotransposons have this type of non-autonomous derivative (Kojima, 2018b). Another group includes derivatives of non-LTR retrotransposons that can encode one structural protein. They are represented by *HeT-A* and *HAL1* (Pardue et al., 1996; Bao and Jurka, 2010). The third group is represented by *SVA*; its members are composite, but their transcription depends on RNA polymerase II, unlike SINEs (Wang et al., 2005). *Sadhu* (from *Arabidopsis*) is another example (Rangwala et al., 2006). Processed pseudogenes are mobilized by non-LTR retrotransposons (Esnault et al., 2000), although they are not usually considered to be TEs because their copies have no capacity to transpose again.

FANZOR: A HITCHHIKING DOMAIN OF TRANSPOSITION

Fanzor is a unique mobile element that is associated with various TE superfamilies (Bao and Jurka, 2013b). Recent bioinformatics studies revealed that Fanzor, TnpB encoded by the IS605 family of ISs, the only protein encoded by the non-autonomous IS family IS1341, IscB encoded by ISC, and Cas9 in the CRISPR-Cas system, are all RuvC-like nucleases (Majorek et al., 2014; Kapitonov et al., 2015). Fanzor can be classified into two lineages, Fanzor1 and Fanzor2. Fanzor2 is associated with serine recombinase and is phylogenetically close to TnpB (Bao and Jurka, 2013b). Thus, Fanzor2 is likely a horizontally transferred IS607 family of ISs in eukaryotes. Fanzor1 is associated with *Mariner*, *Helitron*, *ISL2EU*, *MuDR*, *Sola2*, *Harbinger* and possibly also *Crypton*. The association of Fanzor1 with various TEs indicates that Fanzor1 is a helper for TEs, cleaving one strand of DNA during transposition. The endonuclease (APE) in one group of *Helitron*, *Helentron* (Poulter et al., 2003), may be an analog of Fanzor.

CONCLUDING REMARKS

Rapid progress in eukaryotic genome sequencing has revealed TEs that are diverse in sequence, structure and encoded protein composition. These bioinformatic findings have led to the discovery of new mechanisms for transposition, as well as of genome dynamics, like the case of *Helitron*s, and their contribution to gene shuffling. Many TEs with new sets of protein combinations are, undoubtedly, still waiting to be identified.

I thank Dr. Weidong Bao for discussion and critical reading of the manuscript.

REFERENCES

Abad, J. P., de Pablos, B., Osoegawa, K., de Jong, P. J., Martín-Gallardo, A., and Villasante, A. (2004) *TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. Mol. Biol. Evol. **21**, 1620–1624.

Akasaki, T., Nikaido, M., Nishihara, H., Tsuchiya, K., Segawa, S., and Okada, N. (2010) Characterization of a novel SINE superfamily from invertebrates: "Ceph-SINEs" from the genomes of squids and cuttlefish. Gene **454**, 8–19.

Aksoy, S., Williams, S., Chang, S., and Richards, F. F. (1990) SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs. Nucleic Acids Res. **18**, 785–792.

Arensburger, P., Piégu, B., and Bigot, Y. (2016) The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? Mob. Genet. Elements **6**, e1256852.

Arkhipova, I. R. (2017) Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob. DNA **8**, 19.

Arkhipova, I. R., Pyatkov, K. I., Meselson, M., and Evgen'ev, M. B. (2003) Retroelements containing introns in diverse invertebrate taxa. Nat. Genet. **33**, 123–124.

Babu, M. M., Iyer, L. M., Balaji, S., and Aravind, L. (2006) The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Res. **34**, 6505–6520.

Bao, W., and Jurka, J. (2010) Origin and evolution of LINE-1 derived "half-L1" retrotransposons (HAL1). Gene **465**, 9–16.

Bao, W., and Jurka, J. (2013a) DNA transposons from the red seaweed. Repbase Reports **13**, 2546–2720.

Bao, W., and Jurka, J. (2013b) Homologues of bacterial TnpB_ *IS605* are widespread in diverse eukaryotic transposable elements. Mob. DNA **4**, 12.

Bao, W., and Jurka, J. (2014) DNA transposons from the red seaweed. Repbase Reports **14**, 2–290.

Bao, W., Jurka, M. G., Kapitonov, V. V., and Jurka, J. (2009) New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol. Biol. Evol. **26**, 983–993.

Bao, W., Kapitonov, V. V., and Jurka, J. (2010) *Ginger* DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob. DNA **1**, 3.

Bao, W., Kojima, K. K., and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**, 11.

Biedler, J., and Tu, Z. (2003) Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. Mol. Biol. Evol. **20**, 1811–1825.

Blackman, R. K., Koehler, M. M., Grimaila, R., and Gelbart, W. M. (1989) Identification of a fully-functional *hobo* transposable element and its use for germ-line transformation of *Drosophila*. EMBO J. **8**, 211–217.

Böhne, A., Zhou, Q., Darras, A., Schmidt, C., Schartl, M., Galiana-Arnoux, D., and Volff, J. N. (2012) *Zisupton*--a novel superfamily of DNA transposable elements recently active in fish. Mol. Biol. Evol. **29**, 631–645.

Bureau, T. E., Ronald, P. C., and Wessler, S. R. (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, W. D., Malik, H. S., Rich, S. M., and Eickbush, T. H.

(2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia*. Mol. Biol. Evol. **19**, 619–630.

Burke, W. D., Müller, F., and Eickbush, T. H. (1995) R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. Nucleic Acids Res. **23**, 4628–4634.

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., Jurka, J., Michelmore, R. W., and Dubcovsky, J. (2011) Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. PLoS One **6**, e24230.

Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E., and Fraser, M. J. (1989) Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology **172**, 156–169.

Cavalli, G., and Paro, R. (1998) Chromo-domain proteins: linking chromatin structure to epigenetic regulation. Curr. Opin. Cell Biol. **10**, 354–360.

Chandler, M., de la Cruz, F., Dyda, F., Hickman, A. B., Moncalian, G., and Ton-Hoang, B. (2013) Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat. Rev. Microbiol. **11**, 525–538.

Chong, A. Y., Kojima, K. K., Jurka, J., Ray, D. A., Smit, A. F., Isberg, S. R., and Gongora, J. (2014) Evolution and gene capture in ancient endogenous retroviruses - insights from the crocodilian genomes. Retrovirology **11**, 71.

Claudianos, C., Brownlie, J., Russell, R., Oakeshott, J., and Whyard, S. (2002) *maT*--a clade of transposons intermediate between *mariner* and *Tc1*. Mol. Biol. Evol. **19**, 2101–2109.

Coy, M. R., and Tu, Z. (2005) Gambol and Tc1 are two distinct families of DD34E transposons: analysis of the *Anopheles gambiae* genome expands the diversity of the IS630-Tc1-mariner superfamily. Insect Mol. Biol. **14**, 537–546.

Curcio, M. J., and Derbyshire, K. M. (2003) The outs and ins of transposition: from mu to kangaroo. Nat. Rev. Mol. Cell Biol. **4**, 865–877.

Daniels, G. R., and Deininger, P. L. (1985) Repeat sequence families derived from mammalian tRNA genes. Nature **317**, 819–822.

de Chastonay, Y., Felder, H., Link, C., Aeby, P., Tobler, H., and Müller, F. (1992) Nucleotide sequence of PAT, a retroid element with unusual DR organization, isolated from *Panagrellus redivivus*. DNA Seq. **3**, 251–255.

de la Chaux, N., and Wagner, A. (2011) BEL/Pao retrotransposons in metazoan genomes. BMC Evol. Biol. **11**, 154.

Doak, T. G., Doerder, F. P., Jahn, C. L., and Herrick, G. (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA **91**, 942–946.

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.

Eickbush, T. H., and Malik, H. S. (2002) Origins and evolution of retrotransposons. *In*: Mobile DNA II. (eds.: Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M.), pp. 1111–1144. American Society of Microbiology Press, Washington DC.

Eisen, J. A., Benito, M. I., and Walbot, V. (1994) Sequence similarity of putative transposases links the maize *Mutator* autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res. **22**, 2634–2636.

Esnault, C., Maestre, J., and Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. **24**, 363–367.

Evgen'ev, M. B., Zelentsova, H., Shostak, N., Kozitsina, M., Barskyi, V., Lankenau, D. H., and Corces, V. G. (1997) *Penelope*, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. Proc. Natl. Acad. Sci. USA **94**, 196–201.

Farkašová, H., Hron, T., Pačes, J., Hulva, P., Benda, P., Gifford, R. J., and Elleder, D. (2017) Discovery of an endogenous Deltaretrovirus in the genome of long-fingered bats (Chiroptera: Miniopteridae). Proc. Natl. Acad. Sci. USA **114**, 3145–3150.

Felger, I., and Hunt, J. A. (1992) A non-LTR retrotransposon from the Hawaiian *Drosophila*: the LOA element. Genetica **85**, 119–130.

Feschotte, C. (2004) *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS*1016* insertion sequences. Mol. Biol. Evol. **21**, 1769–1780.

Fillingham, J. S., Thing, T. A., Vythilingum, N., Keuroghlian, A., Bruno, D., Golding, G. B., and Pearlman, R. E. (2004) A non-long terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan *Tetrahymena thermophila*. Eukaryot. Cell **3**, 157–169.

Finnegan, D. J. (1989) Eukaryotic transposable elements and genome evolution. Trends Genet. **5**, 103–107.

Fraser, M. J., Smith, G. E., and Summers, M. D. (1983) Acquisition of host cell DNA sequences by baculoviruses: relationship between host DNA insertions and FP mutants of *Autographa californica* and *Galleria mellonella* nuclear polyhedrosis viruses. J. Virol. **47**, 287–300.

Fu, Y., Kawabe, A., Etcheverry, M., Ito, T., Toyoda, A., Fujiyama, A., Colot, V., Tarutani, Y., and Kakutani, T. (2013) Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. EMBO J. **32**, 2407–2417.

Gabriel, A., Yen, T. J., Schwartz, D. C., Smith, C. L., Boeke, J. D., Sollner-Webb, B., and Cleveland, D. W. (1990) A rapidly rearranging retrotransposon within the miniexon gene locus of *Crithidia fasciculata*. Mol. Cell. Biol. **10**, 615–624.

Geering, A. D., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., McTaggart, A. R., Scalabrin, S., Vezzulli, S., Wing, R. A., et al. (2014) Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat. Commun. **5**, 5269.

Gilbert, C., and Feschotte, C. (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol. **8**.

Gilbert, N., and Labuda, D. (1999) CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc. Natl. Acad. Sci. USA **96**, 2869–2874.

Gladyshev, E. A., and Arkhipova, I. R. (2007) Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. Proc. Natl. Acad. Sci. USA **104**, 9352–9357.

Glöckner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M. A., Cox, E., Eichinger, L., Noegel, A. A., and Rosenthal, A. (2001) The complex repeats of *Dictyostelium discoideum*. Genome Res. **11**, 585–594.

Goodwin, T. J. D., Butler, M. I., and Poulter, R. T. M. (2003) Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology **149**, 3099–3109.

Goodwin, T. J. D., and Poulter, R. T. M. (2004) A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. **21**, 746–759.

Grabundzija, I., Messing, S. A., Thomas, J., Cosby, R. L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda,

A., et al. (2016) A *Helitron* transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat. Commun. **7**, 10716.

Han, G. Z., and Worobey, M. (2012) An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog. **8**, e1002790.

Han, M. J., Xiong, C. L., Zhang, H. B., Zhang, M. Q., Zhang, H. H., and Zhang, Z. (2015) The diversification of PHIS transposon superfamily in eukaryotes. Mob. DNA **6**, 12.

Han, M. J., Xu, H. E., Zhang, H. H., Feschotte, C., and Zhang, Z. (2014) *Spy*: a new group of eukaryotic DNA transposons without target site duplications. Genome Biol. Evol. **6**, 1748–1757.

Hehl, R., Nacken, W. K., Krause, A., Saedler, H., and Sommer, H. (1991) Structural analysis of Tam3, a transposable element from *Antirrhinum majus*, reveals homologies to the Ac element from maize. Plant Mol. Biol. **16**, 369–371.

Hua-Van, A., and Capy, P. (2008) Analysis of the DDE motif in the Mutator superfamily. J. Mol. Evol. **67**, 670–681.

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., et al. (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell **166**, 102–114.

Inoue, Y., Saga, T., Aikawa, T., Kumagai, M., Shimada, A., Kawaguchi, Y., Naruse, K., Morishita, S., Koga, A., and Takeda, H. (2017) Complete fusion of a transposon and herpesvirus created the *Teratorn* mobile element in medaka fish. Nat. Commun. **8**, 551.

Iyer, L. M., Zhang, D., de Souza, R. F., Pukkila, P. J., Rao, A., and Aravind, L. (2014) Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. Proc. Natl. Acad. Sci. USA **111**, 1676–1683.

Jacobs, M. E., Sánchez-Blanco, A., Katz, L. A., and Klobutcher, L. A. (2003) Tec3, a new developmentally eliminated DNA element in *Euplotes crassus*. Eukaryot. Cell **2**, 103–114.

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., and Wessler, S. R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature **431**, 569–573.

Jurka, J., and Bao, W. (2008) A distinct subgroup of Kolobok-type DNA transposons. Repbase Reports **8**, 168–174.

Jurka, J., and Kapitonov, V. V. (2001) *PIFs* meet *Tourists* and *Harbingers*: a superfamily reunion. Proc. Natl. Acad. Sci. USA **98**, 12315–12316.

Jurka, J., Walichiewicz, J., and Milosavljevic, A. (1992) Prototypic sequences for human repetitive DNA. J. Mol. Evol. **35**, 286–291.

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., and Schulman, A. H. (2008) *Cassandra* retrotransposons carry independently transcribed 5S RNA. Proc. Natl. Acad. Sci. USA **105**, 5833–5838.

Kapitonov, V. V., and Jurka, J. (1999) Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica **107**, 27–37.

Kapitonov, V. V., and Jurka, J. (2001) Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA **98**, 8714–8719.

Kapitonov, V. V., and Jurka, J. (2003) A novel class of SINE elements derived from 5S rRNA. Mol. Biol. Evol. **20**, 694–702.

Kapitonov, V. V., and Jurka, J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biol. **3**, e181.

Kapitonov, V. V., and Jurka, J. (2006) Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. USA **103**, 4540–4545.

Kapitonov, V. V., and Jurka, J. (2007a) Chapaev - a novel superfamily of DNA transposons. Repbase Reports **7**, 774–781.

Kapitonov, V. V., and Jurka, J. (2007b) IS4EU, a novel superfamily of eukaryotic DNA transposons. Repbase Reports **7**, 143–147.

Kapitonov, V. V., and Jurka, J. (2007c) Kolobok, a novel superfamily of eukaryotic DNA transposons. Repbase Reports **7**, 111–122.

Kapitonov, V. V., and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. **9**, 411–412; author reply 414.

Kapitonov, V. V., and Jurka, J. (2009) Proto1 non-LTR retrotransposons from the Naegleria gruberi amoeboflagellate genome. Repbase Reports **9**, 1144–1148.

Kapitonov, V. V., and Jurka, J. (2010) Ambal, a novel clade of non-LTR retrotransposons from diatoms. Repbase Reports **10**, 102–108.

Kapitonov, V. V., and Jurka, J. (2012) CryptonV, a group of target-site specific Crypton DNA transposons from cnidarians. Repbase Reports **12**, 2034.

Kapitonov, V. V., and Jurka, J. (2014) piggyBacA - a novel group of piggyBac transposons. Repbase Reports **14**, 2322–2325.

Kapitonov, V. V., and Koonin, E. V. (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol. Direct **10**, 20.

Kapitonov, V. V., Makarova, K. S., and Koonin, E. V. (2015) ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. J. Bacteriol. **198**, 797–807.

Kapitonov, V. V., Tempel, S., and Jurka, J. (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene **448**, 207–213.

Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. P., and Pybus, O. G. (2009) Macroevolution of complex retroviruses. Science **325**, 1512.

Katzourakis, A., Tristem, M., Pybus, O. G., and Gifford, R. J. (2007) Discovery and analysis of the first endogenous lentivirus. Proc. Natl. Acad. Sci. USA **104**, 6261–6265.

Kawakami, K., and Shima, A. (1999) Identification of the *Tol2* transposase of the medaka fish *Oryzias latipes* that catalyzes excision of a nonautonomous *Tol2* element in zebrafish *Danio rerio*. Gene **240**, 239–244.

Kojima, K. K. (2015) A new class of SINEs with snRNA gene-derived heads. Genome Biol. Evol. **7**, 1702–1712.

Kojima, K. K. (2018a) Human transposable elements in Repbase: genomic footprints from fish to humans. Mob. DNA **9**, 2.

Kojima, K. K. (2018b) LINEs contribute to the origins of middle bodies of SINEs besides 3' Tails. Genome Biol. Evol. **10**, 370–379.

Kojima, K. K., and Fujiwara, H. (2003) Evolution of target specificity in R1 clade non-LTR retrotransposons. Mol. Biol. Evol. **20**, 351–361.

Kojima, K. K., and Fujiwara, H. (2004) Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. Mol. Biol. Evol. **21**, 207–217.

Kojima, K. K., and Fujiwara, H. (2005a) An extraordinary retrotransposon family encoding dual endonucleases. Genome Res. **15**, 1106–1117.

Kojima, K. K., and Fujiwara, H. (2005b) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. Mol. Biol. Evol. **22**, 2157–2165.

Kojima, K. K., and Jurka, J. (2011a) *Crypton* transposons: identification of new diverse families and ancient domestication events. Mob. DNA **2**, 12.

Kojima, K. K., and Jurka, J. (2011b) Kiri non-LTR retrotransposons from the southern house mosquito. Repbase Reports **11**, 120–129.

Kojima, K. K., and Jurka, J. (2011c) A lineage of non-LTR retrotransposons encoding an OTU cysteine protease from the yellow fever mosquito. Repbase Reports **11**, 1124–1128.

Kojima, K. K., and Jurka, J. (2011d) Sagan, a new group of DNA transposons belonging to the Mariner/Tc1/IS630 superfamily. Repbase Reports **11**, 2305–2314.

Kojima, K. K., and Jurka, J. (2013a) DNA transposons from the red seaweed. Repbase Reports **13**, 2551–2657.

Kojima, K. K., and Jurka, J. (2013b) A superfamily of DNA transposons targeting multicopy small RNA genes. PLoS One **8**, e68260.

Kojima, K. K., and Jurka, J. (2014a) CryptonH DNA transposons from zebrafish. Repbase Reports **14**, 1414.

Kojima, K. K., and Jurka, J. (2014b) HarbingerS, a novel clade of Harbinger DNA transposons encoding a SET domain histone lysine methyltransferase. Repbase Reports **14**, 2243–2250.

Kojima, K. K., and Jurka, J. (2015) Ancient origin of the U2 small nuclear RNA gene-targeting non-LTR retrotransposons *Utopia*. PLoS One **10**, e0140084.

Kojima, K. K., Kapitonov, V. V., and Jurka, J. (2011) Recent expansion of a new *Ingi*-related clade of *Vingi* non-LTR retrotransposons in hedgehogs. Mol. Biol. Evol. **28**, 17–20.

Koonin, E. V., Zhou, S., and Lucchesi, J. C. (1995) The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res. **23**, 4229–4233.

Kordis, D., and Gubensek, F. (1999) Horizontal transfer of non-LTR retrotransposons in vertebrates. Genetica **107**, 121–128.

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet. **23**, 158–161.

Krupovic, M., Bamford, D. H., and Koonin, E. V. (2014a) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biol. Direct **9**, 6.

Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D., and Koonin, E. V. (2014b) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. **12**, 36.

Krupovic, M., Yutin, N., and Koonin, E. V. (2016) Fusion of a superfamily 1 helicase and an inactivated DNA polymerase is a signature of common evolutionary history of Polintons, polinton-like viruses, Tlr1 transposons and transpovirons. Virus Evol. **2**, vew019.

Kusumoto, M., Ooka, T., Nishiya, Y., Ogura, Y., Saito, T., Sekine, Y., Iwata, T., Akiba, M., and Hayashi, T. (2011) Insertion sequence-excision enhancer removes transposable elements from bacterial genomes and induces various genomic deletions. Nat. Commun. **2**, 152.

Lai, J., Li, Y., Messing, J., and Dooner, H. K. (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. Proc. Natl. Acad. Sci. USA **102**, 9068–9073.

Lescot, M., Hingamp, P., Kojima, K. K., Villar, E., Romac, S., Veluchamy, A., Boccara, M., Jaillon, O., Iudicone, D., Bowler, C., et al. (2016) Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. ISME J. **10**, 1134–1146.

Liu, W., Pan, S., Yang, H., Bai, W., Shen, Z., Liu, J., and Xie, Y. (2012) The first full-length endogenous hepadnaviruses: identification and analysis. J. Virol. **86**, 9510–9513.

Llorens, C., Futami, R., Covelli, L., Dominguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., et al. (2011) The *Gypsy* Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. **39**, D70–D74.

Longo, M. S., Brown, J. D., Zhang, C., O'Neill, M. J., and O'Neill, R. J. (2015) Identification of a recently active mammalian SINE derived from ribosomal RNA. Genome Biol. Evol. **7**, 775–788.

Lorenzi, H. A., Robledo, G., and Levin, M. J. (2006) The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. Mol. Biochem. Parasitol. **145**, 184–194.

Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72**, 595–605.

Lyozin, G. T., Makarova, K. S., Velikodvorskaja, V. V., Zelentsova, H. S., Khechumian, R. R., Kidwell, M. G., Koonin, E. V., and Evgen'ev, M. B. (2001) The structure and evolution of *Penelope* in the *virilis* species group of *Drosophila*: an ancient lineage of retroelements. J. Mol. Evol. **52**, 445–456.

Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J., and Belshaw, R. (2012) *Env*-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. USA **109**, 7385–7390.

Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., and Bujnicki, J. M. (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res. **42**, 4160–7941.

Majumdar, S., Singh, A., and Rio, D. C. (2013) The human THAP9 gene encodes an active *P*-element DNA transposase. Science **339**, 446–448.

Malik, H. S., Burke, W. D., and Eickbush, T. H. (1999) The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. **16**, 793–805.

Malik, H. S., and Eickbush, T. H. (2000) NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. Genetics **154**, 193–203.

Malik, H. S., Henikoff, S., and Eickbush, T. H. (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. **10**, 1307–1318.

Marín, I., and Llorens, C. (2000) *Ty3/Gypsy* retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol. Biol. Evol. **17**, 1040–1049.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA **36**, 344–355.

McClintock, B. (1954) Mutations in maize and chromosomal aberrations in Neurospora. Carnegie Inst. of Wash. Year Book **53**, 254–260.

Morrish, T. A., Garcia-Perez, J. L., Stamato, T. D., Taccioli, G. E., Sekiguchi, J., and Moran, J. V. (2007) Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature **446**, 208–212.

Nishihara, H., Plazzi, F., Passamonti, M., and Okada, N. (2016) MetaSINEs: broad distribution of a novel SINE superfamily in animals. Genome Biol. Evol. **8**, 528–539.

Nishihara, H., Smit, A. F., and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. **16**, 864–874.

Novikova, O., Fet, V., and Blinov, A. (2009) Non-LTR retrotransposons in fungi. Funct. Integr. Genomics **9**, 27–42.

O'Brochta, D. A., Warren, W. D., Saville, K. J., and Atkinson, P. W. (1996) *Hermes*, a functional non-Drosophilid insect gene vector from *Musca domestica*. Genetics **142**, 907–914.

O'Hare, K., and Rubin, G. M. (1983) Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. Cell **34**, 25–35.

Ogiwara, I., Miya, M., Ohshima, K., and Okada, N. (1999) Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Mol. Biol. Evol. **16**, 1238–1250.

Ogiwara, I., Miya, M., Ohshima, K., and Okada, N. (2002) V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. Genome Res. **12**, 316–324.

Okada, N., and Hamada, M. (1997) The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. J. Mol. Evol. **44 Suppl 1**, S52–S56.

Pardue, M. L., Danilevskaya, O. N., Lowenhaupt, K., Wong, J., and Erby, K. (1996) The *gag* coding region of the *Drosophila* telomeric retrotransposon, *HeT-A*, has an internal frame shift and a length polymorphic region. J. Mol. Evol. **43**, 572–583.

Peacock, C. S., Seeger, K., Harris, D., Murphy, L., Ruiz, J. C., Quail, M. A., Peters, N., Adlem, E., Tivey, A., Aslett, M., et al. (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat. Genet. **39**, 839–847.

Peterson, P. A. (1953) A mutable pale-green locus in maize. Genetics **38**, 682–683.

Piégu, B., Bire, S., Arensburger, P., and Bigot, Y. (2015) A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol. **86**, 90–109.

Piskurek, O., and Jackson, D. J. (2011) Tracking the ancestry of a deeply conserved eumetazoan SINE domain. Mol. Biol. Evol. **28**, 2727–2730.

Poulter, R. T. M., and Goodwin, T. J. D. (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet. Genome Res. **110**, 575–588.

Poulter, R. T. M., Goodwin, T. J. D., and Butler, M. I. (2003) Vertebrate helentrons and other novel *Helitrons*. Gene **313**, 201–212.

Pritham, E. J., Putliwala, T., and Feschotte, C. (2007) *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene **390**, 3–17.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317**, 86–94.

Pyatkov, K. I., Arkhipova, I. R., Malkova, N. V., Finnegan, D. J., and Evgen'ev, M. B. (2004) Reverse transcriptase and endonuclease activities encoded by *Penelope*-like retroelements. Proc. Natl. Acad. Sci. USA **101**, 14719–14724.

Rangwala, S. H., Elumalai, R., Vanier, C., Ozkan, H., Galbraith, D. W., and Richards, E. J. (2006) Meiotically stable natural epialleles of *Sadhu*, a novel Arabidopsis retroposon. PLoS Genet. **2**, e36.

Robertson, D. S. (1978) Characterization of a mutator system in maize. Mutat. Res. **51**, 21–28.

Sakano, H., Hüppi, K., Heinrich, G., and Tonegawa, S. (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature **280**, 288–294.

Sarkar, A., Sim, C., Hong, Y. S., Hogan, J. R., Fraser, M. J., Robertson, H. M., and Collins, F. H. (2003) Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Mol. Genet. Genomics **270**, 173–180.

Shao, H., and Tu, Z. (2001) Expanding the diversity of the *IS630-Tc1-mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics **159**, 1103–1115.

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., and Chandler, M. (2015) Everyman's Guide to Bacterial Insertion Sequences. Microbiol. Spectr. **3**, MDNA3-0030-2014.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. **34**, D32–D36.

Sinzelle, L., Kapitonov, V. V., Grzela, D. P., Jursch, T., Jurka, J., Izsvák, Z., and Ivics, Z. (2008) Transposition of a reconstructed *Harbinger* element in human cells and functional homology with two transposon-derived cellular genes. Proc. Natl. Acad. Sci. USA **105**, 4715–4720.

Smit, A. F. (1996) The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. **6**, 743–748.

Starnes, J. H., Thornbury, D. W., Novikova, O. S., Rehmeyer, C. J., and Farman, M. L. (2012) Telomere-targeted retrotransposons in the rice blast fungus *Magnaporthe oryzae*: agents of telomere instability. Genetics **191**, 389–406.

Suh, A., Witt, C. C., Menger, J., Sadanandan, K. R., Podsiadlowski, L., Gerth, M., Weigert, A., McGuire, J. A., Mudge, J., Edwards, S. V., et al. (2016) Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat. Commun. **7**, 11396.

Tellier, M., Bouuaert, C. C., and Chalmers, R. (2015) Mariner and the ITm superfamily of transposons. Microbiol. Spectr. **3**, MDNA3-0033-2014.

Tracey, A. (2010) Unclassified repeat from zebrafish. Repbase Reports **10**, 241.

Tudor, M., Lobocka, M., Goodell, M., Pettitt, J., and O'Hare, K. (1992) The pogo transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet, **232**, 126–134.

Turcotte, K., and Bureau, T. (2002) Phylogenetic analysis reveals *Stowaway*-like elements may represent a fourth family of the IS*630*-Tc1-*mariner* superfamily. Genome **45**, 82–90.

Ullu, E., and Tschudi, C. (1984) *Alu* sequences are processed 7SL RNA genes. Nature **312**, 171–172.

Ustyantsev, K., Novikova, O., Blinov, A., and Smyshlyaev, G. (2015) Convergent evolution of ribonuclease h in LTR retrotransposons and retroviruses. Mol. Biol. Evol. **32**, 1197–1207.

Volff, J. N., Hornung, U., and Schartl, M. (2001) Fish retroposons related to the *Penelope* element of *Drosophila virilis* define a new group of retrotransposable elements. Mol. Genet. Genomics **265**, 711–720.

Volff, J. N., Lehrach, H., Reinhardt, R., and Chourrout, D. (2004) Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. Mol. Biol. Evol. **21**, 2022–2033.

Walsh, A. M., Kortschak, R. D., Gardner, M. G., Bertozzi, T., and Adelson, D. L. (2013) Widespread horizontal transfer of retrotransposons. Proc. Natl. Acad. Sci. USA **110**, 1012–1016.

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., and Batzer, M. A. (2005) SVA elements: a hominid-specific retroposon family. J. Mol. Biol. **354**, 994–1007.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007) A unified classification system for eukary-

otic transposable elements. Nat. Rev. Genet. **8**, 973–982.
Witte, C. P., Le, Q. H., Bureau, T., and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc. Natl. Acad. Sci. USA **98**, 13778–13783.
Yuan, Y. W., and Wessler, S. R. (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc. Natl. Acad. Sci. USA **108**, 7884–7889.
Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W. B., and Wessler, S. R. (2001) *P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. Proc. Natl. Acad. Sci. USA **98**, 12572–12577.
Zimmerly, S., Guo, H., Perlman, P. S., and Lambowitz, A. M. (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. Cell **82**, 545–554.