Structural and sequence diversity of eukaryotic transposable elements

Kenji K. Kojima

1Genetic Information Research Institute, 465 Fairchild Drive, Suite 201, Mountain View, CA 94043, USA
2Department 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...
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 critics (Piégut 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>APE</td>
<td>L1, Proto1, Tx1, Proto2, RTE, RTEX, RTETP, I, Nimb, Ingi, Vingi, Tod1, Loa, R1, Outcast, Jockey, CR1, L2, L2A, L2B, Kiri, Rex1, Crack, Daphne, Ambal</td>
<td>RI_</td>
<td>RA</td>
<td>Non-LTR retrotransposon</td>
</tr>
<tr>
<td>RT</td>
<td>APE + RLE</td>
<td>Dualen</td>
<td>n/d</td>
<td>n/d</td>
<td>Dualen</td>
</tr>
<tr>
<td>RT</td>
<td>RLE</td>
<td>CRE, NeSL, R4, R2, HERO</td>
<td>RIR</td>
<td>RP</td>
<td>Non-LTR retrotransposon</td>
</tr>
<tr>
<td>RT</td>
<td>GIY-YIG</td>
<td>Penelope</td>
<td>RPP</td>
<td>RG</td>
<td>Penelope-like element</td>
</tr>
<tr>
<td>RT</td>
<td>–</td>
<td>Penelope</td>
<td>n/d</td>
<td>RO</td>
<td>Athena, Coprina</td>
</tr>
<tr>
<td>RT</td>
<td>DDE</td>
<td>Copia, BEL, Gypsy, endogenous retrovirus</td>
<td>RL_</td>
<td>RD</td>
<td>LTR retrotransposon</td>
</tr>
<tr>
<td>RT</td>
<td>YR</td>
<td>DIRS</td>
<td>DY_</td>
<td>DD</td>
<td>DNA transposon</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>–</td>
<td>DDE</td>
<td>Academ, Dada, EnSpn, Ginger1, Ginger2, Harbinger, hAT, IS3EU, ISL2EU, Kolobok, Mariner, Merlin, MuDR, Novosib, P, piggyBac, Sola, Transib, Zator, Zisupton</td>
<td>DT_</td>
<td>DD</td>
<td>DNA transposon</td>
</tr>
<tr>
<td>–</td>
<td>YR</td>
<td>Crypton</td>
<td>DYC</td>
<td>DY</td>
<td>Crypton</td>
</tr>
<tr>
<td>–</td>
<td>HUH</td>
<td>Helitron</td>
<td>DHH</td>
<td>DH</td>
<td>Helitron</td>
</tr>
<tr>
<td>PolB</td>
<td>DDE</td>
<td>Pointon</td>
<td>DMM</td>
<td>BD</td>
<td>Polinton</td>
</tr>
</tbody>
</table>

‘_’ 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.

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 DDE/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 Sireivirus. The representatives of the three ICTV genera are: Ty1 from the budding yeast Saccharomyces cerevisiae for Pseudovirus; SIRE from the soybean Glycine max for Sireivirus; 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, Ty1 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” (Marin 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.

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 env-

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

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copia</td>
<td>10,595</td>
</tr>
<tr>
<td>Gypsy</td>
<td>6,694</td>
</tr>
<tr>
<td>BEL</td>
<td>1,855</td>
</tr>
<tr>
<td>ERV</td>
<td></td>
</tr>
<tr>
<td>ERV1</td>
<td>1,967</td>
</tr>
<tr>
<td>ERV2</td>
<td>1,266</td>
</tr>
<tr>
<td>ERV3</td>
<td>657</td>
</tr>
<tr>
<td>ERV4</td>
<td>187</td>
</tr>
<tr>
<td>Lentivirus</td>
<td>4</td>
</tr>
<tr>
<td>Unclassified ERV</td>
<td>325</td>
</tr>
<tr>
<td>Unclassified LTR</td>
<td>719</td>
</tr>
<tr>
<td>DIRS</td>
<td>418</td>
</tr>
</tbody>
</table>
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 genome 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).

**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, Cavenovirus, Petuvirus, Rosadnavirus, Solendovirus, Soymovirus and Tungrovirus. In addition,

---

**Table 3. Relationships between endogenous and infectious retroviruses**

<table>
<thead>
<tr>
<th>Endogenous retrovirus</th>
<th>Infectious retrovirus (genus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERV1</td>
<td>Gammaretrovirus, Epsilonretrovirus</td>
</tr>
<tr>
<td>ERV2</td>
<td>Alpharetrovirus, Betaretrovirus</td>
</tr>
<tr>
<td>ERV3</td>
<td>n/d</td>
</tr>
<tr>
<td>ERV4</td>
<td>n/d</td>
</tr>
<tr>
<td>Endogenous deltaretrovirus (EDV)</td>
<td>Deltaretrovirus</td>
</tr>
<tr>
<td>Endogenous lentivirus (ELV)</td>
<td>Lentivirus</td>
</tr>
<tr>
<td>Endogenous foamy virus (EFV)</td>
<td>Spumavirus</td>
</tr>
</tbody>
</table>

n/d, not defined.
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, RTEP, I, Nimb, IngI, VingI, TadI, 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.

**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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Clade</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRE</td>
<td>CRE</td>
<td>43</td>
</tr>
<tr>
<td>R2</td>
<td>R4</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Hero</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>NeSL</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>R2</td>
<td>159</td>
</tr>
<tr>
<td>Dualen</td>
<td>RandI/Dualen</td>
<td>13</td>
</tr>
<tr>
<td>L1</td>
<td>Proto1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>1,690</td>
</tr>
<tr>
<td></td>
<td>Tx1</td>
<td>273</td>
</tr>
<tr>
<td>RTE</td>
<td>RTEP</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Proto2</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>RTEX</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td>RTE</td>
<td>487</td>
</tr>
<tr>
<td>I</td>
<td>Outcast</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>IngI</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>VingI</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Nimb</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>Tad1</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>Loa</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>R1</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td>Jockey</td>
<td>243</td>
</tr>
<tr>
<td>CR1</td>
<td>Rex1</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>CR1</td>
<td>803</td>
</tr>
<tr>
<td></td>
<td>Kiri</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td>L2A</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>L2B</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Crack</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Daphne</td>
<td>227</td>
</tr>
<tr>
<td>Ambal</td>
<td>Penelope</td>
<td>8</td>
</tr>
<tr>
<td>Penelope</td>
<td>Penelope</td>
<td>477</td>
</tr>
<tr>
<td>SINE</td>
<td>SINE1/7SL</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>SINE2/tRNA</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>SINE3/5S</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>SINEU</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Unclassified SINE</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Unclassified non-LTR retrotransposon</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>7,341</td>
</tr>
</tbody>
</table>
is now clear that many independent group (Putnam et al., 2007), or as a part of sequence-specific. The only clade that encodes both RLE and APE simultaneously, even though some is the only clade that encodes both RLE and APE simultaneously, even though some lack RLE. The clade can be considered an independent group (Putnam et al., 2007), or as a part of the R2 group (Eickbusch and Malik, 2002).

R2 group The R2 group is one of the five original groups (Eickbusch and Malik, 2002). The R2 group includes the clades R2, R4, NeSL and Hero. The R2 clade (or super-clade) 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 motifs 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 Eickbusch, 2000; Kojima and Fujiwara, 2005b; Kojima and Jurka, 2015).

Duali group The Duali group includes only one clade, Duali (from dual endonucleases), also called RandI (Kojima and Fujiwara, 2005a). The Duali clade is the only clade that encodes both RLE and APE simultaneously, even though some Duali families, such as RandI-1_ACas, lack RLE. Duali 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 Duali families such as Duali-5_CCu and Duali-1_GCcr encode a protein longer than 5,000 residues (Lescot et al., 2016). The structure and phylogenetic position of Duali 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 (Eickbusch 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 (Eickbusch 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 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 is from the bovine Bos taurus, Expander from the pufferfish T. rubripes, and SR2 from the blood-fluke 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) (Eickbusch 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 Duali, 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, TadI, 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 mosquitos, 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). LITc, 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 TadI, 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 those found in other families in the I group. Elements belonging to the TadI 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 HeTA) 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 (Vollff 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 (Lyzin et al., 2001; Vollff 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 (Gladyashev 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 TRANSPONSONS (DNA TRANSPONSONS)

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 tertiary 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, MuT, 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.
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

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

**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/Do) from maize, and Tam3 from the snapdragon *Antirrhinum majus* (McClintock, 1950; Blackman et al., 1989; Hohl 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, stramenopiles, 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
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 Alloherpesviridae.

Harbinger The Harbinger superfamily, or the PIF/En/Spm 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 non-autonomous 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 (Zhang et al., 2001). HarbingerS is a group of Harbinger families that encode three proteins: DDD/E transposase, SET histone methyltransferase and an unknown protein (Kojima and Jurka, 2007a). HarbingerS families generate 2-bp TSDs. The transposase proteins encoded by Harbinger 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(DJ) 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(DJ) 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 TIRs and the other is the target.

Crypton A long-standing debate regarding the origin of TEs encoding a YR, it is likely that one of these repeats is the terminus of the TE and the other is the target.

A long-standing debate regarding the origin of TEs encoding a YR, it is likely that one of these repeats is the terminus of the TE and the other is the target. A long-standing debate regarding the origin of TEs encoding a YR, it is likely that one of these repeats is the terminus of the TE and the other is the target. A long-standing debate regarding the origin of TEs encoding a YR, it is likely that one of these repeats is the terminus of the TE and the other is the target.

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). Helitron is a group of Helitron families that encode an APE (Poulter et al., 2003). APEs of Helitrons 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 Helitrons. 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. Trl1 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 (TRMs), 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 itsLTRs (Kalender 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’-terminus, 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 (Akasaski 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; Vingt-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 HeTA 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 IS607 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 Helitrons, 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**


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.


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.


Burke, W. D., Malik, H. S., Rich, S. M., and Eickbush, T. H.


Diversity of eukaryotic transposable elements


Diversity of eukaryotic transposable elements


