1

Transposable Elements in Eukaryotes

Aurélie HUA-VAN
Évolution, Génomes, Comportement et Écologie (EGCE), CNRS, IRD, Université Paris-Saclay, Gif-sur-Yvette, France

COPYRIGHTED MATERIAL

1.1. Introduction

While transposable elements (TEs) are found in all three domains of life, they have expanded the most over time in eukaryotes, such that they sometimes reach 80% of the genome, with copy number above millions. It is also in eukaryotes that their diversity is the highest. Different kinds of transposition strategies have been developed by these selfish DNAs. Each species can host thousands of different TE families, and even closely related species may have specific TE families. Their evolutionary history is quite complicated, as is their relationship with their respective host. At the mercy of their host's silencing strategy, they can manage to escape repression and subsequent extinction by invading new species, where they evolve and amplify freely until defenses are put in place by this new host. The old companionship they maintain with their host is still visible, unlike in prokaryotes, where the turnover is very rapid in order to keep their genomes small.

In this chapter, I will review the basics of structure and transposition mechanisms of eukaryotic TEs, their origins, what we know about their distribution and abundance in eukaryotes, and finally the impact of such selfish elements in genome size, structure, function and evolution.

Transposable Elements and Genome Evolution,
coordinated by Aurélie HUA-VAN and Pierre CAPY. © ISTE Ltd 2024.

1.2. Classification, structure and transposition mechanism

The first attempt at TE classification was by Finnegan (1989). At this time, three types of TEs were known, called LTR retroelements, non-LTR retroelements (LINEs) and DNA transposons. Two classes were defined, based on the nature of the transposition intermediate, with Class I (LTR and non-LTR retroelements) transposing through an RNA intermediate and Class II (DNA transposons) directly from DNA to DNA (Figure 1.1).

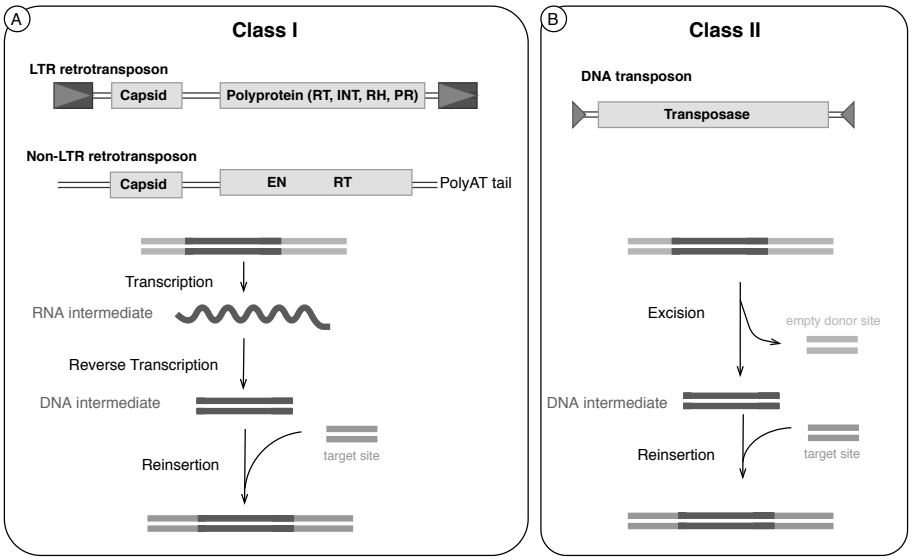

Figure 1.1. *The two classes of transposable elements as defined by Finnegan (1989). (A) Class I is represented by LTR retrotransposons and non-LTR retrotransposons, which transpose through a copy-and-paste mechanism. An inserted copy is transcribed, then reverse transcribed into DNA and integrated at a new genomic location. RT: reverse transcriptase; INT: integrase; RH: ribonuclease H; PR: protease; EN: endonuclease. (B) Class II transposons transpose through a cut-and-paste mechanism: The copy is excised from the genome and reinserted into the target site. For a color version of this figure, see www.iste.co.uk/huavan/transposable.zip*

For most non-specialist researchers, TEs classically belong to one of these three types. However, in the last few decades, a huge diversity of TEs has been uncovered in eukaryotes, which makes this classification too simplistic. Most of the new TE groups were identified through genome mining. These correspond to less active groups, and deciphering their transposition mechanism remains challenging. However, refinement of the former classification or new classification scheme was proposed.

In 2003, a classification based on the types of enzymes responsible for integration was proposed (Curcio and Derbyshire 2003). Interestingly, similar catalytic activities (represented by catalytic motifs) could be found independently in the two main classes. The classification integrated both eukaryotic and prokaryotic elements. However, this classification did not catch all of the diversity that can be found in TE transposition, since it focused mainly on the integration process.

A major revision was proposed by Wicker et al. (2007). While the two main classes remain valid, new hierarchical subdivisions such as subclasses, orders and superfamilies were added to take into account variation in the transposition mechanism (number of cut strands), the structure and finally the phylogeny/sequence homology. The classification system included the most recently discovered types of TEs, but remained focused on eukaryotic TEs; it did not integrate some particular elements, once considered as not belonging to the TEs, such as group II introns. It was then challenged (Piégu et al. 2015; Kapitonov and Jurka 2008), but still currently remains the most widely used.

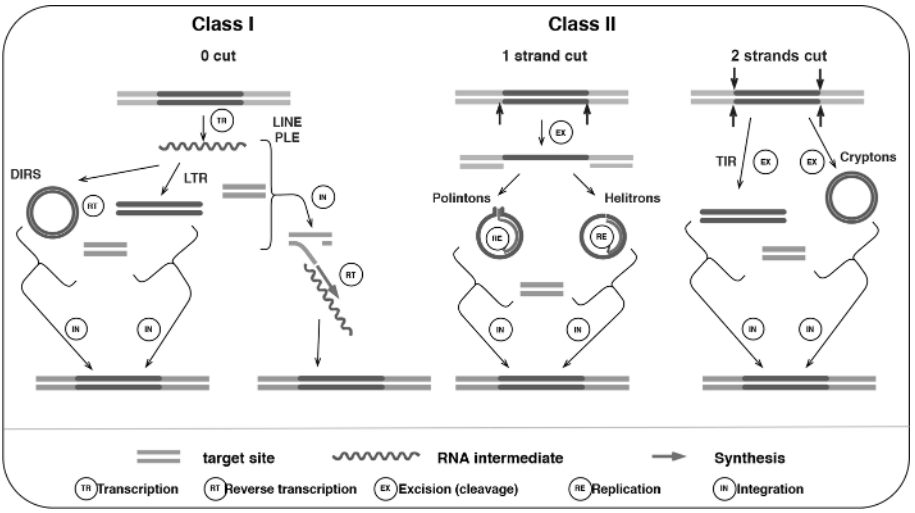

Figure 1.2. *The different transposition mechanisms of eukaryotic transposable elements. For a color version of this figure, see www.iste.co.uk/huavan/transposable.zip*

COMMENTS ON FIGURE 1.2.– *Class I TEs use an RNA intermediate, so no cleavage occurs at the donor site, unlike Class II TEs, for which one strand (Polintons, Helitrons) or both strands (TIR, Cryptons) are excised from the donor site. Whatever the number of cut, intermediates can be found as linear or circular. If DNA synthesis*

has to occur, it can be done on the extrachromosomal intermediate before integration, coupled with reintegration, or after integration. Linear intermediates usually reinsert after staggered cut of the target site, which creates small direct duplications at both ends of the insertion, called the TSD (target site duplication). Circular intermediates (DIRS, Helitrons, Cryptons) usually do not.

Recently, Arkhipova (2017) proposed a revised classification that attempted to integrate all previous classification systems, and also to include prokaryotic elements, as well as other mobile genetic elements (see Introduction). Figure 1.2 illustrates the various transposition mechanisms that characterize the main groups of eukaryotic TEs.

1.2.1. *Class I*

This class is likely the most abundant in eukaryotes, due to the replicative nature of the transposition process. All TEs in this class transpose via a copy-and-paste mechanism, which involves the production of an RNA intermediate that is then reverse transcribed into a new DNA copy, ready to insert at another location in the genome. The initial copy at the donor site is always left intact. According to Wicker et al. (2007), this class is divided into several orders, which are characterized by different structures and transposition mechanisms. The two most famous are the LTR and the LINE orders, which correspond to the initial A and B groups defined by Finnegan. However, some other orders exist, showing a less typical structure, the diversity of which is less known.

1.2.1.1. *LTR retroelements*

LTR retroelements share their structure with LTR retroviruses. Their mechanism of transposition is well known. They are characterized by two long terminal repeats: LTRs that are in direct orientation and that frame a piece of DNA several kb long, which comprise of one or more open reading frames (ORFs) that encode proteins necessary for the transposition process. The first ORF usually encodes a structural protein that associates to form viral-like particles (VLP). A second gene encodes several proteins translated in a row and later cut by a protease. The reverse transcriptase (RT) will generate the cDNA complementary of the transcribed genomic RNA, synthetizing the new copies, while an RNAse H has the role of destroying the RNA in the hybrid DNA-RNA duplex. Reverse transcription usually takes place after formation of the VLP that encapsulates the genomic RNA and needed proteins, along with the two short RNAs needed for initiation of the cDNA. Finally, the integrase will perform the integration to the new site, through strand transfer, after target site cleavage. Several superfamilies have been recognized, based on gene order and on the phylogeny on the RT conserved domains: Copia, Gypsy/Ty3, ERV (Endogenous retrovirus) and BEL/Pao. These superfamilies have been included in the International

Virus Taxonomy, within the order of Ortervirales. For example, ERVs are classified in the class Retroviridae, which also includes retroviruses such as HIV. In fact, the main distinctions between LTR retroviruses and LTR retrotransposons is that the former are infectious, able to exit the cell/individual before entering a new one, while LTR retrotransposons realize all their cycles within the cell. Some of these contain a third ORF that encodes an envelope gene, theoretically allowing them to leave the cell. The *Gypsy* element in *D. melanogaster* was first considered as a retrotransposon but is actually able to change cells and can then be functionally considered as a virus. The distribution of Env genes in the phylogenetic tree (based on RT) is quite patchy. Different envelope genes have been acquired independently and then been subsequently lost. Those LTR retroelements are found in large numbers in all eukaryotic genomes. *Gypsy* and *Copia*-like elements are particularly frequent.

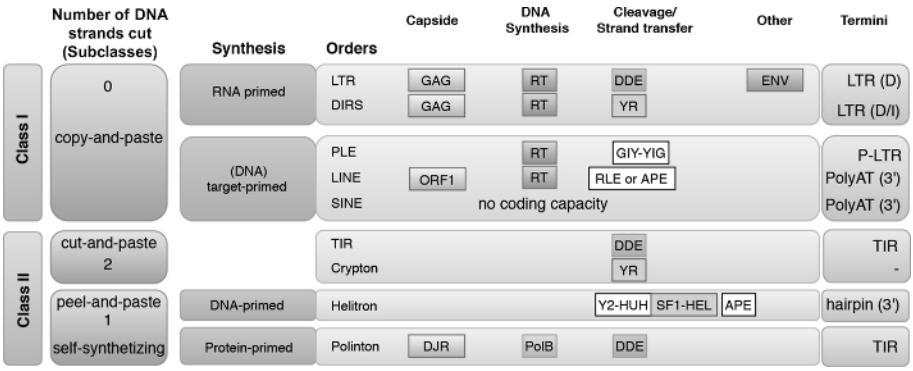

Figure 1.3. *Classification of eukaryotic transposable elements showing their main features and catalytic domains. GAG: capsid-like protein; HUH: Rep endonuclease domain; Hel: helicase domain; APE: apurinic/apyrimidinic endonuclease; RLE: restriction-like endonuclease; DDE: DDE integrase; YR: tyrosine recombinase; SF1-HEL: helicase (subfamily 1); T: reverse transcriptase; DJR: double jelly roll; PolB: DNA-dependant-DNA-polymerase B. D: direct; I: inverted. P-LTR: pseudo-LTR. For a color version of this figure, see www.iste.co.uk/huavan/transposable.zip*

1.2.1.2. *DIRS*

DIRS elements also possess LTRs, but present some specific features, and are clearly different from classical LTR retroelements. Their LTRs can be direct of inverted, depending on the DIRS family. An internal/subterminal region, homologous to the LTR sequence, is present, at least in the DIRS-like family; this appears to be crucial for the transposition mechanism. DIRS encode a reverse transcriptase, which seems to be phylogenetically close to that of the LTR, as well as an RNAse H and a GAG-like gene; however, they typically contain a gene coding for a tyrosine recombinase that replaces the DDE integrase. They also sometimes carry more

specific genes (methylase, or hydrolase, of unknown function). The DIRS order remains a poorly characterized order among Class I. Four different superfamilies have been recognized (Malicki et al. 2020).

1.2.1.3. *Non-LTR retroelements*

Non-LTR retrotransposons lack LTRs and have few structural characteristics at their ends, including a TA rich stretch at the 3' end, which is necessary for transposition. Non-LTR retrotransposons include LINEs for long interspersed elements, which form a large order of Class I elements. They nevertheless share catalytic activities with LTR elements: a GAG-like protein (usually called ORF1), a reverse transcriptase and an endonuclease, which can be of various types, restriction-like endonuclease (RLE) or Apurinic-like endonuclease (APE), defining two subgroups. The transposition mechanism is quite different from the one from LTR elements. After transcription (and translation of proteins), transposition occurs directly at the reintegration site through a mechanism called TPRT (target-primed reverse transcription). The RNA hybridizes through the TA rich tail to the complementary single-strand DNA obtained after the staggered cut by the endonuclease. The DNA is then synthetized by elongation of the staggered cut. One feature of this transposition mechanism is that if the reverse transcription is aborted prematurely, a 5' truncated element will be inserted. Since signal for transcription is usually located in 5', those copies are themselves unable to transpose again and are called dead-on-arrival (DOA). Note that 5' truncated copies are frequent in genomes. LINEs are classified into five big superfamilies based on the specificity of the endonuclease domain and RT phylogeny (Arkhipova 2017; Kojima 2020). They are present in all genomes and can reach huge copy numbers (about 500,000 for L1 in the human genome).

Tightly associated with LINEs are the SINEs (short interspersed elements). These are highly abundant in some genomes (for example, in some mammals). They have no coding capacity and rely entirely on LINEs for their transposition. SINEs are derived from non-mobile short RNA sequences such as tRNA or 7SL transcripts. They do sometimes carry in 3' some short regions homologous to LINEs, suggesting de novo emergence by template switching during reverse transcription. In humans, the SINEs Alu have amplified up to 1 million copies.

1.2.1.4. *PLEs*

Penelope-like element (PLEs) are a particular subclass among Class I elements. They form an ancient group characterized by a reverse transcriptase, which is more related to the eukaryotic telomerase reverse transcriptase (TERTs). They contain atypical LTRs, called pseudo-LTRs, that can be direct or inverted. Among other strange features, we find sometimes the presence of introns and hammerhead-like ribozymes in the pseudo-LTRs. An endonuclease may sometimes be present. Absence of endonuclease correlates with a preference for insertion near telomeres,

taking advantage of the short RNA sequences from which originate the telomeres repeats. The absence of integrase or recombinase suggests they may reinsert using a TPRT process. Recent studies uncovered a distribution larger than expected, with elements present in all major eukaryotic kingdoms (Craig et al. 2021).

1.2.2. *Class II*

1.2.2.1. *TIR*

TIRs elements are the most known TEs among Class II. While the first TIR elements had been identified after their insertion into a gene, triggering a phenotypic change, most recent elements have been exclusively identified through genome sequence analysis. About 20 different superfamilies are now recognized (Yuan and Wessler 2011; Kojima 2020) and are quite universal (present in almost every sequenced genome). All TIR TEs share characteristic features such as the presence of terminal inverted repeats (TIR), ranging from 11 to several hundreds of bp, that surround one or two genes encoding – at least for a DDE transposase. The DDE motif takes its name from three important residues in the catalytic core of the transposase. Initially identified in *Tc1-mariner* elements only, variant DDE motifs have finally been recognized in all known TIR superfamilies. DDE transposases are part of the large RNAse H fold superfamilies of protein. DDE-containing proteins are also found in Polintons, as well as in LTR retroelements (integrase). Interestingly, most of the bacterial TEs also contain DDE transposases, and several eukaryotic TIRs superfamilies share more homologies with their bacterial counterpart than with other eukaryotic TIRs superfamilies (Yuan and Wessler 2011). This is in favor of an ancient diversification, prior to the split between eubacteria and archaea/eukaryotes.

The mechanism of transposition is referred to as the “cut-and-paste” mechanism. It is obvious then that this mechanism per se is not responsible for the amplification of copies. However, TIRs elements are sometimes present in a large number of copies. In eukaryotes, two main processes have been proposed to explain this apparent contradiction. In prokaryotes, in which similar elements exist, transposition figures between two DNA molecules are usually resolved by replication. This is a common mechanism (see Chapter 3). In eukaryotes, the fate of a transposing copy depends partly on genome replication. Indeed, if a copy present in an already replicated site transposes into an unreplicated site, this copy will be duplicated in one of the daughter cells (Figure 1.4). Of course, if the contrary occurs (transposition from unreplicated to replicated site), one daughter cell will lose the copy. Hence, in order to be efficient, transposition of Class II elements should be synchronized with replication. In any case, there is another way for a TIR element to amplify. The excision from the donor site leaves a double strand DNA break (DSB) that must be repaired by the cellular machinery for the cell to survive. From this point of view,

Class II are quite harmful! Several ways exist in eukaryotic cells to repair a DSB (Figure 1.4). One is homology-dependant template repair, which uses an homologous DNA sequence to restore the intact site. Homologous sites are found in sister chromatids after replication, or in homologous chromosomes when the cell is at the diploid stage. If this homologous locus carries the TE copy, a full copy can be restored at the empty site. It often happens that the template-dependent repair is prematurely interrupted; the DSB is then repaired by an alternative mechanism known as non-homologous end-joining (NHEJ), which uses microhomology for direct ligation of the two broken ends. In such cases, an internally deleted copy can be generated, as shown for the P element in *Drosophila*. For other elements, NHEJ could be the main solution for repairing and this explains the excision footprint usually found at the empty donor site.

1.2.2.2. *Cryptons*

Cryptons were discovered in 2003 in several pathogenic fungi (Goodwin et al. 2003), and more recently in animals and some stramenopiles (Kojima and Jurka 2011). They have no repeats at the end and transpose using a tyrosine recombinase (YR), which supposedly excises them as a circular DNA intermediate and reinserts them in target sites with some homology. Indeed, short direct repeats are found at both termini. This group is poorly characterized, sharing relatively low similarity. They have been proposed to be at the origin of DIRS retroelements (Class I), and some domesticated versions have been identified in humans (Kojima and Jurka 2011).

1.2.2.3. *Helitrons*

Helitrons were discovered in silico in 2001, as a new group of Class II elements, characterized by the presence of a rolling-circle replication initiator protein (Rep) present in some prokaryotic TEs (such as IS91) and other mobile genetic elements that replicate through a rolling-circle mechanism (Kapitonov and Jurka 2001). The Rep domain corresponds to a HUH endonuclease and is associated with a C-terminal helicase domain. Helitrons have been found in almost all eukaryotic supergroups and are particularly prevalent in some species. Two groups are recognized that differ in their termini. Unlike TIR elements, Helitrons from group 1 do not have TIRs. They end with short sequences (LTS and RTS, for left and right terminal sequences, respectively) with several characteristics (5'TC, and 3'CTRR with a subterminal hairpin structure at the RTS which functions as a termination signal). In the second group, Helentrons contain an extra endonuclease domain similar to the one found in LINEs, and possess small subterminal inverted repeats in addition to the hairpin structures that can be present at both ends (Thomas and Pritham 2015).

Some aspects of the supposed transposition mechanism have been recently experimentally verified (Grabundzija et al. 2018; Kosek et al. 2021). The top strand of the donor DNA (coding strand) is cut at the LTS and displaced by the helicase

activity, while synthesis of a new strand is initiated. The 3' cut at the RTS is elicited by a pause of the protein at the hairpin structure and allows the formation of a single-stranded circular intermediate that theoretically can undergo rolling circle replication. However, it seems probable that the reinsertion concerns single-stranded DNA (ssDNA), the double strand at the target site being reconstituted during genome replication. This mechanism is referred to as "peel-and-paste". More details are provided in Chapter 3 for prokaryotic TEs.

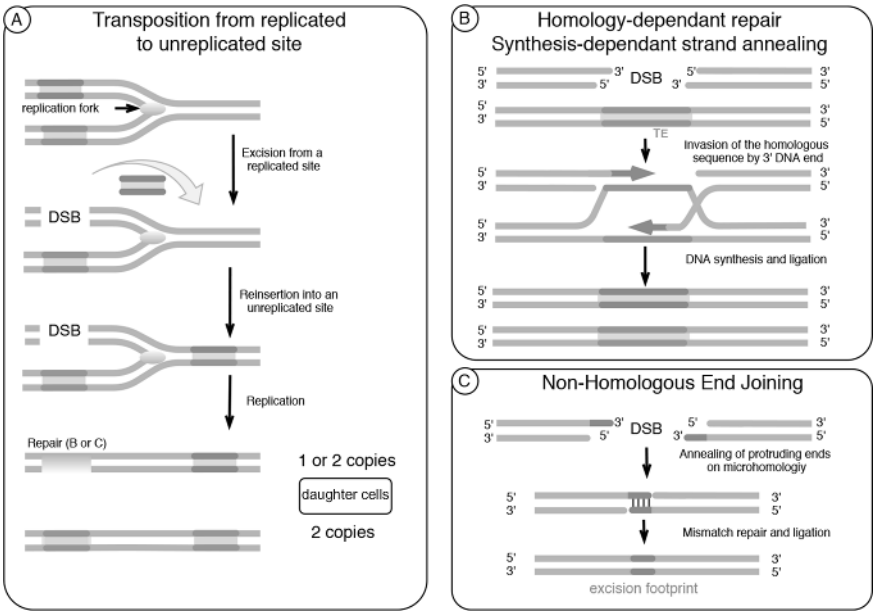

Figure 1.4. *Different ways to increase copy number and to repair the excision site. (A) Transposition can proceed from a replicated site to an unreplicated one. After replication is finished, at least one daughter cell will have a extra copy. (B) The excision leaves a double strand break (DSB) that can be repaired by using an homologous template. If this template contains the insertion, the TE is restored at the excision site. Processes A and B can combine. (C) Another way to repair the DSB is by direct ligation using a non-homologous end joining pathway. This relies on microhomology at both ends that allows annealing of the two protruding ends, followed by ligation. This repair mechanism can be used to repair aborted SDSA (Single-Dependant Strand Annealing), leading to internally deleted elements, and can result in an excision footprint, depending on the way the element is excised. For a color version of this figure, see www.iste.co.uk/huavan/transposable.zip*

The 3' termination of the strand displacement at the donor site is loosely defined by the short 3' hairpin, and adjacent sequences are frequently included in the

transposition intermediate; therefore, these elements are able to capture genes, although this transduction-like mechanism may not account for all cases of gene capture observed. Several other models have been proposed but will not be discussed here (Kapitonov and Jurka 2007).

1.2.2.4. *Polintons*

Polintons (aka Mavericks) were discovered relatively recently (Kapitonov and Jurka 2006). They are long (around 20 kb) and possess TIRs as TIR elements. Between TIRs, several genes are present, including a DDE transposase gene and a DNA polymerase B gene. Hence, those elements are supposed to be replicative and have been called self-synthetizing. However, the precise mechanism of transpositions is not yet fully deciphered. Polintons may not be as active as TIRs, nor as universally spread, although they have been identified in several protists, fungi and animals, with the exception of plants (Haapa-Paananen et al. 2014). However, they can be abundant in some species (30% of the genome in *Trichomonas vaginalis*) (Pritham et al. 2007).

1.2.3. *Autonomous, non-autonomous and relics*

Most of these groups also contain shorter versions of TEs that contain *cis* sequences necessary for transposition, but lack coding capacity. These elements cannot transpose by themselves but can still move and amplify using the transposition machinery of the autonomous element. They are called non-autonomous and can be very short in size (between 200 bp and 1 kb). Non-autonomous elements derived from TIRs are very frequent and often outnumber the autonomous copies. They have been called MITEs (miniature inverted transposable elements); often, they have only kept the TIRs and the short central region appears unrelated or highly degenerated, compared to the autonomous. Every TIR superfamily has MITEs. Non-autonomous short versions also exist for Helitrons, for example. the DINE element of *Drosophila*, and for LTR retroelements (LARDs, TRIMs), LINEs and PLEs. In humans, the most numerous TEs are Alu elements that hijack the transposition machinery of L1 elements for transposing. Recently, a new group of non-autonomous elements have been identified in plants. Their peculiarities are the presence within LTRs of a hammerhead-like ribozyme. Such ribozymes are also found in non-LTR retroelements from animals. Those elements have been called retrozymes (Cervera and de la Peña 2020).

1.3. Abundance, diversity and distribution

Eukaryotic TEs are found everywhere but each species has a unique TE landscape. Abundance may vary from a few percent of the genome to more than

80%, and these characteristics are independent of the eukaryotic groups (Figure 1.5). Class I is usually more abundant, either because of LTR retroelements or of LINEs, but species with a majority of Class II elements can be found in any group. Hence, TE contents (proportion of the different types of TEs) are also quite variable (Wells and Feschotte 2020). Closely related species tend to have similar landscapes which can be explained by phylogenetic inertia (the TE landscape is ancestral). Shared TEs can also be more abundant in closely related species because of increased rates of horizontal transfer (Gilbert et al. 2021). However, very different patterns are sometimes observed between species belonging to a same phylogenetic clade. The huge increase in sequenced genomes should allow better deciphering of tendencies associated with phylogenetic clades. However, drawing an exhaustive picture is certainly impossible. The TE annotations can be performed using different tools, on assemblies of very variable quality, using a TE database more or less complete, which create lots of biases (see Chapter 11). Some recent analyses that combine homology-based de novo searches sometimes exhibit up to 50% of unclassified repetitive sequences. In brief, there are too many genomes and not enough comparative studies.

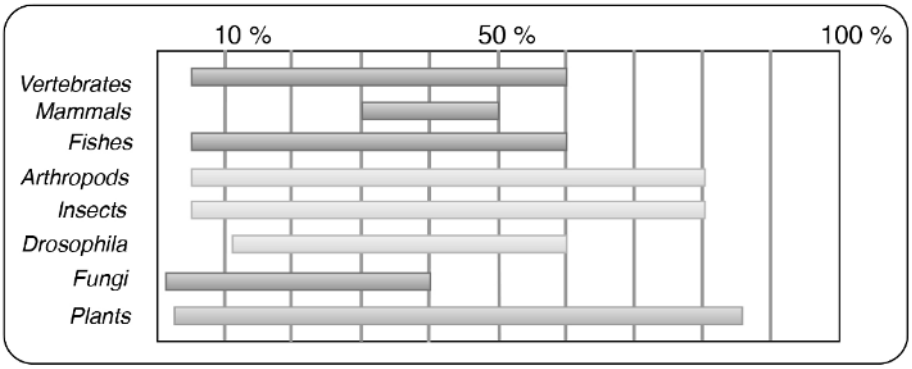

Figure 1.5. *Variation of TE contents (in percentage of the genome) in major Eukaryotic groups*

The diversity of TEs increases with the number of sequenced species. There is no exhaustive database but the most widespread superfamilies are divided into thousands of different TE families. For example, *Ty3/Gypsy* (Class I) and *Tc1-mariner* (Class II) are almost universal, and tens or hundreds are related; however, different families (lineages) often coexist in the same genome for each TE superfamily. Usually, only a few families are active and the remaining correspond to ancient, extinct (or little active) families. Why TEs have exploded in some genomes and not in others and why the success of any given superfamily is so variable are still unresolved questions.

Population size, lifestyle and demography are thought to impose some constraints, or at least have an influence. Genome ability to control TE transposition or to purge efficiently non-essential sequences or genome permissivity to horizontal transfer are likely important factors as well.

Within genomes, TEs are not randomly distributed. Some genomic regions, such as heterochromatin, are enriched in TEs. This may be due to a harsher selection on euchromatic arms rich in genes. Hence, TE will tend to accumulate in regions less important for functions or in specific regions (centromeres, telomeres, sexual chromosomes, dispensable chromosomes, low recombination regions). This also depends on the TE insertional preferences, which range from no preference (TEs insert almost randomly) to highly specific (like some LINE elements do by targeting a specific sequence in the rDNA genes – which are nevertheless repeated!).

1.4. Origins of transposable elements and evolutionary relationships with other genetic elements

TEs are very old components of genomes that might have emerged at the same time as the genes involved in nucleic acid metabolism and have since co-evolved with the genes. Basic catalytic activities involved in DNA metabolism include endonuclease, strand transferase and DNA/RNA-dependent DNA/RNA polymerase, all of which are found in TEs. We have to remember that TEs are parasites. Parasites cannot exist as parasitic entities before host appearance, and hosts obviously need parasitic sequences to evolve. As soon as these activities appeared, that is, very early, they must have been hijacked by sequences that could benefit, and became what we call parasites. In the long term, there must have been continuous back and forth exchanges between hosts and parasites, since the beginning of life. Hence, it is not surprising that links are detectable among TEs, and with all other mobile genetics entities. The previous descriptions draw the landscape of the major eukaryotic TEs based on Wicker et al.'s classification (2007). It does not consider all of Arkhipova's classification (the most recent to-date), which is more complicated, and noticeably has the merit to highlight links between TEs (eukaryotic and prokaryotic), as well as links with other genetic elements (mobile or not). Retracing the phylogenies of TEs is not possible because TEs are not a monophyletic group; however, it is still possible to infer the evolution at the protein domain level. The comparison of protein sequences shared between TEs can give insights on the evolution of TEs and their relationships. Such studies have mainly revealed that TEs evolve through recurrent module (protein domains) exchanges and the exact suite of events that have sequentially shaped the TE landscape cannot be determined.

While two classes of TEs coexist in eukaryotic genomes, prokaryotic TEs are mainly Class II elements. Reverse transcriptasea are nevertheless found in bacteria,

for example, in group II introns (retroelements that can invade homologous empty loci, after self-splicing from the RNA by a ribozyme activity and reverse transcription by the self-encoded reverse transciptase). Reverse transcriptases are also at the heart of the mechanism described in retrons and retroplasmids. Reverse transcriptase could be one of the oldest activities, derived from an ancestral RNA-recognition motif (RRM) that may also be at the origin of polymerase B or rolling-circle replication endonuclease (Krupovic et al. 2019).

The DDE domain exists in the transposase of all TIR DNA elements (Yuan and Wessler 2011), Polintons (Class II), as well as in integrases of LTR retroelements. DDE domains are gathered into structurally distinct 3D structures that share some common motifs. A large number of prokaryotic TEs also rely on DDE transposases. Interestingly, several DDE domains are shared by eukaryotic and prokaryotic elements, suggesting an emergence prior to the last universal common ancestor. DDE domains belong to the RNAse H-like superfamilies of proteins that also comprise exonucleases (Majorek et al. 2014).

The recent discoveries of retrozymes elements, or ribozymes embedded in TEs, are additional arguments in favor of a very early emergence of TEs and other mobile genetic parasites that could be as early as during the RNA world.

In fact, protein module exchanges are observed between TEs and non-TEs entities, such as viruses. Based on the structure and the reverse transcriptase phylogeny, LTR retroelements and ERV are closely related to vertebrate retroviruses and caulimoviruses, and have integrated into the virus classification of the International Committee for Taxonomy of Viruses. Previous analysis showed that vertebrate LTR retroviruses are in fact a clade of LTR retroelements that have acquired an envelope gene before diversification. Other retroviruses have been related to various other TEs and have acquired env genes independently from several viral sources (Malik et al. 2000). However, things become complicated when analyzing the RNAse H domain. Vertebrate retroviruses have independently acquired a second RNAse H domain, and the initial gene has been subfunctionalized (to regulate the RNAse H activity). Several independent acquisitions may have occurred, from non-LTR retroelement, or from unknown origin. Furthermore, the same scenario seems to have occured for plant retrotransposons (Ustyantsev et al. 2015). In any case, it is usually thought that LTR retroelements (and then retroviruses) emerged relatively recently compared to DNA and non-LTR retroelements.

It has been hypothesized that LTR retroelements emerged from a fusion between a non-LTR and a DNA transposon. In particular, the Gypsy superfamily DDE integrase is phylogenetically closely related to the one from the Ginger superfamily (TIR) (Bao et al. 2010). However, a clear picture cannot be drawn because several events or gene

replacements seem to have taken place. Other examples of shared genes include the tyrosine recombinases found in both Class I (DIRS) and Class II (Cryptons) elements; and it has been proposed that DIRS have emerged from a combination of a Crypton with the RH/RT of a Gypsy LTR retroelement (Poulter and Butler 2015). However, those elements are not as frequent as LTR retroelements, and alternative scenarios are also quite possible. In this case, the relationship with bacterial tyrosine recombinase is not very clear.

Finally, Polintons are remarkable for the links they share with several other mobile entities, which gives some clues about their origin. They are actually quite close to virophages, those viruses able to infect giant viruses, that themselves parasite unicellular eukaryotes, such as amoeba. Virophages, Polintons and related Polintons-like viruses (PLVs) all share several common genes and a common origin. Since they also possess capsid genes, Polintons may well have a dual life style (virus and TE), although Polinton virions have never been detected. Polintons are found in eukaryotes, while PLV, virophages and shorter versions of Polintons (called transpovirons, considered as linear integrative plasmids) are specific to eukaryotic giant viruses. On the one hand, giant viruses are supposed to have emerged before LECA (last common eukaryotic ancestor). On the other hand, Polintons are suspected to be at the origin of giants viruses, adenoviruses by virophages and linear cytoplasmic plasmids. This incredible network illustrates the extraordinary entangled relationships between all mobile genetic elements (Koonin and Krupovic 2017).

When speaking of shared domains, numerous exchanges have also occurred between TEs and host genes. In particular, DNA binding domains of Class II elements are recurrently identified within genes (Feschotte 2008). We can wonder about the direction of the transfer. It is always difficult to decide between the two alternatives: which is the one derived from the other one? It is usually thought that it is from the TEs toward the genes. However, recent TE families may have emerged after acquiring gene or viral domains (de Mendoza et al. 2018).

1.5. Genomic impact

Eukaryotic TEs are diversified, abundant and have a huge impact in the genome that hosts them. They greatly participate to its evolution, whether its structure or its function. The activity of TEs make genomes very dynamic. Hence, insertion polymorphism of recent copies increases a lot of the genetic diversity observed among populations, and hence the adaptive capacity of the species.

1.5.1. *Genome size*

Of course, TE copies are finally not that big (usually) but they may be numerous; this impacts the genome size. In eukaryotes, genome size is highly variable ranging for a few Mb for organisms such as obligate parasites, that relies on their host genes for many functions, to several hundreds of gigabases. Genome size is not correlated to the complexity of the organisms, and some unicellular organisms exhibit genomes much bigger than most multicellular complex organisms. Among multicellular organisms, plants are the ones that contain more genes; we also know that gene number is not either correlated to complexity, and more surprisingly, gene number is not correlated to genome size. This has led to what is called the C-value paradox, the C-value being the amount of (haploid) DNA in an organism's cell. What has been found is that genome size correlates extremely well with the genomic TE content. Hence, in eukaryotes, the bigger the genome, the higher the TE contents! Why do some species allow such accumulation of TEs in their genomes? It is still debated: a larger genome will take longer to replicate, and the nucleus, and then the cell, will be larger. These characteristics are likely to influence the organism's lifecycle.

TE accumulation could be a consequence of populational parameters. It has been suggested that the effective population size of a species (Ne), which determines the force of the natural selection relative to the genetic drift, could be a key factor for putting the limit to the genome size (Lynch and Conery 2003).

1.5.2. *Genome structure*

Genome was once seen as a static, unmovable suite of genes that has to be transmitted intact to the next generation. In fact, the genome is quite fluid, and allows for many changes, even if related species usually display a good synteny (gene colinearity) between their genomes. However, major rearrangements (chromosomal inversion, translocation) are observed between closely related species. Within genome, genes also evolve, by substitution or small indels, but also by more important changes such as exon shuffling, or changes in the promoter region. Mutations are the initial evolutionary force that create the diversity necessary for natural selection to act. Mere substitutions are finally quite rare, especially in species with large genomes, and a good part of mutations come from small or big indels and rearrangements. By intrinsic nature (mobile, and repeated), TEs are a major source of mutation: by transposing, they trigger insertions, deletions and recombination between similar copies settled at different loci (a process known as ectopic recombination); they trigger large rearrangements (Figure 1.6). Hence, TEs are major actors in genome structure evolution.

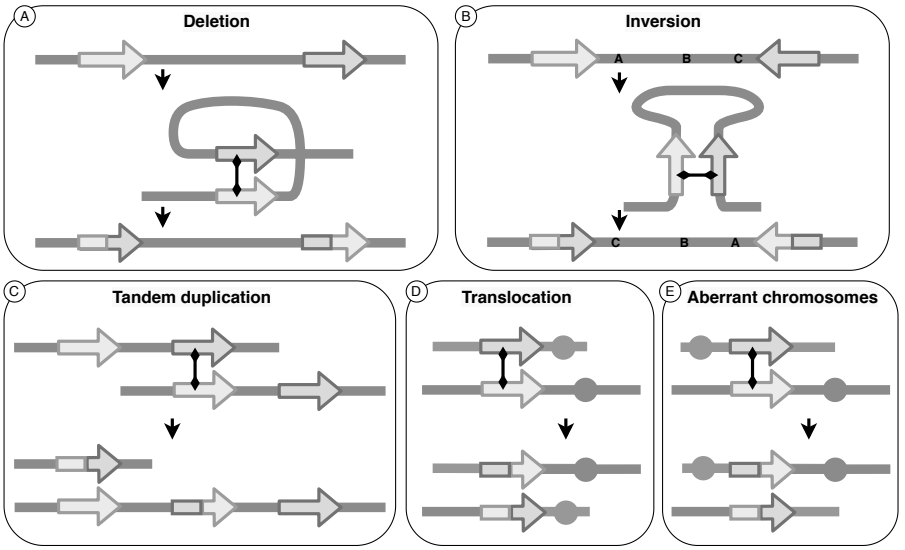

Figure 1.6. *Chromosomal rearrangements triggered by ectopic recombination between TEs. TEs copies can be on the same DNA molecules, on sister chromatids, homologous chromosomes or different chromosomes. Depending on the orientation between the two repeats, or relative to the centromere, different outputs are expected: deletion (A), inversion (B), tandem duplication (C), translocation (D) and aberrant chromosomes (E). For a color version of this figure, see www.iste.co.uk/huavan/transposable.zip*

1.5.3. *Genome function and evolution*

Inevitably, TEs interact with the part of the genome that insures cell and organism survival, which is the genes. More details are provided in other chapters, but TEs have been involved in gene evolution (exon shuffling, exonization) and gene regulation (rewiring of regulation networks by exaptation of TEs short sequences). Numerous processes have emerged in response to TEs (silencing mechanisms via small interfering RNA, or deposition of epigenetics marks). Finally, TE protein domains (such as the DNA binding domains) from numerous elements are found in genes, coding for transcription factors, or proteins that interact with DNA. Finally, TEs are at the origin of several evolutionary innovations such as mass excision of short DNA fragments at the origin of the macronuclei of paramecia, V(D)J recombination in jawed vertebrates, or placenta emergence in mammals, which we call molecular domestication (see Chapter 6).

1.6. References

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

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

Cervera, A. and de la Peña, M. (2020). Small circRNAs with self-cleaving ribozymes are highly expressed in diverse metazoan transcriptomes. *Nucleic Acids Res.*, 48(9), 5054–5064.

Craig, R.J., Yushenova, I.A., Rodriguez, F., Arkhipova, I.R. (2021). An ancient clade of *Penelope*-like retroelements with permuted domains is present in the green lineage and protists, and dominates many invertebrate genomes. *Mol. Biol. Evol.*, 38(11), 5005–5020.

Curcio, M.J. and Derbyshire, K.M. (2003). The outs and ins of transposition: From mu to kangaroo. *Nat. Rev. Mol. Cell Biol.*, 4(11), 865–877.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, 9(5), 397–405.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet.*, 5(4), 103–107.

Gilbert, C., Peccoud, J., Cordaux, R. (2021). Transposable elements and the evolution of insects. *Annu. Rev. Entomol.*, 66, 355–372.

Goodwin, T.J.D., Butler, M.I., Poulter, R.T.M. (2003). Cryptons: A group of tyrosine-recombinase-encoding dna transposons from pathogenic fungi. *Microbiology (Reading)*, 149(Pt 11), 3099–3109.

Grabundzija, I., Hickman, A.B., Dyda, F. (2018). Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. *Nat. Commun.*, 9(1), 1278.

Haapa-Paananen, S., Wahlberg, N., Savilahti, H. (2014). Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. *Mol. Phylogenet. Evol.*, 78, 271–274.

Kapitonov, V.V. and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, 98(15), 8714–8719.

Kapitonov, V.V. and Jurka, J. (2006). Self-synthesizing dna transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, 103(12), 4540–4545.

Kapitonov, V.V. and Jurka, J. (2007). Helitrons on a roll: Eukaryotic rolling-circle transposons. *Trends Genet.*, 23(10), 521–529.

Kapitonov, V.V. and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in repbase. *Nat. Rev. Genet.*, 9(5), 411–412.

Kojima, K.K. (2020). Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet. Syst.*, 94(6), 233–252.

Kojima, K.K. and Jurka, J. (2011). Crypton transposons: Identification of new diverse families and ancient domestication events. *Mob. DNA*, 2(1), 12.

Koonin, E.V. and Krupovic, M. (2017). Polintons, virophages and transpovirons: A tangled web linking viruses, transposons and immunity. *Curr. Opin. Virol.*, 25, 7–15.

Kosek, D., Grabundzija, I., Lei, H., Bilic, I., Wang, H., Jin, Y., Peaslee, G.F., Hickman, A.B., Dyda, F. (2021). The large bat helitron DNA transposase forms a compact monomeric assembly that buries and protects its covalently bound 5'-transposon end. *Mol. Cell*, 81(20), 4271–4286.

Krupovic, M., Dolja, V.V., Koonin, E.V. (2019). Origin of viruses: Primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.*, 17(7), 449–458.

Lynch, M. and Conery, J.S. (2003). The origins of genome complexity. *Science*, 302(5649), 1401–1404.

Majorek, K.A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., Bujnicki, J.M. (2014). The RNase H-like superfamily: New members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.*, 42(7), 4160–4179.

Malicki, M., Spaller, T., Winckler, T., Hammann, C. (2020). DIRS retrotransposons amplify via linear, single-stranded cDNA intermediates. *Nucleic Acids Res.*, 48(8), 4230–4243.

Malik, H.S., Henikoff, S., Eickbush, T.H. (2000). Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.*, 10(9), 1307–1318.

de Mendoza, A., Bonnet, A., Vargas-Landin, D.B., Ji, N., Li, H., Yang, F., Li, L., Hori, K., Pflueger, J., Buckberry, S. et al. (2018). Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nat. Commun.*, 9(1), 1341.

Piégu, B., Bire, S., Arensburger, P., 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.

Poulter, R.T.M. and Butler, M.I. (2015). Tyrosine recombinase retrotransposons and transposons. *Microbiol. Spectr.*, 3(2), MDNA3–0036–2014.

Pritham, E.J., Putliwala, T., Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*, 390(1-2), 3–17.

Thomas, J. and Pritham, E.J. (2015). Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol Spectr*, 3(4), 1–32.

Ustyantsev, K., Novikova, O., Blinov, A., Smyshlyaev, G. (2015). Convergent evolution of ribonuclease H in LTR retrotransposons and retroviruses. *Mol. Biol. Evol.*, 32(5), 1197–1207.

Wells, J.N. and Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.*, 54, 539–561.

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 eukaryotic transposable elements. *Nat. Rev. Genet.*, 8(12), 973–982.

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(19), 7884–7889.