Molecular Phylogenetics and Evolution 86 (2015) 90–109

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Review

A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity

Benoît Piégu [a], Solenne Bire [a,b], Peter Arensburger [a,c,*], Yves Bigot [a,*]

[a] UMR INRA-CNRS 7247, PRC, Centre INRA de Nouzilly, 37380 Nouzilly, France
[b] Institute of Biotechnology, University of Lausanne, Center for Biotechnology UNIL-EPFL, 1015 Lausanne, Switzerland
[c] Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, United States

ARTICLE INFO

Article history:
Received 24 December 2014
Revised 11 March 2015
Accepted 12 March 2015
Available online 20 March 2015

Keywords:
Transposon
Mobility
Host range
Ribozyme
Nuclease
Recombinase

ABSTRACT

The increase of publicly available sequencing data has allowed for rapid progress in our understanding of genome composition. As new information becomes available we should constantly be updating and reanalyzing existing and newly acquired data. In this report we focus on transposable elements (TEs) which make up a significant portion of nearly all sequenced genomes. Our ability to accurately identify and classify these sequences is critical to understanding their impact on host genomes. At the same time, as we demonstrate in this report, problems with existing classification schemes have led to significant misunderstandings of the evolution of both TE sequences and their host genomes. In a pioneering publication Finnegan (1989) proposed classifying all TE sequences into two classes based on transposition mechanisms and structural features: the retrotransposons (class I) and the DNA transposons (class II). We have retraced how ideas regarding TE classification and annotation in both prokaryotic and eukaryotic scientific communities have changed over time. This has led us to observe that: (1) a number of TEs have convergent structural features and/or transposition mechanisms that have led to misleading conclusions regarding their classification, (2) the evolution of TEs is similar to that of viruses by having several unrelated origins, (3) there might be at least 8 classes and 12 orders of TEs including 10 novel orders.

In an effort to address these classification issues we propose: (1) the outline of a universal TE classification, (2) a set of methods and classification rules that could be used by all scientific communities involved in the study of TEs, and (3) a 5-year schedule for the establishment of an International Committee for Taxonomy of Transposable Elements (ICTTE).

Contents

* Corresponding authors at: Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, United States (P. Arensburger). UMR INRA-CNRS 7247, PRC, Centre INRA de Nouzilly, 37380 Nouzilly, France (Y. Bigot).
E-mail addresses: parensburger@csupomona.edu (P. Arensburger), yves.bigot@tours.inra.fr (Y. Bigot).

http://dx.doi.org/10.1016/j.ympev.2015.03.009
1055-7903/

0. Introduction

Any collection of objects, including biological entities, may be classified in multiple ways in order to create groups based on phenotypic features (Mayr and Bock, 2002). Within scientific disciplines examples of such classifications range from the periodic table to the Enzyme Commission number system (Webb, 1992). Within the biological sciences the principle of shared common ancestry is so widely used as a classification criterion that many classifications in this field are assumed (sometimes incorrectly) to incorporate this criterion. Biological classification, a subfield of the study of systematics, is the grouping of species on the basis of evolutionary relationships (Daly et al., 2012). Mayr and Bock (2002) defined species classification as "The arrangement of entities in a hierarchical series of nested classes, in which similar or related classes at one hierarchical level are combined comprehensively into more inclusive classes at the next higher level". The classification of most living organisms has been codified by four international codes of nomenclature: one for animals (Ride et al., 2000); one for algae, fungi and plants (McNeill et al., 2012); one for prokaryotes (Lapage et al., 1992) and one for viruses (King et al., 2011). All four codes share several organizational levels including kingdom, phylum/division, class, order, family, genus, and species. The placement of individuals within these levels implies a series of evolutionary relationships that will often be used as a basis for subsequent research. For the working scientist a well organized biological classification provides the following four advantages: (1) it simplifies the identification of unknown organisms, (2) it reveals connections between groups of closely related organisms, (3) it indicates evolutionary relationship, and (4) it allows the integration of data from a few representatives from distinct groups into a web connection of all living organisms. Methods and criteria used to establish biological classifications have changed over time as evolutionary concepts and technical innovations have progressed. Most recently, phylogenetic analyses from protein and DNA sequences have had a significant impact on classification schemes. Over the last decade debates regarding the classification of some groups, such as viruses, have been the subject of passionate exchanges of views. Indeed, within the virus community discussions range from the definition of what a virus species is (van Regenmortel et al., 2013) to the possibility that certain viruses might represent a distinct fourth domain of life (Bandea, 2009; Williams et al., 2011; Philippe et al., 2013; Pennisi, 2013; Raoult, 2013). Finally, these discussions are complicated by the connection between viruses and a number of mobile genetic elements (more commonly referred to as transposable elements, TEs) that have been characterized in both prokaryotic and eukaryotic genomes (Weiss, 2006; Stoye et al., 2012; Desnues et al., 2012; Yutin et al., 2013). The classification of these TEs is the subject of this review and we begin by examining how to define a TE species.

TEs represent most of the interspersed repeats in the genomes of prokaryotes and eukaryotes. It is therefore striking that a simple but comprehensive definition of what constitutes a TE is not easily found in the literature. Haren et al. (1999) proposed that "TEs are discrete segments of DNA capable of moving from one locus to another in their host genome or between different genomes". Similarly, Kidwell and Lisch (2001) stated that "TEs are DNA sequences that have the capacity to change genomic locations". Since their publication, the above definitions have been widely used in the literature. Currently, as knowledge of the diversity of the TE and virus worlds has grown extensively we would suggest that these definitions could rephrased as (based in part on the evidence we present below) "TEs are discrete segments of DNA capable of moving within a host genome from one chromosome or plasmid location to another and which do not use a specific molecular machinery that they encode to infect the genome of new hosts by lateral transfer". An important aspect of any TE definition is that it includes mobile DNA sequences that are primarily maintained by vertical transmission as copies integrated into the chromosomes or plasmids of their hosts. Therefore, our amended definition considers that viruses, phages, and integrative conjugative elements (ICE) have similar features to TEs but they are not considered TEs since they are able to move between hosts independent of transmission vectors. To round out our proposed TE definition, it should also be understood that the state of TE copies within a host genome varies depending on the age and activity of the element. Autonomous TEs encode the enzymes required for their mobility while non-autonomous elements depend for their mobility on enzymes supplied by autonomous elements belonging to the same or a related element. TE sequences in a genome accumulate mutations over time which will most often inactivate the ability of these sequences to mobilize further. This ageing process has led to the presence, in most genomes, of many fossil TE sequences alongside a few active copies (Kidwell and Lisch, 2001).

In this report we review various TE classification systems, with particular attention to how each system has affected the development of TE biology. We also examine how these systems have held up in light of the exponential increase in genomic data.

This manuscript is organized into three sections. We begin by reviewing existing TE classifications and outline their respective strengths and weaknesses. Next we describe a number of TE sequences that have not been included in some TE classification systems. Finally, we outline a proposal for an international committee to help draft a unified TE classification.

1. History of existing TE classifications

1.1. TE classification pioneer

Finnegan (1989, 1992) launched the field of TE systematics based largely on what was previously proposed in this field from human, drosophila, and yeast models (Singer, 1982; Boeke, 1989; Finnegan and Fawcett, 1986). His proposal was that TEs could be

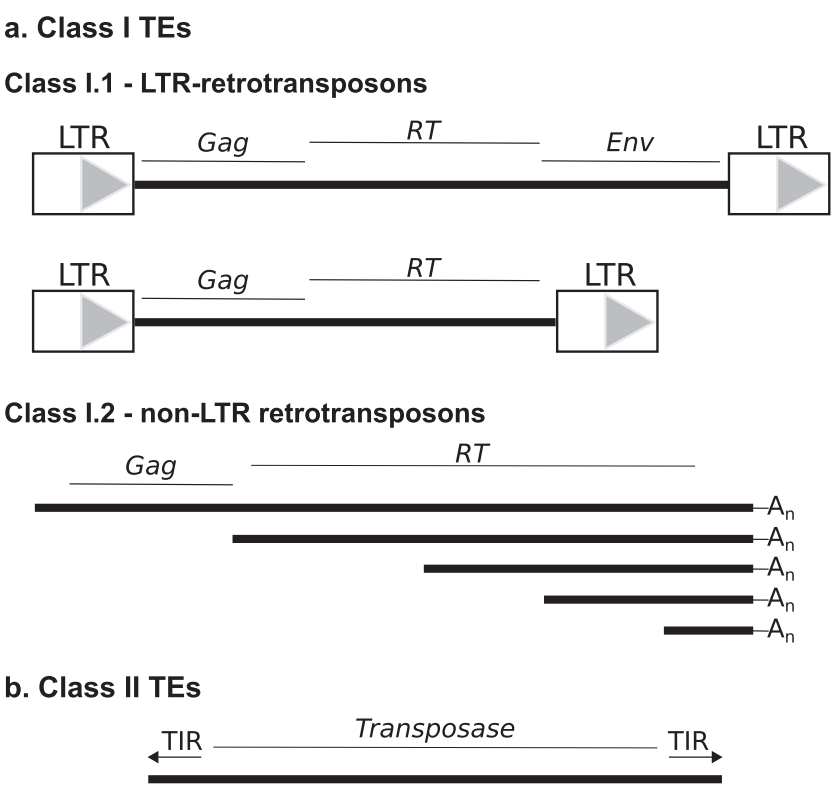

Fig. 1. Reproduction of Fig. 1 from Finnegan (1992). This TE classification proposal was based on TE DNA sequence features. (a) Class I is composed of elements that transpose by reverse transcription of an RNA intermediate and includes two sub-groups. Class I.1 TEs have all the signatures of an endogenous retrovirus-like element, including long terminal repeats at both ends and open reading frames (ORFs) coding for, a group antigen (Gag), a reverse transcriptase (RT), and in some case an envelope protein. Class I.2 elements only have Gag and RT ORFs. Class I.2 is composed of elements that look like long retro-inserted messenger RNA (mRNA) with an A-rich tail at their 3′ end. Within a species of such elements many copies are truncated at their 5′ ends. (b) Class II elements that transpose directly from DNA to DNA and have short terminal inverted repeats (arrowed) at both ends. They contain a gene coding a transposase, an enzyme required for their own transposition.

classified into two classes based on their presumed mechanism of transposition: class I elements that transpose by reverse transcription of an RNA intermediate using a DNA–RNA–DNA mechanism, and class II elements (commonly referred to as DNA transposons) that transpose directly from DNA to DNA (Fig. 1). Class I elements were further divided into two types using structural features. Class I.1 included TEs resembling retroviruses with long terminal repeats (LTR) and containing three open reading frames (ORFs) that code for: (1) a group specific antigen (Gag), (2) a reverse transcriptase (RT) and (3) an unknown protein that was subsequently identified as a retroviral envelope protein (Env). This TE type is now more commonly known as an LTR retrotransposon or endogenous retrovirus. Class I.2 included TEs with no terminal repeats (later research showed the presence of short terminal repeats) and two ORFs coding for proteins similar to Gag and RT and a poly-A tail at the 3′ end. These TEs are now known as non-LTR retrotransposons or retroposons. The classification of short interspersed elements (SINEs) was not addressed in Finnegan (1989) even though some authors already considered them as TEs (Haynes et al., 1981; Schmid and Jelinek, 1982; Britten et al., 1989). The existence of such sequences was suggested in Finnegan (1992) but their lack of capacity to encode enzymes required for their transposition prevented their integration as TEs in his manuscript. Class II TEs were described as elements containing terminal inverted repeats (TIRs) as well as a gene encoding a transposase, an enzyme required for the TE's own transposition (Finnegan, 1992). Finnegan (1989) described two different types of class II elements but this was dropped in the 1992 publication. This work, which became the basis for TE classification, will be referred below as the Finnegan proposal.

1.2. Updates to the Finnegan proposal

While Finnegan (1989) defined his class II elements as transposing directly from DNA to DNA this definition was subtly modified in the following decade by rephrasing the definition as elements that moved from DNA to DNA via a DNA intermediate using a "cut and paste" mechanism (Capy et al., 1996; Lerat et al., 1999). This change, that came about as the mechanics behind transposition became better understood, is problematic because of the clearly demonstrated links between prokaryotic DNA transposons, such as certain insertion sequence elements (IS) and phage Mu, with certain eukaryotic Class II elements (Mahillon and Chandler, 1998). Specifically, this rephrasing is inconsistent with studies done, among others, on phage Mu which showed that it transposes from one locus to another using a "copy and paste" (rather than a "cut and paste") mechanism which does not involve an intermediate (Mizuuchi, 1992).

The Finnegan proposal was the subject of two major updates that have been actively debated and are referred to below as the "Repbase" and "Wicker" proposals (Jurka et al., 2005; Wicker et al., 2007, 2008, 2009; Kapitonov and Jurka, 2008; Seberg and Petersen, 2009). The ambition of these updates was to outline a unified classification system with a series of classification criteria that was more in depth than those used in the original Finnegan proposal. While both updates retained the notion that all eukaryotic transposons could be classified as retrotransposons or DNA transposons they differ in their classification, naming systems and the number of TE classes.

In the Repbase proposal (Fig. 2) the most basic criterion is the suspected mechanism of transposition dividing all TEs into Type

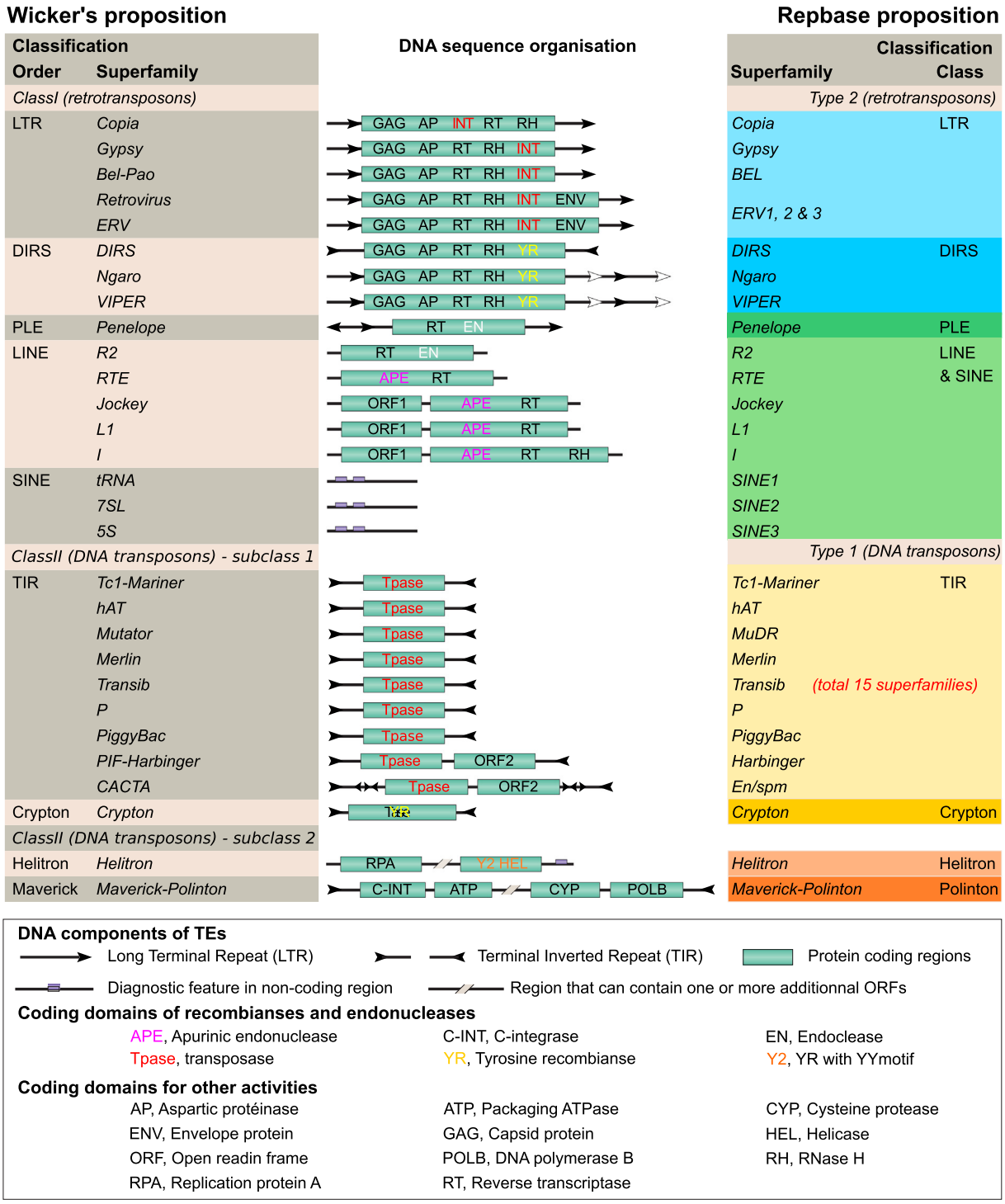

Fig. 2. Comparison and content from two proposals for the classification and annotation of eukaryotic TEs. The Repbase proposal is shown on the right (Jurka et al., 2005; Kapitonov and Jurka, 2008) and the Wicker proposal on the left (Wicker et al., 2007). Both proposals are based on DNA and amino acid sequence features. Both proposals divide all TEs into two groups, the retrotransposons and the DNA transposons. This basal division is called the "type" level in the Repbase proposal and a "class" level in the Wicker proposal. Each of these two classes or types is then subdivided into "classes" in the Repbase proposal or into "orders" in the Wicker proposal. Overall, Repbase "classes" and Wicker "orders" are very similar and each group contains the same TE superfamilies. A schematic representation of the DNA sequence organization within each TEs superfamily is supplied in the middle of the figure. A symbol legend is provided at the bottom of the figure.

1 (DNA transposons) or Type 2 (retrotransposons) elements. Division by this initial criterion is then followed by criteria related to the types of enzymes involved in transposition/retrotransposition, structural similarities, and sequence similarities (Jurka et al., 2005; Kapitonov and Jurka, 2008). The first classification criterion (suspected transposition mechanism) led the authors to propose 7 TE classes. Three classes are DNA transposons and correspond to: (1) the cut and paste DNA transposons using a [DDE/D] transposase, (2) a rolling-circle DNA transposon (*Helitrons*; Kapitonov and Jurka, 2007), and (3) the self-synthesizing DNA transposons (*Polintons/Mavericks*; Kapitonov and Jurka, 2006; Pritham et al., 2007). A review of the literature and more recent

research suggests that two more classes could be added to this proposal. The first would gather the tyrosine-recombinase-encoding transposons (Goodwin et al., 2003; Kojima and Jurka, 2011). The second would include the recently discovered *Zisupton* TEs in which transposition is catalyzed by a protein with no similarity with the known [DDE/D], tyrosine or serine transposases (Böhne et al., 2012). Within the retrotransposons (Type 2), the Repbase proposal contains two classes of non-LTR retrotransposons which includes both the long interspersed nuclear elements (LINEs) and SINEs as well as the *Penelope*-like elements (Arkhipova, 2006; Gladyshev and Arkhipova, 2007). There are also two classes of LTR retrotransposons, including LTRs and LTR-like elements related to the *DIRS* element which contain a tyrosine recombinase instead of an integrase fused to an RT (Poulter and Goodwin, 2005). Each TE class is then divided into superfamilies that can contain several families. In addition to the existing name given to each transposon when it is discovered the Repbase proposal suggests a universal nomenclature system that allows both naming and describing each TE sequence quickly.

The Wicker proposal (Fig. 2) is more in line with the original Finnegan scheme and preserves its basic structure. The Wicker proposal's basal criterion divides TEs into two classes based on the presence or the absence of an RNA transposition intermediate (i.e. class I and class II respectively). Class I is further divided into five orders based on mechanistic features, organization, and RT phylogeny. Class II TEs continue to be elements that transpose without using an RNA molecule as an intermediate. Class II elements are divided into two subclasses, subclass 1 are elements that use a cut and paste transposition mechanism, while subclass 2 elements transpose using a copy and paste transposition mechanism. Subclass 1 TEs are characterized by TIRs at their extremities and are subdivided into two orders depending on the recombinase used for their transposition: a [DDE/D] transposase in the TIR order (Yuan and Wessler, 2011) and a tyrosine recombinase for the Crypton order. Two TE types are rather mysteriously placed into Wicker's Class II-subclass 2, the rolling-circle DNA transposons called *Helitrons* (Kapitonov and Jurka, 2007) and the *Polintons*/*Mavericks* (Kapitonov and Jurka, 2006; Pritham et al., 2007). A characteristic that is unique to this subclass 2 (but not one highlighted by Wicker et al., 2007) is that their origin seems to be related to those of certain virus families including Geminivirus and Maviruses (Murad et al., 2004; Fischer and Suttle, 2011; Desnues et al., 2012; Chandler et al., 2013; Yutin et al., 2013). However, the molecular mechanisms behind their mobility remain to be elucidated.

Similar to the Repbase proposal a new universal nomenclature system is proposed which provides guidelines both on how to name each TE sequence and how to supply its accession number into databases. A unique feature of the Wicker proposal is that it attempts to deal with the classification of non-autonomous TEs such as LARDs (large LTR retrotransposon derivatives), MITEs (miniature inverted–repeat transposable elements), SNACs (small non-autonomous CACTA; class II TEs in the TIR order and the CACTA superfamilies) and TRIM (terminal-repeat retrotransposons in miniature) by giving these elements a "structural description" that places them in one of these categories.

1.3. Critical analysis of the Wicker and Repbase proposals

The Repbase and Wicker classification proposals are pragmatic means of organizing existing DNA sequences annotations of the most representative eukaryotic TEs. Established classifications of animals, plants, viruses, etc. (see introduction) achieve this as well but go a step further. A biological classification is commonly expected to form groups on the basis of evolutionary relationships or, when such relationships cannot be found for some entities, to separate these into distinct and unrelated groups that may be assumed (pending further evidence) to have independent origins. The Repbase and Wicker proposals do not fulfil this expectation. Indeed these proposals include groups of phylogenetically unrelated classes or subclasses of TEs within the retrotransposon and DNA transposon phenotypes. There exist biological classifications, such as the Enzyme Commission number (EC number; Webb, 1992), that group biological entities based on their phenotype (activity in the case of enzymes) without taking into account their evolutionary relationships. If the objective of TE classifications was merely as a basis for sequence annotation, then using phenotypic characters regardless of shared ancestry would be sufficient. However, such a classification should not be used for research that takes evolutionary context into account. Unfortunately, the Wicker and Repbase classifications are widely used in many TE studies where it is assumed (most often implicitly) that the classification reflects evolutionary history. This widespread misunderstanding regarding the Wicker and Repbase proposals emphasizes the importance of establishing a TE classification taking into account the evolutionary principles.

Since 1995 several reviews of TE mobility machineries have highlighted a number of weaknesses in the Repbase and Wicker proposals similar to those outlined above (Craig, 1995; Capy et al., 1996; Capy and Maisonhaute, 2002; Curcio and Derbyshire, 2003; Biémont and Vieira, 2006; Roberts et al., 2008; Hickman et al., 2010). For example, LTR retrotransposon transposition appears to have strong similarities with the mechanisms used by most cut and paste DNA transposons using a [DDE/D] transposase. Indeed, LTR mobility results from an RNA copy that is reverse transcribed into a double-stranded DNA intermediate followed by a transposition event. From the standpoint of their transposition mechanism, LTR retrotransposons could be viewed as class II TEs encoding a [DDE/D] transposase/integrase that have acquired a reverse transcription machinery to transform their RNA transcript into a DNA intermediate. Such associations between both enzymatic machineries within a natural TE should only be viewed as a proposed a model. This model has recently found support in a bacterial TE, ISLdr1, that was found to encode both a RT and a mutator-related transposase (Guérillot et al., 2014). Although less supporting data are available, a similar hypothesis might be considered for DIRS elements that use a tyrosine recombinase similar to that of certain DNA transposons (Curcio and Derbyshire, 2003). Analyses in the above cited reviews suggest that classification criteria based on the RT phenotype, which is an enzyme not directly involved in DNA cleavage and strand transfer reactions of eukaryotic retrotransposons, are given too much importance in the Repbase and Wicker proposals. Indeed, a group of related genes coding for RTs (see Gladyshev and Arkhipova, 2011) may have provided components that were co-opted during evolution by three different transposition machineries; LTR retrotransposons, LINE non-LTR retrotransposons and the DIRS-like elements (Curcio and Derbyshire, 2003).

A second weakness of these proposals comes from the choice of focusing only on eukaryotic TEs in order to achieve a universal classification system. The first consequence is that the most abundant TEs in eukaryotic genomes, the retrotransposons, have been considered as a single group of related TEs because they have genes encoding an RT that are phylogenetically related. A second consequence is that this choice has limited the diffusion and the assimilation of transposition concepts developed by biologists working with prokaryotic TEs. This is particularly unfortunate because some transposition mechanisms were first discovered in prokaryotes and then later extended to eukaryotes, as was recently exemplified with IS605/Fanzor elements (Bao and Jurka, 2013).

Finally, it should be noted that the Repbase proposal tries to escape these mechanistic weaknesses by dividing transposition

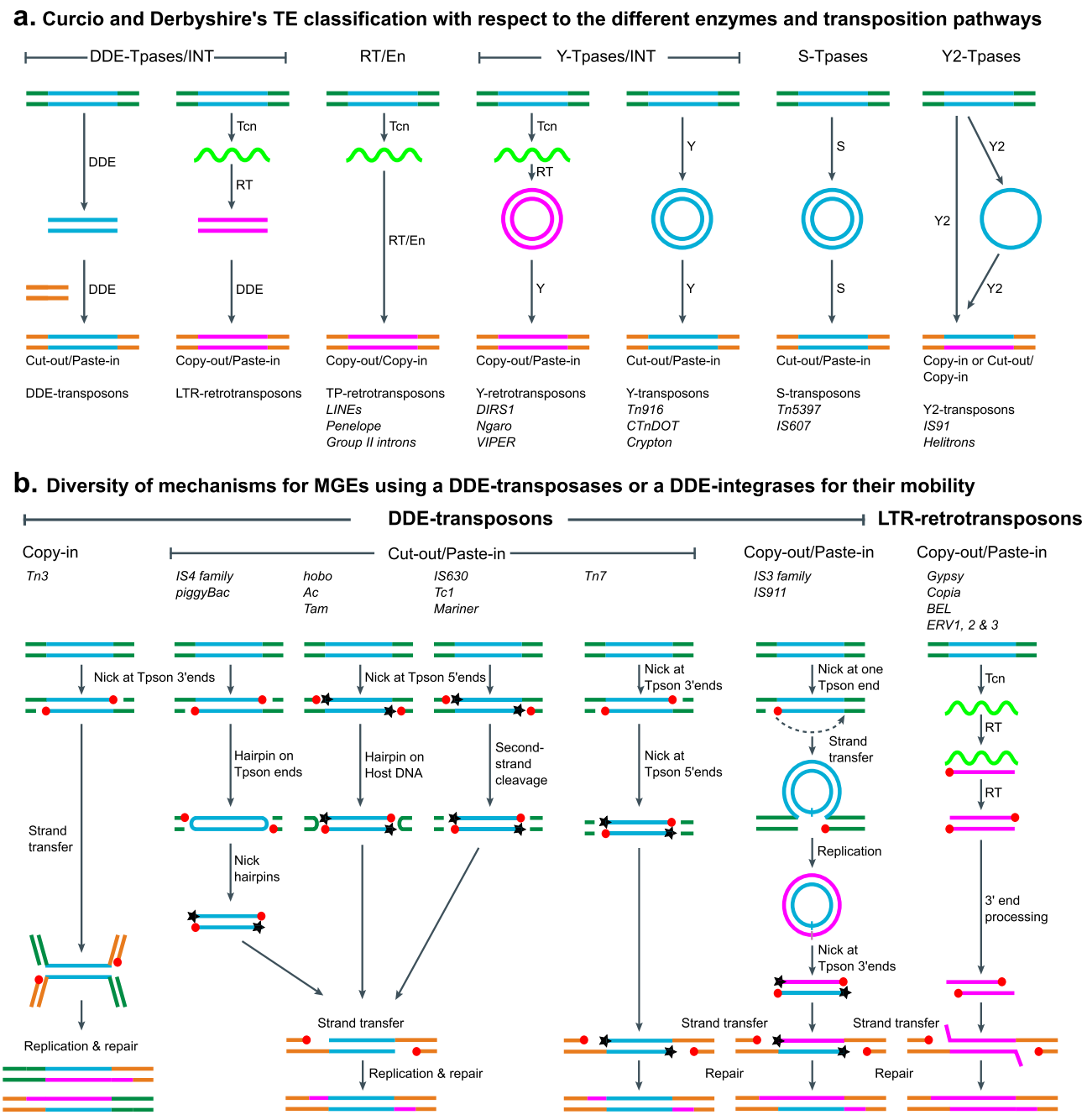

Fig. 3. Curcio and Derbyshire proposal schematic representation. This figure and legend below are slight modifications of previously published figures as indicated, and are shown here for ease of reference. These representations and this legend took their inspiration from similar diagrams presented in three papers and updated over time (Turlan and Chandler, 2000; Curcio and Derbyshire, 2003; Hickman et al., 2010). (a) Nearly unaltered Fig. 1 from Curcio and Derbyshire (2003). Five protein families that dictate transposition pathways: DDE-transposases/integrase (INT), reverse transcriptase/endonucleases (RT/En), tyrosine (Y)-transposases, serine (S)-transposases and Y2-transposases. TEs (blue) can be either 'cut-out' or 'copied-out' of the flanking donor DNA (green). Representatives of each type of transposon are listed below each pathway. At the end of the transposition process, they can be either 'paste-in' or 'copied-in' for integration into a host DNA target (orange). Most DDE-transposons excise from the flanking DNA to generate an excised linear transposon, which is the substrate. Retrotransposons with a DDE-integrase copy-out by reverse-transcribing (RT) a full-length copy of their RNA (apple green) that is generated by transcription (Tcn). Long-terminal repeat (LTR)-retrotransposons make a full-length cDNA copy (pink represents newly replicated DNA) from their RNA and integrate this into a target using a DDE-integrase. TP-retrotransposons use reverse transcriptase (RT) to copy their RNA directly into a target that has been nicked by an element-encoded nuclease (En). Y-retrotransposons are thought to generate a circular cDNA intermediate by reverse transcription. A Y-transposase integrates the element into the target. Y- and S-transposons encode either a tyrosine or serine transposase, which mediates excision of the transposon to form a circular intermediate. A reversal of the catalytic steps results in transposon insertion. Y2-transposons 'paste' one strand of the transposon into a target and use it as a template for DNA replication. Two models have been proposed for Y2-transposition. (b) View of Curcio and Derbyshire (2003), updated using Hickman et al. (2010), showing the seven transposition pathways among DDE-transposons and DDE-retrotransposons mediated by evolutionary related DDE-transposases/IN. Color symbols used to represent the origins of nucleic acids involved in the processes are the same as in (a). This figure was modified by adding information regarding the ends of the transposition intermediate and the ends of the host DNA at the excision and insertion sites: black stars represent phosphates at DNA intermediate ends; filled red circle show free 3′OH groups that are released by cleavages at host excision sites, transposons ends and host insertion sites; hairpins at the transposon ends or at host DNA ends in excision site are represented by a semi-circular line joining the two DNA strands. Representatives of each type of transposons are listed below each pathway. The seven different pathways are discriminated mainly by (1) the enzymatic reactions that cleave both DNA strands at excision, (2) the presence or absence of a replication step, and (3) the production or absence of hairpins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mechanisms into seven different classes. Unfortunately, this proposal remains fundamentally anchored in the retrotransposons versus DNA transposons basal split which prevents it from achieving a satisfactory classification scheme.

1.4. The Curcio and Derbyshire proposal

Nearly a dozen years have passed since the first publication critical of the basal split of TEs into retrotransposons versus DNA transposons, and advocating instead that focus should be shifted to homologous mechanisms of transposition. Curcio and Derbyshire (2003) did not propose a new TE classification per se, they simply reviewed the state of knowledge of TE transposition mechanisms to date and how similar the pathways for related elements found in different organisms were. However, after reading their review one is led to clearly appreciate a new way of classifying TEs. Therefore, while their publication was not originally set out for the purpose of reclassifying TEs we will use it as such below and refer to it as the Curcio and Derbyshire proposal.

Based on Curcio and Derbyshire (2003), TEs could be categorized by the mechanism they use for moving from one locus to another, a criterion which is linked to the origin of each transposase/integrase family (Fig. 3a). They described four classes of TEs based on the four families of proteins described as mediating transposition, namely (1) the DDE-transposases, (2) rolling-circle (RC) or Y2-transposases, (3) tyrosine (Y1)-transposases and (4) the serine (S)-transposases. Furthermore, they proposed a fifth class, the target-primed retrotransposons (TP retrotransposons), that gathers the eukaryotic non-LTR retrotransposons and the prokaryotic group II introns, both of which encode a combination of RT and endonuclease activities (RT/En). Within each of these five classes they proposed a further series of classification criteria that are based on particular features of transposition mechanisms. Furthermore, they investigated whether TEs are duplicated during transposition. Finally they examined possible classification criteria based on the number of transposition intermediates, their nature (RNA, double stranded-DNA [dsDNA] or single stranded-DNA), and their configuration (linear versus circular). The Curcio and Derbyshire proposal continues to use the terms "DNA transposon" and "retrotransposon" but this may be misleading because they used these terms to link unrelated TEs. For example, there are three kinds of DNA transposons that transpose using a cut and paste mechanism and have a double stranded (dsDNA) intermediate: (1) TEs that use a DDE transposase, (2) TEs that use a Y1 transposase and (3) those that use an S transposase. The first two have TIRs while the last one has short dyadic structures at both ends; these would not have been placed together in earlier proposals. Therefore, we would like to emphasize that in the Curcio and Derbyshire proposal the term "DNA transposon" describes TE phenotypes with convergent characters but with different evolutionary origins. Similarly, endogenous retroviruses and DIRS-like elements have a phenotype of LTR retrotransposons with convergent features at levels of their LTRs and integrases that differ in their origins. Further complicating classification of these elements, it should be noted that both DIRS-like and LTR retrotransposons have co-opted related RT genes early in their evolution (Gladyshev and Arkhipova, 2011).

The Curcio and Derbyshire proposal defines a series of possible classification criteria that are specific to each TE class. These criteria are based on cleavage at the transposon ends (namely on the configuration of both ends of the TE) as well as on the configuration of the host DNA at the excision site, and on the involvement of the host replication machinery for those TEs that do not move using a dsDNA intermediate. Using these criteria, TEs with a DDE transposase/integrase can be sorted into five groups (Fig. 3b) that use: (1) copy-in/paste-in mechanisms (e.g. Tn3 and Mu), (2) cut-out/paste-in mechanisms with hairpins at both end of the transposon intermediate (e.g. IS10 and IS50), (3) cut-out/paste-in mechanisms with hairpins at the excision site of the host DNA (e.g. *hAT* and *Transib* elements), (4) cut-out/paste-in mechanisms without hairpins (IS630-Tc1-mariner elements), and (5) cut-out and paste-in mechanisms via a ssDNA transposition intermediate (e.g. IS3-like elements). Recently, a sixth group with a cut-out/paste-in mechanism and a dsDNA intermediate was added (Fig. 3b; Hickman et al., 2010). This group is characterized by the absence of a hairpin on either the transposition intermediate or the host DNA at the excision site (e.g. *Tn7*). Several published review articles have adopted the Curcio and Derbyshire (2003) proposal and have deepened and updated their criteria in the light of more recent discoveries including TEs with a DDE (Hickman et al., 2010), Y1, Y2 (Chandler et al., 2013) and S-transposases (Boocock and Rice, 2013).

1.5. Critical analysis of the Curcio and Derbyshire proposal

Of the TE classifications reviewed here the Curcio and Derbyshire proposal is the only one that may be considered to be properly rooted in an evolutionary context. Indeed, their criteria lead to arrangements of TEs into a hierarchical series of classification groups in which similar or related groups at one hierarchical level are combined into more inclusive groups at the next higher level. These criteria have three important features: (1) they highlight that two TEs can have similar transposition phenotypes but different origins (a feature notably lacking in earlier proposals), (2) they draws attention, for the first time in this field, on the impact of evolutionary convergences in the context of transposition, (3) they consider that TEs and their mobility mechanisms have several independent evolutionary origins and have, in some cases, co-opted similar replication mechanisms such as retrotransposition or the involvement of host DNA replication machineries. The multiplicity of mechanisms and proteins outlined by Curcio and Derbyshire clearly shows that TEs cannot be understood as organisms with a single common ancestor, but instead have several independent origins. This differentiates TE evolution from that of their hosts. At the same time, it brings TEs closer to what is known about the evolution and classification of viruses, that is numerous classes with different origins. We will note that while the Curcio and Derbyshire proposal has failed to penetrate the eukaryotic TE community it has been fully accepted by the microbiologists. However, while this proposal may be very appealing for classifying TE classes, its reliance on transposition pathways may not make it suitable at all classification levels. Specifically, a literature review shows at least two examples that demonstrate some of the pitfalls of relying exclusively on such criteria.

The first example concerns the origins of different kinds of non-LTR retrotransposons. In Curcio and Derbyshire (2003), target-primed retrotransposons (TP-retrotransposons) include three TE types, LINEs, *Penelope*-like elements and group II introns. However, TEs in this group transpose using a combination of RT and endonuclease activities (RT/En) that have different origins. This observation is supported by the fact that their RT moieties are not directly related (Gladyshev and Arkhipova, 2011). Within the LINEs RTs are apurinic/apyrimidinic (AP) endonucleases (Feng et al., 1996; Yang et al., 1999; Eickbush and Jamburuthugoda, 2008), while *Penelope*-like elements have a thumb-like domain fused to a GIY-YIG domain (Arkhipova, 2006). Group II introns, which encode an RT, have endonuclease requirements for their integration that are monitored by the ribozyme activity of their lariat RNA intermediate which can be completed, in some of them, by a LAGLIDADG or a HNH endonuclease encoded as a fusion to the RT (Lambowitz and Zimmerly, 2011; Edgell et al., 2011; Marcia et al., 2013). It would therefore be expected that these three types

of TP-retrotransposons should be classified into three groups of unrelated TEs that use three transposition mechanisms with different origins (but having numerous convergent features). Recently, it has been highlighted that even within LINEs two types of endonucleases (EN) with different origins are encountered: AP and a nuclease similar to PD-(D/E)XK (Stoddard, 2005; Mukha et al., 2013). Within certain LINE species, such as the *Dualen* elements (Kojima and Fujiwara, 2005), the situation is even more complex since they encode both EN types. The second example comes from two DDE transposon superfamilies, *hAT* and *Transib*. Phylogenetic analyses have shown that the catalytic domains of transposases from these two superfamilies are not directly related and are more closely related to the catalytic domain of *piggyBac* than to each other (Yuan and Wessler, 2011). Because *piggyBac* has been shown to have a very different transposition pathway (Mitra et al., 2008) than members of the *hAT* and *Transib* superfamilies (Zhou et al., 2004; Kapitonov and Jurka, 2005; Hencken et al., 2012) the similarity between *hAT* and *Transib* transposition pathways is likely the result of evolutionary convergence.

These examples lead us to the following two conclusions. First, our current understanding of non-LTR retrotransposons puts too much emphasis on the transposition phenotype rather than on their molecular features and the origin of their enzymatic transposition machinery. Second, these examples highlight the importance, when selecting classification criteria, of finding a balance between criteria associated with mechanistic features which can capture aspects of TE evolution at a basal level, and those criteria associated with sequence features that may be more efficient at capturing evolutionary relationships at higher levels.

2. TEs that have not received proper attention

2.1. SSEs: Self-splicing elements, the ugly ducklings set aside by classification systems for TEs

Although their discovery dated from the early 80's (Kruger et al., 1982; Garriga and Lambowitz, 1983; Waring and Davies, 1984; Kane et al., 1990; Hirata et al., 1990) self-splicing elements (SSEs), which are DNA segments inserted into expressed host regions, have been virtually ignored by the eukaryotic TE community, despite the fact that SSEs are consistently treated as true TEs in a large body of scientific literature associated with prokaryotes and certain eukaryotic nuclear genomes (Belfort, 2003; Stoddard, 2005; Dassa et al., 2009; Pietrokovski, 2001; Gogarten and Hilario, 2006; Barzel et al., 2011). To date, this has had two major consequences. The first is that eukaryotic TE researchers may have been blind to the possibility of a third basal TE type, in addition to retrotransposons and DNA transposons. Indeed, a great deal of data is available demonstrating that SSEs transpose by a different mobilization mechanism from other TEs. This mechanism does not involve an RNA or a DNA intermediate but the host machinery of gene conversion instead. The second consequence of the SSEs low visibility is that research into them has been limited, particularly by the lack of high throughput bioinformatics tools to study the extent of their host ranges among prokaryotes, eukaryotes and viruses. Thus, it would appear appropriate, and timely to integrate SSEs into a TE classification as one or several novel classes of TEs.

SSEs are simple genetic entities that can be divided into two components. The first is a splicing mechanism that averts deleterious effect caused by the SSE insertion into the host genes (Belfort, 2003). This property is essential for their host since SSEs specifically insert into conserved regions of housekeeping genes. The second is a DNA gene that encodes a cleavage specific nuclease, most of these are homing endonucleases (HEN; Stoddard, 2005). This enzyme ensures the dissemination of SSEs by horizontal transfer using homologous recombination between occupied and unoccupied gene alleles (also called "homing"). Currently, two main families have been characterized: self-splicing introns and inteins. Self-splicing introns consist of a self-splicing ribozyme transcribed within a host RNA molecule while inteins generate a self-splicing peptide when transcribed from a host gene and translated into a protein (Fig. 4a and b). Although their organization looks simple, the pair association from at least six HEN types (LAGLIDADG, HNH, His-Cys box, GIY-YIG, PD-[D/E]XK (Stoddard, 2005) and Vsr (Dassa et al., 2009)) and two splicing mechanisms (protein splicing and RNA splicing; Belfort, 2003) leads to a large diversity of SSEs among various branches of the tree of life. Like other TEs, autonomous SSEs encode the HEN required for their mobility, while elements that depend on HENs encoded by related autonomous SSEs are described as non-autonomous SSEs. Furthermore, similar to other TEs, SSEs can accumulate mutations over time that will inactivate them. Depending on the insertion site of the SSE this ageing process leads to the presence of remnant and fossil SSEs in chromosomes alongside active copies, or to their fast elimination due to the dramatic lethal effect on their host (Pietrokovski, 2001). The two types of SSEs, inteins and group I introns, are described next.

2.1.1. Inteins

Intein DNA sequences are in-frame insertions into protein coding genes, typically genes that are important for the host genome's survival (Gogarten and Hilario, 2006; Barzel et al., 2011). These sequences insert into highly conserved gene motifs and code for a protein that is able to catalyze both the excision of the intein amino acid sequence post-translation and the ligation of the host protein (the extein, Fig. 4a). In addition, many inteins are able to reintegrate themselves back into specific sites by using homing endonucleases (HENs), all those so far described belong to the LAGLIDADG and HNH families. These dual functions allow the intein to avoid the deleterious effects of inserting into host protein coding genes and to spread via horizontal gene transfer by using its homing capacities. Inteins can increase their copy number by gene conversion, integration into specific gene sites followed by host DNA repair using the newly integrated intein as a template (the intein at the original donor site is maintained). Inteins are therefore TEs that are found in a limited number of sites in the genome. So far these elements have only been found in certain species of bacteria, archaea, fungi, viruses of unicellular algae and amoebozoa. Recent data confirmed that they also occur in some metazoan viruses (Pietrokovski, 1998; Bigot et al., 2013). Three criteria can be used to classify inteins: (1) the insertion site of the host gene that they parasitize, (2) the sequence relationships between the sequences of their intein and (3) the HEN moieties. Inbase (http://tools.neb.com/inbase/) is the main database that explicitly gathers information about the features of inteins.

2.1.2. Group I introns

Group I introns are ribozymes found in a range of different genes (messenger RNA, transfer RNA and ribosomal RNA) that have the ability to catalyze their own splicing reactions. These are found in many different genomes including chloroplast and mitochondrial genomes of lower eukaryotes and higher plants, as well as in archaebacterial and eubacterial genomes (Hasselmayer et al., 2004; Haugen et al., 2005; Vicens and Cech, 2006; Raghavan and Minnick, 2009; Ton-Hoang et al., 2010; McManus et al., 2012). Furthermore, they are also found in certain nuclear genes of unicellular eukaryotes such as diatoms, euglenoids, green and red algae, and in some viral genomes. Most group I introns range in size from 250 to 500 nucleotides. Following transcription the core of a typical group I intron consists of approximately nine paired

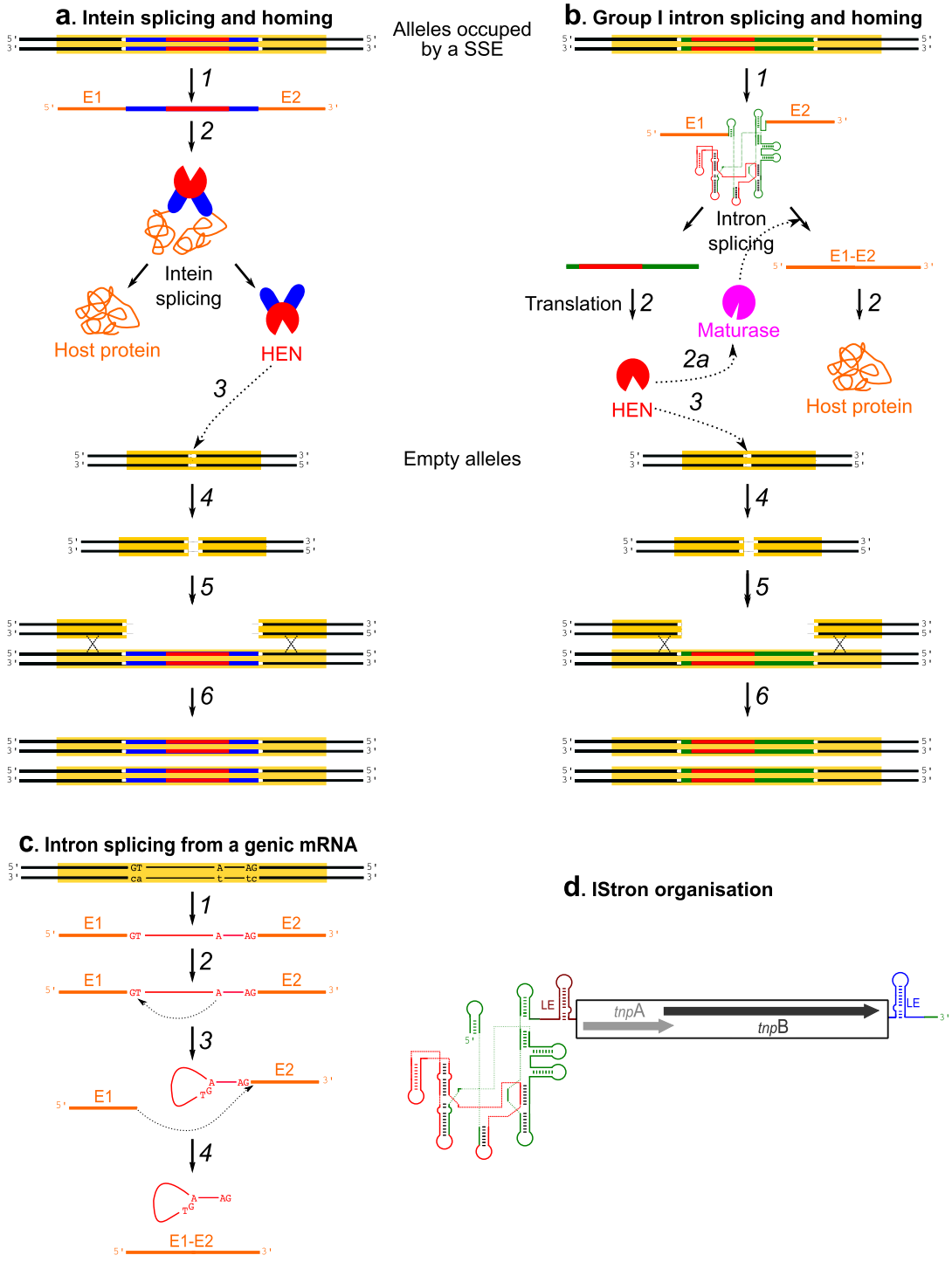

Fig. 4. Features of SSEs, spliceosomal introns and IStrons. (a) Organization, expression, splicing and mobility from one allele occupied by an intein to an unoccupied allele. (b) Organization, expression, splicing and mobility from an allele occupied by a group I intron to an unoccupied allele. (c) Model of splicing from a messenger RNA transcript of a spliceosomal intron contained in a eukaryotic gene. (d) Sequence organization of an IStron. In (a) and (b) the main steps of the SSE splicing and homing are numbered. Step 1 is RNA transcription. Black lines indicate DNA strands of the occupied and unoccupied alleles, yellow bars indicate the DNA region transcribed for each allele, blue and red lines indicated regions coding for the intein (blue) and HEN moieties (red) in the DNA and the corresponding RNA transcripts, green and red lines are regions coding the HNH and the group I intron ends, orange lines are exons E1 and E2 in the RNA transcripts. The intrastrand dyadic structure of the group I intron in RNA transcripts is shown in part (c). Step 2 is translation into protein. "pacmans" HEN moieties are shown in red, blue ellipses are N- and C-terminal intein moieties, in pink "pacman" HENs refolded into maturase (step 2a). Step 3 is site-specific recognition of an unoccupied allele by HEN. Blank spaces indicated insertion site specifically cleaved by the HEN in both alleles. Step 4 is specific cleavage by the HEN. Finally steps 5 and 6 are SSE invasion from an occupied allele toward an unoccupied one. In (c) the main steps are also numbered. Step 1 is mRNA transcription, step 2 is initiation of the hydrolytic cleavage by the A nucleotide of the branch point motif to the phosphodiester upstream to the GT intron boundary to assemble the lariat intermediate, step 3 is the second splicing cleavage, step 4 is the release of the lariat RNA intron and the spliced host mRNA. In (c) the thin black line indicates the region corresponding to the spliceosomal intron, conserved nucleotides located at the intron boundaries (GT/AG) and within the branch point motif (A) are indicated. Red lines indicate the spliceosomal intron between both exons in RNA transcripts or in a lariat RNA conformer post-splicing. In (d) the intrastrand dyadic structure of the group I intron within the IStron is shown using the same colors as in (c). The two open reading frames contained within the IS605-like elements inserted into the 3′ end of the group I intron are boxed and its dyadic ends, LE and RE are in red and blue respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

elements assembled in an intra-strand dyadic structure that is organized into three domains at the tertiary structure level (Haugen et al., 2005; Vicens and Cech, 2006; Raghavan and Minnick, 2009; Fig. 4b). This specific three-dimensional architecture is necessary for their self-splicing activity; the intron-containing RNA molecule can rearrange its own covalent structure to precisely remove itself and paste the exons together in the correct order. Splicing is initiated by placing an atypical guanosine nucleotide (typically GTP) on the 5′ end of the excised intron. Autonomous group I introns harbor ORFs related to mobility or splicing such as HENs or a maturase (i.e. a HEN having acquired the ability to chaperone RNA splicing). Their mobility from one occupied allele to an unoccupied one is consistent with classifying them with other SSEs in a new basal class of TE (Fig. 4b). Indeed, their transposition does not involve a transposition intermediate and follows a scheme similar to that of inteins, with a HEN that generates a double-strand chromosomal break within the unoccupied allelic target then repaired by the host DNA repair machinery using the allele containing the group I intron as a template. Group I intron "species" can be distinguished from each other using the features of their host site, but also those of their nine paired elements assembled in the ribozyme moiety (Haugen et al., 2005). The HEN that they carry can also be used for their identification. Indeed, although LAGLIDADG HENs are frequently found in these SSEs, four other HENs are found in group I introns (HNH, His-Cys box, GIY-YIG, PD-(D/E)XK; Stoddard, 2005 and Vsr; Dassa et al., 2009), thus defining at least 6 basic types of group I introns.

2.2. Introns

To date, five main types of introns have been characterized based on their splicing mechanism (Bruce, 2008). The first type is regular spliceosomal introns (RSI) that are located in protein-coding nuclear genes of eukaryotes and which are removed by spliceosomes. These are the most familiar introns to biologists, involving the formation of a lariat during the splicing process (Fig. 4c). The second type of introns is transfer RNA (tRNA)/archaeal introns that splice with the help of specialized proteins (tRNA splicing enzymes). The removal of tRNA introns occurs by a tRNA splicing endonuclease in precursor sequences followed by ligation by the tRNA splicing ligase enzyme. The last three intron types are the self-splicing group I introns (mentioned above), group II introns that are removed by RNA autocatalysis and intron-like elements, and group III introns that appear to be related to group II introns (Copertino and Hallick, 1991; Maier et al., 1995; Scamborova et al., 2004).

Finally, we will note that a sixth intron type, the IStrons (Hasselmayer et al., 2004), might also exist. These are group I introns that have inserted into the 3′ end of an IS605-like element (Fig. 4d). Due to their ability to splice into a group I intron, IStron transposition is typically harmless to the interrupted gene. To date, details regarding the molecular mechanism of IStron mobility remain to be elucidated and it is not yet verified whether proteins encoded by the two ORFs of IS605-like element, tnpA and tnpB, are able to affect the mobility of these TEs. It remains to be determined whether IStrons are able to alternatively use both transposition machineries (i.e. those of the group I intron and of the IS605; Ton-Hoang et al., 2010). If this is the case, they should be considered as composite elements rather than typical group I introns.

2.2.1. Group II introns

Group II introns are found in protein coding genes, transfer RNA, ribosomal RNA and in many bacterial genomes. Within eukaryotes their distribution is restricted to the mitochondrial and chloroplast DNA of lower eukaryotes, higher plants, and certain annelid species (Lambowitz and Zimmerly, 2011; Edgell et al., 2011; Marcia et al., 2013). To date, they are believed to be absent from eukaryotic nuclear genes. However, there are interesting structural similarities between non-LTR retrotransposons and group II introns. Indeed, non-LTR retrotransposons belonging to the *Penelope* family (Arkhipova, 2006) contain a polyprotein coding gene that is composed of a reverse transcriptase (RT) fused at its C-terminal end to a thumb-like domain and a GIY-YIG HEN. This protein organization is similar to that encountered in certain group II intron polyproteins. However, it has been reported that the similarity in RTs might be the result of evolutionary convergence (Gladyshev and Arkhipova, 2011). RNA transcripts containing group II introns are characterized by a conserved secondary structure originating from a span of 400 to 800 nucleotides that is different from that found in group I introns. It is organized into 6 domains interacting to form a conserved tertiary structure that is necessary for ribozyme activity (Lambowitz and Zimmerly, 2011; Edgell et al., 2011; Marcia et al., 2013). In some cases, particular intron-binding proteins assist in correctly folding the intron into a three-dimensional structure. Splicing of RNA molecules containing group II introns generates branched introns with a lariat configuration similar to that of spliceosomal RNAs (Fig. 5).

Group II introns are TEs that are able to invade DNA, their transposition can occur from one allele to another but also to non-allelic sites. They could be classified with inteins and group I introns because their mobility involves an RNA molecule as a transposition intermediate that originates from a spliced RNA transcript. This intermediate inserts into an unoccupied DNA locus by reverse-splicing or by partial reverse-splicing before being reverse transcribed into a DNA target. Depending on the configuration of the RNA intermediate and the integration target this retrohoming process can proceed in one of at least three ways (see Fig. 5; Edgell et al., 2011). Group II introns encode a reverse transcriptase (RT) but the endonuclease-type requirements of their integration are monitored by the ribozyme activity of the RNA intermediate lariat and are possibly completed by a HEN encoded as a fusion to the RT.

Some studies (Edgell et al., 2011) define group II introns as mobile ribozymes or non-LTR retrotransposable elements. As noted above, the origins of their RT and EN (HEN) moieties support them as members of one of the three classes of non-LTR retrotransposable elements. There are currently no data that would exclude the origin of group II introns as the result of evolutionary convergences with LINE and *Penelope* elements.

2.2.2. Introners: group I, II or III introns?

Over the ten last years, genome sequencing of a wide range of species has deeply modified our understanding of TE diversity. The concept of what a TE is has been strongly impacted by the emergence of new bioinformatic tools including our understanding of TE abundance and distribution in various genomes including humans (de Koning et al., 2011). Nevertheless, a number of unknowns remain amid this flow of new data. One of them is the presence of new interspersed repeats called introners. These were first described in the genome of the marine picoeukaryotic algae *Micromonas pusilla* (NCBI assembly CCMP1545; Worden et al., 2009), where they occupy 15% of the genome. A search of an NCBI database (the marine metagenome database under "whole genome shotgun contigs" category) revealed that introner-like elements (ILEs) related to those of *M. pusilla* CCMP1545 occur in other plankton species (Supplementary Fig. 1, Supplementary material online). More distantly related ILE (van der Burgt et al., 2012; Collemare et al., 2013) were also discovered in the genomes of a microcrustacean (*Daphnia pulex*), a urochordate (*Oikopleura dioica*) and in at least six different fungi (*Cladosporium fulvum*, *Dothistroma septosporum*, *Mycosphaerella graminicola*, *Mycosphaerella finjiensis*, *Hysterium pulicare* and *Stagonospora nodorum*). The size of ILEs ranges from 10 bps to about 700 bps, most being between 100 to

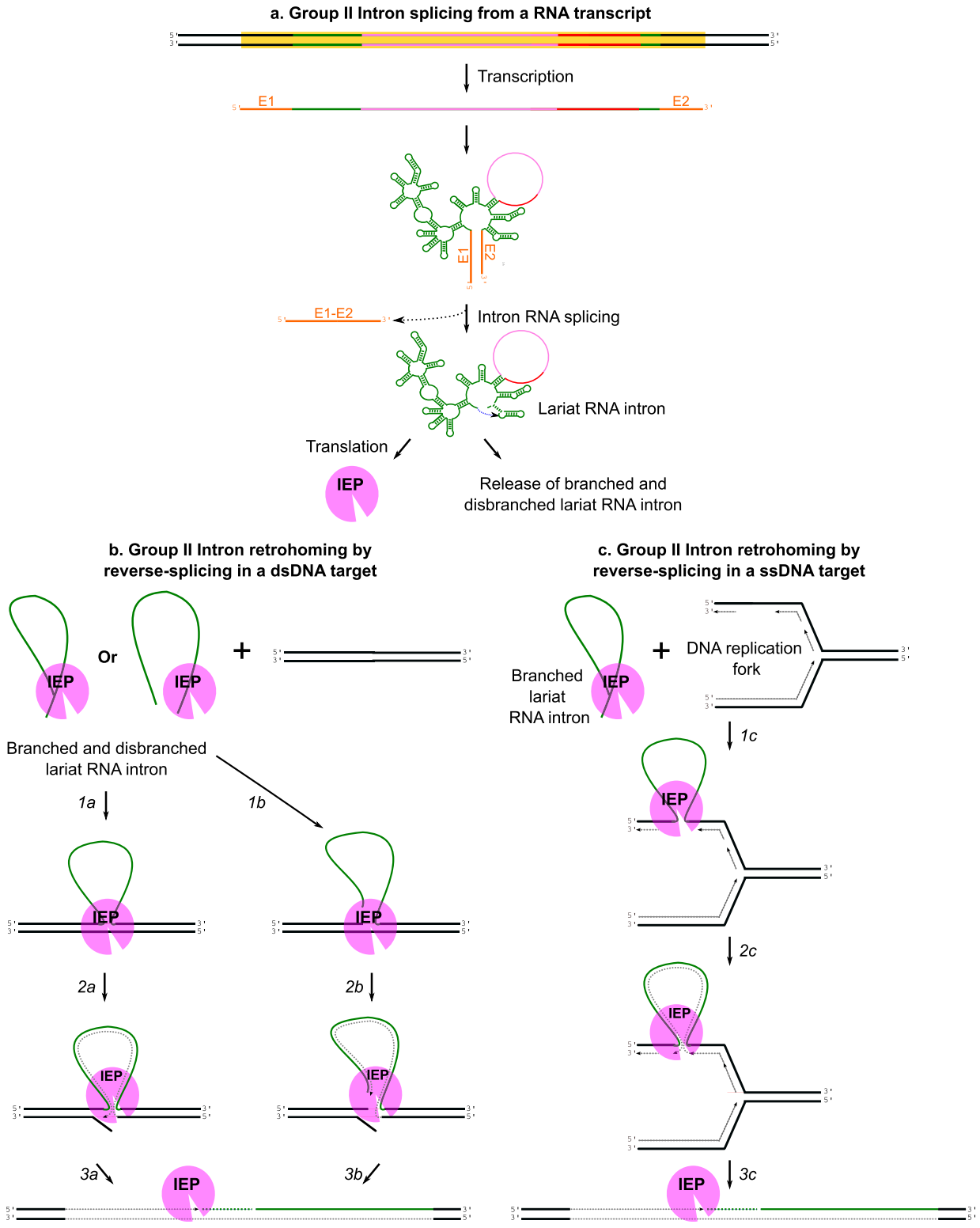

Fig. 5. Features of group II introns. (a) Organization, transcription, splicing from mRNA and protein expression of group II introns. (b) Retrohoming of a group II intron into a double-stranded DNA (dsDNA) target from a branched or a disbranched lariat intron. (c) Retrohoming of a group II intron into a single-stranded DNA (ssDNA) target located within a DNA replication fork from a branched lariat intron. In (b) and (c) the main steps of the group II intron retrohoming are numbered and correspond to: 1a and 1c, full reverse-splicing of a lariat RNA resulting in the insertion of intron RNA into one DNA strand; 1b, partial reverse splicing of a disbranched lariat RNA resulting in the ligation of the 3′ end of the intron RNA to the 5′ end of a cleaved DNA strand; 2a, 2b and 2c, reverse transcription of the fully or partly inserted RNA intron by the intron encoded protein (IEP) and bottom strand cleavage (2a and 2b); 3a, 3b and 3c, 5′ overhang resection, cDNA ligation, RNA degradation and finally two strand DNA synthesis. Rules used to represent the DNA allele and the RNA transcript are identical to those used in Fig. 1a and c. Pink line: ORF coding IEP. Pink "pacman": IEP that contains the RT activity and the HEN in some species. Dotted gray arrows: DNA strands that are under extension in the DNA replication fork or reverse transcribed. Dotted green lines: sections of RNA strand hydrolysed in DNA/RNA duplex. Activity involved at each step and products are indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

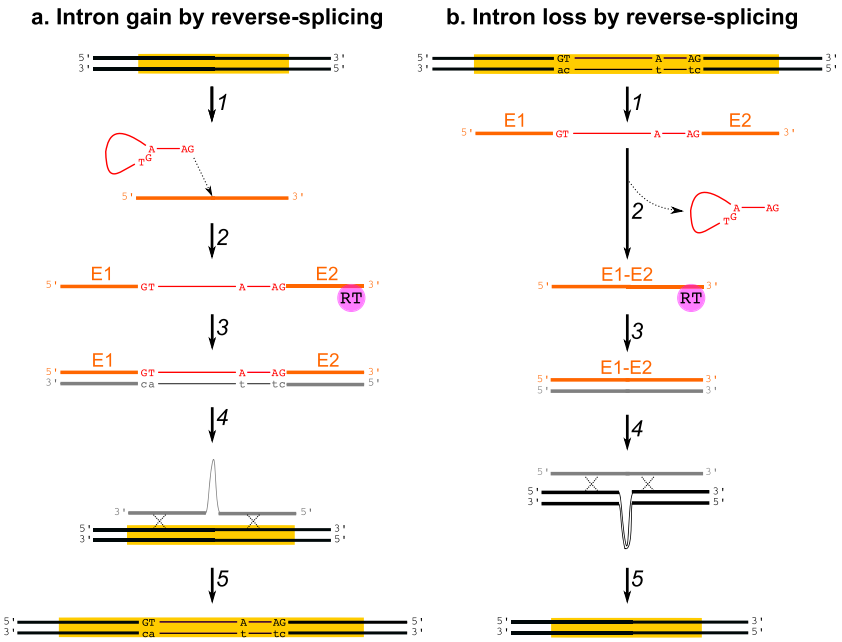

Fig. 6. Intron gain (a) and loss (b) by reverse-splicing. The main steps in both pathways are numbered and correspond to: 1, transcription of an intron-free allele (a) or an allele occupied by intron (b); 2a, integration of a lariat RNA intron into the transcript of the intron-free allele; 2b, excision of the spliceosomal intron; 3, cDNA synthesis by reverse transcription of the invaded allelic transcript (a) or the spliced transcript (b); 4, homologous recombination between the cDNA and an empty DNA allele (a) or an allele occupied by an intron (b). In both schematic representations, symbols used to represent DNA, RNA, exons, allele and transcribed regions are identical to those used in Fig. 1b. Pink balls: reverse transcriptase (RT). Gray lines: DNA strands reverse transcribed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

300 bps. A few bigger ILEs (>2000 bps) are also present (Verhelst et al., 2013). These sequences have no strict insertion target sequence. Instead, they are only found in transcribed gene regions inserted in a single orientation, either within the coding frame, upstream or downstream of untranslated regions (UTRs), or within introns. In some genomes, like that of *M. pusilla*, several introner "species" were characterized on the basis of their sequence similarities. ILEs have most of the characteristics of RSIs. In order to allow splicing most of them have GT-AG donor and an acceptor site ends as well as a predicted branch point motif followed by a polypyrimidine tract (van der Burgt et al., 2012). Some display unusual GT-TG ends that have been reported in other species including human (Verhelst et al., 2013). These hallmark features support the conclusion that ILEs splice using a branched lariat RNA intermediate. However, they also have features that distinguish them from RSIs. The most striking is that they are able to form dyadic intra-strand nucleic acid folds, which suggests that they have stable secondary structures mostly resulting from the annealing of complementary regions in three branches (van der Burgt et al., 2012).

The availability of a sibling genome to *M. pusilla* CCMP1545 that is devoid of introner elements (genome RCC299; 47) has revealed that ILEs are an excellent model to study intron gains and losses. Indeed, variations in ILE number can be quite rapid when examined at the scale of the host species evolution. There is molecular evidence supporting the hypothesis that ILE gains and losses originate from homologous recombination events between chromosomal genes and cDNA copies of reverse transcribed mRNAs that may contain reverse spliced introns (Fig. 6; Roy and Irimia, 2009). However, two points weaken this hypothesis. The first concerns the efficiency of intron gains. Indeed, intron gains are thought to be less efficient than intron losses because they require one extra molecular step, the reverse splicing of the intron into the mRNA. Alternative theoretical models have been proposed to circumvent this issue, (i.e. spliceosomal retrohoming and the reverse transcription template switching, Roy and Irimia, 2009) but no molecular evidence has been provided so far to support these models. The second weakness concerns the identification of genes involved in intron gains. The nature of these unidentified proteins depends on the underlying assumptions regarding ILE origin: an HNH for an origin among group I introns, an RT with a maturase and/or an HNH domain for an origin among group II introns, or an origin among RT related cellular genes (RVT; Gladyshev and Arkhipova, 2011). We used HMMER3 software as well as pfam probabilistic models of conserved domains for RT (pfam00078), maturase (pfam01348) and HNH (pfam00961, pfam01844, pfam13391, pfam13395, pfam01541, pfam07453, and pfam07460) to search for protein homologs in the six translated frames of the nine ILE-containing genomes cited above. Our results show that in the surveyed genomes no gene code an RT related to those found in group II introns and RVT genes, and we could not identify any HNH genes flanked by ILE ends. However, we do find that all surveyed genomes contain several genes coding for putatively active RTs of non-LTR retrotransposons. Based on what is currently known, these RTs are the most likely candidates that ILEs could use in *cis* to amplify their copy number in their respective host genomes. However, even if these enzymes perform the reverse transcription into DNA of mRNA containing newly inserted ILE they cannot be considered as being SINEs. Indeed, their features (absence of similarity of their 3′ end with that of a known LINE; absence of target site duplication at the insertion site; insertion in a single orientation only into transcribed gene regions) indicate a putative mechanism of integration into genes different from that of known SINEs that use LINE machinery to transpose into metazoan chromosomes (Vassetzky and Kramerov, 2013; Collemare et al., 2013; Verhelst et al., 2013).

It appears therefore that group II introns and ILEs are two types of TEs that might use similar integration pathways (Yenerall and Zhou, 2012; Collemare et al., 2013; Verhelst et al., 2013). This conclusion is supported even if: (1) the RT used for their mobility is unrelated, (2) there are no similarities between the intrastrand dyadic structures of their ends, and (3) their ability to splice from

an mRNA via a branched lariat RNA intermediate raises questions about their evolutionary relationships. The history of their discovery also questions their reported abundance (or lack thereof) in eukaryotic genomes. Indeed, their overrepresentation in host genomes was the signature of their presence. If only a few copies were present, they would more likely be considered as unknown retrotransposed SINEs than as a novel ILE species. This highlights the point that the development of sophisticated TE discovery and annotation tools will depend on our understanding of TE diversity.

2.3. Other understudied TEs

In addition to the above elements, eukaryotic TEs with a transposition mechanism undescribed in Curcio and Derbyshire (2003) have also been discovered. These are the putatively self-synthesizing *Polintons/Mavericks* (Kapitonov and Jurka, 2006; Pritham et al., 2007) and the *Zisuptons* (Böhne et al., 2012). These TEs appear to be restricted to eukaryotes. Reciprocally, there are TEs within prokaryotic genomes that do not, so far, have relatives within the eukaryotic world such as the retrons and msDNA. These can transpose using a retrotransposition mechanism different from those used by eukaryotic transposons (Singer, 1995; Lampson et al., 2005). There are also composite TEs using only part of transposition machineries described in Curcio and Derbyshire (2003) (see Roberts et al. (2008) for a review). While these elements have been identified, evidence of their mobility remains a subject in need of further study.

2.4. Placing understudied TEs into existing classification proposals

SSEs are not currently classified in the Finnegan, Repbase, or Wicker proposals. If they were to be placed into one of these classifications we would expect them to be placed into a new class or order, since they transpose without an intermediate and use the host genome's homologous recombination machinery. Similarly, these proposals would classify mobile introns simply as novel non-LTR retrotransposons. However, as we have demonstrated above, this would radically oversimplify their complexity. In the Curcio and Derbyshire proposal, which is based on the way TEs move from one locus to another, the classification of SSEs would be more easily accommodated. However, it would also lead to an explosion of the number of classes since there are at least five different HENs occurring both in inteins and group I introns. Furthermore, the dynamics of HEN evolution among SSEs, as well as the evolution of cleavage specificity of HEN proteins is not yet fully elucidated using proposed evolutionary models (Belfort, 2003; Bigot et al., 2013). Therefore, we do not advocate for the classification of SSEs using any one of these proposals, instead we suggest that the issues raised here justify a debate by SSE specialists in order to determine how to place them within a universal TE classification.

For group II introns, the situation is similarly complex because they have a non-LTR retrotransposon phenotype and in some cases use a HEN for transposition. Because their transposition mechanism does not involve homologous recombination (even though they use the host's replication and repair machineries) they should be distinguished from eukaryotic non-LTR retrotransposons and SSEs. ILEs and some lariat introns mobilize in ways similar to other TEs, they do not encode enzymes required for their mobility, using instead available host enzymes for reverse transcription followed by integration or excision by homologous recombination. However, their ability to reverse splice within and between RNA molecules gives them a transposition phenotype that is similar to that of the group I introns but is mechanistically very different. Indeed, group I intron insertions occur by homologous recombination from DNA to DNA (Fig. 4b) while lariat introns and ILEs excise or integrate into genomic DNA by reverse transcription followed by homologous recombination (Fig. 6).

Finally, in the Finnegan, Repbase or Wicker proposal the *Ginger*, *Polintons/Mavericks* and *Zisuptons* would be classified as DNA transposons. However, in the Curcio and Derbyshire proposal a "DNA transposon" refers to a transposition phenotype that can be applied to elements with similar enzymatic transposition machineries but with different origins. Even though *Ginger* and *Polintons/Mavericks* were found to have DDE integrases (Bao et al., 2010; Yuan and Wessler, 2011), the data are not yet consistent enough to include these TEs in an existing class or to classify them in a novel class. Similarly, before classifying *Zisuptons* in a novel class because of the unique protein that they encode, it will be necessary to demonstrate that this protein is involved in *Zisupton* mobility and does not correspond to an accessory protein with a status similar to that of the TnpB-like protein encoded by *Fanzor* elements (Bao and Jurka, 2013).

3. A call for an international committee

During the period from the original Finnegan proposal until the publication of the Curcio and Derbyshire proposal the amount of knowledge regarding TEs, their transposition mechanisms, diversity, etc. has increased dramatically, a trend that continues to this day. At the same time the microbial and eukaryotic communities have approached the study of TEs from fundamentally different perspectives. Microbiologists have favored the study of TEs as DNA segments that may not necessarily be present in many copies but which are good models for the study of the molecular mechanisms behind recombination and transposition, as well as for the diversity of these mechanisms. This has led microbiologists to view relationships between TEs as defined by modes of transposition with different origins within a set of transposition phenotypes. This probably explains the easy acceptance of the Curcio and Derbyshire proposal by the microbiology community. For scientists working on eukaryotic genomes the challenge was very different. Indeed, the large number and density of TEs in these genomes forced them to invest more time in the identification and characterization of TE sequences (i.e. their phenotype) than into understanding and describing the functional relationships of their mobility. During the last two decades this trend has also been reinforced by the need to annotate sequenced genomes as precisely as possible. This may help explain why the Curcio and Derbyshire proposal has not gained wide acceptance among scientists working on eukaryotic TEs.

Such historical vagaries regarding the evolution of ideas and concepts in TE classification is not unique in the history of the life sciences. The concept of TEs held by many in the eukaryotic community is reminiscent of the history of Lamarkist and neo-Lamarkist ideas that pervaded French scientific research for one and half centuries (nineteenth and first half of the twentieth) rather than those of Darwin and the neo-Darwinists (Loison, 2010). However, the major difference between that historical precedent and today's TE classification debate is that no one proposal provides a satisfactory implementation of a proper scientific classification at all levels. While the Curcio and Derbyshire proposal may be better suited for understanding TE origins we have shown above that it is not well suited at all classification levels. The Finnegan, Wicker, and Repbase proposals with their emphasis on sequence similarities may be much better suited for classification at and below the superfamily level.

3.1. A universal TE classification system

As mentioned above, TEs have evolved from numerous transposition mechanisms with independent origins. Therefore, their

Table 1a
Proposal of TE classes with some members having a DNA transposon phenotype.

Class *Nuclease/Recombinase*	**Order** *Transposition mechanism*	**Superfamilies** *Phylogenetic relationships between Nuclease/Recombinase*
DDE-transposons	*DDE transposons with no DNA-transposition intermediate* (Copy-in)	*Mu* *Tn3*
	DDE/D transposons with a linear dsDNA transposition intermediate * (Cut-out/Paste in)	*IS1, IS3, IS4, IS701, ISH3, IS1634, IS1182, IS6, IS21, IS30, IS66, IS110, IS630, IS982, IS1380, ISAs1, ISL3* *IS630/Tc1/mariner (ITm)/Zator* *IS1595-Merlin,* *IS5/PIF/Harbinger,* *IS256/MuDR/Mutator/Rehavkus* *IS1380/PiggyBac,* *Academ, CACTA/Mirage/Chapaev (CMC), Dada, Hobo/Ac/Tam (hAT), Kolobok, P(?),Sola, Transib,*
	DDE/D transposons with a linear dsDNA transposition intermediate and using a heteromeric transposase (Cut-out/Paste in)	*Tn7*
	DDE transposons with a circular dsDNA transposition intermediate (Copy-out /Paste-in)	*IS3*
	LTR retrotransposons ** (Copy-out/Paste-in)	*Copia* *Gypsy* *BEL* *ERV1* *ERV2* *ERV3*
Y1-transposons	*Y1 transposons with a circular dsDNA transposition* (Cut-out/Paste in)	*IS200/IS605* *Tn916* *CTnDOT* *Crypton*
	Y1 retrotransposons with a circular dsDNA transposition (Copy-out/Paste-in)	*DIRS* *Ngaro* *VIPER*
Y2-transposons	*Y2 transposons with a circular ssDNA transposition* (Copy-in or -out/ Copy-in)	*IS91* *Helitrons*
S-transposons	*S transposons with a circular dsDNA transposition* (Cut-out/paste-in)	*IS607* *Tn5397*
TEs pending classification	*?*	*ISAs1*
	?	*Fanzor*
	Polintons/Mavericks *DDE integrase* (Copy-in or -out/ Copy-in)	*Mavirus (?)* *Polintons/Mavericks* *Tlr1*
	Transposase putatively related to integrases of LTR retrotransposons	*Ginger1* *Ginger2*
	DDE-transposons with a DDE-transposase having another origin	*P(?)*
	Zisupton (Unknown transposition depending on a "Zisuptase")	*Zisupton*

*Superfamily inventory was synthesized from Yuan and Wessler (2011), Siguier et al. (2012), Kojima and Jurka (2013), and Guérillot et al. (2014).
**Relationships with *Retroviridae* need to be further clarified since endogenous elements belonging to at least four *Orthoretrovirinae* genus occur in eukaryotic genomes (Weiss, 2006; Blomberg et al., 2009). Similar investigations will also be required with respect to Caulimoviridae and Hepdnaviridae (Gladyshev and Arkhipova, 2011).

evolution cannot be understood as a tree with a single origin. TE evolutionary concepts are closer to what we know about the evolution and classification of viruses, that is numerous classes with different origins. However, it should be remembered that the ecosystems in which TEs have co-evolved with their hosts as well as their strategy for invasion and maintenance are not the same as viruses, despite sharing some nuclear ecological niches with certain viruses (viruses may also serve as vectors of horizontal transfer for TEs). These nuclear ecological niches are likely the reason why TEs have evolved their own particular survival solutions, as exemplified by certain TE species with strict requirements for integration into host chromosomes (Tn7, Pokey, non-LTR retrotransposons; Mukha et al., 2013), or into non-conventional introns (Milanowski et al., 2013). Such features might have sped up TE evolution and their coevolution with host factor that drives the evolution of their transposition mechanism. A different but similar issue was encountered with viruses regarding the coevolution of virus proteins with host factors (Cardone et al., 2012; Taylor et al., 2013).

In the light of the history of TE classification proposals and their associated problems we have four recommendations for the development of a universal TE classification system. Two of these recommendations are social in nature, urging open communication between scientific communities, the other two are scientific. The first recommendation would be to avoid the temptation of proposing a "perfect" system that would classify all currently known TEs into one proposal. Indeed, no one group of researchers may be expected to master all the required knowledge for such a scheme. Numerous questions will need to be validated and/or corrected by groups of specialists in prokaryotic, viral and eukaryotic TEs. Even if the nature of TEs is not identical within each community, data and concepts that emerge within one community often reemerges in another and is reinforced later by other groups of researchers. Therefore, our proposal (see "proposal outline" below) has no other ambition than to be a starting point for discussion. We believe that such a starting point is the best way to gather all TE communities for a universal TE classification. The second recommendation that logically follows from first, will be for the establishment of an

Table 1b
Proposal of TE classes for TEs with a non-LTR retrotransposon phenotype.

Class *non-LTR retrotransposons*	**Order** *Endonuclease (En)*	**Superfamilies** *Phylogenetic relationships between endonuclease, then RT*
Retroposons	*LINEs*	*LINEs with an AP EN* *LINEs with a PD-(D/E)XK EN* *LINEs with both AP and PD-(D/E)XK EN**
	Penelope-like elements (PLE)	*Athena, no GIY-YIG domain* *Coprina, no GIY-YIG domain* *Neptune, GIY-YIG domain* *Penelope, GIY-YIG domain*
	Group II introns	*Group II introns*** *Mobile lariat introns* *Introners-like elements*

∗LINEs are then organized into 6 families as described (Kapitonov et al., 2009): R2, Randl, L1, RTE, I And Jockey.
∗∗Families of Group II introns can be sorted depending on the presence or not of an LAGLIDADG or HNH endonuclease within their IEP protein and-or as described (Lambowitz and Zimmerly, 2011).

Table 1c
Proposition of TE classes for TEs with an SSE phenotype.

Class *Machinery for excision of host genes*	**Order** *Transposition mechanism*	**Superfamilies** *Phylogenetic relationships between HEN, and-or site into host genes*
Intein	*LAGLIDADG inteins* (HEN dependent HR)	*Host genes in which each intein specifically inserted could be used, as proposed in InBase (http://tools.neb.com/inbase/)*
	HNH inteins (HEN dependent HR)	
Group I intron (G1i)	*LAGLIDADG G1i* (HEN dependent HR)	*Host sites in which each group I intron specifically inserted could also be used*
	HNH G1i (HEN dependent HR)	
	His-Cys G1i (HEN dependent HR)	
	GIY-YIG G1i (HEN dependent HR)	
	PD-(D/E)XK G1i (HEN dependent HR)	
	Vsr G1i (?) (HEN dependent HR)	

organization to allow revisions to TE classification, data dissemination, and active debate concerning TE classification proposals. Our third recommendation would be to attempt to use all useful concepts from former classification proposals. For example, we do not propose a new naming system since several already exist, names assigned to TEs when they are discovered, Repbase names, Wicker's names, IS Finder, Inbase, etc. (Siguier et al., 2012). We believe that it would be more efficient to reach a naming consensus that uses one or several of these existing systems. Our fourth recommendation is to be on guard against using classification criteria that are linked to phenotypes that may be the results of evolutionary convergences. In this report we have shown that using such criteria can be misleading at three levels: (1) describing the overall phenotype of transposition (e.g. retrotransposon versus DNA transposon), (2) assigning a single origin to groups of TEs only because they share a similar sequence organization and transposition machinery (e.g. non-LTR retrotransposons gathering LINEs, *Penelope*-like elements and Group II introns), and (3) not grouping together TEs that likely share synapomorphic excision pathways (e.g. *hAT* and *Transib*). Special attention should also be taken when analyzing patterns of amino acid residues so as not to be misled by evolutionary convergences. This last point was well illustrated by Ekici et al. (2008) and their work on a family of serine proteases from different origins that share a common catalytic Ser/His/Asp triad configuration. Fortunately, 3D protein structures may be particularly useful for identifying inaccurately grouped sequences (Russell, 1998; Schaeffer and Daggett, 2011; Mackin et al., 2014; Fajardo and Fiser, 2013). In addition to the serine protease example there are an increasing number of cases that demonstrate repeated structural evolutionary convergence during catalytic domain evolution (Bork et al., 1993; Russell, 1998; Paiardini et al., 2003; Havrylenko et al., 2010; Schaeffer and Daggett, 2011; Fajardo and Fiser, 2013; Jeoung et al., 2013; Mackin et al., 2014). In the absence of 3D structures these concerns should not be ignored. For example, *P*-like elements have primary and secondary sequences as well as phylogenetic data that support them as containing a catalytic DD[E/D] triad, but ambiguities within these data are also consistent with a paraphyletic origin for P elements (Yuan and Wessler, 2011).

Table 1d
Proposition of TE classes for rare prokaryotic TEs with a retroposon phenotype.

Class *RT features*	Order *Transposition mechanism*	Superfamilies *Phylogenetic relationships between RT*
Retron/msRNA	Retron/msRNA (retrotransposition)	*msRNA*

With the exception of the Intein and Group I intron Classes in Table 1c and 1d, names of superfamilies found in prokaryotes are typed in black, those in eukaryotes being in blue in Table 1a and 1b. Both colors are used for mixed superfamilies. The criteria used are indicated in italics just below the levels of Class, Order and Superfamilies.

3.2. A proposal outline

The origins and evolution of TEs shows similarities with that of viruses, including several TEs having direct evolutionary links with viruses (DDE retrotransposons and Retroviridae, *Marverick/Polinton* and Maviruses, *Helitrons* and Geminivirus, etc.). A universal TE classification system might therefore benefit from using similar hierarchical levels as those used by the universal virus classification system: order, family, subfamily, genus, and species (Adams et al., 2013). Another concept that could be borrowed from the

Table 2
Examples of links between both kinds of non-autonomous TE (NTE) and Superfamilies: internally deleted (ID-NTE) and composite (C-NTE) TEs.

Order	Superfamilies	ID-NTE	C-NTE
DDE transposons with no DNA-transposition intermediate	*Mu, Tn3*	*neh*	*neh*
DDE/D transposons with a linear dsDNA transposition intermediate	*IS1, IS3, IS4, IS701, ISH3, IS1634, IS1182, IS6, IS21, IS30, IS66, IS110, IS630, IS982, IS1380, ISAs1, ISL3, IS630/Tc1/mariner (ITm)/ Zator, IS1595-Merlin, IS5/PIF/Harbinger, IS256/MuDR/Mutator/Rehavkus, IS1380/piggyBac, Academ, CMC, Dada, hAT, ISL2EU, Kolokok, P (?), Sola, Transib,*	*Prokaryotic MITEs (1)* *Eukaryotic MITEs (2) and* SNACs	*CTn, MTn, IMe, MGI (3)* *Chicken hAT-mariner fusion (4)*
DDE transposons with a circular dsDNA transposition intermediate	*IS3*	*neh*	*neh*
LTR retrotransposons	*Copia, Gypsy, BEL, ERV1, ERV2, ERV3*	LARDs, TRIMs	LARDs
Y1 transposons with a circular dsDNA transposition	*IS200/IS605,Tn916, CTnDOT* *Crypton*	*REPtron (REP-BIME; 5)* *CryptonI-1N1_RPro, CryptonS-N1_PS to CryptonS-N6_PS (6)*	*REPtron (REP-BIME; 5)* *BTMR1 (?)*
Y1 retrotransposons with a circular dsDNA transposition	*DIRS, Ngaro, VIPER*	*neh*	*neh*
Y2 transposons with a circular ssDNA transposition	*IS91* *Helitrons*	*Examples in 7*	*Examples in 7*
Fanzor	*Fanzor*	*IDC-Fanzor (8)*	*CC-Fanzor (8)*
Polintons	*Polintons, Tlr1*	*IDC-polintons (9)*	*CC-polintons (9)*
LINEs	*LINEs with an AP EN* *LINEs with a PD-(D/E)XK EN* *LINEs with both AP and PD-(D/E)XK EN*	*Bov-A (10), RIME (11), Vingi-1N1_EE , Vingi-1N11_EE (12), HAL1 (13), HeT-A (14)**	*tRNA derived SINEs (15)* *7SL derived SINEs (15)* *5S derived SINEs (15)* *SVA (16)*
Penelope-like elements (PLE)	*PLE with a GIY-YIG endonuclease*	neh	neh
Group II introns	*Group II introns*	*Examples in 17*	neh
	Mobile lariat introns	neh	neh
	Introners-like elements	*Examples in 18*	neh
Inteins		*Examples in 19*	neh
Group I introns		*Examples in 20*	*ISTron (21)*
Retron/msRNA	*msRNA*	neh	neh

neh, non-exemplified herein; ICE, integrative conjugative element; CTn, conjugative transposon; MTn, mobilisable transposon; IMe, integrative mobilisable element; MGI, mobile genomic island; ∗, all these LINE ID-NTE are derivatives of LINEs with an AP EN. Reviewed or e.g. in references: 1, Delihas (2008); 2, Wicker et al. (2007); 3, Roberts et al. (2008); 4, Wicker et al. (2005), Ton-Hoang et al. (2012); 6, Kojima and Jurka (2011); 7, Tempel et al. (2007); 8, Bao and Jurka (2013); 9, Pritham et al. (2007); 10, Onami et al. (2007); 11, Bringaud et al. (2002); 12, Kojima et al. (2011); 13, Bao and Jurka (2010); 14, Kahn et al. (2000); 15, Vassetzky and Kramerov (2013); 16, Hancks and Kazazian (2010); 17, Dai and Zimmerly (2003); 18, Worden et al. (2009); 19, Bigot et al. (2013); 20, Jackson et al. (2006); 21, Hasselmayer et al. (2004). With the exception of the Intein and Group I intron Classes, names of superfamilies found in prokaryotes are typed in black, those in eukaryotes being in blue. Both colors are used for mixed superfamilies.

virus classification system would be for a group of TEs, whatever its hierarchical level, to be labelled either as an "accepted classification", a "proposed classification", or as "unclassified TEs". These last two labels indicating that further data is needed before either validating the proposed label or changing its status to "evolution" (e.g. its status remains open to modification).

Given the complexity and diversity of TE origins we would like to initially propose that a universal TE classification could be composed of 8 classes (Table 1). Here, the "class" level would be similar to that previously proposed in the Baltimore (1971) virus classification, that is, a grouping of entities with common biological characteristics but not necessarily requiring that they have to share a common origin. While we recognize that this "class" level might not fully reflect the diversity of TE origins we suggest it as practical solution in order to avoid having a too large number of new classes.

Beside the four classes corresponding to the DDE-, Y1-, Y2 and S-transposons, we suggest another four classes. The first would be the retroposon class (including various non-LTR retrotransposons) that would include three orders: LINEs, *Penelope*-like elements and group II introns. Within this class it might be useful to debate whether mobile lariat introns and ILE belong within the group II intron order or whether a new class should be created for these elements (Table 1b). SSEs, we believe, should be composed of two classes, inteins and group I introns. These are grouped based on their protease or ribozyme machineries and on their HEN moieties (Table 1c). This class gathers the Retrons/msRNA (Table 1d). When too little is known regarding a group of TEs to properly classify it we propose keeping it among TEs pending classification (Table 1a). Further details regarding our proposed content within each order and superfamily are provided in Table 1. Within orders, an important challenge for the future will be to elucidate the phylogenetic relationships between prokaryotic and eukaryotic enzymes, as exemplified by DDE transposases within each host type (Siguier et al., 2012; Yuan and Wessler, 2011).

A strength of the Wicker proposal is that it addresses the classification of non-autonomous eukaryotic TEs such as the LARDs, the MITEs, the SNACs and the TRIMs. In addition to these internally deleted TEs there are also non-autonomous eukaryotic TEs that mobilize by using the transposition machinery of other unrelated TEs and from composite elements such as SINEs, SVA, or BTMR1 elements in animal genomes (Hancks and Kazazian, 2010; Casteret et al., 2011). Such non-autonomous TEs which co-opt the transposition machinery of other TEs are also found in prokaryotic genomes. These are MITEs (Delihas, 2008) or composite elements (Hickman et al., 2010). These features could be used to integrate non-autonomous TEs into each order and superfamily as one of two categories: (1) internally deleted non-autonomous elements (ID-NTE) or (2) composite non-autonomous elements (C-NTE) (Table 2). Because the origins of the RT used by ILEs remain to be elucidated and no direct demonstration of their mobility mechanism is yet available, we have placed these TEs into a third superfamily of group II introns in Table 1. However, these might alternatively be classified as C-NTE in Table 2 if it is confirmed that they hijack the RT of a LINE or a group II intron for their mobility.

3.3. Concluding remarks

The proposed backbone for a universal TE classification as well as suggested classification criteria for placing them into taxonomic groups is only a starting point. These and other criteria will need to be evaluated, updated and possibly rejected by groups of specialists from relevant scientific communities. Using this initial proposal as a basis of discussion, we propose that a series of connected classification criteria should be established that will define the groups within lower taxonomic levels, from superfamily to genus and species. Afterwards, using a bottom-up process, the consistency of these criteria should be tested for their ability to identify and annotate TEs from DNA sequences. This work might be carried out within a new organization called the International Committee for the Taxonomy of Transposable Element (ICTTE). This organization would gather TE specialists from prokaryotic, viral and eukaryotic fields. We recognize that such an organization might seem redundant since in the recent past several similar projects have been initiated. Among prokaryotes, such organizations include *ISfinder* (https://www-is.biotoul.fr//) and *ISsaga* (http://issaga.biotoul.fr/ISsaga/issaga_index.php) for prokaryotic IS elements, *Tn Number Registry* (http://www.ucl.ac.uk/eastman/research/departments/microbial-diseases/tn), Inbase (http://tools.neb.com/inbase/) for Inteins, INTEGRALL (http://integrall.bio.ua.pt/) for integrons and integron cassettes, ACLAME (http://aclame.ulb.ac.be/) for ICEs, CTn, MTn, IMe, and MGI. Among eukaryotes REPBASE (http://www.girinst.org/repbase/) is the main TE reference database and its founders have made significant efforts between 2000 and 2006 to create the International Committee on Classification of Transposable Elements. While all of these groups have had more or less success mobilizing a part of the scientific community and in improving global understanding of TE evolution, this multiplicity of organizations has resulted in at least two distinct problems.

The first is that all of the above organizations are impeded by their lack of intercommunication. The slow dissemination of the Curcio and Derbyshire proposal within the eukaryotic TE community is probably the best illustration of this phenomenon. This lack of communication and data exchange between organizations extends, for some of them, to large scale DNA and protein sequence databases as well. The second problem revolves around the friction between the academic need for data sharing and the desire to apply this knowledge for commercial purposes. Specifically, the ownership of TE sequences and TE derived products may be unclear. While most researchers are familiar with databases such as Genbank, ENA, and DDBJ that are members of the International Nucleotide Sequence Database Collaboration (INSDC) with well-established ownership rules (Cochrane, 2010), many of the above mentioned TE databases are not affiliated with the INSDC. Indeed, TE databases are hosted by a mixture of public, private, and other organizations. Because of this, ownership of intellectual property of data entrusted to these organizations and the extent to which these data can be used for downstream purposes is sometimes ambiguous, even when exemptions for research purposes have been added to sequence deposition agreements. For example, it is not clear how consensus sequences or conserved motifs may operate as part of a business plan, or how sequences or applications derived from the content of one of these databases may or may not be included as part of a copyright or a patent application. It is likely that such considerations have limited the enthusiasm of some academic researchers to submit sequences to these databases. The recent US Supreme Court decision in Association for Molecular Pathology et al. V. Myriad Genetics, Inc., et al. (No. 12-398. Argued April 15, 2013 – Decided June 13, 2013) claiming that naturally occurring sequences from the human genome are not patentable may provide a precedent (see Jefferson et al., 2013 for a discussion of the extension of this decision). However, it is not clear how this decision will apply to TEs that may have nearly identical sequences between genomes and may in some cases even move between genomes by horizontal transfer.

We believe that it is time to group disparate TE researchers and databases under an ICTTE organization, not only to share scientific knowledge but also to get ahead of the brewing debate regarding TE data ownership. A workshop to debate and promote these questions appears to us the best approach. The dual aims of this

workshop would be to gather representatives from a variety of TE communities to investigate a universal TE classification and to set up a scientific consortium to gather the resources and the tools over the next 3–5 years to allow the emergence of a functional ICTTE.

Acknowledgments

We thank Aurélie Hua-Van, Pierre Capy, Jonathan Filée, Arnaud Lerouzic (CNRS, Gif sur Yvette, France) and Mike Chandler (CNRS, Toulouse, France) for the quality of our active debate during the development of our investigations. We thank Dr Fabien Palazzoli (IP Studies SARL, Switzerland) for fruitful discussion about copyrights and ownership of public data and their derivatives. This work was funded by the C.N.R.S., the I.N.R.A., the Groupement de Recherche CNRS 2157, and the Ministère de l'Education Nationale, de la Recherche et de la Technologie. Solenne Bire holds a post-doctoral fellowship from the overheads of the European Project SyntheGeneDelivery (FP6 N° 18716). Peter Arensburger holds a senior researcher fellowship from the STUDIUM.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.03.009.

References

Adams, M.J., Lefkowitz, E.J., King, A.M., Carstens, E.B., 2013. Recently agreed changes to the International Code of Virus Classification and Nomenclature. Arch. Virol. 158, 2633–2639.

Arkhipova, I.R., 2006. Distribution and phylogeny of *Penelope*-like elements in eukaryotes. Syst. Biol. 55, 875–885.

Baltimore, D., 1971. Expression of animal virus genomes. Bacteriol. Rev. 35, 235–241.

Bandea, C.I., 2009. The Origin and Evolution of Viruses as Molecular Organisms. Available from *Nature Precedings* <http://hdl.handle.net/10101/npre.2009.3886.1>.

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

Bao, W., Jurka, J., 2013. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. Mob. DNA 4, 12.

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, 3.

Barzel, A., Naor, A., Privman, E., Kupiec, M., Gophna, U., 2011. Homing endonucleases residing within inteins: evolutionary puzzles awaiting genetic solutions. Biochem. Soc. Trans. 39, 169–173.

Belfort, M., 2003. Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. Genes Dev. 17, 2860–2863.

Biémont, C., Vieira, C., 2006. Genetics: junk DNA as an evolutionary force. Nature 443, 521–524.

Bigot, Y., Piégu, B., Casteret, S., Gavory, F., Bideshi, D.K., Federici, B.A., 2013. Characteristics of inteins in invertebrate iridoviruses and factors controlling insertion in their viral hosts. Mol. Phylogenet. Evol. 67, 246–254.

Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G., Mayer, J., 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 448, 115–123.

Boeke, J.D., 1989. Transposable elements in *Saccharomyces cerevisiae*. In: Berg, D.L., Howe, M. (Eds.), Mobile DNA. ASM Press, Washington (DC), pp. 335–374.

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

Boocock, M.R., Rice, P.A., 2013. A proposed mechanism for *IS607*-family serine transposases. Mob. DNA 4, 24.

Bork, P., Sander, C., Valencia, A., 1993. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Sci. 2, 31–40.

Bringaud, F., García-Pérez, J.L., Heras, S.R., Ghedin, E., El-Sayed, N.M., Andersson, B., Baltz, T., Lopez, M.C., 2002. Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 124, 73–78.

Britten, R.J., Stout, D.B., Davidson, E.H., 1989. The current source of human *Alu* retroposons is a conserved gene shared with Old World monkey. Proc. Natl. Acad. Sci. USA 86, 3718–3722.

Bruce, A., 2008. Molecular Biology of the Cell. Garland Science, New York.

Capy, P., Maisonhaute, C., 2002. Acquisition and loss of modules: the construction set of transposable elements. Genetika 38, 719–726.

Capy, P., Vitalis, R., Langin, T., Higuet, D., Bazin, C., 1996. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J. Mol. Evol. 42, 359–368.

Cardone, G., Heymann, J.B., Cheng, N., Trus, B.L., Steven, A.C., 2012. Procapsid assembly, maturation, nuclear exit: dynamic steps in the production of infectious herpesvirions. Adv. Exp. Med. Biol. 726, 423–439.

Casteret, S., Moiré, N., Aupinel, P., Tasei, J.N., Bigot, Y., 2011. Profile of the mosaic element BTMR1 in the genome of the bumble bee *Bombus terrestris* (Hymenoptera: Apidae). Insect Mol. Biol. 20, 153–164.

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

Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y., and on behalf of the International Nucleotide Sequence Database Collaboration, 2010. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 39, no. Database (November 23, 2010), D15–D18.

Collemare, J., van der Burgt, A., de Wit, P.J., 2013. At the origin of spliceosomal introns: is multiplication of introner-like elements the main mechanism of intron gain in fungi? Commun. Integr. Biol. 6, e23147.

Copertino, D.W., Hallick, R.B., 1991. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. EMBO J. 10, 433–442.

Craig, N.L., 1995. Unity in transposition reactions. Science 270, 253–254.

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

Dai, L., Zimmerly, S., 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. RNA 9, 14–99.

Daly, M., Herendeen, P.S., Gurlanick, R.P., Westneat, M.W., McDade, L., 2012. Systematic Agenda 2020: the mission evolves. Syst. Zool. 61, 549–552.

Dassa, B., London, N., Stoddard, B.L., Schueler-Furman, O., Pietrokovski, S., 2009. Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. Nucleic Acids Res. 37, 2560–2573.

de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7, e1002384.

Delihas, N., 2008. Small mobile sequences in bacteria display diverse structure/function motifs. Mol. Microbiol. 67, 475–481.

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E.V., Raoult, D., 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc. Natl. Acad. Sci. USA 109, 18078–18083.

Edgell, D.R., Chalamcharla, V.R., Belfort, M., 2011. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol. 9, 22.

Eickbush, T.H., Jamburuthugoda, V.K., 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 134, 221–234.

Ekici, O.D., Paetzel, M., Dalbey, R.E., 2008. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. Protein Sci. 17, 2023–2037.

Fajardo, J.E., Fiser, A., 2013. Protein structure based prediction of catalytic residues. BMC Bioinf. 14, 63.

Feng, Q., Moran, J.V., Kazazian Jr, H.H., Boeke, J.D., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905–916.

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

Finnegan, D.J., 1992. Transposable elements. Curr. Opin. Genet. Dev. 2, 861–867.

Finnegan, D.J., Fawcett, D.H., 1986. Transposable elements in Drosophila melanogaster. Oxf. Surv. Eukaryot. Genes 3, 1–62.

Fischer, M.G., Suttle, C.A., 2011. A virophage at the origin of large DNA transposons. Science 332, 231–234.

Garriga, G., Lambowitz, A.M., 1983. RNA splicing in *Neurospora* mitochondria. The large rRNA intron contains a noncoded, 5′-terminal guanosine residue. J. Biol. Chem. 258, 14745–14748.

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

Gladyshev, E.A., Arkhipova, I.R., 2011. A widespread class of reverse transcriptase-related cellular genes. Proc. Natl. Acad. Sci. USA 108, 20311–20316.

Gogarten, J.P., Hilario, E., 2006. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol. Biol. 6, 94.

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

Guérillot, R., Siguier, P., Gourbeyre, E., Chandler, M., Glaser, P., 2014. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. Genome Biol. Evol. 6, 260–272.

Hancks, D.C., Kazazian Jr., H.H., 2010. SVA retrotransposons: evolution and genetic instability. Semin. Cancer Biol. 20, 234–245.

Haren, L., Ton-Hoang, B., Chandler, M., 1999. Integrating DNA: transposases and retroviral integrases. Annu. Rev. Microbiol. 53, 245–281.
Hasselmayer, O., Braun, V., Nitsche, C., Moos, M., Rupnik, M., von Eichel-Streiber, C., 2004. Clostridium difficile IStron *CdISt1*: discovery of a variant encoding two complete transposase-like proteins. J. Bacteriol. 186, 2508–2510.
Haugen, P., Simon, D.M., Bhattacharya, D., 2005. The natural history of group I introns. Trends Genet. 21, 111–119.
Havrylenko, S., Legouis, R., Negrutskii, B., Mirande, M., 2010. Methionyl-tRNA synthetase from *Caenorhabditis elegans*: a specific multidomain organization for convergent functional evolution. Protein Sci. 19, 2475–2484.
Haynes, S.R., Toomey, T.P., Leinwand, L., Jelinek, W.R., 1981. The Chinese hamster *Alu*-equivalent sequence: a conserved highly repetitious, interspersed deoxyribonucleic acid sequence in mammals has a structure suggestive of a transposable element. Mol. Cell. Biol. 1, 573–583.
Hencken, C.G., Li, X., Craig, N.L., 2012. Functional characterization of an active *Rag*-like transposase. Nat. Struct. Mol. Biol. 19, 834–836.
Hickman, A.B., Chandler, M., Dyda, F., 2010. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit. Rev. Biochem. Mol. Biol. 45, 50–69.
Hirata, R., Ohsumk, Y., Nakano, A., Kawasaki, H., Suzuki, K., Anraku, Y., 1990. Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. J. Biol. Chem. 265, 6726–6733.
Jackson, S.A., Koduvayur, S., Woodson, S.A., 2006. Self-splicing of a group I intron reveals partitioning of native and misfolded RNA populations in yeast. RNA 12, 2149–2159.
Jefferson, O.A., Köllhofer, D., Ehrich, T.H., Jefferson, R.A., 2013. Transparency tools in gene patenting for informing policy and practice. Nat. Biotechnol. 31, 1086–1093.
Jeoung, J.H., Bommer, M., Lin, T.Y., Dobbek, H., 2013. Visualizing the substrate-, superoxo-, alkylperoxo-, and product-bound states at the nonheme Fe(II) site of homogentisate dioxygenase. Proc. Natl. Acad. Sci. USA 110, 12625–12630.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.
Kahn, T1., Savitsky, M., Georgiev, P., 2000. Attachment of HeT-A sequences to chromosomal termini in Drosophila melanogaster may occur by different mechanisms. Mol. Cell. Biol. 20, 7634–7642.
Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Goebl, M., Stevens, T.H., 1990. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science 250, 651–657.
Kapitonov, V.V., Jurka, J., 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biol. 3, e181.
Kapitonov, V.V., Jurka, J., 2006. Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. USA 103, 4540–4545.
Kapitonov, V.V., Jurka, J., 2007. *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trends Genet. 23, 521–529.
Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9, 411–412.
Kapitonov, V.V., Tempel, S., Jurka, J., 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448, 207–213.
Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55, 1–24.
King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2011. Viral taxonomy. In: King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J. (Eds.), Virus Taxonomy, IXth Report of the International Committee on the Taxonomy of Viruses, third ed. Elsevier – Academic Press, London.
Kojima, K.K., Fujiwara, H., 2005. An extraordinary retrotransposon family encoding dual endonucleases. Genome Res. 15, 1106–1117.
Kojima, K.K., Jurka, J., 2011. *Crypton* transposons: identification of new diverse families and ancient domestication events. Mob. DNA 2, 12.
Kojima, K.K., Jurka, J., 2013. A superfamily of DNA transposons targeting multicopy small RNA genes. PLoS One 8, e68260.
Kojima, K.K., Kapitonov, V.V., Jurka, J., 2011. Recent expansion of a new Ingi-related clade of Vingi non-LTR retrotransposons in hedgehogs. Mol. Biol. Evol. 28, 17–20.
Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., Cech, T.R., 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell 31, 147–157.
Lambowitz, A.M., Zimmerly, S., 2011. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. 3, a003616.
Lampson, B.C., Inouye, M., Inouye, S., 2005. Retrons, msDNA, and the bacterial genome. Cytogenet. Genome Res. 110, 491–499.
Lapage, S.P., Sneath, P.H.A., Lessel, E.F., Skerman, V.B.D., Seeliger, H.P.R., Clark, W.A., 1992. In: Lapage, S.P., Sneath, P.H.A., Lessel, E.F., Skerman, V.B.D., Seeliger, H.P.R., Clark, W.A. (Eds.), International Code of Nomenclature of Bacteria, Bacteriological Code, 1990 Revision. ASM Press, Washington (DC).
Lerat, E., Brunet, F., Bazin, C., Capy, P., 1999. Is the evolution of transposable elements modular? Genetica 107, 15–25.
Loison, L., 2010. In: Loison, Laurent (Ed.), Qu'est ce que le néolamarkisme? Les biologistes français et la question de l'évolution des espèces. Vuibert, Paris.
Mackin, K.A., Roy, R.A., Theobald, D.L., 2014. An empirical test of convergent evolution in rhodopsins. Mol. Biol. Evol. 31, 81–95.
Mahillon, J., Chandler, M., 1998. Insertion sequences. Microbiol. Mol. Biol. Rev. 62, 725–774.
Maier, U.G., Rensing, S.A., Igloi, G.L., Maerz, M., 1995. Twintrons are not unique to the Euglena chloroplast genome: structure and evolution of a plastome cpn60 gene from a cryptomonad. Mol. Gen. Genet. 246, 128–131.
Marcia, M., Somarowthu, S., Pyle, A.M., 2013. Now on display: a gallery of group II intron structures at different stages of catalysis. Mob. DNA 4, 14.
Mayr, E., Bock, W.J., 2002. Classifications and other ordering systems. J. Zool. Syst. Evol. Res. 40, 169–194.
McManus, H.A., Lewis, L.A., Fučíková, K., Haugen, P., 2012. Invasion of protein coding genes by green algal ribosomal group I introns. Mol. Phylogenet. Evol. 62, 109–116.
McNeill, J., Barrie, F.R., Buck, W.R., Demoulin, V., Greuter, W., Hawksworth, D.L., Herendeen, P.S., Knapp, S., Marhold, K., Prado, J., Prud'homme van Reine, W.F., Smith, G.F., Wiersema, J.H., Turland, N.J., 2012. International Code of Nomenclature for algae, fungi and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011. Koeltz Scientific Books, Koenigstein, Germany.
Milanowski, R., Karnkowska, A., Ishikawa, T., Zakrys, B., 2013. Distribution of conventional and nonconventional introns in tubulin (α and β) genes of Euglenids. Mol. Biol. Evol. 31, 584–593.
Mitra, R., Fain-Thornton, J., Craig, N.L., 2008. *PiggyBac* can bypass DNA synthesis during cut and paste transposition. EMBO J. 27, 1097–1109.
Mizuuchi, K., 1992. Transpositional recombination: mechanistic insights from studies of mu and other elements. Annu. Rev. Biochem. 61, 1011–1051.
Mukha, D.V., Pasyukova, E.G., Kapelinskaya, T.V., Kagramanova, A.S., 2013. Endonuclease domain of the Drosophila melanogaster R2 non-LTR retrotransposon and related retroelements: a new model for transposition. Front. Genet. 4, 63.
Murad, L., Bielawski, J.P., Matyasek, R., Kovarík, A., Nichols, R.A., Leitch, A.R., Lichtenstein, C.P., 2004. The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. Heredity 92, 352–358.
Onami, J., Nikaido, M., Mannen, H., Okada, N., 2007. Genomic expansion of the Bov-A2 retroposon relating to phylogeny and breed management. Mamm. Genome 18, 187–196.
Paiardini, A., Contestabile, R., D'Aguanno, S., Pascarella, S., Bossa, F., 2003. Threonine aldolase and alanine racemase: novel examples of convergent evolution in the superfamily of vitamin B6-dependent enzymes. Biochim. Biophys. Acta 1647, 214–219.
Pennisi, E., 2013. Ever-bigger viruses shake tree of life. Microbiology 341, 226–227.
Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., Abergel, C., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341, 281–286.
Pietrokovski, S., 1998. Identification of a virus intein and a possible variation in the protein-splicing reaction. Curr. Biol. 8, R634–R635.
Pietrokovski, S., 2001. Intein spread and extinction in evolution. Trends Genet. 17, 465–472.
Poulter, R.T., Goodwin, T.J., 2005. *DIRS-1* and the other tyrosine recombinase retrotransposons. Cytogenet. Genome Res. 110, 575–588.
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, 3–17.
Raghavan, R., Minnick, M.F., 2009. Group I introns and inteins: disparate origins but convergent parasitic strategies. J. Bacteriol. 191, 6193–6202.
Raoult, D., 2013. TRUC or the need for a new microbial classification. Intervirology 56, 349–353.
Ride, W.D.L., Cogger, H.G., Dupuis, C., Kraus, O., Minelli, A., Thompson, F.C., Tubbs, P.K., 2000. In: Ride, W.D.L., Cogger, H.G., Dupuis, C., Kraus, O., Minelli, A., Thompson, F.C., Tubbs, P.K. (Eds.), International Commission on Zoological Nomenclature, fourth ed. electronic ed. <http://www.nhm.ac.uk/hosted-sites/iczn/code/>.
Roberts, A.P., Chandler, M., Courvalin, P., Guédon, G., Mullany, P., Pembroke, T., Rood, J.I., Smith, C.J., Summers, A.O., Tsuda, M., Berg, D.E., 2008. Revised nomenclature for transposable genetic elements. Plasmid 60, 167–173.
Roy, S.W., Irimia, M., 2009. Mystery of intron gain: new data and new models. Trends Genet. 25, 67–73.
Russell, R.B., 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J. Mol. Biol. 279, 1211–1227.
Scamborova, P., Wong, A., Steitz, J.A., 2004. An intronic enhancer regulates splicing of the twintron of *Drosophila melanogaster* prospero pre-mRNA by two different spliceosomes. Mol. Cell. Biol. 24, 1855–1869.
Schaeffer, R.D., Daggett, V., 2011. Protein folds and protein folding. Protein Eng. Des. Sel. 24, 11–19.
Schmid, C.W., Jelinek, W.R., 1982. The Alu family of dispersed repetitive sequences. Science 216, 1065–1070.
Seberg, O., Petersen, G., 2009. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat. Rev. Genet. 10, 276.
Siguier, P., Varani, A., Perochon, J., Chandler, M., 2012. Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. Methods Mol. Biol. 859, 91–103.
Singer, M.F., 1982. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28, 433–434.
Singer, M.F., 1995. Unusual reverse transcriptases. J. Biol. Chem. 270, 24623–24626.
Stoddard, B.L., 2005. Homing endonuclease structure and function. Q. Rev. Biophys. 38, 49–95.
Stoye, J.P., Blomberg, J., Coffin, J.M., Fan, H., Hahn, B., Neil, J., Quackenbush, S., Tristem, M., 2012. Family *Retroviridae*. In: King, M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J. (Eds.), Virus Taxonomy, IXth Report of the International Committee on the Taxonomy of Viruses, third ed. Elsevier – Academic Press, London.

Taylor, D.J., Ballinger, M.J., Bowman, S.M., Bruenn, J., 2013. Virus-host co-evolution under a modified nuclear genetic code. Peer J 1, e50.
Tempel, S., Nicolas, J., El Amrani, A., Couée, I., 2007. Model-based identification of Helitrons results in a new classification of their families in Arabidopsis thaliana. Gene 403, 18–28.
Ton-Hoang, B., Pasternak, C., Siguier, P., Guynet, C., Hickman, A.B., Dyda, F., Sommer, S., Chandler, M., 2010. Single-stranded DNA transposition is coupled to host replication. Cell 142, 398–408.
Ton-Hoang, B., Siguier, P., Quentin, Y., Onillon, S., Marty, B., Fichant, G., Chandler, M., 2012. Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. Nucleic Acids Res. 40, 3596–3609.
Turlan, C., Chandler, M., 2000. Playing second fiddle: second-strand processing and liberation of transposable elements from donor DNA. Trends Microbiol. 8, 268–274.
van der Burgt, A., Severing, E., de Wit, P.J., Collemare, J., 2012. Birth of new spliceosomal Introns in fungi by multiplication of Intron-like elements. Curr. Biol. 22, 1260–1265.
van Regenmortel, M.H., Ackermann, H.W., Calisher, C.H., Dietzgen, R.G., Horzinek, M.C., Keil, G.M., Mahy, B.W., Martelli, G.P., Murphy, F.A., Pringle, C., Rima, B.K., Skern, T., Vetten, H.J., Weaver, S.C., 2013. Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. Arch. Virol. 158, 1115–1119.
Vassetzky, N.S., Kramerov, D.A., 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 41, D83–D89.
Verhelst, B., Van de Peer, Y., Rouzé, P., 2013. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. Genome Biol. Evol. 5, 2393–2401.
Vicens, Q., Cech, T.R., 2006. Atomic level architecture of group I introns revealed. Trends Biochem. Sci. 31, 41–51.
Waring, R.B., Davies, R.W., 1984. Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing – a review. Gene 28, 277–291.
Webb, E.C., 1992. Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, San Diego. <http://www.chem.qmul.ac.uk/iubmb/enzyme/>.
Weiss, R.A., 2006. The discovery of endogenous retroviruses. Retrovirology 3, 67.
Wicker, T., Robertson, J.S., Schulze, S.R., Feltus, F.A., Magrini, V., Morrison, J.A., Mardis, E.R., Wilson, R.K., Peterson, D.G., Paterson, A.H., Ivarie, R., 2005. The repetitive landscape of the chicken genome. Genome Res. 15, 126–136.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2008. Reply: a universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9, 414.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2009. Reply: a unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat. Rev. Genet. 10, 276.
Williams, T.A., Embley, T.M., Heinz, E., 2011. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. PLoS One 6, e21080.
Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., Foulon, E., Grimwood, J., Gundlach, H., Henrissat, B., Napoli, C., McDonald, S.M., Parker, M.S., Rombauts, S., Salamov, A., Von Dassow, P., Badger, J.H., Coutinho, P.M., Demir, E., Dubchak, I., Gentemann, C., Eikrem, W., Gready, J.E., John, U., Lanier, W., Lindquist, E.A., Lucas, S., Mayer, K.F., Moreau, H., Not, F., Otillar, R., Panaud, O., Pangilinan, J., Paulsen, I., Piegu, B., Poliakov, A., Robbens, S., Schmutz, J., Toulza, E., Wyss, T., Zelensky, A., Zhou, K., Armbrust, E.V., Bhattacharya, D., Goodenough, U.W., Van de Peer, Y., Grigoriev, I.V., 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324, 268–272.
Yang, J., Malik, H.S., Eickbush, T.H., 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc. Natl. Acad. Sci. USA 96, 7847–7852.
Yenerall, P., Zhou, L., 2012. Identifying the mechanisms of intron gain: progress and trends. Biol. Direct 7, 29.
Yuan, Y.W., Wessler, S.R., 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc. Natl. Acad. Sci. USA 108, 7884–7889.
Yutin, N., Raoult, D., Koonin, E.V., 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol. J. 10, 158.
Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F., Craig, N.L., 2004. Transposition of *hAT* elements links transposable elements and *V(D)J* recombination. Nature 432, 995–1001.