GUIDELINES

A unified classification system for eukaryotic transposable elements

Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel and Alan H. Schulman

Abstract | Our knowledge of the structure and composition of genomes is rapidly progressing in pace with their sequencing. The emerging data show that a significant portion of eukaryotic genomes is composed of transposable elements (TEs). Given the abundance and diversity of TEs and the speed at which large quantities of sequence data are emerging, identification and annotation of TEs presents a significant challenge. Here we propose the first unified hierarchical classification system, designed on the basis of the transposition mechanism, sequence similarities and structural relationships, that can be easily applied by non-experts. The system and nomenclature is kept up to date at the WikiPoson web site.

Our knowledge of the structure and composition of genomes has progressed in pace with their sequencing. As was expected early on[1], TEs have been found in virtually all eukaryotic species investigated so far[2–4]. The only known exceptions are *Plasmodium falciparum* and probably several closely related species. The TEs affect the genome by their ability to move and replicate, thereby generating plasticity. Although the main TE groups are ancient and are present in all kingdoms, TEs display extreme diversity: there are thousands or even tens of thousands of different TE families in plants[5–8], constituting 80% or more of the total genomic DNA. Although they are less abundant than those in plants, TEs in fungi and metazoans also represent a substantial part of their genomes (3–20% in fungi and 3–45% in metazoans), and include members of most superfamilies[9,10].

More and more large and complex eukaryotic genomes are being selected for sequencing, and they too are expected to be rich in TEs. Given the abundance and diversity of TEs, these genomes will present an enormous TE identification and annotation problem. Hence, there is an increasing demand for tools to handle these rapidly emerging sequences effectively. This requires straightforward and efficient nomenclature keys and classification strategies that can help non-specialists to easily annotate TEs.

The sequencing of the genomes of multiple related species will facilitate comparative studies and provide insights into genome evolution. However, comparisons will be compromised unless TE annotation is consistent. Until now, TE analysis and classification have generally been carried out on a species-by-species basis, and comparative studies of TEs between taxa have been rare. Thus, no unified system has been applied across species and, therefore, across kingdoms. Accordingly, TE databases are generally restricted to individual or closely related species and tend to lack systematic structure.

Here we propose a common TE classification system that can be easily handled by non-specialists during annotation. It provides a consensus between the various conflicting classification and naming systems that are currently in use. A key component of this system is a naming convention: a three-letter code with each letter respectively denoting class, order and superfamily; the family (or subfamily) name; the sequence (database accession) on which the element was found; and the 'running number', which defines the individual insertion in the accession. The unified system is also intended to facilitate comparative and evolutionary studies on TEs from different species.

Outline of the classification system

In 1989, Finnegan proposed the first TE classification system, which distinguished two classes by their transposition intermediate: RNA (class I or retrotransposons) or DNA (class II or DNA transposons). The transposition mechanism of class I is commonly called 'copy-and-paste', and that of class II, 'cut-and-paste'[11]. The discovery of bacterial[12] and eukaryotic[13,14] TEs that copy and paste but without RNA intermediates, and of highly reduced non-autonomous TEs called miniature inverted repeat transposable elements (MITEs), has challenged the two-class system. These findings led to schemes that either invoke a third class or jettison the two-class system in favour of enzymological categories[15].

Our hierarchical TE classification system (FIG. 1) reconciles both approaches, maintaining two classes while applying mechanistic and enzymatic criteria. It includes (in hierarchical order) the levels of class, subclass, order, superfamily, family and subfamily. These terms were chosen to mirror those used in organismal phylogenetics and to recognize the fact that some terms (such as superfamily) are already in widespread use for TE groups. The highest level (class) divides TEs by the presence or absence of an RNA transposition intermediate, as before[11]. Subclass, previously used to separate long terminal repeat (LTR) from non-LTR (long and short interspersed nuclear element (LINE and SINE)) class I TEs, here is used to distinguish elements that copy themselves for insertion from those that leave the donor site to reintegrate elsewhere. It concomitantly reflects the number of DNA strands that are cut at the TE donor site. A new taxon — order — marks major differences in the insertion mechanism and, consequently, overall organization and enzymology, thereby replacing subclass for type I TEs.

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)		存在RT酶			
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR 缺少整合酶INT	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF– Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Figure 1 | **Proposed classification system for transposable elements (TEs).** The classification is hierarchical and divides TEs into two main classes on the basis of the presence or absence of RNA as a transposition intermediate. They are further subdivided into subclasses, orders and superfamilies. The size of the target site duplication (TSD), which is characteristic for most superfamilies, can be used as a diagnostic feature. To facilitate identification, we propose a three-letter code that describes all major groups and that is added to the family name of each TE. DIRS, *Dictyostelium* intermediate repeat sequence; LINE, long interspersed nuclear element; LTR, long terminal repeat; PLE, *Penelope*-like elements; SINE, short interspersed nuclear element; TIR, terminal inverted repeat.

Superfamilies within an order share a replication strategy, but are distinguished by uniform and widespread large-scale features, such as the structure of protein or non-coding domains. Furthermore, superfamilies differ in the presence and size of the target site duplication (TSD), a short direct repeat that is generated on both flanks of a TE upon insertion. There is virtually no sequence conservation at the DNA level and only limited similarities at the protein level between superfamilies.

Superfamilies are subdivided into families, which are defined by DNA sequence conservation. Similarity at the protein level is generally high between different families that belong to the same superfamily; DNA sequence conservation, however, is minimal and restricted to highly conserved parts of coding regions. Although TEs can be classified into relatively few orders and superfamilies, a single large genome can contain hundreds or thousands of diverse TE families. Subfamilies are defined on the basis of phylogenetic data, and might in specific cases serve to distinguish internally homogeneous autonomous and non-autonomous populations (see below).

The lowest taxon, insertion, describes one particular individual copy, corresponding to a specific transposition and insertion event, and is of particular relevance to genome annotation. At all levels above insertion, an element can be temporarily assigned the appellation 'unknown' as its classification. For example, a novel TE could be identified as an LTR retrotransposon on the basis of the presence of characteristic LTRs, but the classification into one of the superfamilies (for example, *Gypsy* or *Copia*) may not be possible owing to the lack of a coding sequence. In this way, the complex task of classification is left to specialists, and is not required of annotators. For consistency of style, we propose that designations at all taxonomic levels below order should be written in italics.

Class I elements

Class I TEs all transpose via an RNA intermediate. This class needs no subclasses — no members cleave or transfer DNA strands at the donor site. Instead, the RNA intermediate is transcribed from a genomic copy, then reverse-transcribed into DNA by a TE-encoded reverse transcriptase (RT). Each complete replication cycle produces one new copy. Consequently, retrotransposons are often the major contributors to the repetitive fraction in large genomes[16–19]. Retrotransposons can be divided into five orders (FIG. 1) on the basis of their mechanistic features, organization and reverse transcriptase phylogeny: LTR retrotransposons, *DIRS*-like elements, *Penelope*-like elements (PLEs), LINEs and SINEs.

The LTR retrotransposons are less abundant in animals, but are the predominant order in plants. They range from a few hundred base pairs up to, exceptionally, 25 kb (*Ogre*[20]). LTRs flanking the elements range from a few hundred base pairs to more than 5 kb, and start with 5′-TG-3′ and end with 5′-CA-3′. Upon integration, LTR retrotransposons produce a TSD of 4–6 bp. They typically contain ORFs for GAG, a structural protein for virus-like particles, and for POL. *Pol* encodes an aspartic proteinase (AP), reverse transcriptase, RNase H (RH), and DDE integrase (INT). Occasionally, there is an additional ORF of unknown function[20]. LTR retrotransposons also contain specific signals for packaging, dimerization, reverse transcription and integration (see below). The two main superfamilies outside metazoans, *Gypsy* and *Copia*, differ in the order of RT and INT in the POL (FIG. 1). All superfamilies in the order LTR use a similar transposition mechanism.

Many LTR retrotransposon families are not restricted to any one species, at least in grasses. Those with large genomes, such as maize, wheat or barley can contain thousands of families. However, despite their enormous diversity, relatively few families comprise the bulk of the repetitive fraction in those large genomes. Examples of such families are *Angela* in wheat[21], *BARE1* in barley[22], *Opie* in maize[23] and *Retrosor6* in sorghum[24]. Numerous families of LTR retroelements are present in humans, but most of them are no longer active[25] and are present in only moderate numbers.

Evolutionarily, retroviruses and LTR retrotransposons are closely related. Retroviruses might have evolved from *Gypsy* LTR retrotransposons that adopted a viral lifestyle through acquisition of an envelope protein (ENV) and a set of additional proteins and regulatory sequences[26,27]. Retroviruses are largely restricted to vertebrates, although some members of the *Drosophila melanogaster Gypsy* family are able to infect new individuals[28]. The retroviruses can therefore be placed within our classification scheme, although they have long been classified as viruses[29] and differ in key aspects from retroelements. For example, their evolution is influenced by epidemiology rather than organismal phylogeny.

A retrovirus can also be transformed into an LTR retrotransposon through inactivation or deletion of the domains that enable extracellular mobility[30]. No longer infectious, they rely on vertical transmission through the germ line for propagation. Thus, we have placed these so-called endogenous retroviruses (ERVs)[31] into our system as a superfamily within the LTR subclass. Although their *pol* ORF is similar to that of *Gypsy* TEs, their *gag* ORF encodes both capsid and matrix functions (*Gypsy gag* encodes only the matrix). Moreover, they still possess an ENV that harbours both surface and transmembrane units, which is characteristic of retroviruses.

Elements from the fifth superfamily, *BEL–Pao*[32] are structurally similar to *Gypsy* or *Copia* elements. They contain bona fide LTRs, encode GAG and POL proteins, and create a 4–6 bp TSD upon insertion. However, they form a distinct clade based on RT phylogenies[33]. So far, they have been detected only in metazoans.

In addition to the classical class I TEs, two new groups have recently been described: the *DIRS*-like and the *Penelope*-like (PLE) elements (FIG. 1). Members of the *DIRS*[34] order contain a tyrosine recombinase gene instead of an INT, and therefore do not form TSDs. Their termini are unusual, resembling either split direct repeats (SDR) or inverted repeats. These features indicate a mechanism of integration that is different from the LTR and LINE TEs. Nevertheless, their RT places them in class I. Members of this order have been detected in diverse species, ranging from green algae to animals and fungi[35].

The PLE order was first found in *Drosophila virilis*[36]. Its absence in other sequenced genomes reflects its patchy distribution, although it has been detected in more than 50 species, including unicellular animals, fungi and plants[37]. PLE encodes an RT that is more closely related to telomerase than to the RT from LTR retrotransposons or LINEs, and an endonuclease that is related both to intron-encoded endonuclease and to the bacterial DNA repair protein UvrC. Some members contain a functional intron. Members also have LTR-like sequences that can be in a direct or an inverse orientation.

The LINEs lack LTRs, can reach several kilobases in length and are found in all eukaryotic kingdoms. They have been separated in five major superfamilies[38]: *R2*, *L1*, *RTE*, *I* and *Jockey* (FIG. 1). Each superfamily is then subdivided into many families. Historically, the LINE superfamilies have

been subdivided into 17 'clades' and only then into families[38,39]. Although a clade taxon is not included in our system, it is widely used among LINE researchers.

Autonomous LINEs encode at least an RT and a nuclease in their *pol* ORF for transposition[38,40]; superfamily *RTE* members, with only this ORF, resemble archaic LINEs. The nuclease can be either an endonuclease (C-terminal in RT for the superfamily *R2*) or an apuric or apyrimidic endonuclease (N-terminal in the RT for the superfamilies *L1*, *RTE*, *I* and *Jockey*). A *gag*-like ORF is sometimes found 5′ to *pol*, but its role remains unclear. Only members of superfamily *I* contain RNaseH. Although LINEs generally form TSDs upon insertion, truncated 5′ ends make them difficult to find[38]. The truncations probably result from premature termination of reverse transcription[38,41]. At their 3′ end, they can display either a poly(A) tail, a tandem repeat or merely an A-rich region.

LINEs vary in prevalence and diversity in eukaryotes, but predominate over the LTR retrotransposons in many animals. The *L1* family numbers about 10^5 copies in mammals, or about 20% of the human genome. By contrast, in the malaria mosquito *Anopheles gambiae*, around 100 divergent families compose only 3% of the genome[39]. In plants, LINEs (for example, *Cin4* in maize and *Ta11* in *Arabidopsis thaliana*) seem rare compared with LTR retrotransposons, with notable exceptions (for example, *Del2* in *Lilium* spp.[42]). Most known plant LINEs are from the *L1* and *RTE* superfamilies[38,43]. However, given the lack of systematic investigation in plants, a far larger diversity is likely to exist.

The SINE order lies in class I, but is distinct in origin. Although non-autonomous (see below), they are not deletion derivatives of autonomous class I elements; instead, they originate from accidental retrotransposition of various polymerase III (Pol III) transcripts[44]. Unlike retroprocessed pseudogenes, they possess an internal Pol III promoter, allowing them to be expressed. They rely on LINEs for *trans*-acting transposition functions such as RT[44–46]. Some 'stringent' SINEs have a unique, obligatory partner[45] whereas others are generalists[46].

SINEs are small (80–500 bp) and generate TSDs (5–15 bp). The 'head', which harbours the Pol III promoter, defines SINE superfamilies and reveals their origin: tRNA, 7SL RNA and 5S RNA. SINE internal regions (50–200 bp) are family-specific and of variable origin, sometimes deriving from SINE dimerization or trimerization. The origin of the 3′ region, although it can sometimes be a LINE, is generally obscure. It can be either A- or AT-rich, harbour 3–5-bp tandem repeats, or contain a poly(T) tail, the Pol III termination signal[44]. The best known SINE is the *Alu* element, which is present in at least 500,000 copies in the human genome[47].

Class II elements

Class II TEs, like class I, are ancient and found in almost all eukaryotes (FIG. 1). Usually present in low to moderate numbers, some, such as *Pogo–Fot1* in the fungi[9] or *CACTA* in wheat and its relatives[48], have nevertheless been successful colonizers. Class II elements are also found in prokaryotes in simple forms called insertion sequences (*IS*) or as part of more complex structures[49]. Class II contains two subclasses, which are distinguished by the number of DNA strands that are cut during transposition, but neither moves via an RNA intermediate.

Subclass 1 comprises classical 'cut-and-paste' TEs of the order TIR, characterized by their terminal inverted repeats (TIRs) of variable length. The nine known superfamilies are distinguished by the TIR sequences and the TSD size (FIG. 1). These TEs can increase their numbers by transposing during chromosome replication from a position that has already been replicated to another that the replication fork has not yet passed[50]. Alternatively, they can exploit gap repair following excision to create an extra copy at the donor site[51,52]. Transposition is mediated by a transposase enzyme that recognizes the TIRs and cuts both strands at each end. Previous classifications were based on the presence of a DDE catalytic core in the transposase[51]; however, we do not use this criterion because it is currently in flux: DDE and DDE-like motifs have been found in previously non-DDE groups[53]. For the remaining non-DDE superfamilies (*P*, *piggyBac* and *CACTA*) the catalytic domains have yet to be well established. They might contain catalytic aspartate or glutamate residues, but it is difficult to identify these residues in the absence of structure-based alignments.

The *Tc1–Mariner* superfamily, which is ubiquitous in eukaryotes, possesses a simple structure of two TIRs and a transposase ORF[54]. Its numerous families all strongly prefer to insert adjacent to TA, generating TA TSDs. Members of the *hAT* superfamily have TSDs of 8 bp, relatively short TIRs of 5–27 bp[55] and overall lengths of less than 4 kb. The name derives from three well-described TE families: *hobo* from *Drosophila*[56], *Ac-Ds* from maize[57] and *Tam3* from snapdragons[58].

The diverse *Mutator* superfamily also occurs in all eukaryotic kingdoms[59]. Although its TIRs can extend to several hundred base pairs, they are sometimes either very short or undetectable. *Mutator* TEs produce 9-11 bp TSDs. The TIRs of the superfamily *Merlin* likewise range from a few dozen to several hundred base pairs; these TEs are flanked by 8-9 bp TSDs. Although fully functional *Merlin* elements encoding a DDE transposase are larger than 10 kb, deletion derivatives may be only few hundred base pairs. Members of this superfamily have been described only in animals and eubacteria[60].

The transposase of the superfamily *Transib* contains the DDE motif as described above; moreover, it is related to the RAG1 protein involved in V(D)J recombination[61]. *Transib* TEs were found so far in *D. melanogaster* and mosquitoes[62]. The *P* superfamily, initially found in insect genomes, is now known to be present also in metazoans[63] and the alga *Chlamydomonas reinhardtii*[64]. *P* elements generate 8-bp TSDs. The superfamily *piggyBac*, which is primarily[65] but not exclusively[66] found in animals, favours insertion adjacent to TTAA.

The superfamily *PIF–Harbinger* likewise displays a target-site preference, but for TAA[67]. These TEs contain two ORFs, one encoding a DNA binding protein, the other encoding a DDE transposase. The TEs of the superfamily *CACTA* carry both a transposase and a second ORF of unclear function[48]. In plants the short TIRs terminate in highly conserved CACTA (sometimes CACTG) motifs and flank 3-bp TSDs, whereas in animals and fungi CCC replaces the CACTA motif and 2-bp TSDs are generated[68]. The TIRs often flank complex arrays of subterminal repeats.

We include the poorly known *Crypton*[69] TEs, which have so far been found only in fungi, as a second order in subclass 1. They contain a tyrosine recombinase, as do some phages, *IS* and *DIRS*-like retrotransposons, but lack an RT domain, which suggests that they transpose via a DNA intermediate. Their transposition, which is demonstrated by the appearance of empty sites, might involve recombination between a circular molecule and the DNA target[35,69], requiring cleavage of both DNA strands (hence their inclusion in subclass 1). Accordingly, they lack TIRs, but seem to generate TSDs as a result of recombination and integration.

Subclass 2 holds DNA TEs that undergo a transposition process that entails replication without double-stranded cleavage, sharply different from that of subclass 1.

These copy-and-paste TEs transpose by replication involving the displacement of only one strand. Their placement within class II reflects the common lack of an RNA intermediate, but not necessarily common ancestry.

Elements in the order *Helitron* appear to replicate via a rolling-circle mechanism, with only one strand cut[70], and do not generate TSDs. *Helitron* ends are defined only by TC or CTRR motifs (where R is a purine) and a short hairpin structure lying a few nucleotides before the 3′ end, although this does not seem to be a true diagnostic feature. Autonomous *Helitrons* encode a Y2-type tyrosine recombinase such as that found in the bacterial *IS91* rolling-circle transposons, with a helicase domain and replication initiator activity. They can also encode a single-strand binding protein or other proteins. *Helitrons* have been best characterized in maize, in which most are non-autonomous derivatives[70]. Curiously, many maize *Helitrons* carry gene fragments that have been captured from the host genome[13]. Although *Helitrons* have been described mainly in plants, they also exist in animals and fungi[71,72], constituting 2% of the *Caenorhabditis elegans* genome[70] and at least 3% of the genome of the bat species *Myotis lucifugus*[73].

TEs of the order *Maverick* (also known as *Polintons*) are large, reaching 10–20 kb[74], and are bordered by long TIRs. They encode up to 11 proteins, but these vary in number and order. Some show limited homology to proteins of various DNA viruses. *Mavericks* encode DNA polymerase B and an INT (c-int type) that is related to those found in some class I TEs, but they do not contain RT, suggesting that they undergo replicative transposition without RNA intermediates. This is proposed to proceed via excision of a single strand followed by extrachromosomal replication, then integration to a new site[75]. So far, *Mavericks* have been found sporadically in diverse eukaryotes, but not in plants[76].

Definition of family and subfamily

The precise definition of a family is problematic because groups of TEs with similar features sometimes form a continuum of sequence homology; elements from one end of the spectrum have little DNA sequence identity with those at the other end. Nevertheless, it seems that evolutionary lineages are sufficiently distinct to allow the borders of such a continuum to distinguish a family. Furthermore, the differential success in the replication of some groups in various genomes has led to relatively large numbers of highly similar copies, simplifying the task of identifying them as members of a family.

With the aim of creating a classification that is not only relevant to TE specialists but also to genome annotators, we define family as a group of TEs that have high DNA sequence similarity in their coding region (if present) or internal domain, or in their terminal repeat region. For practical reasons, we specify strong sequence similarity as 80% or more in at least 80% of the aligned sequence, because this level of homology produces strong BLASTN hits at default settings (as are commonly used to identify and classify TEs; see below). Thus, two elements belong to the same family if they share 80% (or more) sequence identity in at least 80% of their coding or internal domain, or within their terminal repeat regions, or in both (FIG. 2).

To avoid misclassification of short and possibly random stretches of homology, we recommend analysing only segments of longer than 80 bp. TEs that are smaller than 80 bp require specialized analyses, and are beyond the scope of the basic classification and nomenclature proposed here. For abundant elements, consensus sequences should be constructed, deposited in public databases and used to test family membership. In genomes for which only limited sequence data are available, all individual TE sequences must serve for comparison.

The terminal repeat regions and other non-coding regions are the fastest evolving parts of TEs. Therefore, they offer the most specificity in defining families. Allowing the 80–80–80 rule for DNA sequence identity in either the internal domain or in the terminal regions, or both, also addresses the problem caused by frequent TE truncations. In some cases only terminal repeats and non-coding regions may be present, whereas in other cases only parts of the coding region but no terminal repeats may be available for analysis (FIG. 2a).

In some cases, it may be necessary to add the subfamily taxon, depending on the population structure of a TE family. Subfamilies can be populations of non-autonomous deletion derivatives (see below) or distinct subpopulations in large families that can be clearly segregated. The similarity threshold can differ between subfamilies, depending on the number and homogeneity of elements described. However, such distinctions are matters for TE specialists and should not be a burden for annotators (see the WikiPoson web site for further discussion). Importantly, the term should not be used for groupings above the family level.

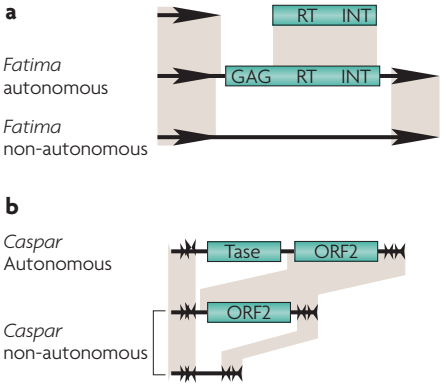

Figure 2 | **Examples of transposable elements (TEs) that are classified as members of one family on the basis of their sequence homology.** Green areas indicate regions of 80% or more DNA sequence identity. **a** | A full-length autonomous *Fatima* element is used to define fragments in long terminal repeat (LTR) and coding regions (top) or a non-autonomous derivative (bottom). **b** | *CACTA* elements of the *Caspar* family are found in various autonomous (top) and non-autonomous deletion derivatives (middle and bottom). GAG, a structural protein for virus-like particles; INT, integrase; ORF, open reading frame; RT, reverse transcriptase; Tase, transposase.

Autonomous and non-autonomous TEs

Our definition of family also addresses another problem: many autonomous TEs give rise to non-autonomous deletion derivatives (FIG. 2). Here we define an element as autonomous simply if it appears to encode all the domains that are typically necessary for its transposition, without implying that the element is either functional or active. A family can contain various elements that have been rendered defective from point mutations or small insertions or deletions (indels), but that nevertheless retain not only all the coding regions of an active TE, but also sufficiently high DNA identity for family membership. Such elements that are defective owing to their limited mutations were probably functional in the recent past[18].

Autonomous but defective elements stand in contrast to non-autonomous TEs, which we define as any group of elements that lacks some (or all) of the domains found in autonomous elements[18]. Usually, non-autonomous elements have a highly degenerate coding region, or even completely lack coding capacity. Occasionally, non-autonomous TEs lack some genes but still contain others. For example, members of the *Caspar* family[48] (superfamily *CACTA*) often lack the transposase gene but still contain the second ORF (FIG. 2b), whereas the *BARE2* elements in the Triticeae have a conserved deletion that inactivates *gag*[77]. Nevertheless, non-autonomous

and autonomous elements usually still share strong sequence conservation and specific characteristics within their termini and in the 5′ UTR (LTR retrotransposons), because these are required for transposition.

A group of non-autonomous elements may be abundant and contain sufficient sequence homology between them to qualify as a subfamily. Autonomous and non-autonomous partners[18] of the same family are therefore classified as different subfamilies in the proposed system. However, some non-autonomous elements might be cross-activated by autonomous partners from different families. This is suspected for many MITEs, as few show sequence similarity to known *PIF–Harbinger* or *Tc1–Mariner* families[78]. SINEs, an extreme case, are non-autonomous but are not derived from TEs. Therefore, cross-activation must be addressed on a case-by-case basis.

The large retrotransposon derivatives (LARDs), terminal repeat retrotransposons in miniature (TRIMs), MITEs and SINEs (discussed above) are groups of TEs that are clearly non-autonomous (TABLE 1). LARDs[79] and TRIMs[80] describe large (>4 kb) and small (<4 kb) non-autonomous LTR retrotransposon derivatives, respectively, that harbour an internal domain with no coding capacity or any sequence that is reminiscent of the *gag*, *rt* or *int* genes of autonomous elements. An autonomous counterpart for most of them is currently unknown, preventing their precise classification. For example, until their autonomous partner is discovered, *Sukkula* elements[79] will be classified as class I, order LTR, superfamily unknown, family unknown, subfamily *Sukkula*. However, *Dasheng*[81] from rice — non-autonomous elements of the *RIRE2* family — can be classified as a subfamily of *RIRE2*.

MITEs[82] are a heterogeneous group of small non-autonomous elements of a few dozen to a few hundred base pairs in size, which are flanked by TIRs and are frequently found in or close to genes. In several species, a few autonomous *Tc1–Mariner* TEs are responsible for the origin and activation of large populations of non-autonomous elements, such as of tens of thousands of *Stowaway* MITEs in rice[83] or more than a thousand non-autonomous *Galluhop* elements in chicken[84]. Likewise, *PIF–Harbinger* TEs control the activation of non-autonomous *Tourist* MITEs in plants, nematodes, insects and fish[78].

Designations such as LARD, TRIM and MITE have no descriptive power in a taxonomic sense, because we cannot attribute shared structural features to a common origin at a particular taxonomic level. Therefore, they are not used in our proposed system. However, they do have practical value in describing a 'way of life' that is common to many elements from different families. Furthermore, the similarities between members of these groups might reflect shared adaptations or functions. Therefore, we do not discourage their continued use. A conceptually similar term is that of 'grain', which refers to morphology, structure or use, rather than comprising a taxon (for example, *Poaceae*). The term is nevertheless useful, and used, to describe cases such as 'grain legumes' that are not taxonomically close to grains of the grasses.

Naming system and name format

Currently, there are no clear rules for choosing names for new families — the discoverer of a new TE family should have the privilege of naming it — but we propose a few guidelines. A name can include letters (for example, *Angela*) and numbers (for example, *TOS17*), but should not contain hyphens or underscores so as not to interfere with the name format described below. Family names should be no longer than five or six syllables and should be easily pronounceable, at least in English (facilitating use in symposia). If an element is isolated from a different species than that in which it was first described, but nevertheless fulfils the criteria of the original family, the established name is used. This practice will help clarify TE and genome evolution across taxa. Therefore, reference to species in which the TE was initially discovered should be avoided in the family name. This recommendation represents a change from past practice (for example, *BARE1* for 'barley retrotransposon').

Table 1 | **Structural descriptions of non-autonomous transposable elements**

Abbreviation	Description	Refs
LARD	Large retrotransposon derivative	71
MITE	Miniature inverted-repeat transposable element	7
SNAC	Small non-autonomous *CACTA* transposon	46
TRIM	Terminal repeat retrotransposon in miniature	72

Genome projects, particularly in plants, require an automated approach to identify each of the potentially hundreds of thousands of TEs in any genome. To facilitate a fast and accurate assessment of the main classification of a particular insertion, we propose a name format that contains: a three-letter code with each letter respectively denoting class, order and superfamily; the family (or subfamily) name; the ID (database accession) of the sequence on which the element was found; and the running number of that element on the sequence. Ignoring subclass in the code gives brevity without a loss of specificity. Thus, the first element of the *Angela* family (class I, order LTR, superfamily *Copia*) to be described on sequence AA123456 will be annotated as RLC_*Angela*_*AA123456-1*. Similarly, the third identified copy of *Caspar* (class II, subclass 1, order TIR, superfamily *CACTA*) in the same sequence will be designated as DTC_*Caspar*_*AA123456-3*. The running numbers need not be in linear order (that is, parallel to the nucleotide numbering in a database accession) but can reflect the annotation order. Thus, if a deposited sequence or its annotation is updated (for example, after a new BAC is added to a contiguous set), the previously identified TEs will conserve their numbering and the new ones will be assigned subsequent consecutive numbers. This convention is already used for gene annotation in many species[85].

In genomes that can be described as highly advanced drafts (pseudomolecules representing entire chromosomes), the chromosome number could replace the BAC address or accession number. For example, the 314th *RIRE2* element (*Gypsy* LTR retrotransposon) on rice (*O. sativa* ssp. *japonica*) chromosome 5 would be named RLG_*RIRE2*_*Osjn5-314*. Traceability to the original BAC or database address could be maintained as part of the genome database. The three-letter codes are summarized in FIG. 1. If classification at any level is uncertain, an X must be used. Four examples of classifications including structural descriptions are given in TABLE 2.

Protocol for TE classification

In this hierarchical system, a TE can be classified in a few steps (FIG. 3). The first step involves a BLASTN (DNA versus DNA) search of the element against a TE database (such as TREP). If this search produces strong hits to a known family of elements (that is, meets the 80–80–80 rule), the new element is classified as belonging to that family. If no strong DNA similarity

Table 2 | **Examples of the application of the proposed classification system**

	Barbara	*Sukkula*	*Thalos*	*Xithos*
Class	Retrotransposon	Retrotransposon	DNA transposons	Unknown
Subclass	N/A	N/A	1	Unknown
Order	LTR retrotransposon	LTR retrotransposon	TIR	Unknown
Superfamily	*Copia*	Unknown	*Mariner*	Unknown
Family	*Barbara*	Unknown	*Stowaway*	*Xithos*
Subfamily	Not defined	*Sukkula*	*Thalos*	Not defined
Insertion	RLC_*Barbara*_AF1234-2	RLx_*Sukkula*_AF1234-1	DTT_*Thalos*_AF1234-8	xxx_Xithos_AF1234-1
Structural description	Autonomous retrotransposon	LARD	MITE	-

'Unknown' can be used at any taxonomic level other than subfamily. Unless detailed analyses are made, the subfamily level is not applied. Structural descriptions are independent of a taxonomic classification system. LARD, large retrotransposon derivative; LTR, long terminal repeat; TIR, terminal inverted repeat ; MITE, miniature inverted-repeat transposable element; N/A, not applicable.

can be detected, the TE might represent a new family, and should be given a new name only if the sequence has sufficient size to avoid ambiguity, for example, 1.5 kb for classical LTR retrotransposons (a Perl script for naming new TEs can be found at the web site for the Element Namer[86]).

In the second step, a BLASTX (translated DNA versus protein) search against a database containing predicted protein sequences should indicate to which class, order and superfamily the TE belongs. For example, a novel family of *Gypsy* retrotransposons will produce strong sequence alignments with virtually all other *Gypsy* families at the protein level. If the protein similarity search still yields no results, either the new TE represents a new (unknown) superfamily or class, or, alternatively, it might be a non-autonomous deletion derivative or recombinant with no coding capacity. The threshold for assigning new superfamily or class status must necessarily be high, and requires, at least, reference to a peer-reviewed, accepted publication that supports this decision.

In the third step, the element must be examined for the presence of terminal direct or inverted repeats, primer binding site (PBS) and polypurine tract (PPT) motifs, and TSDs (see below). This analysis may allow classification into one of the known groups. If all tests fail to identify known groups, or if in doubt, the family of the element should be assigned as unknown.

属于同一超家族的不同家族之间在蛋白质水平上的相似性通常很高。

Classifying non-autonomous elements

Several features can be used to classify elements that lack clear internal similarity with known families. For LTR retrotransposons, they include the PBS, packaging signal (PSI), dimerization signal (DIS), PPT and INT signal. For class II elements, the TIR motifs are generally informative. These motifs and signals should be used for classification only if the analysed element does not have a high level of similarity to any previously identified element.

Non-autonomous class I elements. In all LTR retrotransposons, the PBS is a 20–25-nucleotide sequence (FIG. 4a) that is complementary to a structural RNA (most commonly a ^{met}tRNA) and located adjacent to the 3′ end (or 2–3 bp downstream) of the 5′ LTR. It serves as the priming site for synthesis of minus-strand cDNA by RT. The second, plus strand is primed by the PPT. Composed mainly of purines, the PPT is 20–25 nucleotides long and found just upstream of the 5′ end of the 3′ LTR. The PBS and PPT are conserved primarily within families, but can also be highly similar among closely related families[87]. The integrase signal, a ~25-nucleotide domain located at the 3′ end of the LTR (FIG. 4a), is specifically recognized by INT and required for the correct genomic integration of the new cDNA copy[88]. The signal is conserved within a family, but the level of conservation decreases rapidly between representatives of higher taxonomic levels, making it valuable for classifying a particular element into a specific family.

The PSI motif (studied mainly in retroviral RNAs) forms an RNA secondary structure (FIG. 4b), allows a specific RNA recognition by the GAG protein of a particular family, and is located in the same region as

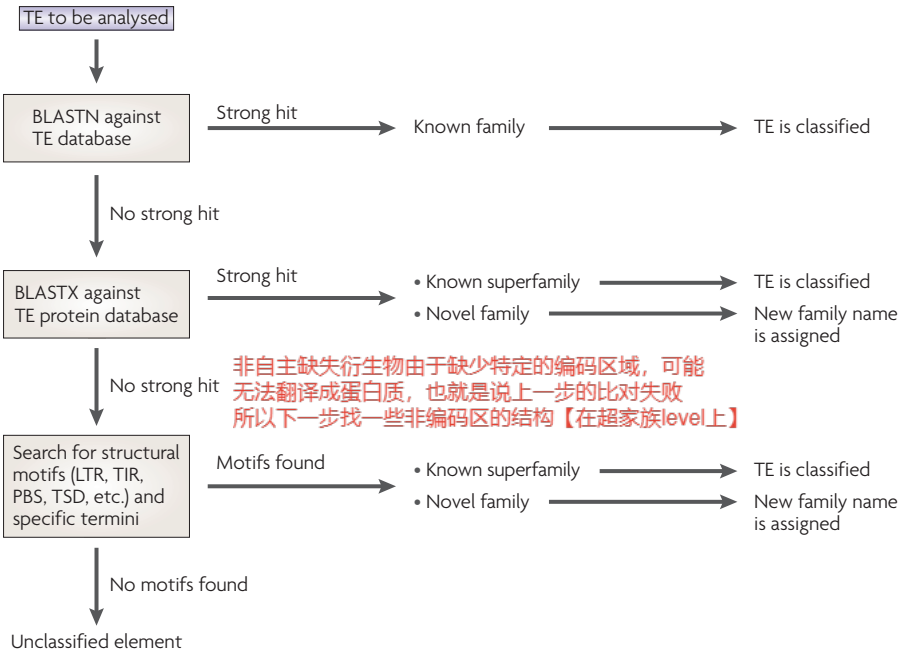

Figure 3 | **Step by step transposable element (TE) classification.** LTR, long terminal repeat; PBS, primer binding site; TIR, terminal inverted repeat; TSD, target site duplication.

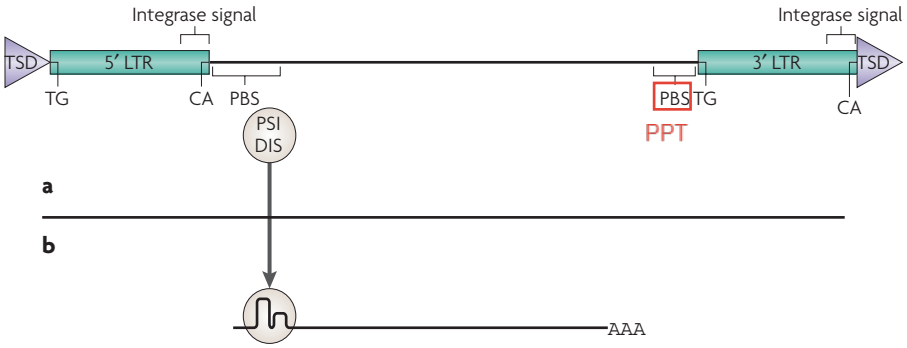

Figure 4 | **Motifs and signals that are present in long terminal repeat (LTR) retrotransposons.** **a** | The positions of those signals (see text for further information) are indicated on the DNA sequence of the element. The primer binding site (PBS), polypurine tract (PPT), integrase signal and terminal inverted repeat (TIR) can be identified with standard sequence alignment programs. **b** | The packaging signal (PSI) and dimerization signal (DIS) motifs form secondary RNA structures that can be identified with secondary-structure prediction software, and thus may be analysed on the fold of the predicted RNA transcript of the element. TSD, target site duplication.

the PBS. In LTR retrotransposons, the PSI sequence has been clearly identified only in yeast *Ty1* (REF. 89). However, conserved and specific secondary structures in the 5′ UTR indicate that the PSI probably functions also in plants (F.S. & A.S., unpublished observations).

The DIS motif allows specific dimerization of retroviral or retrotransposon RNA by the formation of particular RNA secondary structures[90]. This RNA motif lies in the same region as PBS and PSI in the 5′ UTR of LTR retrotransposons (FIG. 4b). The secondary structure can be predicted with software such as Mfold[91]; a high level of conservation in the secondary structure within a subfamily or family indicates likely functionality (F.S. & A.S., unpublished observations).

For the LINEs, few markers can be used to classify non-autonomous elements. A link between the family and the type of 3′ tail (poly(A) signal, tandem repeat or A-rich) has been proposed but is so far unconfirmed[39]. Moreover, the length of LINE TSDs is not regular or conserved.

Non-autonomous class II elements. TIR motifs and the size of the TSD are the most reliable characteristics for classifying class II elements of subclass 1 that lack internal sequence similarities. For example, plant *Mariner* elements of the *Stowaway* family are flanked by TIRs with the almost invariant 5′-CTCCTCCC…GGGAGGAG-3′ motif and, like all *Tc1–Mariner* elements, always produce a TA TSD upon insertion. However, classification can be complicated by the diversity within some superfamilies. Elements from the superfamily *CACTA* terminate in the well conserved 5′-CACTA…TAGTG-3′ motif before a 3-bp TSD in plants, and in the 5′-CCC…GGG-3′ motif with a 2-bp TSD in animals. *Mutator* TIRs are usually long but are also highly divergent — sharing only terminal G…C nucleotides — or are absent. In this extreme case, the length of the TSD (usually 9 bp) remains probably the most useful criterion.

The most difficult to identify TEs in class II are *hAT* elements, which usually have very short TIRs that lack diagnostic sequence motifs. This superfamily is best identified by its 8-bp TSD, although other elements like *P* or *Merlin* also duplicate 8 nonspecific base pairs. Thus, for some non-autonomous elements, it might not be possible to clearly assign an element to a superfamily.

Identifying derivatives of subclass 2 is more complex because they are poorly characterized. *Mavericks* are large and have long TIRS, which can a problem for recognition. The *Helitrons* also are highly heterogeneous, and the short termini and possibly the hairpin are the only structural features that can be applied in identification at present.

Classifying compound or hybrid elements

Complex or hybrid TEs are commonly seen in genomic sequences, and may cause confusion in annotation. Some arise from nested TE integration or by intrachromosomal recombination[23,92]; others result from variant replication[77,93]. Old insertions have often suffered mutations and subsequent insertions and rearrangements that both reduce their similarity to known families and interrupt their continuity. These must be annotated segmentally, using the rules of assignment described above over a minimum stretch of 80 nucleotides. New but highly disrupted elements are classified as unknown. Functional hybrid TEs (for example, *BARE*2 of barley[77]) that have emerged as a repetitious group with shared, conserved features can be assigned a name on a taxonomic level that is commensurate with their distinctiveness.

Concluding remarks

The proposed, deliberately simple system is intended as a general guideline for researchers who work on eukaryotic genomes, in particular for those faced with the task of annotation. TEs are extremely diverse because of their multiple means of replication, mutability and abundance. This diversity, combined with an earlier lack of consensus in the naming of TEs, presents a confusing landscape to newcomers. The complexity was not problematic when most of those analysing TE sequences were doing so because they were specifically interested in TEs themselves. The ever-increasing amount of long genomic sequences has changed that situation. The aim of this system of TE nomenclature and classification is to both simplify annotation and clarify future detailed structural, functional and evolutionary analyses.

With these goals in mind, we have focused on TEs that have been described in eukaryotes. Bacterial elements comprise primarily *IS* TEs, all belonging to class II, divided into several families[49]. Most encode a DDE transposase, whereas others possess a serine recombinase (S), or tyrosine recombinase (Y and Y2). Bacterial *IS* transposition mechanisms are diverse and can be either conservative (*IS50*) or replicative, the latter by either cointegrate (*Tn3*) or rolling-circle mechanisms (*IS91*)[15]. Although mainly focused on eukaryotes, our classification system can be extended to include prokaryotic families. For example, *IS91* would be grouped with DHH. Moreover, several *IS* families are related to eukaryotic counterparts: *IS256* and *Mutator*; *IS630* and *Tc1–Mariner*; *IS5* and *PIF–Harbinger*.

A unified TE classification system would, additionally, be strengthened by two developments. First, a well curated and taxonomically organized TE database that spans all kingdoms is needed. This would allow the proposed classification system to be consistently applied and could encourage comparative, cross-species analyses. Such an effort will eventually improve our understanding of this abundant and ever-present but poorly characterized fraction of the eukaryotic

genome. Second, a flexible forum devoted to TEs and their behaviour, origin and impact will allow the classification to be kept up to date. With this in mind, we have established WikiPoson as a resource and forum for non-specialists and specialists alike.

The development of TE nomenclature or taxonomy should go hand-in-hand with functional and evolutionary analyses. This is implicit in basing TE classification on the means of replication and transposition at the higher taxonomic levels, before moving to gene order and sequence similarity on the lower levels. It conceptually parallels organismal nomenclature because TEs are themselves conceptually parallel: self-replicating systems have derived from ancestral forms through descent with modification[51]. Likewise, the use of sequence similarity *per se* is open to the risk of confusing similarity or convergence with homology through common origin. A natural taxonomic system for TEs therefore requires their continuing evolutionary and functional analyses. At present, even if such analyses can be made within superfamilies, they remain difficult for the upper levels of the classification (classes and subclasses). In other words, the question of a common origin of all classes, subclasses and superfamilies remains open.

Thomas Wicker is at the Institute of Plant Biology, University Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland.*

François Sabot and Alan H. Schulman are at MTT/BI Plant Genomics Laboratory, Institute of Biotechnology, Viikki Biocenter, University of Helsinki, P.O. BOX 56, FIN-00014 Helsinki, Finland.*

Aurélie Hua-Van and Pierre Capy are at Laboratoire Evolution, Génomes et Spéciation, UPR 9034, CNRS 91198 Gif-sur Yvette Cedex, France and Université Paris-Sud 11, 91405 Orsay Cedex, France.*

Jeffrey L. Bennetzen is at the Department of Genetics, University of Georgia, Athens, Georgia, 30602-7223, USA.

Boulos Chalhoub is at Unité de Recherche en Génomique Végétale (URGV/URGI), Organization and Evolution of Plant Genomes, 2 rue Gaston Crémieux CP 5708, FR-91057 Evry Cedex, France.

Andrew Flavell is at the Plant Research Unit, University of Dundee at SCRI, DD2 5DA Invergowrie, Dundee, United Kingdom.

Philippe Leroy amd Etienne Paux are at INRA-Université Blaise-Pascal, UMR 1095, 234 avenue du Brézet, FR-63100 Clermont-Ferrand Cedex, France.

Michele Morgante is at Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, Via delle Scienze 208, I-33100 Udine, Italy.

Olivier Panaud is at Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS-IRD-Université de Perpignan, 52 Avenue Paul Alduy, F-66860 Perpignan, France.

Phillip SanMiguel is at Purdue Genomics Core Facility, 170 South University Street, West Lafayette, IN, 47907-2072, USA.

Alan H. Schulman is also at Plant Genomics, Food and Biotechnology, MTT Agrifood Research Finland, Myllytie 10, FIN-31600 Jokioinen, Finland.

**These authors contributed equally to this work.*
Correspondence to A.H.S.
e-mail: alan.schulman@helsinki.fi

doi:10.1038/nrg2165
Published online 6 November 2007

1. Flavell, R. B., Rimpau, J. & Smith, D. B. Repeated sequence DNA relationships in four cereal genomes. *Chromosoma* **63**, 205–222 (1977).
2. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
3. Adams, M. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Morgante, M. Plant genome organisation and diversity: the year of the junk! *Curr. Opin. Biotechnol.* **17**, 168–173 (2005).
6. Bennetzen, J. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* **15**, 621–627 (2005).
7. Feschotte, C., Jiang, N. & Wessler, S. Plant transposable elements: where genetics meets genomics. *Nature Rev. Genet.* **3**, 329–341 (2002).
8. SanMiguel, P. & Bennetzen, J. L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* **82**, 37–44 (1998).
9. Daboussi, M. & Capy, P. Transposable elements in filamentous fungi. *Annu. Rev. Microbiol.* **57**, 275–299 (2003).
10. Hua-Van, A., Le Rouzic, A., Maisonhaute, C. & Capy, P. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet. Genome Res.* **110**, 426–440 (2005).
11. Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**, 103–107 (1989).
12. Duval-Valentin, G., Marty-Cointin, B. & Chandler, M. Requirement of *IS911* replication before integration defines a new bacterial transposition pathway. *EMBO J.* **23**, 3897–3906 (2004).
13. Morgante, M. *et al.* Gene duplication and exon shuffling by *Helitron*-like transposons generate intraspecies diversity in maize. *Nature Genet.* **37**, 997–1002 (2005).
14. Lai, J., Li, Y., Messing, J. & Dooner, H. K. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc. Natl Acad. Sci. USA* **102**, 9068–9073 (2005).
15. Curcio, M. & Derbyshire, K. The outs and ins of transposition: from mu to kangaroo. *Nature Rev. Mol. Cell Biol.* **4**, 865–877 (2003).
16. Kumar, A. & Bennetzen, J. Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532 (1999).
17. Han, J. S. & Boeke, J. D. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *BioEssays* **27**, 775–784 (2005).
18. Sabot, F. & Schulman, A. H. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* **97**, 381–388 (2006).
19. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
20. Neumann, P., Pozarkova, D. & Macas, J. Highly abundant pea LTR retrotransposon *Ogre* is constitutively transcribed and partially spliced. *Plant Mol. Biol.* **53**, 399–410 (2003).
21. Wicker, T. *et al.* Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum*) reveals multiple mechanisms of genome evolution. *Plant J.* **26**, 307–316 (2001).
22. Vicient, C. M., Kalendar, R., Anamthawat-Jonsson, K. & Schulman, A. H. Structure, functionality, and evolution of the *BARE-1* retrotransposon of barley. *Genetica* **107**, 53–63 (1999).
23. SanMiguel, P., Gaut, B. S., Tikhoniv, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons in maize. *Nature Genet.* **20**, 43–45 (1998).
24. Peterson, D. *et al.* Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12**, 795–807 (2002).
25. Eickbush, T. & Furano, A. Fruit flies and humans respond differently to retrotransposons. *Curr. Opin. Genet. Dev.* **12**, 669–674 (2002).
26. Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Ann. Rev. Biochem.* **67**, 1–25 (1998).
27. Seelamgari, A. *et al.* Role of viral regulatory and accessory proteins in HIV-1 replication. *Front. Biosci.* **9**, 2388–2413 (2004).
28. Bucheton, A. The relationship between the *flamenco* gene and *gypsy* in *Drosophila*: how to tame a retrovirus. *Trends Genet.* **11**, 349–353 (1995).
29. International Committee on Taxonomy of Viruses. The Universal Virus Database. [online], < http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm > (2007).
30. Capy, P. Classification and nomenclature of retrotransposable elements. *Cytogenet. Genome Res.* **110**, 457–461 (2005).
31. Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149–173 (2006).
32. Xiong, Y., Burke, W. & Eickbush, T. *Pao*, a highly divergent retrotransposable element from *Bombyx mori* containing long terminal repeats with tandem copies of the putative *R* region. *Nucleic Acids Res.* **21**, 2117–2123 (1993).
33. Cook, J., Martin, J., Lewin, A., Sinden, R. & Tristem, M. Systematic screening of *Anopheles* mosquito genomes yields evidence for a major clade of *Pao*-like retrotransposons. *Insect Mol. Biol.* **9**, 109–117 (2000).
34. Cappello, J., Handelsman, K. & Lodish, H. Sequence of *Dictyostelium DIRS-1*: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell* **43**, 105–115 (1985).
35. Goodwin, T. & Poulter, R. A new group of tyrosine recombinase-encoding retrotransposons. *Mol. Biol. Evol.* **21**, 746–759 (2004).
36. Evgen'ev, M. *et al.* Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. *Proc. Natl Acad. Sci. USA* **94**, 196–201 (1997).
37. Evgen'ev, M. & Arkhipova, I. Penelope-like elements — a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet. Genome Res.* **110**, 510–521 (2005).
38. Eickbush, T. H. & Malik, H. S. in *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1111–1146 (ASM, Herndon, 2002).
39. Biedler, J. & Tu, Z. Non-LTR retrotransposons in the african malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. *Mol. Biol. Evol.* **20**, 1811–1825 (2003).
40. Ostertag, E. M. & Kazazian, H. H. Genetics: LINEs in mind. *Nature* **435**, 890–891 (2005).
41. Petrov, D. A. & Hartl, D. L. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol. Biol. Evol.* **15**, 293–302 (1998).
42. Leeton, P. R. & Smyth, D. R. An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Mol. Gen. Genet.* **237**, 97–104 (1993).
43. Zupunski, V., Gubensek, F. & Kordis, D. Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. *Mol. Biol. Evol.* **18**, 1849–1863 (2001).
44. Kramerov, D. & Vassetzky, N. Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.* **247**, 165–221 (2005).
45. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked *Alu* sequences. *Nature Genet.* **35**, 41–48 (2003).
46. Kajikawa, M. & Okada, N. LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell* **111**, 433–444 (2002).
47. Rowold, D. J. & Herrara, R. J. *Alu* elements and the human genome. *Genetica* **108**, 57–72 (2000).
48. Wicker, T., Guyot, R., Yahiaoui, N. & Keller, B. CACTA transposons in *Triticeae*. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132**, 52–63 (2003).
49. Chandler, M. & Mahillon, J. in *Mobile DNA II* (eds Craig, N., Craigie, R., Gellert, M. & Lambowitz, A.) (ASM, Washington D.C., 2002).

50. Greenblatt, I. M. & Brink, R. A. Twin mutations in medium variegated pericarp maize. *Genetics* **47**, 489–501 (1962).
51. Capy, P., Bazin, C., Higuet, D. & Langin, T. (eds) *Dynamics and evolution of transposable elements* (Library of Congress, Austin, 1998).
52. Nassif, N., Penney, J., Pal, S., Engels, W. & Gloor, G. Efficient copying of nonhomologous sequences from ectopic sites via *P*-element-induced gap repair. *Mol. Cell. Biol.* **14**, 1613–1625 (1994).
53. Hickman, A. *et al.* Molecular architecture of a eukaryotic DNA transposase. *Nature Struct. Biol.* **12**, 715–721 (2005).
54. Shao, H. & Tu, Z. Expanding the diversity of the *IS630–Tc1–mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics* **159**, 1103–1115 (2001).
55. Kempken, F. & Windhofer, F. The *hAT* family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma* **110**, 1–9 (2001).
56. Calvi, B. R., Hong, T. J., Findley, S. D. & Gelbart, W. M. Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: *hobo*, *Activator*, and *Tam3*. *Cell* **66**, 465–471 (1991).
57. Courage, U. *et al.* Transposable elements *Ac* and *Ds* at the *shrunken*, *waxy*, and alcohol dehydrogenase 1 loci in *Zea mays* L. *Cold Spring Harb. Symp. Quant. Biol.* **49**, 329–338 (1984).
58. Hehl, R., Nacken, W. K., Krause, A., Saedler, H. & Sommer, H. Structural analysis of *Tam3*, a transposable element from *Antirrhinum majus*, reveals homologies to the *Ac* element from maize. *Plant Mol. Biol.* **6**, 369–371 (1991).
59. Pritham, E. J., Feschotte, C. & Wessler, S. R. Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. *Mol. Biol. Evol.* **22**, 1751–1763 (2005).
60. Feschotte, C. *Merlin*, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial *IS1016* insertion sequences. *Mol. Biol. Evol.* **21**, 1769–1780 (2004).
61. Kapitonov, V. V. & Jurka, J. *RAG1* core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biol.* **3**, e181 (2005).
62. Kapitonov, V. V. & J., J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).
63. Hammer, S. E., Strehl, S. & Hagemann, S. Homologs of *Drosophila P* transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. *Mol. Biol. Evol.* **22**, 833–844 (2005).
64. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
65. Sarkar, A. *et al.* Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related 'domesticated' sequences. *Mol. Genet. Genomics* **270**, 173–180 (2003).
66. Jiang, R. *et al.* Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. *Mol. Genet. Genomics* **273**, 20–32 (2005).
67. Jurka, J. & Kapitonov, V. V. *PIFs* meet *Tourists* and *Harbingers*: a superfamily reunion. *Proc. Natl Acad. Sci. USA* **98**, 12315–12316 (2001).
68. DeMarco, R., Venancio, T. & Verjovski-Almeida, S. *SmTRC1*, a novel *Schistosoma mansoni* DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily *BMC Evol. Biol.* **6**, 89 (2006).
69. Goodwin, T., Butler, M. I., Poulter, R. T. *Cryptons*: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology* **149**, 3099–3109 (2003).
70. Kapitonov, V. & Jurka, J. Rolling-circle transposons in eukaryotes. *Proc. Natl Acad. Sci. USA* **98**, 8714–8719 (2001).
71. Poulter, R. & Goodwin, T. *DIRS-1* and the other tyrosine recombinase retrotransposons. *Cytogenet. Genome Res.* **110**, 575–588 (2005).
72. Hood, M. Repetitive DNA in the automictic fungus *Microbotryum violaceum*. *Genetica* **124**, 1–10 (2005).
73. Pritham, E. & Feschotte, C. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc. Natl Acad. Sci. USA* **104**, 1895–1900 (2007).
74. Feschotte, C. & Pritham, E. J. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet.* **21**, 551–552 (2005).
75. Kapitonov, V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl Acad. Sci. USA* **103**, 4540–4545 (2006).
76. Pritham, E., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007).
77. Tanskanen, J. A., Sabot, F., Vicient, C. & Schulman, A. H. Life without GAG: The *BARE-2* retrotransposon as a parasite's parasite. *Gene* **390**, 166–174 (2006).
78. Jiang, N., Feschotte, C., Zhang, X. & Wessler, S. R. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* **7**, 115–119 (2004).
79. Kalendar, R. *et al. LARD* retroelements: novel, non-autonomous components of barley and related genomes. *Genetics* **166**, 1437–1450 (2004).
80. Witte, C. P., Le, Q. H., Bureau, T. & Kumar, A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl Acad. Sci. USA* **98**, 13778–13783 (2001).
81. Jiang, N., Jordan, I. K. & Wessler, S. R. *Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiol.* **130**, 1697–1705 (2002).
82. Bureau, T. E. & Wessler, S. R. *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**, 907–916 (1994).
83. Feschotte, C., Swamy, L. & Wessler, S. R. Genome-wide analysis of *Mariner*-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163**, 747–758 (2003).
84. Wicker, T. *et al.* The repetitive landscape of the chicken genome. *Genome Res.* **15**, 126–136 (2005).
85. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
86. SanMiguel, P., Ramakrishna, W., Bennetzen, J., Busso, C. & Dubcovsky, J. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5Am. *Funct. Integr. Genomics* **2**, 70–80 (2002).
87. Wicker, T. & Keller, B. Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* **17**, 1072–1081 (2007).
88. Bugreev, D. *et al.* Dynamic, thermodynamic, and kinetic basis for recognition and transformation of DNA by human immunodeficiency virus type 1 integrase. *Biochemistry* **42**, 9235–9247 (2003).
89. Luschnig, C. & Bachmair, A. RNA packaging of yeast retrotransposon *Ty1* in the heterologous host, *Escherichia coli*. *Biol. Chem.* **378**, 39–46 (1997).
90. Feng, Y. X., Moore, S. P., Garfinkel, D. J. & Rein, A. The genomic RNA in *Ty1* virus-like particles is dimeric. *J. Virol.* **74**, 10819–10821 (2000).
91. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
92. Shirasu, K., Schulman, A. H., Lahaye, T. & Schulze-Lefert, P. A contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10**, 908–915 (2000).
93. Sabot, F. & Schulman, A. Template switching can create complex LTR retrotransposon insertions in *Triticeae* genomes. *BMC Genomics* **8**, 247 (2007).

Acknowledgements

The authors want to thank J. Estill (University of Georgia, Athens, USA) for very useful scientific discussions. We are particularly grateful to C. Feschotte (University of Texas, Austin, USA) and two other anonymous reviewers for their constructive comments and suggestions. J. W. Bizzaro and all the bioinformatics.org team are thanked for their hosting of WikiPoson and helping with its release. This work was supported by GDR 2157 of the Centre National de la Recherche Scientifique (CNRS; A.H.V., P.C & O.P.), by a University of Helsinki, Finland, Postdoctoral Fellowship (F.S.) and by the Institute of Plant Biology, Zurich, Switzerland (T.W.).

FURTHER INFORMATION

Element Namer: http://www.genomics.purdue.edu/~pmiguel/name_elements/

WikiPoson: http://www.wikiposon.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

ONLINE CORRESPONDENCE

Nature Reviews Genetics publishes items of correspondence online. Such contributions are published at the discretion of the Editors and can be subject to peer review. Correspondence should be no longer than 500 words with up to 15 references and should represent a scholarly attempt to comment on a specific Review or Perspective article that has been published in the journal. To view this correspondence, please go to our homepage at http://www.nature.com/nrg and follow the link from the current table of contents.

The following correspondence has recently been published:

On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics

Stefan Viktor Vormfelde and Jürgen Brockmöller

Reply

David J. Balding

This correspondence relates to the article:

A tutorial on statistical methods for population association studies

David J. Balding

Nature Rev. Genet. **7**, 781–791 (2006)