Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 147–162

# Exaptation of Protein Coding Sequences from Transposable Elements

*N.J. Bowen, I.K. Jordan*

School of Biology, Georgia Institute of Technology, Atlanta, Ga., USA

## Abstract

The activity of transposable elements (TEs) has had a profound impact on the evolution of eukaryotic genomes. Once thought to be purely selfish genomic entities, TEs are now recognized to occupy a continuum of relationships, ranging from parasitic to mutualistic, with their host genomes. One of the many ways that TEs contribute to the function and evolution of the genomes in which they reside is through the donation of host protein coding sequences (CDSs). In this chapter, we will describe several notable examples of eukaryotic host CDSs that are derived from TEs. Despite the existence of a number of such well-established cases, the overall extent and significance of this phenomenon remains a matter of controversy. Genome-scale computational analyses have yielded vastly different estimates for the fraction of host CDSs that are derived from TEs. We explain how these seemingly contradictory findings are the result of specific ascertainment biases introduced by the different methods used to detect TE-related sequences. In light of this problem, we propose a comprehensive and systematic framework for definitively characterizing the contribution of TEs to eukaryotic CDSs.



## Transposable Elements Defined

Transposable elements (TEs) are DNA sequences that can move (transpose) from one chromosomal location to another within the genome. Along with the capacity to move around the genome, TEs can replicate themselves and accumulate over time. The transpositional activity of TEs has had a substantial effect on the structure and function of eukaryotic genomes. For instance, TEs can cause phenotypically relevant mutations by inserting in or around genes, and ectopic recombination, mediated by homologous element sequences dispersed throughout the genome, can lead to large scale chromosomal re-arrangements. In addition, TEs are both abundant and ubiquitous. Remnants of TE insertions

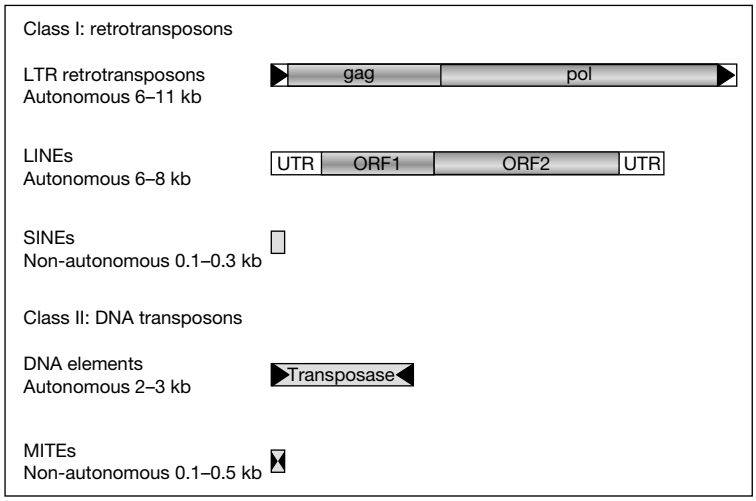


***Fig. 1.*** Classification and structure of TEs. Long terminal repeat (LTR) retrotransposons are flanked by two direct repeat sequences. They typically possess the gag and pol ORFs, which encode structural (gag) and enzymatic (pol) proteins involved in reverse transcription. LINE elements have untranslated regions (UTRs) of variable length on either side of two ORFs. ORF1 often encodes a nucleic acid binding protein, while ORF2 encodes reverse transcriptase. SINEs are much shorter non-autonomous retroelements that lack coding capacity. Autonomous DNA-type elements contain two terminal inverted repeats (TIRs) surrounding a single open reading frame that encodes the transposase enzyme. Non-autonomous DNA elements, such as MITEs, possess TIRs as well but lack coding capacity.

often make up more than half of any given eukaryotic genome sequence, and TEs have been found in all three domains of life for virtually every organism with characterized genomic sequences. In this chapter, we will focus exclusively on eukaryotic TEs, with an emphasis on human elements that have contributed to the evolution of protein coding sequences (CDSs).

Eukaryotic TEs are categorized into two broad classes [1] (fig. 1). Class I elements, or retroelements, transpose via the reverse transcription of an RNA intermediate. Retroelements include long- and short-interspersed nuclear elements, known as LINEs and SINEs respectively, as well as long terminal repeat (LTR) containing retrotransposons. LINEs encode all the enzymatic machinery necessary for their retrotransposition, while SINEs are non-autonomous elements that lack coding capacity. SINEs are thought to be retrotransposed in trans using enzymes encoded by LINEs [2]. LINEs are the single most abundant class of elements in the human genome making up more than 20% of the sequence [3]. A single family of LINE elements alone, LINE1, has amplified to

more than half a million copies in the human genome. In addition to catalyzing their own transcription, as well as that of SINEs, LINE encoded reverse transcriptase enzymes are probably responsible for generating the majority of processed pseudogenes in the human genome [4].

SINEs originate from non protein-coding RNAs transcribed by RNA polymerase III, tRNAs for the most part, that have been amplified by reverse transcription [5]. SINEs are actually the most numerous class of elements in the human genome, with more than 1.5 million copies identified, but make up a slightly lower overall fraction of the human genome (~13%) than LINEs due to their smaller size [3]. Most of the human genome SINEs are Alu elements. Alus are relatively young elements found exclusively within the primate evolutionary lineage [6]; they originated from 7SL RNA transcripts, which make up part of the ribosomal signal recognition particle [7]. Alus are particularly interesting because, unlike most other human TEs, they have accumulated in relatively GC-rich sequences found in close proximity to genes. An analysis of the age distribution of Alus revealed that this phenomenon is not due to any insertion site preference [3]. Rather it appears that Alus have been preferentially retained at gene encoding loci. This has been taken to suggest that Alus are genomic symbionts that play some beneficial role for the genomes in which they reside [3].

LTR retrotransposons are closely related to retroviruses [8]. They have open reading frames (ORFs) that encode capsid-like proteins, as well as the enzymes involved in retrotransposition, but lack the envelope ORF that confers intercellular infectivity to retroviruses [9]. In fact, LTR retrotransposons in humans are referred to as endogenous retroviruses, and many of them probably evolved from retroviruses that infected primate germline sequences and subsequently lost their infectivity [10]. Most of the LTR retrotransposon sequences in the human genome are found as solo LTRs, which are the result of intraelement LTR-LTR recombination events that excise the internal element sequences. LTR elements make up just under 10% of the human genome [3].

Class II, or DNA-type, elements transpose from DNA-to-DNA via a cut-and-paste mechanism catalyzed by the enzyme transposase. These elements generally contain inverted terminal repeats (TIRs) recognized by the DNA-binding domain of transposase and a single ORF encoding the transposase. Non-autonomous DNA elements containing only TIRs may be transposed in trans by related full length autonomous elements. Miniature inverted-repeat elements (MITEs) are a group of small high copy number non-autonomous DNA-type elements originally discovered in plants [11] and subsequently found in a wide variety of eukaryotic genomes [12]. Like Alus, MITEs are often found in close association with gene sequences and are thus thought to play some beneficial role related to gene regulation. DNA-type elements are the most common class of bacterial transposons, where they are known as IS

elements. DNA-type elements are also particularly abundant among insect and plant genomes, but less so in the human genome where they make up ~3% of the sequence. Despite their relatively low numbers in the human genome, DNA-type elements make up most of the known cases of TE-derived human CDSs [3]. The reason for this over-representation is currently unknown. It might be due to the fact that the transposase ORF provides a ready-made protein with DNA-binding properties that are particularly useful for the host [13]. For instance, domesticated transposase ORFs could play a role in mitigating the harmful effects of TEs by repressing transposition and/or they could influence the expression of host genes by acting as novel transcription factors.

There are also scattered examples of anomalous TEs that do not fit neatly into either of the two aforementioned broad classes I or II. For instance, DIRS1-like elements encode reverse transcriptase enzymes that are similar to those of LTR retrotransposons, but they lack integrase coding capacity as well as LTRs [14]. An even more unusual class of elements found in insects and plants possesses similarities to both non-LTR and LTR retrotransposons [15]. Some of these so-called Penelope-like elements do have LTRs, but they may be inverted in orientation. Many are 5′ truncated like non-LTR elements, and the reverse transcriptases of these elements are interrupted by an intron and most similar to the enzyme telomerase. The increasing appearance of such difficult to classify elements suggests the need for a revised classification scheme for TEs, an issue which has been addressed recently [1].

## Selfish DNA Theory of TEs

The recognition of TEs' broad distribution and high copy numbers, i.e. their evolutionary success, in eukaryotic genomes posed an explanatory challenge to biologists. Many wondered which attributes of the elements could best explain their ubiquity. This line of inquiry was based on a classic neo-Darwinian mode of thought, which held that the success of a gene must be predicated upon the utility that it provided for the organisms that encoded it. If this paradigm held for TEs, then it follows that they must be playing important and demonstrable roles for the genomes in which they reside. Thus, the first impulse for investigators interested in explaining the presence of TEs was to posit potential adaptive benefits that they may provide to their host organisms. In 1980, two seminal papers, published back-to-back in Nature [16, 17], completely inverted this explanatory paradigm for the presence and success of TEs.

These two papers laid the foundation for what is known as the selfish DNA theory of TEs. The selfish DNA theory holds that TEs are essentially genomic parasites, which serve only to increase their own abundance even at the expense

of their host genomes. The authors pointed out that existence of TEs can be explained solely by virtue of their ability to out-replicate their host genomes. This is because, in addition to being transmitted vertically like standard host genes, TEs are also replicated within the genome when they transpose. This replicative component of their life cycle gives TEs an inherent fitness advantage relative to host genes transmitted in a strict Mendelian fashion. This replicative advantage alone, with no regard whatsoever to any functional role or adaptive benefit that they may provide to their hosts, is sufficient to explain their evolutionary success. Later it was shown that TEs can spread within genomes and populations even in the face of deleterious effects that they may exert, via insertional mutations for example, on their host genomes [18]. This finding further emphasized the potentially parasitic nature of TEs.

The selfish DNA theory of TEs represented a true paradigm shift and continues to play an important role as a null hypothesis for the understanding of TEs' evolutionary significance. This is important in the sense that it helps to avoid the kind of tautological adaptationist thinking whereby the mere presence of a biological feature demands a plausible adaptive explanation. On the other hand, this new paradigm for TEs, while logically unassailable, also had a chilling effect on investigations into any functional role, or adaptive benefit, that TEs may play for their host genomes. In retrospect, it is interesting to note the divergent tacks taken by the authors of the two selfish DNA papers with regard to the potential functional roles played by TEs. One group applied a more cautious and measured approach being careful to point out that the selfish nature of TEs would not necessarily preclude them from being co-opted to play some beneficial role for their hosts [17]. However, the second group advocated a much harder line pointing out that the selfish nature of TEs rendered the consideration of any functional role that they may play ultimately futile [16]. Of course these two aspects of TE biology – selfish versus adaptive – are not mutually exclusive, and in recent years a more balanced perspective, which holds that TEs exist on a continuum from strict parasitism to mutualism, has emerged [19].

### Molecular Domestication

While the selfish DNA theory can still be considered as the null hypothesis by which the presence of TEs is explained, there are by now many exceptions to this view on TEs' evolutionary and functional significance, or lack thereof [19]. Wolfgang Miller coined the phrase 'molecular domestication' to describe the process whereby a formerly selfish TE is co-opted to perform a function that benefits its host genome [20]. Molecular domestication of TEs is

an example of the more general phenomenon of exaptation. The term exaptation was introduced by Gould and Vrba to account for a biological adaptation that plays a current role distinct from the original function that was selected for [21]. In the case of TEs, exaptations result from selection pressures exerted at different levels of biological organization. On the one hand, TE CDSs originally evolve under selection pressure at the genomic level. While on the other, the selection that governs the evolution of host gene sequences is exerted at the organismic level. In both cases, the selection is based on differential reproductive success. Organismic level selection is based on differential reproductive success of individuals in a population. In order to get established within a genome, TEs face selection pressure to transpose, i.e. reproduce in the genome, efficaciously and thus out-reproduce both their host genomes and other competing elements. However, this evolutionary mode does not represent an effective long-term strategy since transposition is a highly mutagenic process that often causes deleterious changes to the host genome. Unchecked transposition and accumulation of element sequences could ultimately lead to the extinction of the TEs' host evolutionary lineage, which in turn would mean extinction of the elements themselves. To counter this possibility, TEs have evolved a number of mechanisms that mitigate the harmful effects of transposition. For instance some TEs, such as LINEs in human and mouse, confine their expression to germline tissues [22]. This helps to ensure transmission of newly replicated elements to future generations while minimizing the harmful effects of somatic mutations caused by transposition. As another example of host-element co-evolution, P-elements in *Drosophila* have evolved a strategy of self-regulation by encoding a repressor protein that blocks transposition in already infected genomes [23]. Of course, the ultimate co-evolutionary strategy exhibited by TEs is molecular domestication whereby the formerly selfish element sequences make themselves indispensable to their host genome by taking on some essential functional role.

There is a growing body of evidence that demonstrates a number of different ways that TEs have evolved from strictly parasitic elements to mutualistic sequences that benefit their host genomes. For instance, there are numerous documented cases where TEs have been shown to donate regulatory sequences that control the expression of nearby host genes [24–26]. For the rest of this chapter though, we will focus exclusively on the cases where formerly selfish TE sequences have been domesticated to provide CDSs for their eukaryotic host genomes. The extent and overall significance of this phenomenon is currently a matter of some debate. As genome sequences accumulate, more and more examples of TE-derived host CDSs are posited. However, some of these cases have proven illusory and different methods of detection of TE-derived CDSs often yield vastly different results. In addition to providing a few canonical

examples of TE-derived host CDSs, we will explore issues pertaining to ascertainment biases at play in their discovery.

### Host CDSs from TEs

In this section, we will briefly outline several canonical examples of how TE ORFs have been exapted as host gene CDSs – Telomerase, RAG recombinase and SETMAR. Telomerase is an enzyme that helps to replace DNA residues that are inevitably lost from the ends of eukaryotic chromosomes during their replication [27]. Telomerase uses RNA oligonucleotides as templates for the addition of DNA sequences to chromosome ends; in other words, it is a reverse transcriptase (RT). Sequence analysis of telomerase revealed that it shares a common evolutionary origin with the RTs of retrotransposable elements [28]. Phylogenetic comparison of telomerase with the RT domains of TEs indicates that the telomerase RT probably diverged from LINE-like elements in early eukaryotic evolution, which points to the exaptation of this critical cellular activity from a class of formerly selfish elements [28].

The RAG recombinase enzymes – RAG1 and RAG2 – are together responsible for catalyzing V(D)J recombination in vertebrate genomes [29]. V(D)J recombination is the mechanism by which vertebrates generate immunological diversity in antibody and T-cell receptor molecules [30]. The breaking and rejoining of chromosomal segments catalyzed by the RAG enzymes allows for the production of a vastly diverse array of immunoglobulin encoding sequences, which is capable of countering the diversity of pathogens that challenge the immune system. The striking similarity between the processes of RAG catalyzed V(D)J recombination and transposition of DNA-type elements led to the suggestion of an evolutionary relationship between the two [31]. This proposition was later confirmed by experimental work showing that the RAG recombinases can catalyze transposition within and between chromosomes [32]. Recently, a direct evolutionary link between the RAG1 sequence and a family of DNA-type elements has been established [33]. Thus, it appears likely that the ability of the vertebrate immune system to generate immunogenic diversity evolved from TEs as well. The implications of this particular exaptation event for the survival of the vertebrate evolutionary lineage are striking.

The human SETMAR protein provides a more recent example of the exaptation host CDSs from a TE [34, 35]. Because this particular domestication event took place in the more recent evolutionary past, investigators were able to more definitively characterize the relationship between the TE and its related host gene. Indeed, SETMAR was originally characterized as a chimeric transcript that combined a SET methyltransferase domain with the transposase domain

from a specific family of DNA-type element named Hsmar1 [35]. Comparative analysis of corresponding genomic regions cloned from related primates revealed that this particular domestication event occurred between 40–58 million years ago after the Hsmar1 element inserted downstream from the SET domain encoding exons [34]. The function of this particular domesticated gene remains less well understood than is the case for the more ancient exaptation events that led to evolution of telomerase and RAG recombinase from TE sequences. Investigators were able to demonstrate that the catalytic activity of the transposase derived domain of SETMAR has been lost while the DNA binding activity remains [34]. This raised the intriguing possibility that the evolution of SETMAR may have facilitated the de novo emergence of complex regulatory network involving the binding of SETMAR to numerous Hsmar1 derived TIR binding sites dispersed throughout the genome.

The three selected examples of TE-CDS domestication described here represent only a fraction of the known cases. Another noteworthy example is the case of Daysleeper, a DNA-type element that has been domesticated in *Arabidopsis* and shown to be essential for plant development by virtue of the regulatory effects that it exerts on numerous genes [36]. In humans, both the centromere binding protein gene (*CENBP*) and the Jerky gene are derived from related DNA-like elements [37–39]. An exhaustive description of all such cases is outside the scope of this manuscript. However, several other reports do provide a more in depth accounting of protein coding sequences exapted from TEs [3, 37, 39, 40].

## Genome Wide Analyses

Despite these solid examples of TE contributions to CDSs, the extent of this phenomenon remains a matter of substantial controversy. With the accumulation of eukaryotic genome sequences, a number of attempts have been made to exhaustively characterize instances of TE-derived host CDSs [3, 37, 39, 41, 42]. One of the open questions that such studies address is the proportion of host CDSs that can be demonstrated to have evolved from TEs. Prior to the completion of the human genome sequence there were 20 known cases of human CDSs derived from TEs [37, 39]. Analysis of the draft sequence of the human genome found 27 additional cases yielding a total of 47 distinct human TE-derived CDSs [3]. This figure represents a fairly negligible fraction (~0.16%) of all human genes, given the lower bound estimate for the human gene count (30,000) reported at that time [3]. The same year however, using similar detection techniques on a set of human gene sequences from a different source, Nekrutenko and Li published their own genome-scale analysis where they reported that ~4%

of analyzed human genes had CDSs that were derived from TEs [42]. Clearly such vastly different estimates call for some sort of reconciliation.

Pavlicek and colleagues took another look at the findings of Nekrutenko and Li and pointed out several caveats that should be taken into consideration when trying to determine the extent of CDSs derived from TEs [43]. First of all, they found that, of the set of genes identified by Nekrutenko and Li as TE-derived, 30% were annotated as hypothetical and 63% were annotated as predicted. In other words, there was no experimental evidence that supported these particular genes as being bona fide functional CDSs. In addition, the majority of human CDSs that Nekrutenko and Li found with similarity to TEs were derived from Alu (SINE) elements that lack protein coding capacity. Pavlicek et al. found this suspicious as well, since the vast majority of CDSs previously reported to contain Alu related sequences have only been detected at the mRNA level. In light of these issues, Pavlicek et al. took a far more conservative approach to identifying TE-derived CDSs; specifically, they analyzed CDSs taken only from representative 3D structures. The 3D structures represent the most accurate source evidence for the actual existence of the proteins under consideration. When these CDSs were analyzed using the same detection technique as Nekrutenko and Li no evidence of TE-derived sequences was found. A slightly more sensitive technique did reveal 28 cases of TE-derived CDSs but all of these came from TEs that are known to encode proteins and none were from Alu elements, which lack protein coding capacity.

Despite the apparent absence of SINE related sequences among CDSs with representative 3D structures, the facts remain that Alu elements are both highly prevalent in human gene-rich regions [3] and harbor numerous potential splice sites that can facilitate their incorporation into mRNA transcripts [44]. Thus, Alus would appear to be ideally positioned to be integrated into the CDSs of host genes. Indeed, numerous alternatively spliced human exons (~5%) were found to contain Alu sequences [44]. Several individual cases of how Alu sequences have become 'exonized' have been explicated in detail, revealing the evolutionary timing of these events as well as the specific mutations that led to their incorporation into gene transcripts [45]. However, the actual protein coding potential and biological function of these sequences is still an open question. Comparative sequence analysis, establishing both the conservation of Alu-derived open reading frames and a conservative pattern of sequence substitution [46] should help to settle this matter.

Consistent with the conservative approach of Pavlicek et al., a more recent publication from the Nekrutenko group refuted one of their own earlier discoveries of a mouse CDS that appeared to be derived almost entirely from SINEs [47]. Comparative sequence analysis with other *Mus* species, as well as the rat, did not find any evidence for the conservation of the ORF of this TE-derived

gene. The authors concluded that the original discovery represented an artifact and emphasized the importance of validation of computationally predicted TE-derived CDSs. This report also underscored the paucity of examples of non-artifactual validated cases where non-coding TEs, such as SINEs, contribute CDSs to host genes [48]. Rather, it appears that the vast majority of well-supported cases of TE-derived host CDSs come from TEs that already encode proteins.

## A New Framework is Needed

The contradictory results outlined in the previous section illustrate the confusion and controversy that still surround the issue of TE-derived CDSs [49]. We would like to close this chapter by arguing that a new framework, one that is both comprehensive and systematic, is needed to understand the extent and biological significance of TE-derived CDSs. It is worth noting here that the extent of TE-derived CDSs may not necessarily be directly related to its evolutionary significance. For instance, even if the contribution of TEs to host CDSs turns out to be low in terms of overall numbers, its impact in terms of biology and evolution may still be substantial. The single case of the RAG-recombinases and vertebrate specific immunity underscores this point. It is also tempting to speculate as to whether differing extents of domestication between evolutionary lineages may be responsible for driving increases in complexity that mark the eukaryotic crown group. In any case, a rigorous elucidation of the extent of host protein coding sequences that are derived from TEs will be critical for our understanding of eukaryotic genome structure, function and evolution.

It has occurred to us and others that a substantial problem lies in the differences in sensitivity of the methods used to detect TE-related sequences among protein coding genes. Most studies rely on the widely used RepeatMasker program [50] to detect TE related sequences. RepeatMasker works at the DNA-level and compares genomic sequences to a library of consensus sequences that represent previously characterized TE families. This approach has two distinct disadvantages. First of all it can only detect TEs that are already known. This is not a big problem for well-characterized species such as human, but it may represent a substantial limitation for less well characterized evolutionary lineages. Perhaps even more importantly, RepeatMasker can only detect TE sequences that have diverged relatively recently from other members of the same family. This is partly due to the use of consensus sequences but more so to the reliance on DNA-DNA sequence comparisons. With only four different residues to compare, substitutions between related DNA sequences quickly become saturated and their evolutionary relationships are obscured. Protein sequences, on the other hand, retain the signal of common ancestry for much longer periods of time.

The ascertainment bias that results from the reliance on DNA-DNA sequence comparisons was driven home by a recent re-analysis of TE-derived sequences among human protein coding genes [41]. Roy Britten took the same TE consensus sequence libraries used for RepeatMasker and translated them in all reading frames. These conceptual translations were then used as queries in protein-protein BLAST searches against all human proteins. The protein sequence comparison resulted in a more than two-fold increase from 814 detected TE-derived CDSs to 1,950 such cases. These newly detected cases represent more ancient associations between TE-derived sequences and human genes and in that sense may be even more likely to be validated with experimental and/or 3D structural information. In addition, while protein-protein sequence comparisons are more sensitive than DNA-DNA comparisons, there are even more sensitive ways to search for common ancestry between sequences such as profile comparisons, using position specific score matrices (PSSMs) or hidden Markov models (HMMs), and direct comparisons between 3D protein structures, which are the most sensitive methods of all. The use of such techniques would undoubtedly turn up additional cases of TE-derived, or at the very least TE-related, host CDSs.

A favorite example of ours can serve to further illuminate the challenges for uncovering relationships between TEs and host CDSs. PAX8 is a nuclear transcription factor that is involved in thyroid and kidney development and implicated in the etiology of several different cancers. PAX8 is a member of the paired box (PAX) family of transcription factors that are expressed in cell specific patterns during metazoan development [51, 52]. PAX proteins contain an amino-terminal, sequence specific DNA binding domain known as the paired box, which consists of tandem helix-turn-helix (HTH) motifs [53, 54]. Protein sequence-based homology searches have uncovered a highly significant similarity between the paired box domain and the transposase present in members of the Tc1 family of TEs [55, 56]. Structural modeling has likewise shown the presence of two HTH motifs in the Tc1 transposase sequences [55, 57]. The similarity between Tc1 transposase and the paired box domain is so reliable that transposase sequences are now routinely used as an outgroup to root phylogenetic comparisons of within and between species comparisons of PAX proteins [55, 58].

There are 9 PAX genes in the human genome and many more genes that encode domains with HTH motifs that may be distantly related to transposase domains. However, the PAX genes in particular are present only in the animal lineage of eukaryotes; they have not been found in unicellular eukaryotes, fungi, plants nor in prokaryotes [59]. This lineage-specificity of PAX genes stands in stark contrast to the widespread distribution of the Class II family of DNA-type elements to which Tc1 elements belong. Based on these disparate

phyletic distributions, a transposase origin of PAX genes is the most likely evolutionary scenario that explains their sequence similarity. Despite their robust and well-characterized relationship, when the PAX8 DNA CDS is run through RepeatMasker no evidence of a TE origin is uncovered. Clearly, a strict DNA-centric genome-scale approach to uncovering TE-derived CDSs will only tell part of the story.

Finally, we would like to emphasize here that the challenge of ascertainment biases inherent to the different methods also presents an important opportunity. Once the particular strengths and weaknesses of different approaches are recognized and considered, a more systematic approach to the detection of TE-derived CDSs can be devised. Specifically, we would like to propose that any genome-scale computationally based attempt to uncover TE-derived host CDSs must involve the use of numerous complementary approaches, each of which is appropriate to its own level of evolutionary relatedness between the TEs and CDSs (fig. 2): (i) DNA-DNA sequence comparison methods can be used to detect recent putative exaptation events followed successively by (ii) protein-protein sequence comparisons, (iii) profile-protein comparisons and (iv) structure-protein or structure-structure comparisons, each of which in turn may reveal more ancient associations between TEs and CDSs. Detection of such relationships must only represent the first step in the process though. Further confidence in the validity of TE-gene associations can be achieved by comparative sequence analyses aimed at detecting both conservation of TE-derived ORFs as well as conservative DNA substitution patterns that are indicative of purifying selection based on protein function. Finally, the kind of phyletic distribution comparison described earlier for the case of PAX8 can be used to definitively establish the evolutionary directionality of the relationship between the TE and host CDSs. The donor sequence set (the transposase in the case of PAX encoding genes) should be characterized by a broader distribution among more distantly related species than the acceptor group of sequences. The breadth of sequence distribution can also be used to inform the direction of the relationship between TE and host gene. In the case of the telomerase for example, its RT represents only a fraction of the sequence diversity of all retrotransposon RTs, which is consistent with its origin from one particular lineage along the RT phylogenetic tree. While seemingly exhaustive, this kind of comprehensive approach that we propose is ideally suited to the computational approach. In fact a recently published paper from the group of Peer Bork proposed an analogous, if more narrow in scope, algorithmic approach aimed at discovering and then validating cases of host CDSs that may be derived from TEs [60]. Hopefully, the problem of the extent and significance of TE-derived host CDSs will yield to such a systematic approach, and in so doing, help us to better understand the ancient and ongoing evolutionary dynamic between TEs and their host genomes.

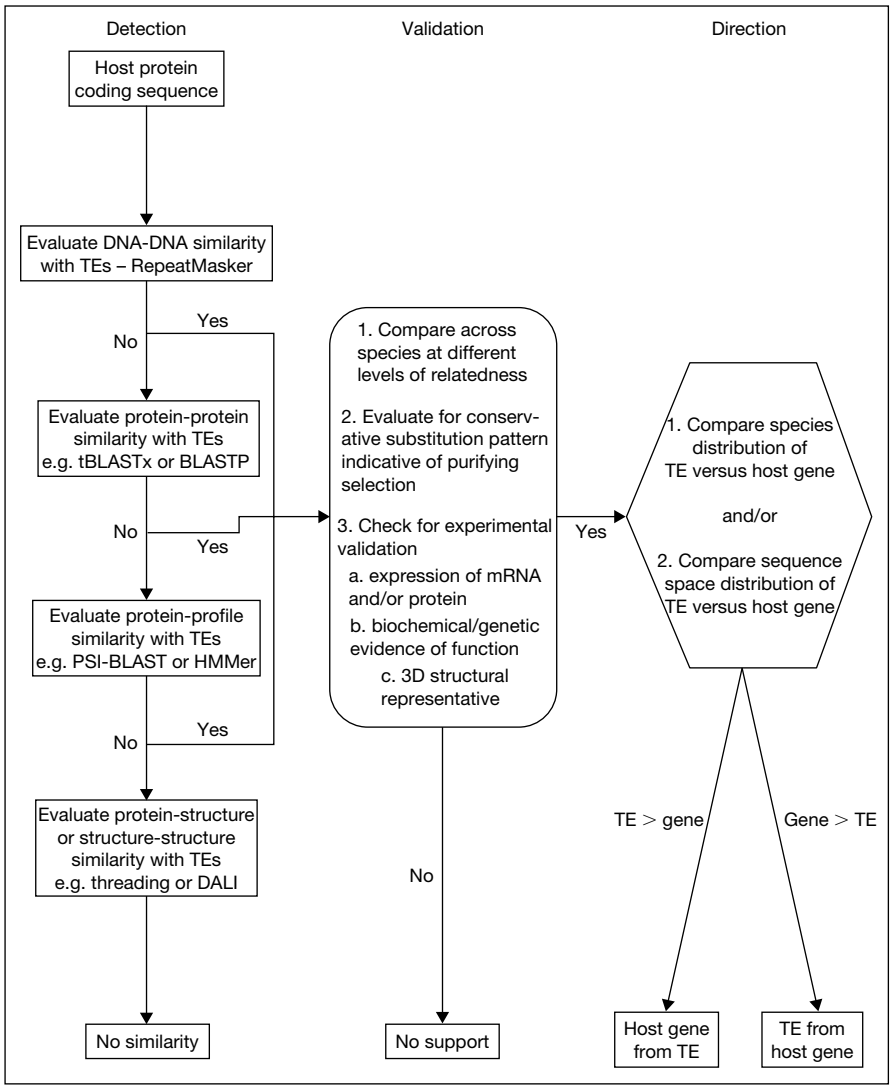


***Fig. 2.*** Scheme for the detection, validation and characterization of TE-derived host CDSs.

## References


1 Capy P: Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 2005;110:457–461.

2 Dewannieux M, Esnault C, Heidmann T: LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003;35:41–48.

3 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al: Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
4 Esnault C, Maestre J, Heidmann T: Human LINE retrotransposons generate processed pseudogenes. Nat Genet 2000;24:363–367.
5 Daniels GR, Deininger PL: Repeat sequence families derived from mammalian tRNA genes. Nature 1985;317:819–822.
6 Deininger PL, Daniels GR: The recent evolution of mammalian repetitive DNA elements. Trends Genet 1986;2:76–80.
7 Ullu E, Tschudi C: Alu sequences are processed 7SL RNA genes. Nature 1984;312:171–172.
8 Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 1990;9:3353–3362.
9 Inouye S, Yuki S, Saigo K: Sequence-specific insertion of the *Drosophila* transposable genetic element 17.6. Nature 1984;310:332–333.
10 Bannert N, Kurth R: Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci USA 2004;101(suppl 2):14572–14579.
11 Bureau TE, Wessler SR: Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 1992;4:1283–1294.
12 Feschotte C, Zhang X, Wessler SR: Miniature inverted-repeat transposable elements and their relationship to established DNA transposons; in Craig NL, Craigie R, Gellert M, Lambowitz A (eds): Mobile DNA II. (ASM Press, Washington, DC 2002).
13 Jordan IK: Evolutionary tinkering with transposable elements. Proc Natl Acad Sci USA 2006;103:7941–7942.
14 Cappello J, Handelsman K, Lodish HF: Sequence of *Dictyostelium* DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell 1985;43: 105–115.
15 Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB: Retroelements containing introns in diverse invertebrate taxa. Nat Genet 2003;33:123–124.
16 Doolittle WF, Sapienza C: Selfish genes, the phenotype paradigm and genome evolution. Nature 1980;284:601–603.
17 Orgel LE, Crick FH: Selfish DNA: the ultimate parasite. Nature 1980;284:604–607.
18 Hickey DA: Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 1982;101:519–531.
19 Kidwell MG, Lisch DR: Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution 2001;55:1–24.
20 Miller WJ, Hagemann S, Reiter E, Pinsker W: P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. Proc Natl Acad Sci USA 1992;89: 4018–4022.
21 Gould SJ, Vrba E: Exaptation – a missing term in the science of form. Paleobiology 1982;8:4–14.
22 Trelogan SA, Martin SL: Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc Natl Acad Sci USA 1995;92: 1520–1524.
23 Robertson HM, Engels WR: Modified P elements that mimic the P cytotype in *Drosophila melanogaster*. Genetics 1989;123:815–824.
24 Britten RJ: DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 1996;93:9374–9377.
25 Jordan IK, Rogozin IB, Glazko GV, Koonin EV: Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 2003;19:68–72.
26 van de Lagemaat LN, Landry JR, Mager DL, Medstrand P: Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 2003;19:530–536.
27 Blackburn EH: Structure and function of telomeres. Nature 1991;350:569–573.
28 Eickbush TH: Telomerase and retrotransposons: which came first? Science 1997;277:911–912.
29 Oettinger MA, Schatz DG, Gorka C, Baltimore D: RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 1990;248:1517–1523.
30 Tonegawa S: Somatic generation of antibody diversity. Nature 1983;302:575–581.

31 Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G: The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. Cell 1996;87:263–276.
32 Agrawal A, Eastman QM, Schatz DG: Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 1998;394:744–751.
33 Kapitonov VV, Jurka J: RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 2005;3:e181.
34 Cordaux R, Udit S, Batzer MA, Feschotte C: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 2006;103:8101–8106.
35 Robertson HM, Zumpano KL: Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene 1997;205:203–217.
36 Bundock P, Hooykaas P: An *Arabidopsis* hAT-like transposase is essential for plant development. Nature 2005;436:282–284.
37 Jurka J, Kapitonov VV: Sectorial mutagenesis by transposable elements. Genetica 1999;107: 239–248.
38 Kipling D, Warburton PE: Centromeres, CENP-B and Tigger too. Trends Genet 1997;13:141–145.
39 Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 1999;9:657–663.
40 Volff JN: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 2006;28:913–922.
41 Britten R: Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci USA 2006;103:1798–1803.
42 Nekrutenko A, Li WH: Transposable elements are found in a large number of human protein-coding genes. Trends Genet 2001;17:619–621.
43 Pavlicek A, Clay O, Bernardi G: Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEBS Lett 2002;523:252–253.
44 Sorek R, Ast G, Graur D: Alu-containing exons are alternatively spliced. Genome Res 2002;12: 1060–1067.
45 Krull M, Brosius J, Schmitz J: Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 2005;22:1702–1711.
46 Nekrutenko A, Chung WY, Li WH: An evolutionary approach reveals a high protein-coding capacity of the human genome. Trends Genet 2003;19:306–310.
47 Wilson C, Goetting-Minesky P, Nekrutenko A: mNSC1 shows no evidence of protein-coding capacity. Gene 2006;370:83–85.
48 Claverie JM, Makalowski W: Alu alert. Nature 1994;371:752.
49 Gotea V, Makalowski W: Do transposable elements really contribute to proteomes? Trends Genet 2006;22:260–267.
50 Smit AFA, Hubley R, Green P: RepeatMasker Open-3.0. 1996–2004.
51 Chi N, Epstein JA: Getting your Pax straight: Pax proteins in development and disease. Trends Genet 2002;18:41–47.
52 Robson EJ, He SJ, Eccles MR: A PANorama of *PAX* genes in cancer and development. Nat Rev Cancer 2006;6:52–62.
53 Bopp D, Burri M, Baumgartner S, Frigerio G, Noll M: Conservation of a large protein domain in the segmentation gene paired and in functionally related genes of *Drosophila*. Cell 1986;47: 1033–1040.
54 Xu W, Rould MA, Jun S, Desplan C, Pabo CO: Crystal structure of a paired domain-DNA complex at 2.5 A resolution reveals structural basis for Pax developmental mutations. Cell 1995;80:639–650.
55 Breitling R, Gerber JK: Origin of the paired domain. Dev Genes Evol 2000;210:644–650.
56 Franz G, Loukeris TG, Dialektaki G, Thompson CR, Savakis C: Mobile Minos elements from *Drosophila hydei* encode a two-exon transposase with similarity to the paired DNA-binding domain. Proc Natl Acad Sci USA 1994;91:4746–4750.
57 Ivics Z, Izsvak Z, Minter A, Hackett PB: Identification of functional domains and evolution of Tc1-like transposable elements. Proc Natl Acad Sci USA 1996;93:5008–5013.

58 Hadrys T, DeSalle R, Sagasser S, Fischer N, Schierwater B: The Trichoplax *PaxB* gene: a putative Proto-*PaxA/B/C* gene predating the origin of nerve and sensory cells. Mol Biol Evol 2005;22: 1569–1578.
59 Vorobyov E, Horst J: Getting the Proto-Pax by the Tail. J Mol Evol 2006;63:153–164.
60 Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P: Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res 2005;33: 946–954.

I. King Jordan
School of Biology
Georgia Institute of Technology
310 Ferst Drive
Atlanta, GA 30332-0230 (USA)
Tel. 404-385-2224, Fax 404-894-0519, E-Mail king.jordan@biology.gatech.edu