Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong?

Flávio S.J. de Souza,*1,2 Lucía F. Franchini,1 and Marcelo Rubinstein*1,2

1Instituto de Investigaciones en Ingeniería Genética y Biología Molecular, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina
2Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina
*Corresponding authors: E-mail: mrubins@dna.uba.ar; fsouza.ingebi@gmail.com.
Associate editor: Jianzhi Zhang

Abstract

Transposable elements (TEs) are mobile genetic sequences that can jump around the genome from one location to another, behaving as genomic parasites. TEs have been particularly effective in colonizing mammalian genomes, and such heavy TE load is expected to have conditioned genome evolution. Indeed, studies conducted both at the gene and genome levels have uncovered TE insertions that seem to have been co-opted—or exapted—by providing transcription factor binding sites (TFBSs) that serve as promoters and enhancers, leading to the hypothesis that TE exaptation is a major factor in the evolution of gene regulation. Here, we critically review the evidence for exaptation of TE-derived sequences as TFBSs, promoters, enhancers, and silencers/insulators both at the gene and genome levels. We classify the functional impact attributed to TE insertions into four categories of increasing complexity and argue that so far very few studies have conclusively demonstrated exaptation of TEs as transcriptional regulatory regions. We also contend that many genome-wide studies dealing with TE exaptation in recent lineages of mammals are still inconclusive and that the hypothesis of rapid transcriptional regulatory rewiring mediated by TE mobilization must be taken with caution. Finally, we suggest experimental approaches that may help attributing higher-order functions to candidate exapted TEs.

Key words: exaptation, enhancer, mobile element, gene expression.

Introduction

An important observation of the genomic era is that a great part of vertebrate genomes is comprised of transposable elements (TEs). In mammals, TEs are classified as 1) autonomous long interspersed nucleotide elements (LINEs), 2) short interspersed nucleotide elements (SINEs), which depend on LINEs for their propagation, 3) endogenous retrovirus (ERV)-like elements with long terminal repeat (LTRs) sequences, and 4) DNA transposons. The first three of these TE classes are the most abundant in mammals and reproduce via a RNA intermediate, being thus called retrotrosposons (see Böhne et al. 2008 for review), whereas DNA transposons jump around the genome by a cut-and-paste mechanism. In mammals, 30–50% of the genome comprised TEs, chiefly retrotrosposons, where most family members have a narrow taxonomic distribution. Alu and B1 elements, for instance, are SINEs only found in primates and rodents, respectively. Although TEs are in essence parasitic DNA elements with no intrinsic function for the host (Doolittle and Sapienza 1980; Orgel and Crick 1980), early studies indicated that particular instances of TE insertions contributed with new genes, exons, or regulatory regions (Brosius 2003). In 1992, Brosius and Gould (1992) suggested the term “exaptation” for the phenomenon of “junk” DNA sequences such as TEs acquiring a novel function in the genome. Because regulatory innovation is expected to drive evolution (Carroll 2008), and considering the huge TE content of genomes, exaptation of TEs into new promoters, enhancers, chromatin barriers, and other regulatory elements is likely to have contributed to the evolution of regulatory networks (Medstrand et al. 2005).

Since the times of Brosius and Gould, much evidence has accumulated suggesting exaptation of TEs as transcriptional regulatory regions; indeed, this idea has reached genomic proportions with several whole-genome studies showing that TEs can bind to thousands of TE instances in the genome. This has fueled the hypothesis that TE insertions are responsible for an extremely rapid transcriptional rewiring in mammals and other vertebrates (Feschotte 2008; Bourque 2009), reminiscent of Barbara McClintock’s original idea that transposons acted as “controlling elements” of several loci in the genome (McClintock 1956). Based on the correlation between past bursts of TE activity and the emergence of certain vertebrate lineages, it has even been hypothesized that TEs have crucially contributed to the origin and radiation of mammals (Okada et al. 2010) and specific mammalian groups such as anthropoid primates, rodents, and microbats (Oliver and Greene 2009, 2011; Britten 2010).

Where do we stand regarding transcriptional regulation by TE insertions? Here, we review some recent developments in the exaptation of TEs into new promoters, enhancers, and insulators from both gene-centered and genome-wide studies. We conclude that, although it can be safely ascertained that TE exaptation has been an important evolutionary drive
in transcriptional networks in the past, genomic searches for transcription factor binding sites (TFBSSs) have been of limited value in uncovering TE exaptations in recent lineages, including humans, and more work is necessary before the global contribution of TE exaptation to transcriptional network evolution can be fully appreciated.

**Promoters Derived from TEs**

LINEs and TEs carrying LTRs have intrinsic ability to recruit RNA polymerase II and thus have great potential to be exapted as promoters. That TE insertions in the near vicinity of genes can be detrimental to their function is evidenced by the lower-than-expected frequency of LINEs and LTR TEs within or close to transcriptional units (Medstrand et al. 2005). In spite of this, whole-genome analyses have shown that up to 25% of genes have TEs into their promoter and/or untranslated regions (Jordan et al. 2003; van de Lagemaat et al. 2003). In addition, transcriptome analyses indicate that 18% and 31% of cap-selected RNA transcripts initiate within repetitive elements in mouse and human cell lines, respectively, showing that transcription start sites (TSS) embedded within TEs are not an uncommon phenomenon (Faulkner et al. 2009).

Even though the results of genome-wide expression surveys are interesting, the functional impact of TE-derived TSSs is still unknown. Faulkner et al. (2009) observed that these RNA variants are rare, with only 2.8% and 5.2% of the most abundant RNAs deriving from retrotransposon TSSs in mice and humans, respectively. The same authors observed that more than half of TE-derived transcripts stay in the nucleus, which may indicate a function as noncoding regulatory RNAs or that most of them are simply nonfunctional (Faulkner et al. 2009). Indeed, detailed, gene-centered studies of protein-encoding genes have shown that TEs usually act as alternative promoters that originate minor mRNA variants of uncertain functional significance (reviewed in Cohen et al. 2009).

For certain genes, however, most or all transcriptions initiate within a TE, a hallmark for authentic exaptation. For instance, mouse Naip genes, which function in apoptosis, are expressed exclusively from an ORR1E LTR-derived promoter with ubiquitous activity (Romanish et al. 2007). The human homolog of mouse Naip harbors an unrelated LTR element (MER21C) that drives expression to testis but in this case as a minor (12%) mRNA species (Romanish et al. 2009). In the teleost fish medaka, the master regulator gene dmrt1b has changed its position in the hierarchy of sex determination genes after the insertion of a novel DNA transposon in its promoter (Herpin et al. 2010). In some mammalian cases, the insertion of a TE drives expression to a novel tissue, usually the placenta, where LTR retroviruses are known to be most active (Cohen et al. 2009). For instance, the insulin-like 4 (INSL4) gene only exists in anthropoid primates and its only known expression site is the placenta, where all of its transcripts are derived from an LTR promoter, a strong indication of exaptation of this TE in humans and our close relatives (Bie\'che et al. 2003). Recently, it has been shown that different TEs have been independently recruited by the prolactin genes of anthropoid primates (LTR MER39), rodents (LTR MER77), and African elephants (LINE L1-2_LA) as promoters with specific activity in the placenta, where prolactin is known to have an effect in sustaining pregnancy, at least in rodents (Emera et al. 2012).

**Enhancers Derived from TEs**

Enhancers are arrays of TFBs that alter transcriptional rates of genes. These regions, usually 150–500 bp in length, may be located far away from the transcriptional start site making their discovery and confirmation more difficult (Visel et al. 2007). Many enhancers are under purifying selection and can be identified by phylogenetic footprinting strategies that search for conserved noncoding regions in related genomes (Visel et al. 2007).

In contrast to enhancers, most TE-derived sequences in vertebrate genomes are expected to have no function and thus to evolve at a neutral rate. Indeed, analysis of ancestral TEs shared by humans and mice using a neutral insertion–deletion model indicate that over 95% of ancient TE insertions in the human genome are evolving at a neutral rate (Lunter et al. 2006; Meader et al. 2010). However, genome-wide analyses have uncovered a large number of TE-derived sequences that are under purifying selection and are located in noncoding, conserved regions of vertebrate genomes (e.g., see table 2). A recent comparison of 29 mammalian genomes found a staggering more than 280,000 conserved, TE-derived noncoding elements with potential regulatory function (Lindblad-Toh et al. 2011; Lowe and Haussler 2012). The high level of conservation of these elements strongly suggests functionality, and many could be enhancers.

Few putative cases of TE exaptation into transcriptional enhancers have been experimentally tested in transgenic mouse models (table 1 and figs. 1 and 2). Bejerano et al. (2006) used transgenesis to show that a conserved noncoding sequence derived from an LF-SINE was in fact a developmental enhancer of the Isl1 gene in mice. LF-SINES are an ancient repeat family, and phylogenetic conservation analysis suggests that this TE-derived sequence works as an Isl1 enhancer in all tetrapods (fig. 1; Bejerano et al. 2006). Sasaki et al. (2008) identified two different AmnSINE1-derived sequences that function as distal transcriptional enhancers in mouse embryos, controlling Fgf8 and SATB2 (Sasaki et al. 2008, Tashiro et al. 2011, Nakanishi et al. 2012). Although AmnSINES are also an ancient TE family, phylogenetic conservation indicates that exaptation apparently happened in a mammalian ancestor (fig. 1). Looking for the evolutionary origin a conserved neuronal enhancer of the POMC gene named nPE2 (de Souza et al. 2005), we found that it was derived from a member of the CORE-SINE family (Santangelo et al. 2007). Recently, we also found that a second neuronal enhancer of POMC—named nPE1—is derived from a mammalian apparent LTR retrotransposon (Franchini et al. 2011). Interestingly, the TEs that gave rise to the POMC enhancers were exapted at different time points: nPE2 originated from an earlier exaptation of in the lineage leading to mammals sometime before the Prototheria/Theria split (~170 Ma), whereas nPE1 is a placental novelty that originated after the Metatheria/Eutheria split (~150 Ma) and before the
### Table 1. Evidence for Conserved and Nonconserved TEs Exapted as Enhancers and Insulators

<table>
<thead>
<tr>
<th>TE Function</th>
<th>Genes Controlled</th>
<th>Technique(s)</th>
<th>Evolutionarily Conserved?</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINE1 Enhancer</td>
<td>Human apo</td>
<td>YAC transgenic mice, reporter assays in cell culture</td>
<td>No (humans, primates?)</td>
<td>Yang et al. (1998)</td>
</tr>
<tr>
<td>Alu Insulator, barrier element</td>
<td>Human K18</td>
<td>Transgenic mice</td>
<td>No (humans, primates?)</td>
<td>Willoughby et al. (2000)</td>
</tr>
<tr>
<td>LF-SINE Enhancer</td>
<td>Mammalian Isl1</td>
<td>Conservation, mouse transgenesis</td>
<td>Mammals</td>
<td>Bejerano et al. (2006)</td>
</tr>
<tr>
<td>CORE-SINE Enhancer</td>
<td>Mammalian Pomc</td>
<td>Conservation, mouse transgenesis</td>
<td>All mammals</td>
<td>Santangelo et al. (2007)</td>
</tr>
<tr>
<td>B2 SINE Insulator, barrier element</td>
<td>Mouse growth hormone</td>
<td>BAC transgenic mice, reporter assays in cell culture</td>
<td>No (only mouse/rodents)</td>
<td>Lunyak et al. (2007)</td>
</tr>
<tr>
<td>MIR SINE Enhancer</td>
<td>“boost”</td>
<td>Conservation, reporter assays stable cell culture, embryoid bodies</td>
<td>Mammals</td>
<td>Smith et al. (2008)</td>
</tr>
<tr>
<td>AmnSINE Enhancer</td>
<td>Mammalian SATB2</td>
<td>Conservation, mouse transgenesis</td>
<td>Mammals</td>
<td>Sasaki et al. (2008) and Tashiro et al. (2011)</td>
</tr>
<tr>
<td>ERV9 (LTR) Enhancer</td>
<td>b-globin locus</td>
<td>BAC transgenic mice, Cre, 3C</td>
<td>Great apes and humans only</td>
<td>Pi et al. (2004)</td>
</tr>
<tr>
<td>MaLR (LTR) Enhancer</td>
<td>Mammalian Pomc</td>
<td>Conservation, mouse transgenesis</td>
<td>Placental mammals</td>
<td>Franchini et al. (2011)</td>
</tr>
<tr>
<td>DNA Enhancer</td>
<td>Maize/teosinte Hopscotch</td>
<td>QTL positional cloning</td>
<td>No (polymorphic in teosinte)</td>
<td>Studer et al. (2011)</td>
</tr>
</tbody>
</table>

**Note:** MaLR, mammalian apparent LTR; QTL, quantitative trait locus.

### Text

Figure 1 shows the distribution of TE insertions that were exapted into enhancers. The TE insertion increases 2-fold the expression of the gene Tal1, another functional TE exaptation into an enhancer comes from maize, where a quantitative trait locus controlling stem branching was mapped to a *Hopscotch* transposon (Studer et al. 2011). The TE insertion increases 2-fold the expression of the gene *tb1* (teosinte branched 1, located over 60 kb away) and is a major factor in the branching pattern of domesticated maize. The TE got inserted into this location approximately 20,000 years ago, and this modified locus was selected by farmers for its ability to enhance stem branching. The evolution of transcriptional enhancers (Franchini et al. 2011, 2012) indicates that the +18-kb element acts as a modulator of another Tal1 enhancer located nearby at +19 kb (Smith et al. 2008).

Although the examples mentioned earlier are from ancient TE insertions, there are a few particularly convincing examples of recent TE insertions that were exapted into enhancers. Near the locus control region of the higher primate b-globin locus, there is an LTR of an ERV9 element that can act as an enhancer in cell culture assays and transgenic zebrafish (Ling et al. 2002; Pi et al. 2004). This ERV9 LTR is found in orangutans, humans, and close relatives but not other mammals (Ling et al. 2002). Recently, Pi et al. (2010) created bacterial artificial chromosomes (BACs) carrying the human b-globin locus with a floxed (flanked with loxP sites) ERV9 LTR allele and inserted the BACs into transgenic mice. Deletion of the ERV9 LTR by Cre recombinase caused a 50~80% drop in b-globin transcription in adult erythroid mouse cells, whereas interestingly, transcription of fetal y-globin gene was spuriously increased in adult erythroid cells (Pi et al. 2010). In addition, chromatin immunoprecipitation (ChIP) experiments with erythroid cells from transgenic mice showed that the ERV9 LTR is responsible for recruiting transcription factors (TFs) and RNA polymerase II to the b-globin promoter, and chromosome conformation capture (3C) assays with transgenic spleen erythroid cells showed that the ERV9 LTR loops with the b-globin promoter and that this looping depends, to a great extent, on the presence of the LTR (Pi et al. 2010). Thus, conversions of TE sequences into transcriptional enhancers have happened at various levels of the vertebrate phylogenetic tree (fig. 1).
when teosinte—the wild ancestor of maize—was domesticated 10,000 years ago (Studer et al. 2011).

**TE-Derived Insulators**

Insulators are DNA sequences that can act either as a barrier element preventing the spread of heterochromatin (i.e., condensed chromatin) from one locus to another or as an enhancer blocker preventing enhancers from spuriously influencing other loci (Gaszner and Felsenfeld 2006). In mammals, a well-known insulator-binding protein is the zinc-finger CCCTC-binding factor (CTCF), which participates in the organization of chromatin loops and in the creation of transcriptional domains (Phillips and Corces 2009). Bourque et al. (2008) found that approximately 33% of CTCF-binding elements present in mouse ES cells are derived from the B2 subfamily of SINEs, which is restricted to a few rodent groups. As expected, these CTCF-binding regions were not conserved in the human genome, suggesting that the chromatin architecture in ES cells may be different from that of other mammals (Bourque et al. 2008). Recently, Román et al. (2011) showed that another mouse-specific SINE subfamily, the B1, has insulator activity in cell culture assays and in transgenic zebrafish also via CTCF binding.

That TEs can become insulators of a nearby transcriptional unit (table 1) is best evidenced by the work of Lunyak et al. (2007) with the murine growth hormone (GH) locus. During the organogenesis of the pituitary, a boundary is established between open and closed chromatin domains 5 kb upstream of the GH locus. Interestingly, this boundary coincides with a B2 SINE, a retroposon restricted to mice, rats, and close relatives (fig. 1; Serdobova and Kramerov 1998; Churakov et al. 2010). Detailed studies showed that the B2 element is transcribed in both orientations by RNA polymerases II and III and that it can act as an enhancer blocker in cell culture. Notably, Lunyak et al. showed using BAC transgenesis that GH expression depends on the presence of the B2 element and that in its absence the GH locus remains transcriptionally silent, presumably because heterochromatin spreads in from neighboring loci. The results of Lunyak et al. are reminiscent of an earlier work by Willoughby et al. (2000) that found an Alu-rich region upstream of the human keratin-18 (K18) gene that has typical insulator properties such as chromatin barrier activity.
although not enhancer blocking activity) in transgene mouse analysis.

**TEs as a Source of New TFBS**

Promoters and enhancers work primarily by the binding of TF proteins, which ultimately interact with the basal transcriptional machinery. For over 20 years, TFBSs have been found within TE sequences (Brosius 2003). In these earlier studies, evidence for functionality of TE insertions and embedded TFBSs was usually assessed by reporter gene expression in transient transfection assays in cell culture and, more recently, by ChIP, which can reveal in vivo binding of TFs to TEs (for recent examples see table 3 and for older examples see Brosius 2003). It is important to note, however, that TF binding by itself does not necessarily mean that a TE should work as a regulatory region.

In the last years, ChIP coupled to microarray hybridization (ChIP-chip), high-throughput sequencing (ChIP-seq) or paired-end diTag (ChIP-PET) have been employed to determine the genome regions bound by TFs in an unbiased way (Farnham 2009). To estimate the prevalence of TF binding to TEs, Bourque et al. (2008) reassessed seven genome-wide ChIP studies of six different TFs and found that these factors frequently bind to sites embedded within TEs. Thus, for instance, approximately 20% of estrogen receptor α (ESR1) binding regions overlap SINEs of the MIR family, 24% of Oct4-Sox2 binding regions overlap endogenous retrovirus K (ERVK) repeats, and approximately 40% of p53 binding occur within ERV1 elements (Bourque et al. 2008). Similarly, Kunarso et al. (2010) found that, in human ES cells, 21% and 14.6% of binding sites for Oct4 and Nanog, respectively, reside in TEs and, interestingly, only approximately 5% of the binding regions for each factor are conserved in mouse ES cells, indicating a possible divergence in the transcriptional network between mammalian ES cells. Xie et al. (2010) found that the nearby presence of murine TEs carrying TFBSs correlates with increased mRNA expression in mouse in relation to human and bovine embryos. Interestingly, this correlation was significant only with TEs harboring two or more TFBSs, suggesting that the TEs with potential to influence gene expression are preferentially those that carry a full cis-regulatory module (Xie et al. 2010). These and other similar studies (Wang et al. 2007; Román et al. 2008, see table 1) led to the idea that TE insertion can rapidly change the regulatory landscape of mammalian genomes during evolution (Bourque 2009).

**TEs and the Birth of New Regulatory Regions**

As evolutionary time goes by, new TF-binding motifs are expected to appear in promoters or enhancers by random mutation. These new motifs can be eventually lost from the population, become fixed by genetic drift or, in case they have a large positive effect on fitness, might be kept by natural selection (Lynch 2007). New motifs may also replace pre-existent ones (turnover) by stabilizing selection, without significantly changing the expression of the gene (Ludwig 2002). In this context, TEs might contribute to the birth of new transcriptional regulatory elements using two alternative routes (fig. 3):

- A new TFBS appears as soon as a TE that inherently carries one or more functional motif(s) is inserted near a gene. In this case, the change in transcription levels is immediate, implying that some TEs may act as “jumping regulatory regions” and might rewire the network of genes controlled by a TF in a short evolutionary time (Bourque et al. 2008).
- New TFBSs appear by random mutation on a TE previously inserted in the vicinity of a transcriptional unit. In
In this case, TEs act as raw material for the appearance of novel active TFBSs, and, given the abundance of TEs, it is a mechanism that deserves serious consideration. It is also possible that some TE classes have “pre-sites” that only require few mutations to originate functional TFBSs.

It must be noted that regulatory regions such as enhancers carry complex arrays of TFBSs, making it unlikely that TEs might work as a full enhancer when inserted near a gene, at least in most cases. Thus, in both scenarios above, the first functional site(s) to appear may work as a “seed” for the appearance of additional functional sites nearby, eventually originating a regulatory block with several TFBSs that works as a bona fide enhancer or promoter (a process called “epistatic capture” by Emera and Wagner [2012]). For regulatory elements that appeared long ago and have changed significantly from their ancestral TEs, it is difficult to reconstruct what was the mechanism of exaptation, but recent exaptation events may be more informative. In a recent work, Emera and Wagner (2012) describe the exaptation of MER20 and MER68 elements as a prolactin (prl) alternative promoter with activity in endometrial cells of the ape uterus. A functional comparison of TE-derived prl promoter from apes with that of nonape primates—which possess the same MER insertions in prl but lacking promoter activity—revealed a ETS1 site that was present in the ancestral MER element, as well as new TFBSs that appeared in the ape lineage (Emera and Wagner 2012). Thus, the creation of an alternative promoter for prl in apes involved both a “ready-to-go” site (ETS1) and new sites that appeared subsequently (Emera and Wagner 2012).

### TE Insertion and Changes in Regulatory Function

Although the examples reviewed above suggest that TEs have contributed to the creation of functional regulatory regions, each example differs in the techniques used and in the depth of experimental evidence collected to assign function. Thus, we believe that it is necessary to compare the relative strength of the different experimental designs to properly evaluate the putative regulatory functions of TE-derived sequences. The ENCODE Project Consortium (2007) uses the idea of 1) biochemical function as the detection of a certain biochemical behavior of a particular sequence, for example, ChIP analyses showing that a TF can bind to a particular TE-derived sequence. A second concept, also used in the ENCODE project, is that of 2) regulatory role, defined as the transcriptional consequence of TE insertions on genes, for instance, showing that a particular TFBS present on a given TE can increase transcription of a nearby gene. Usually, these studies are done by transient transfection assays in cell culture with artificial reporter gene constructs harboring wild-type or mutated forms of TEs. A third level is 3) physiological function, understood as the higher-order physiological and/or morphological changes induced by TE-derived sequences and that may create variation or innovation. One example is the change in branching pattern in maize and teosinte due to

<table>
<thead>
<tr>
<th>TE</th>
<th>Mainly MIR and L1</th>
<th>L1-SINE</th>
<th>Many</th>
<th>Many</th>
<th>Many</th>
<th>CORE-SINE</th>
<th>AntsINE1</th>
<th>MER20 (DNA)</th>
<th>Many</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of TE Instances Found in Intergenic Regions</td>
<td>&gt;200 noncoding conserved instances in human genome</td>
<td>&gt;10,000 noncoding conserved instances in putative regulatory regions (GRMs)</td>
<td>&gt;33,000 conserved instances in putative regulatory regions</td>
<td>&gt;2,500 conserved instances in putative regulatory regions near endometrium-specific expressed genes</td>
<td>&gt;28,000 conserved instances in placentae</td>
<td>&gt;280,000 conserved in placental mammals</td>
<td>&gt;280,000 conserved in placental mammals</td>
<td>&gt;280,000 conserved in placental mammals</td>
<td>&gt;280,000 conserved in placental mammals</td>
</tr>
<tr>
<td>Evolutionary Conservation</td>
<td>Human and mouse</td>
<td>Tetrapods (degree of conservation not specified)</td>
<td>Primates, rodents, and dog</td>
<td>Human, rodents, and dog</td>
<td>Human, rodents, and dog</td>
<td>Marsupial and placental mammals</td>
<td>Marsupial and placental mammals</td>
<td>Marsupial and placental mammals</td>
<td>Marsupial and placental mammals</td>
</tr>
<tr>
<td>TE</td>
<td>TF Bound (Putative) Genes Controlled</td>
<td>Technique(s)</td>
<td>Cell Type(s)</td>
<td>References</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AluSx (primate SINE)</td>
<td>Vitamin D receptor (VDR)</td>
<td>Reporter assay in cell culture</td>
<td>NB4 (human myeloid line)</td>
<td>Gombart et al. (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AluS (primate SINE)</td>
<td>Retinoic acid receptor (RAR)</td>
<td>ChIP</td>
<td>SCC25 (human squamous cell carcinoma)</td>
<td>Laperriere et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTR10, MER61 (class I ERV LTR)</td>
<td>p53</td>
<td>ChIP, reporter assay in cell culture</td>
<td>HCT116 (human colorectal carcinoma)</td>
<td>Wang et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1 (murine SINE)</td>
<td>Dioxin receptor (AhR) and Slug</td>
<td>ChIP (overexpressed TFs), reporter assay in cell culture</td>
<td>Hepa-1 (mouse hepatoma)</td>
<td>Román et al. (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AluS (primate SINE), MIRb (SINE), L2A (LINE), others</td>
<td>Estrogen receptor α (ESR1)</td>
<td>Genome-wide ChIP, reporter assay in cell culture</td>
<td>MCF7 (human breast cancer)</td>
<td>Mason et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alu (primate SINE)</td>
<td>LXR, PPARα, RXR</td>
<td>Myeloperoxidase (MPO)</td>
<td>CV-1 (monkey kidney fibroblasts), transgenic mouse macrophages</td>
<td>Reynolds et al. (2006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AluS (primate SINE), MLT2b2 (ERV LTR)</td>
<td>NF-κB</td>
<td>Interferon λ1</td>
<td>Modified HEK-293 (human embryonic kidney)</td>
<td>Thomson et al. (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mostly MIR (ESR1); ERV1 (p53); ERVK (Oct4-Sox2) and murine B2 SINE (CTCF)</td>
<td>ESR1, p53, c-Myc, RELA, POUSF1 (Oct4), Sox2, CTCF</td>
<td>Genome-wide ChIP-PET and ChIP-seq</td>
<td>MCF7, HCT116, Burkitt's human lymphoma, human leukemia T cells, mouse ES cells</td>
<td>Bourque et al. (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alu (primate SINE), L2 (LINE)</td>
<td>REST</td>
<td>CACNA1A (calcium channel subunit), others</td>
<td>HeLa (human cervical cancer)</td>
<td>Johnson et al. (2006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainly ERV1 (LTR), others</td>
<td>Nanog, Oct4</td>
<td>Many, undetermined</td>
<td>mouse and human ES cells</td>
<td>Kunarso et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainly B2, B4 (SINE), ERVK</td>
<td>Nanog, Oct4, CTCF, others</td>
<td>Genome-wide ChIP-PET</td>
<td>Mouse zygotes, morulae, blastocysts</td>
<td>Xie et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERSB (DNA transposon)</td>
<td>Foxa1, p53, Sma2/4</td>
<td>ω-fetoprotein (AFP)</td>
<td>Mouse ES cells</td>
<td>Taube et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EnSpnN6_DR (DNA transposon)</td>
<td>p53</td>
<td>trim8, others</td>
<td>Reporter assays in human (HeLa) cell culture</td>
<td>Zebrfish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Recent Experimental Evidence for Gene Regulation by TFBSs Present in TEs.
the insertion of a TE-derived enhancer, which increased the morphological variation of teosinte and was later selected by early farmers (Studer et al. 2011). The ultimate functional demonstration is provided by the 4) fitness effect of a TE insertion, which needs an evaluation of the impact of the TE-derived transcriptional regulatory change on the viability and reproductive success (i.e., the fitness) of the organism.

These four functional levels are obviously related, but one should not assume that a TE-mediated effect at a lower level would translate on an impact at a higher level, as the fitness effect of the TE-mediated perturbation may be near neutral. This is a significant concern because it is expected from population genetics that mutations in the genomes of species with small effective population sizes (such as vertebrates and higher plants) will only be selected against if the fitness disadvantage of the mutation is very high (Lynch and Conery 2003; Lynch 2007). Genome-wide studies certainly indicate a potential for TE expansions in particular lineages to change or “rewire” transcriptional networks, but the physiological significance of this rewiring, if any, is still unknown. In this regard, the work of Naito et al. (2009) analyzing the ongoing colonization of a rice strain by the mPing DNA transposon offers a note of caution. mPing mobilization generates approximately 40 new TE insertions per plant per generation, providing a unique opportunity to study the consequences of a TE burst as it happens in a complex organism. Although few inserts (~1%) happen within exons, minimizing deleterious effects, many mPing insertions occur close to genes, and quantitative comparisons between the transcriptomes of different strains revealed that most insertions either upregulate or have no effect on the transcription of nearby genes (Naito et al. 2009). Notably, in some cases, the insertions confer inducibility to cold and high salinity. The work by Naito et al. (2009) shows that a complex genome can be quite resistant to massive TE colonization without an overt loss in fitness and, although many insertions influence transcription of nearby genes, natural selection has had no time to act and no “real” function can be assigned to
them. This illustrates the difficulty of assuming an important regulatory function for the transcriptional influence of recent TE insertions, as suggested in the genome-wide studies listed in table 3.

Although the highest levels of functionality (3 and 4 above) might be considered the most relevant, they are seldom addressed experimentally due to intrinsic technical limitations of most available models. A good proxy for functionality is, nevertheless, evolutionary conservation. Thus, an important fitness effect of some TE-derived mammalian enhancers (Bejerano et al. 2006; Santangelo et al. 2007; Sasaki et al. 2008; Smith et al. 2008; Franchini et al. 2011) is clearly inferred from their deep conservation among mammals, indicating strong purifying selection for over 100 My. In addition, most of these works use transgenic mice to show that these elements display enhancer activity compatible with the expression of nearby genes. The mere existence of thousands of conserved-noncoding elements derived from ancient TE insertions (Lowe et al. 2007; Lindblad-Toh et al. 2011; Lynch et al. 2011) suggests that these well-studied cases are but the tip of an iceberg. Loss of function studies of TE-derived enhancers in knockout mice have not been performed yet and would provide an ultimate experimental proof of their importance in fitness. It must be borne in mind, however, that expression of genes may be controlled by several enhancers with overlapping transcriptional activity, as in the case of TE-derived Pomer enhancers nPE1 and nPE2 (de Souza et al. 2005; Franchini et al. 2011) and that even enhancer knockout studies may fail to show a clear fitness effect (Ahituv et al. 2007), unless the experimental mutants are raised in challenging environmental conditions (Frankel et al. 2010). As for TE-derived promoters, several of them are also clear examples of exaptation, when transcription starts exclusively within a TE-derived sequence (see Cohen et al. 2009 for a review).

In addition to the identification of conserved sequences, it will be interesting to explore the overlap between TEs and active enhancers identified by chromatin marks (e.g., H3K27ac and H3K4me3) or transcription cofactor binding (e.g., p300), which are a better proxy to functionality than TF-binding alone. Recently, Huda et al. (2011) have done just that using data from two human cell lines analyzed by the ENCODE project and were able to find hundreds of regions with epigenetic markers typical for active enhancers that overlap TE insertions. Interestingly, the largest proportion of TEs with enhancer marks are MIR-SINE and LINE subfamilies (Huda et al. 2011), two ancient mammalian TE groups, indicating that exaptation may have happened before the mammalian radiation. Another work linking ancient TE insertions with epigenetic marks and cofactors was done by Lynch et al. (2011), in this case studying the contribution of the DNA-transposon MER20 to the differentiation of the endometrium during mammalian pregnancy. They found that the neighborhood of genes expressed in progesterone-induced, differentiating human endometrium is enriched in MER20 instances that are conserved in placental mammals. With an array of ChIP assays, selected MER20 instances could be roughly divided into two groups, an “enhancer-repressor type,” which binds p300, progesterone receptor, and others, whereas the second group, the “insulator-type,” bound CTCF, USF1, and other insulator proteins (Lynch et al. 2011). Although detailed functional studies are still lacking, the evolutionary conservation and the ChiP data indicate that MER20 exaptation had a role in reprogramming the endometrial gene expression network when pregnancy evolved in placentals, over 100 Ma (Lynch et al. 2011), consistent with the hypothesis that TE mobilization played a role in the early evolution of mammals (Okada et al. 2010).

Detecting Exaptation Events by Sequence Conservation

As discussed earlier, the identification of evolutionary conserved noncoding elements (CNEs) derived from TEs has been successfully used to identify candidates for exaptation events in the mammalian lineage. Although this may seem straightforward, there are some limitations involved. First, it can be difficult to recognize an exapted element if the sequence has evolved too fast since the TE insertion. If a TE-derived sequence retains less than 70% nucleotide identity with the original TE, it is very possible that it will not be recognized by alignment algorithms such as BLASTN (Altschul et al. 1990) or BLAT (Kent 2002). Second, the exapted area of the TE may be too small. In general, searches at the genome level use threshold levels in the length of the overlapping area and retain only elements that cover the majority of the conserved element as a TE (Lowe and Haussler 2012). It has been observed that some regions within TEs are more likely to be exapted than others, either because TE insertion is generally incomplete (e.g., LINEs tend to be truncated at the 5′-region) or because some TE sequences may carry functional binding sites for TFs (Lowe et al. 2007; Lowe and Haussler 2012). Third, the TE that gave origin to a regulatory sequence may have been inactivated long ago, making it very difficult to identify exaptation. Current estimates indicate that repeat families can be recognized if they are younger than 200 My (Brosius 2003) suggesting that some exaptation events cannot be recognized as such. However, several reports have reported ancient exaptation events (Bejerano et al. 2006; Xie et al. 2006; Santangelo et al. 2007). In all these cases, the identification was facilitated because close relatives of the TE that originated the conserved noncoding sequence is active in some other organisms such as the LF-SINE in the coelacanth (Bejerano et al. 2006), the MAR1 in the opossum (Santangelo et al. 2007), and the SINE3 in the zebrafish (Xie et al. 2006). The fact that the element was still active in other organisms provided many instances of the element that allowed the generation of consensus sequences to improve detection based on sequence similarity.

The limitations in the identification of TE exaptation events are exemplified by Lowe and Haussler (2012) who worked with aligned sequences of 29 mammals and compared them with the RepeatMasker annotation in the human genome. Although they show that at least 11% of CNEs in the human genome have most probably originated from TEs, they have only identified 133 exaptation events that
predate the speciation of fishy species from the human lineage from a total set of 284,857 exapted elements (Lowe and Haussler 2012). These 133 exapted elements are identifiable because they have evolved in a very slow rate and are large enough that they can still be aligned to the TE consensus sequences. On the other hand, in this approach, sequences that are not annotated as human repetitive elements are missed. To overcome this difficulty, other researchers (Xie et al. 2006) compared all CNEs not overlapping with known human TEs with the whole set of repeats deposited in the RepBase database (Jurka et al. 2005) that includes TEs from all kinds of organisms. This approach allowed them to discover a family of CNEs derived from a SINE3, an element still active in the zebrafish genome. Thus, the difficulties in identifying ancient exaptation events, together with the problematic identification of lineage or species-specific exaptation events, suggest that the impact of TEs in the evolution of gene regulatory networks in vertebrates is underestimated.

Evaluating the Function of Recent TE Insertions

Although evolutionary conservation is harder to be employed for sequences that are shared between few species, recent TE exaptation might still be identified by comparative genomics provided a good number of sequences of related species is available. For Alu elements, for instance, a possibility is to use "phylogenetic shadowing," which involves the comparison of several primate sequences (Bofelli et al. 2003). Another possibility is to infer selective pressure within human populations using polymorphism variation data from the increasing number of sequenced human genomes, as has been done to evaluate evolutionary pressure in nonconserved parts of the human genome (Ward and Kellis 2012).

One constraint in the study of primate-specific TEs is that the generation of transgenic or knockout animals is unrealistic. An alternative is to study transgenic mouse models, ideally using large BAC transgenes to better simulate the genomic environment of the locus under study. Using BAC constructs, Lunyak et al. (2007) and Pi et al. (2010) provided very convincing evidence for exaptation of two recent TE insertions that act as a barrier element and an enhancer, respectively. In another work, yeast artificial chromosomes (YACs) were engineered to harbor a LINE 1-derived enhancer of the human apolipoprotein a (apoap) gene flanked by cre-loxP sites (Yang et al. 1998; Huby et al. 2003). Deletion of the enhancer in the context of the 270-kb YAC construct in transgenic mice indicated that it was responsible for approximately 30% of the transcriptional activity of the human gene in the mouse liver. The observed effect was relatively small considering previous results obtained in cell culture, showing the importance of confirming transient transfection assays in more physiological contexts (Huby et al. 2003).

Over large evolutionary distances, for example, zebrafish and mice, conserved enhancers may not exhibit the same regulatory activity (see, e.g., Ariza-Cosano et al. 2012), but the examples above indicate that mice are suitable models to analyze TE exaptation events in the primate and human lineage. These few examples show the feasibility of transgenic mice as tools to explore the functional significance of recent TE insertions in mammalian genomes. In addition, traditional transient transgene assays in cell lines, which is a staple in most studies testing the functionality of TFBs in TEs, might be complemented by more sophisticated cell culture experiments. For example, Smith et al. (2008) employed stable transfection assays, which presumably better simulates the genomic environment of the locus under study, to determine whether a TE was exapted as an hematopoietic enhancer of the Tal1 gene. Taking advantage of the possibility of obtaining blood progenitors from ES cells, they further showed the enhancing function of the TE-derived sequence by testing transgenes inserted by homologous recombination into ES followed by differentiation in culture. Although Smith et al. (2008) studied an ancient mammalian TE exaptation, the techniques they used might also be employed to functionally test recent TE insertions. Finally, ChIP studies with several TFs and transcriptional cofactors might also provide a strong indication of recent exaptation, as was done by Lynch et al. (2011) to show functionality of ancient TE insertions.

Final Remarks

Although the results stemming from genome-wide studies are still difficult to be fully interpreted, the available evidence from gene-centered studies indicates that TEs provide essential raw material for the process of regulatory evolution in mammals and other organisms. Even if, as predicted by population genetics, the majority of the transcriptional influence exerted by recent TE insertions in mammals and other complex organisms is nearly neutral from an evolutionary point of view, they must definitely contribute to the array of regulatory variation that allows the exploration of the mutational "neutral network" that drives molecular and morphological evolution (Wagner 2008). Thus, today may be a minor effect of a TE insertion, for instance, a new TFB leading to a small increase in transcription or a minor alternative promoter that drives expression of a gene to a new tissue, may constitute the first step in a major regulatory reshaping and/or in the co-option of that gene to a new crucial function. Although recent TE exaptation is difficult to test experimentally, the remnants of ancient exaptation events show that this area of research bears great promise for illuminating the mechanisms of genome evolution.

While this manuscript was on final revision and proof stage, critiques on the ENCODE project by Niu and Jiang (2013), Graur et al (2013), and Doolittle (2013) have appeared that, together with a previous comment by Eddy (2012), essentially agree with us that functionality should not be lightly attributed to biochemical activities on the genome, including transposable elements, without proper experimental evidence.

Acknowledgments

The authors thank Daniela Orquera for help with figures. This work was supported by NIH grant DK068400 (M.R.), Agencia Nacional de Promoción Científica y Tecnológica, Argentina.
References


Exaptation of TEs into Novel Cis-Regulatory Elements. doi:10.1093/molbev/mst045