OPINION

Transposable elements and the evolution of regulatory networks

Cédric Feschotte

Abstract | The control and coordination of eukaryotic gene expression rely on transcriptional and post-transcriptional regulatory networks. Although progress has been made in mapping the components and deciphering the function of these networks, the mechanisms by which such intricate circuits originate and evolve remain poorly understood. Here I revisit and expand earlier models and propose that genomic repeats, and in particular transposable elements, have been a rich source of material for the assembly and tinkering of eukaryotic gene regulatory systems.

It has been known for some time that eukaryotic genomes, with rare exceptions, are replete with interspersed repetitive DNA[1]. Large-scale DNA sequencing has revealed that most of the repetitive DNA is derived from the activity of transposable elements (TEs) — sequences that are able to move and replicate within the genome. TEs use different replicative strategies, which involve either RNA intermediates (class 1 or retrotransposons) or DNA intermediates (class 2 or DNA transposons)[2]. The broad distribution of all major TE classes across the eukaryotic tree of life indicates that they are long-standing residents of eukaryotic genomes[2]. Unlike other lasting components of the genome, one needs not bestow TEs with adaptive value to account for their evolutionary persistence. Theoretical considerations and empirical studies show that TEs are best viewed as genomic parasites, which essentially owe their survival to their ability to replicate faster than the host that carries them[3,4]. This conjecture, also known as the selfish DNA theory, seems sufficient to explain the maintenance of TEs over a long evolutionary time as well as the wide variations in the amount, diversity and chromosomal location of TEs that is observed between or even sometimes within species[3,5]. In spite of and, to some extent, because of this selfish and parasitic nature, the movement and accumulation of TEs have exerted a strong influence on the evolutionary trajectory of their hosts[3–6]. Here I review recent discoveries supporting early theories that postulate that TEs have had a key role in the evolution of eukaryotic gene regulation. Specifically, I explore the properties of TEs that might facilitate their recruitment as building blocks for the assembly of a diversity of systems to regulate and coordinate eukaryotic gene expression.

> “In spite of and, to some extent, because of this selfish and parasitic nature, the movement and accumulation of transposable elements have exerted a strong influence on the evolutionary trajectory of their hosts.”

Life after death: TE exaptation

The co-option of TEs (or exaptation[7]) to serve cellular function has long been recognized[8–10]. But in recent years, the ability to align large amounts of human genomic sequences to their orthologous regions in widely diverged mammals has provided an opportunity to estimate the amplitude of TE exaptation by revealing fixed TE sequences that have been under functional constraint for an extended period of evolutionary time. A pioneering study[11] comparing a sample of human–mouse orthologous sequences suggested that a substantial fraction of ancestral repeats (that were inserted before the eutherian radiation) have been subject to strong selective constraint since at least the divergence of humans and mice, implying that mammalian TEs frequently acquire beneficial functions for their host.

Recently, this comparative genomics approach was scaled up[12], unveiling at least 10,000 TE fragments in the human genome that have evolved under strong purifying selection throughout the eutherian radiation. Furthermore, comparisons of the genomes of a marsupial (the opossum) with several eutherian species (human, dog, mouse and rat) revealed that at least 16% of eutherian-specific conserved non-coding elements (CNEs) were derived from various kinds of TEs[13]. In addition, sensitive sequence-similarity searches uncovered thousands of deeply conserved human CNEs (many of them pre-dating the mammalian radiation). These CNEs included a number of so-called ultraconserved elements, which can be clustered into hundreds of families, suggesting a distant TE origin[14–16]. So far, only a handful of these CNE families could be unambiguously traced back to TE families[15–19]. Interestingly, three of these are ancient families of short interspersed nuclear elements (SINEs) that had previously escaped detection. The apparent enrichment of exapted SINEs might reflect a proclivity of these elements to be recruited for cellular function. Alternatively, it might simply mirror their preponderance in vertebrate genomes and\ or the fact that their short size and distinctive sequence signatures make them more readily identifiable as fossil SINE families. Dozens of other CNE families have weak, but significant, similarity to known TEs[15], implying that many broadly conserved TE families remain to be characterized.

Individual examples of selectively beneficial TE insertions with apparent regulatory functions have been described

in non-mammalian species, especially in *Drosophila melanogaster*[20–22]. However, there has been no attempt to measure the extent of TE exaptation at a genome-wide scale in non-mammalian lineages. Cross-species genome alignments have revealed an abundance of CNEs in species as diverse as dipteran insects, nematodes, yeasts, grasses and crucifers[23,24]. Unfortunately, the rapid turnover and decay of TEs in these lineages make it extremely difficult, if it is possible at all, to recognize ancient elements and assess their contribution to deeply conserved non-coding sequences. But comparative analysis of more closely related species and improved detection and annotation of TEs might be enlightening.

TEs as a supply of regulatory elements

A large body of studies has illustrated the myriad ways by which TEs can directly influence the regulation of nearby gene expression, both at the transcriptional and post-transcriptional levels (FIG. 1). Initially, these mechanisms were discovered in the laboratory through the analysis of mutations caused by individual TE insertions. But it is now clear that the same changes in gene structure and expression have occurred in the distant past and have been preserved by natural selection[9,25,26].

In a seminal study, Jordan *et al.* reported that nearly 25% of experimentally characterized human promoters contain TE-derived sequences, including empirically defined *cis*-regulatory elements[27]. Further genome-scale analyses showed that many promoters and polyadenylation signals in human and mouse genes are derived from primate-specific and rodent-specific TEs, respectively, for examples see REFS 26,28. Another study found that one-quarter of the DNase I hypersensitive sites identified in human CD4$^+$ T cells overlap with annotated TEs, suggesting that these TEs harbour *cis*-regulatory sequences. Interestingly, the vast majority of these elements are not deeply conserved, but are primate specific. Hence, insertion of these TEs probably contributed to the establishment of lineage-specific patterns of gene expression[29]. Further evidence that TEs commonly acquire regulatory function comes from TE fragments that are deeply conserved among mammals. First, these elements tend to cluster around genes that are involved in development and transcriptional regulation[12]. Second, they are over-represented within predicted *cis*-regulatory modules[30], that is, genomic segments containing dense arrays of transcription factor binding sites (TFBSs). Third, approximately one-fifth of eutherian-specific CNEs, thousands of which are derived from ancient TEs, overlap with DNase I hypersensitive sites that are mapped experimentally in human lymphocytes, which implies that they provide promoter sequences or binding sites for regulatory proteins[13]. Finally, there is a growing number of individual cases of highly conserved TEs documented to act as transcriptional enhancers[16,19] or as alternatively spliced exons that introduce premature stop codons and trigger nonsense mediated decay, thereby contributing to mRNA homeostasis in the cell[16,31] (FIG. 1). Together, these data suggest that TEs have been a profuse source of new regulatory sequences throughout mammalian evolution.

What makes TEs such a rich stock of promoter and *cis*-regulatory elements? One simple explanation is that the pervasive accumulation of TEs creates raw sequence material from which *cis*-regulatory elements evolve *de novo* by point mutations. Most *cis*-regulatory elements, such as TFBSs, tend to be short and degenerate in sequence[32]. Thus, it is conceivable that decaying TE sequences provide an abundant material from which *cis*-regulatory elements emerge *de novo*, through the introduction of a single or a few point mutations — several examples of this have been reported[9,33–35]. Another non-mutually

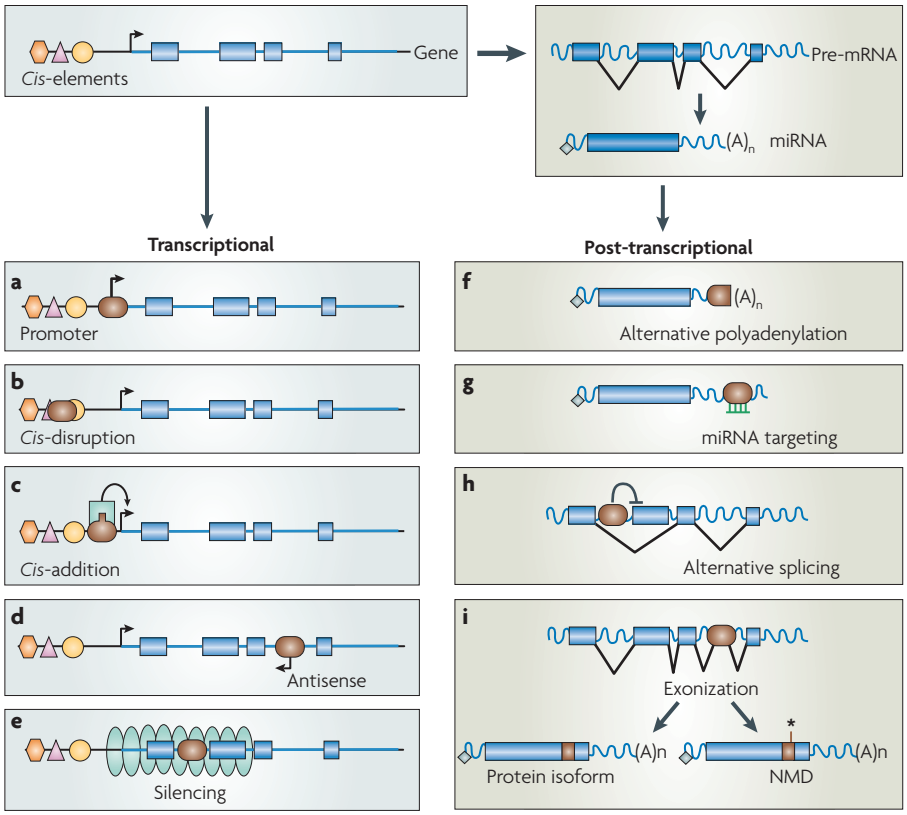

Figure 1 | **Transposable elements can influence gene expression in many ways.** At the transcriptional level, a transposable element (TE) (shown in brown) that has inserted upstream of a gene can insert promoter sequences and introduce an alternative transcription start site (**a**), disrupt existing *cis*-regulatory element or elements (**b**), or introduce a new *cis* element such as a transcription factor binding site (**c**). In addition, a TE that has inserted within an intron can drive antisense transcription and potentially interfere with sense transcription (**d**). Finally, a TE can serve as a nucleation centre for the formation of heterochromatin (green ovals), potentially silencing the transcription of an adjacent gene or genes (**e**). At the post-transcriptional level, a TE that has inserted in the 3′ UTR of a gene can introduce an alternative polyadenylation site (**f**), a binding site for a microRNA (**g**) or for an RNA-binding protein (not shown). A TE that has inserted within an intron can interfere with the normal splicing pattern of a pre-mRNA (**h**), provoking various forms of alternative splicing (for example, intron retention and exon skipping). A TE that has inserted within an intron and contains cryptic splice sites can be incorporated (exonized) as an alternative exon (**i**). This can result in the translation of a new protein isoform, or in the destabilization or degradation of the mRNA by the nonsense-mediated decay (NMD) pathway, especially if the exonized TE introduces a premature stop codon (represented by an asterisk).

exclusive scenario is that *cis*-regulatory elements pre-exist within the TE at the time of its insertion and are co-opted either immediately upon insertion or after modifications of the surrounding environment. A wide variety of regulatory elements have been identified in active TEs or reconstructed consensus sequence of active TEs. These include signals that are normally used by TEs to control their own expression (for example, basal promoters for RNA polymerase II or III, enhancers, insulators, splice sites and polyadenylation signals) and a plethora of TFBSs[36–39]. Many empirical studies have demonstrated how such 'ready to use' *cis*-elements can be incorporated into the 'natural' regulatory apparatus of adjacent genes[9,25,35,40–42], including microRNA (miRNA) genes[43].

TE wiring of genetic networks

Regardless of whether the regulatory elements arise *de novo* by a few mutations or are pre-existing within TE sequences, the dispersal of expanding TE families throughout genomes potentially allows the same regulatory motif or motifs to be recruited at many chromosomal locations, drawing multiple genes into the same regulatory network (FIG. 2). This model, first proposed decades ago by Britten and Davidson[44,45], has recently gained experimental support (for example, REFS 38,46). In a recent study in the human genome, Wang *et al.*[39] found that a set of closely related families of long terminal repeat (LTR) elements that are affiliated to class I endogenous retroviruses have dispersed more than 1,500 near-perfect binding sites for the master regulatory factor p53. These sites encompass 30% of all p53 binding sites that were mapped using genome-wide chromatin immunoprecipitation analysis. In five individual cases that were examined further, the p53 binding site within the LTR could be directly associated with p53-dependent transcriptional activation of the closest adjacent gene[39]. Interestingly, all the LTR families that contain p53 binding sites are primate specific and the p53 binding sites were apparently present at the time of their insertion. These results strongly suggest that the dispersion of the LTR elements promoted the assembly of a primate-specific transcriptional network of p53-regulated genes, similar to that shown in FIG. 2a.

Britten and Davidson had initially formulated their gene-battery model as operating at the RNA level, invoking the co-transcription of *cis*-elements

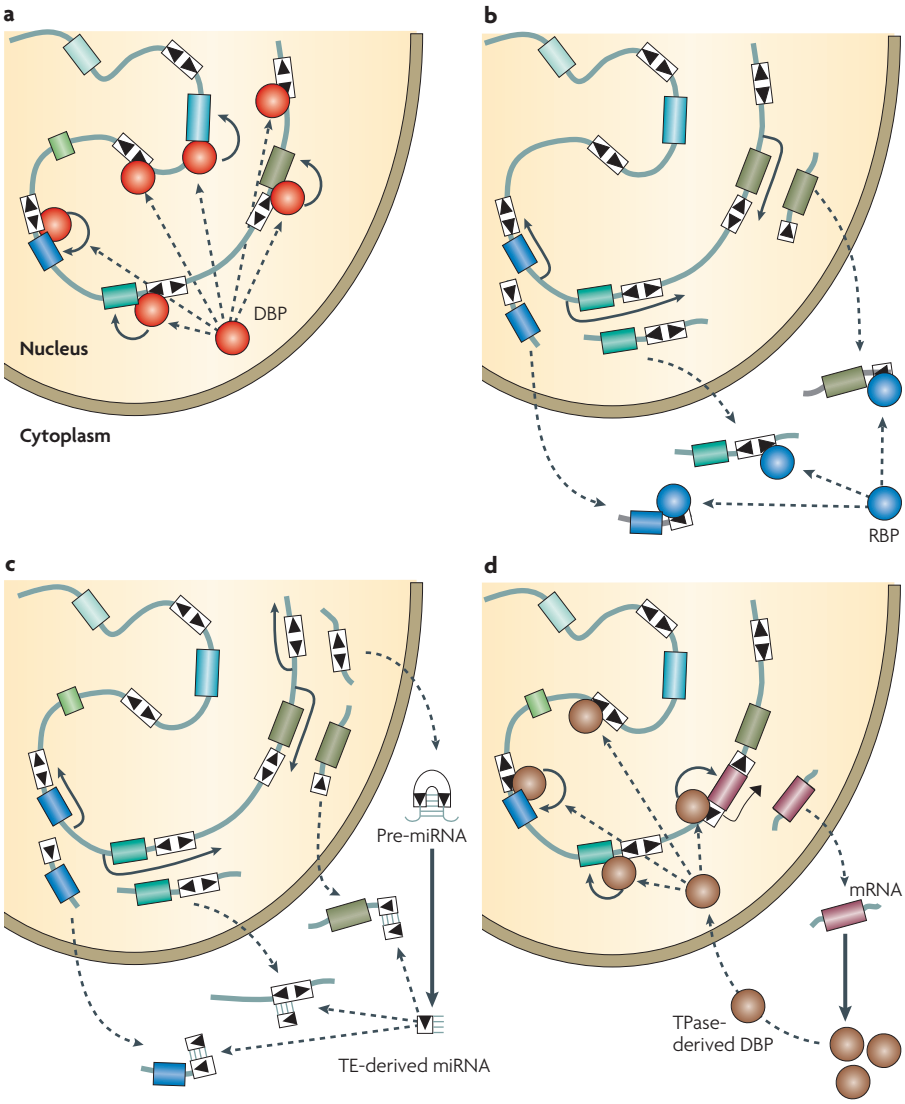

Figure 2 | **Building regulatory systems with transposable elements.** A family of DNA transposons is shown, with its multiple copies (white boxes) delimited by terminal inverted repeats (black triangles) and interspersed with genes (coloured boxes) in the genome. For panels **a** and **b**, the transposable element (TE) family could be also a retrotransposon family. **a** | Wiring of a transcriptional regulatory network by TE-derived *cis*-elements. A binding site for a DNA binding protein (DBP) has been dispersed throughout the genome as part of the TE. If the DBP is a transcription factor, its binding to a TE adjacent to a gene might influence the expression of that gene (solid arrows). Multiple genes are brought simultaneously under the control of the transcription factor through their association with different copies of the same TE family. **b** | Wiring of a post-transcriptional regulatory network by TE-derived *cis*-elements. Several TE copies are co-transcribed along with their neighbouring gene, resulting in the production of different mRNAs containing similar TEs (solid arrows). If the TE contains a binding site for an RNA-binding protein (RBP), it might engage the different mRNAs in the same post-transcriptional pathway of gene regulation. **c** | *De novo* assembly of a microRNA (miRNA) network from a TE family. This model combines the idea of TE and host gene co-transcription, as described in **b**, with the origin of a miRNA precursor containing a TE of the same family. This precursor could arise by transcription and intramolecular folding of a TE with a nearly perfect palindromic structure (for example, a miniature inverted repeat transposable element; MITE). The resulting double-stranded RNA could then be processed into a mature miRNA. The resulting TE-derived miRNA can pair with complementary TE sequences that are embedded within the 3′ UTR of co-transcribed mRNAs. **d** | *De novo* assembly of *cis* and *trans* components of a transcriptional network from a DNA transposon family. In this model, the DBP is derived from a transposase (TPase), and therefore has the potential to bind to a network of sites previously distributed around the genome by related TEs. If the TPase-derived DBP has transcription-factor activity, it might regulate the expression of genes located in proximity to a binding site embedded within a related TE, including its own.

along with the genes that they control[44]. A recent study in the trypanosomatid *Leishmania major*[47] brings support to the idea that the same TE family can be recruited at a genome-wide scale for post-transcriptional regulation (FIG. 2b). In trypanosomatids, most of the protein-coding genes are co-transcribed as large polycistronic transcripts. Individual gene regulation occurs predominantly at the post-transcriptional level, and sequences that are located in 3′ UTRs are known to be important for this process. In *L. major*, almost all of 1,000 copies of *LmSIDER2*, an extinct family of retroposons, are located in the 3′ UTRs of predicted mRNAs[47], a strikingly biased distribution suggesting a global function in post-transcriptional regulation. Consistent with this hypothesis, experimental introduction of an *LmSIDER2* copy in the 3′ UTR of a reporter gene decreased the stability of the resulting mRNA *in vivo*. Furthermore, microarray analyses revealed that *L. major* mRNAs containing *LmSIDER2* in their 3′ UTR are generally expressed at lower levels than other mRNAs[47]. Together, these data support the idea that this TE family has been recruited at the whole-genome level to modulate post-transcriptionally the expression of hundreds of genes.

In addition to donating new *cis*-elements and participating in the *de novo* assembly of regulatory networks, TEs might also contribute to the rewiring of pre-established networks through their movement and through the genomic rearrangements they provoke (BOX 1). It is widely believed that the tinkering and reorganization of pre-existing networks is a prominent mode of regulatory evolution[32,48,49].

Box 1 | **TE-mediated tinkering of *cis*-regulatory networks**

In addition to donating *cis*-elements and creating new regulatory networks, the movement and accumulation of transposable elements (TEs) are likely to participate in the rewiring of pre-established regulatory networks. First, TE insertions can disrupt and effectively eliminate *cis*-regulatory elements, thereby removing some genes from an existing network. Examples of TE insertions altering gene regulation have been described in the context of deleterious mutations causing disease (for example, REF. 98) or mutant alleles recovered in the laboratory[99]. The depletion of TEs in certain regions of the human genome that are enriched in regulatory sequences, such as *HOX* gene clusters and other transposon-free regions[100], attests to the deleterious nature of TE insertions in genomic environments that contain a high density of *cis*-regulatory elements. However, like any mutational event, it is conceivable that these disruptive insertions might be beneficial under some circumstances. Potentially adaptive TE insertions that disrupt promoter function have been identified at the heat-shock protein 70 (*Hsp70*) locus in natural populations of *Drosophila melanogaster*[20,101]. A screen for naturally occurring *P* elements within *Hsp* promoters recovered over 200 independent insertions, suggesting they are hot spots for *P* element insertions[102]. Many of these insertions are present at a high frequency in populations, and some are associated with a decrease in *Hsp* expression and a reduced thermotolerance, suggesting that these insertions are phenotypically consequential[102].

A less destructive route for TEs to modify an existing gene network is through TE-mediated shuffling and duplication of *cis*-regulatory elements[49]. It is well established that TEs can promote the mobilization, rearrangement and duplication of host sequences through various mechanisms, including recombination between TE copies, aberrant transposition events and transduction[5,6]. Thus, TEs have the potential to shuffle regulatory sequence information into new genomic contexts. A concrete example is the acquisition of a functional binding site for the nuclear factor of activated T-cells (NFAT) transcription factor that is required for transcriptional activation of the interferon-γ gene in human lymphocytes[103]. The intact NFAT binding site was introduced into the promoter as part of a short DNA segment that was co-mobilized by an *Alu* element inserted 22–34 million years ago. Furthermore, subsequent nucleotide substitutions next to the NFAT site created another transcription factor binding site (this time for nuclear factor-kappa B) that remains polymorphic in human populations[103].

The preferential insertion and accumulation of some TEs, notably short interspersed nuclear elements and DNA transposons, into the vicinity of genes might further enhance the scrambling of *cis*-regulatory elements. In addition, the excision of cut-and-paste DNA transposons is often imprecise, leaving behind small stretches of sequences (footprints) or re-arranging flanking host sequences, which might offer yet another mechanism for generating subtle alterations of adjacent regulatory sequences. This is well illustrated by allelic series of regulatory mutations, with a range of pigmentation phenotypes, that have been recovered in the laboratory following aberrant transposition and imprecise excision events of the *Tam3* transposon in the *nivea* promoter of snapdragon[104] and of *Tol2* in the promoter of the tyrosinase gene of medaka fish[105].

TEs as miRNAs and their targets

It is now evident that non-coding RNAs are important players in the regulation of eukaryotic gene expression[50]. Several classes of small regulatory RNA, including miRNAs, small interfering RNAs (siRNA), repeat-associated small interfering RNAs (rasiRNAs) and piwi-interacting RNAs (piRNAs) — collectively referred to hereafter as smRNA — use partially overlapping pathways that are akin to RNA interference (RNAi) to silence gene expression by degradation or by translational inhibition of mRNAs containing complementary sites. Thus, the logic of post-transcriptional regulation by smRNAs, whereby a single smRNA species can *trans*-regulate multiple genes through recognition of shared *cis*-elements, is similar to the logic of transcriptional regulation by transcription factors[51]. In addition, there is evidence that some smRNAs are able to mediate homology-dependent transcriptional silencing and participate in the nucleation of heterochromatin[52,53].

Ever since their discovery, the relationship of piRNAs, siRNAs and rasiRNAs to TEs has been apparent. Indeed, the natural, and presumably ancestral, function of these smRNAs is to silence invasive DNA such as viruses and TEs[53,54]. There are now numerous examples illustrating how these genome defence systems and the epigenetic marks that are deposited to silence TEs have been co-opted to control adjacent gene expression[35,53].

The evolutionary origins of miRNAs remain more obscure. Although new miRNA genes can arise through duplication of existing miRNA, it appears that the bulk of miRNAs have originated from sequences that did not previously encode miRNA[51]. One model proposes that new miRNAs arise from pre-existing hairpin structures in the genome that are fortuitously transcribed[55,51]. The influence of TEs in the origin, biogenesis and mode of action of miRNA is increasingly being recognized. Several mammalian miRNA precursors have been found to contain or be derived from TE sequences[56]. Likewise, a substantial amount of predicted miRNA targets map within members of the same TE families[57,58], again pointing at a model whereby large sets of *cis*-regulatory sequences have been seeded by transposition (FIG. 2c). The most recent count[57] shows that 55 out of 452 (12%) experimentally characterized human miRNA genes originated from TEs. Although this proportion seems lower than expected

in relation to the space occupied by TEs in the human genome (~48%), it is a minimal estimation because many of the currently known miRNAs have deeper evolutionary origins than the TEs that are recognizable in the human genome. Also, many uncharacterized miRNAs probably remain to be identified. For example, the same study[57] computationally predicted an additional 85 likely miRNA precursors derived from transcribed TEs.

Several classes and families of TEs show far more overlap with miRNA genes than is expected on the basis of their relative frequency in the genome[57]. This observation suggests that certain TE families possess characteristics that make them prone to give rise to miRNA. For instance, many miniature inverted repeat transposable elements (MITEs) have a palindromic structure, with terminal inverted repeats (TIRs) joined by little or no spacer DNA[59]. Transcription of MITEs is not required for transposition, but because they preferentially integrate in the non-coding portion or the immediate vicinity of genes, MITEs are frequently transcribed[59]. Following transcription, intramolecular folding of the MITE TIR sequences would produce RNA hairpins that, in principle, could be processed into siRNA[60]. Such MITE-derived siRNAs could then act in *trans* to mediate the silencing of multiple genes that are associated with related MITE sequences, providing a stepping-stone towards the emergence of a typical miRNA regulatory circuit (FIG. 2c).

In support of this model, it was recently established[61] that *mir-548*, a small family of human miRNAs, is derived from *Made1* TEs that are spuriously transcribed from adjacent promoters. *Made1* is an anthropoid-specific family of MITEs with an almost perfect palindromic structure. The mRNA targets of *mir-548* have not been defined experimentally, but bioinformatic predictions revealed over 3,500 human genes with putative target sites for *mir-548* (REF. 61). Interestingly, a subset of the predicted *mir-548* targets are also derived from *Made1* sequences that previously inserted within or close to the 3′ UTRs of genes. Some of these *Made1*-containing transcripts are down-regulated in colorectal cancer tissues, where *mir-548* is upregulated, and they fall within the same functional categories, consistent with the idea that they belong to a network of *mir-548*-regulated genes assembled through the past propagation of *Made1* (REF. 61).

Transposases recycled into regulators

The evolution of complex multicellular organisms in several branches of the tree of life was accompanied, and perhaps facilitated, by an expansion and diversification of transcription factors (TFs)[32,48,49,51,62]. It is thought that the emergence of new TFs allowed for the elaboration of increasingly complex networks of genes wired by *cis*-elements that are recognized by different sets of TFs[48,49,62]. Gene duplication and domain shuffling are well-established mechanisms contributing to the emergence of new regulatory proteins[51,62]. In the following sections, I argue that DNA transposons and their cognate transposases (TPases) are another significant, but largely underappreciated, source of the basic components that are necessary for the co-assembly of new TFs and their DNA targets (FIG. 2d).

> “…transposases continue to stand out as a recurrent supply of new proteins in diverse organisms.”

Recurrent domestication of transposases. A particular form of TE exaptation, also known as domestication, occurs when TE-encoded proteins or domains become co-opted into functional host proteins[10,63]. In principle, any of the activities or domains encoded by TE proteins can be domesticated. However, as the list of TE-derived proteins increases, it is becoming evident that TPases are more prone to domestication than other TE proteins[5,63]. The propensity of TPases for domestication was first noticed in the initial analysis of the human genome, which identified 47 genes that are entirely or mostly derived from TE coding sequences[64], with all but 4 of them related to TPases, despite the fact that DNA transposons are a modest fraction of human TEs (7%) and of the genome (3%)[64,65].

Since this influential publication, dozens of additional cases of TE-derived proteins have been identified in animals, fungi and plants, even when stringent criteria were applied to validate the functionality of the TE-derived genes, such as syntenic conservation and evidence of strong purifying selection acting in distantly related species, for example, REFS 66–69. These studies reveal that several categories of TE coding sequences have been domesticated on multiple independent occasions, such as retroviral *gag*-like and envelope proteins[63,68], but TPases continue to stand out as a recurrent supply of new proteins in diverse organisms[5] (FIG. 3).

Transposases as a source of DNA-binding domains. Like TFs, TPases must translocate to the nucleus to recognize specific DNA sites on the chromosomes. To achieve this, most TPases use a nuclear localization signal and an N-terminal DNA-binding domain (DBD) that interacts specifically with a short DNA motif that is located near each of the transposon ends, often within the TIRs[70]. Also, like other DNA-binding proteins (DBPs), TPases must either promote their own access to open chromatin, for example, by recruiting host chromatin remodelling complexes[71], or take advantage of transient relaxing of chromatin at certain regions of the genome[72].

TPase DBDs are structurally diverse and many can be allied to those that are found in established families of TFs and DBPs[67,73–77] (TABLE 1), but originally it was often unclear whether the host's DBD derived from the TPase or vice versa (for example, REFS 78,79). With the accumulation of genomic sequence and the mining of diverse transposons across the eukaryotic tree, it is becoming increasingly clear that these DBDs first originated in TPases[73,77,80–83]. Typically, the TPases have a broader taxonomic distribution than the related host DBPs and phylogenetic reconstructions point to the association of the DBD with the TPase catalytic core as the ancestral state, whereas the host proteins are derived from a subclade of TPases. The origination of DBDs from TPases has involved all major TPase superfamilies (FIG. 3; TABLE 1) and has occurred repeatedly in the evolution of plants, fungi and animals[5,77,81,82,84,85], giving rise to some master developmental regulators (for example, PAX proteins, see FIG. 3 ; TABLE 1).

Transposases possess intrinsic regulatory activities. In many instances, the sequence similarity of TPase-derived proteins with their ancestral TPase is not limited to the DBD but spans their entire sequence, including the catalytic core of the TPase (FIG. 3). However, it is often found that the acidic residues that are essential for catalysis have

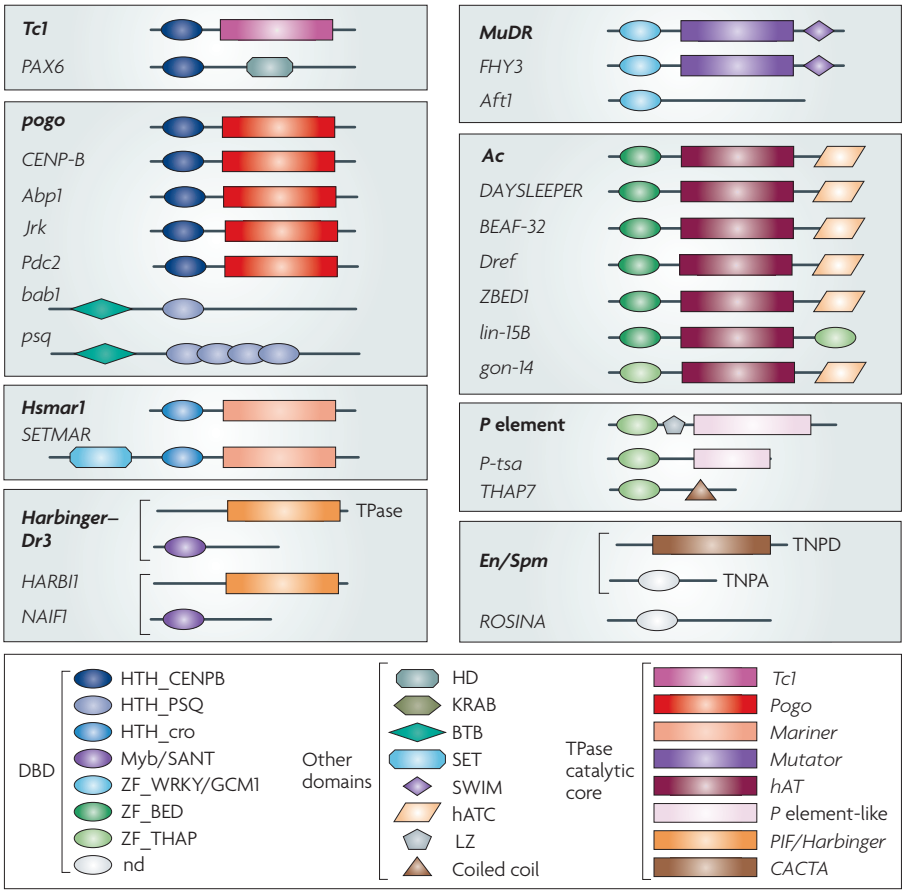

Figure 3 | **DNA-binding proteins and transcription factors derived from transposases.** The domain structure of several well-documented cases of host DNA-binding proteins and transcription factors (TFs) that are derived from transposases (TPases), the names of which are in bold, and their closest TPase relatives. *Tc1*, *pogo* and *Hsmar1* belong to three different subgroups of the *Tc1/mariner* superfamily of TPases; *Harbinger-Dr3* belongs to the *PIF/Harbinger* superfamily; *MuDR* to the *Mutator* superfamily; *Ac*, to the *hAT* (*hobo/Activator/Tam3*) superfamily; *P* element to the *P* superfamily; and *En/Spm* to the *CACTA* superfamily[2]. The gene names, the species where they were originally described and a brief description of their function follows. *PAX6*, Paired box protein Pax-6 (*Homo sapiens*, development of sensory organs and brain); *CENP-B*, Centromere protein B (*H. sapiens*, centromere function); *Abp1*, ARS-binding protein 1 (*Saccharomyces cerevisiae*, centromeric heterochromatin formation, chromosome segregation and retrotransposon repression); *Jrk*, jerky (*Mus musculus*, probable neuronal translational and transcriptional regulator); *Pdc2*, pyruvate decarboxylase 2 (*S. cerevisiae*, transcription factor involved in pyruvate and thiamine metabolism); *bab1*, bric a brac 1 (*Drosophila melanogaster*, transcriptional regulator); *psq*, pipsqueak (*D. melanogaster*, transcriptional repressor, embryonic and adult development); *SETMAR* (*H. sapiens*, histone modification and DNA repair); *HARBI1* (*H. sapiens*, no known function); *NAIF1*, nuclear apoptosis-inducing factor 1 (*H. sapiens*, directly interacts with and mediates nuclear translocation of HARBI1); *FHY3*, far-red elongated hypocotyls 3 (*Arabidopsis thaliana*, far-red light signalling); *Aft1*, activator of ferrous transport 1 (*S. cerevisiae*, transcription factor involved in iron utilization and homeostasis); DAYSLEEPER (*A. thaliana*, probable developmental regulator); *BEAF-32*, boundary element-associated factor of 32 kDa (*D. melanogaster*, insulator function, gene regulation and chromosome organization); *Dref*, DNA replication-related element-binding factor (*D. melanogaster*, regulation of cell proliferation and differentiation); *ZBED1*, Zinc finger BED-domain containing protein 1 (*H. sapiens*, transcriptional activator of cell proliferation and ribosomal proteins); *lin-15B*, abnormal cell LINeage 15B (*Caenorhabditis elegans*, developmental regulator through cell-cycle control); *gon-14*, gonadogenesis deficient 14 (*C. elegans*, required for gonadogenesis and other developmental processes); *P-tsa*, P-neogene (*Drosophila tsacasi*, unknown function); *THAP7*, thanatos-associated protein 7 (*H. sapiens*, transcriptional repressor); *ROSINA* (*Anthirinium majus*, floral organ development). DNA-binding domains (DBDs), alphabetical order: BED, BEAF and DREF; CENPB, centromere-binding protein B; GCM1, glial cell missing 1; HTH, helix-turn-helix; Myb, myeloblastosis; nd, not determined; PSQ, pipsqueak; SANT, Swi3–Ada2–NCoR–TFIIIB; ZF, zinc-finger. Other domains, alphabetical order: BTB, broad-complex, tramtrack, and bric a brac; hATC, *hAT* C-terminal dimerization; HD, homeodomain; KRAB, Kruppel-associated box; LZ, leucine zipper. SET, Su(var)3-9, E(z) and Trithorax; SWIM, SWI2/SNF2 and *MuDR*. TNPA and TNPD, the two proteins encoded by *En/Spm* transposons.

been altered, compromising cleavage and/or strand-transfer activities (for example, REFS 86–88). Nonetheless, the overall conservation of full-length TPase sequence and architecture suggests that biochemical activities other than DNA binding have been co-opted. These might include oligomerization activity, which allows the pairing of DNA sites that are bound simultaneously by different TPase molecules, and looping of the intervening DNA[70]. The inherent ability of TPases to 'loop' and 'bundle' the DNA to which they are attached might predispose them to be recruited as proteins that package and organize the genome into functionally independent chromatin domains, akin to insulator proteins. Indeed, BEAF-32 is a *Drosophila* insulator protein that is entirely derived from an *hAT* TPase[73] (FIG. 3) that binds the *scs* chromatin boundary element and that connects chromatin to the nuclear matrix[89]. In fission yeast, ARS-binding protein 1 (Abp1) and its two paralogues, collectively known as CENP-B homologues, are centromere-binding proteins involved in chromosome segregation that originated from a fungi lineage of *pogo*-like TPases[82]. Recently, it was shown that the fission yeast CENP-Bs can also bind non-centromeric interspersed DNA repeats and promote the bundling of these repeats at the nuclear periphery[90]. Furthermore, Abp1 interacts directly with histone deacetylases and directs them to its associated DNA, thereby triggering a local nucleation of heterochromatin and repressing the transcription of adjacent genes at several chromosomal locations[90] (FIG. 2a). It is unknown whether the ability of CENP-Bs to interact with histone deacetylases was drawn from the progenitor TPase. However, it is evident that the DNA-binding and self-dimerization activities, which are necessary for bundling and tethering of DNA to the nuclear periphery, directly descend from the domesticated TPases.

In *Arabidopsis thaliana,* far-red elongated hypocotyls (FHY3) and far-red impaired response (FAR1) are two closely related proteins that are entirely derived from *Mutator*-like TPases[85,91]. FHY3 and FAR1 are bonafide TFs that bind directly to promoter regions and activate several genes involved in far-red light and circadian signalling[85]. The transcriptional activation domains of FHY3 and FAR1 are physically separable from their DBD, and this activity requires two residues that are highly conserved in *Mutator*-like

Table 1 | **DNA-binding domain (DBD) families probably originated from transposases**

DBD family	Example of DBP and its function	Distribution of host proteins	TE origin (superfamily)	Distribution of the TE superfamily	References
Paired box (HTH)	PAX6, development of sensory organs and brain	Metazoans	*Tc1*	Metazoans Fungi Entamoeba (lobosea)	79,80
CENPB (HTH)	CENP-B, centromere function	Vertebrates* Fungi*	*pogo*	Metazoans Fungi Entamoeba (lobosea) Angiosperms *Phytopthora infestans* (oomycetes)	78,82
PSQ (HTH)	Bric a brac 1, development of ovaries, appendages and abdomen	Insects	*pogo*	Metazoans Fungi	74
Myb-like‡ (HTH)	NAIF1, apoptosis and/or cell-cycle regulation?	Metazoans* Angiosperms*	*PIF/Harbinger*	Metazoans Fungi Angiosperms Diatom (stramenopiles) *Trichomonas vaginalis* (trichomonads)	L. Sinzelle & Z. Ivics, personal communication
WRKY (also known as GCM1) (ZF)	FHY3, far-red light signaling	Angiosperms*, Yeasts* Insects*	*Mutator*	Metazoans Fungi Entamoeba (lobosea) Angiosperms *Phytopthora infestans* (oomycetes) *Trichomonas vaginalis* (trichomonads)	77,85,91
BED (ZF)	DREF, regulation of cell proliferation and differentiation	Metazoans* Angiosperms*	*hAT*	Metazoans Fungi Entamoeba (lobosea) Angiosperms *Chlamydomonas reinhardtii* (green algae) *Phytophthora infestans* (oomycetes) *Trichomonas vaginalis* (trichomonads)	73
THAP (ZF)	THAP1, apoptosis and cell-cycle regulation	Metazoans*	*P* element	Metazoans *Chlamydomonas reinhardtii* (green algae)	75,81

*Indicates multiple independent domestication events within that clade. ‡Contains the DBDs from Myb proteins, as well as the SANT domain family. BED, BEAF and DREF; CENP-B, centromere binding protein B; DBP, DNA-binding protein; DREF, DNA replication-related element-binding factor; FHY3, far-red elongated hypocotyls; HTH, helix turn helix; Myb, myeloblastosis; NAIF1, nuclear apoptosis-inducing factor 1; PAX6, paired box protein Pax-6; PSQ, pipsqueak; TE, transposable element; THAP7, thanatos-associated protein 7; ZF, zinc finger.

TPases[85]. Thus, it is conceivable that *Mutator*-like TPases possess intrinsic TF activity, which might explain repeated domestication events of this superfamily of TPases in plants, fungi and animals[77,84]. Interestingly, the TPase that is encoded by the maize *MuDR* element, and also TNPA, one of the two proteins encoded by the maize *Spm* transposon, can function as transcriptional regulators of their own expression[92,93]. Such transcriptional self-regulation might offer an opportunity for a new TPase-derived TF to instantly acquire a regulatory feedback loop, a characteristic of most regulatory circuits[32,48,49] (FIG. 2d).

Birth of a genetic network

In vitro and *in vivo* studies have shown that TPases often cross-interact with distantly related transposons with similar termini[94–97]. Thus, not only the amplification of an active transposon, but also the past accumulation of evolutionarily related elements, result in a build-up of TPase binding sites throughout the genome. Domestication can occur when one or several TPase–DNA interactions become selectively advantageous for the host (FIG. 2d). This selective benefit might arise as the result of a transposon insertion in the proximity of a host gene, bringing the host gene under the regulatory influence of the TPase. Alternatively, domestication can be initiated by mutational events at a TPase-encoding locus, leading to the emergence of a modified TPase with new *trans*-regulatory activities. This can occur by mutations in the TPase sequence, fusion of flanking exons to the TPase[86] and/or a change in the pattern of TPase expression. Natural selection might further shape and expand this primordial regulatory network by acting on newly arisen or polymorphic transposon insertions to retain those that bring beneficial interactions with the domesticated TPase, while removing those with deleterious consequences[5].

Testing this model is hampered by the fact that most DBPs that are known to derive from TPases are of relatively ancient origin, making it impractical to trace the origin of their binding sites to ancestral transposons because the sequences surrounding the TPase binding site would probably be erased by extended periods of neutral evolution. In addition, TPase binding sites are short and fast-evolving motifs and consequently they tend to be poorly conserved, even among related transposon families[70,96,97]. Therefore, to validate the model it is necessary to study recently emerged TPase-derived proteins that can be tied to transposon families that are still recognizable in the same genome.

One promising candidate is SETMAR, a primate-specific protein that results from the fusion of a pre-existing SET histone methyltransferase gene to the TPase gene of an *Hsmar1* transposon[86]. The amplification of the *Hsmar1* family and its related transposons (*Made1* MITEs) occurred

approximately 45 million years ago[65] and was concomitant to the emergence of *SETMAR*. Evolutionary sequence analyses of *SETMAR* across anthropoid primate species have revealed that the DBD, but not the catalytic region, has been subject to continuous purifying selection since its domestication[86]. Consistent with these observations, *in vitro* experiments demonstrated that the TPase region of SETMAR has retained robust DNA-binding activity, but has lost some of its catalytic abilities[86–88]. The 19-bp binding site that is recognized by SETMAR is reiterated in about 1,500 perfect or nearly perfect copies in the human genome, and almost all of these sites map within the TIRs of *Hsmar1* and *Made1* (REF. 86). Thus, the TPase region of SETMAR might be used to target the SET domain to multiple genomic sites, where it might modify the surrounding chromatin and modulate the transcription of adjacent genes.

Outlook

The recent discoveries and models presented in this Review echo the visionary predictions of McClintock and those of a few pioneers on the significance of TEs for eukaryotic gene regulation. Meanwhile, it is being realized that the most influential contributions of TEs to macroevolution could arise and persist long after transposition activity has ceased and that these contributions typically emerged as a by-product of the selfish and parasitic lifestyle of TEs. It is these characteristics that ensure the long-term survival of TEs and that entail their intimate co-evolutionary relationship with the host genome. Ironically, the breadth and versatility of TE exaptation has become most apparent in mammalian genomes, in which repetitive DNA has traditionally been perceived as a hurdle to geneticists rather than as a valuable source of genomic information. Whereas the wide diversity, large-scale amplification and relatively slow mutational decay of TEs in mammals have probably promoted co-option, these characteristics have also allowed the process to be observable. Quantitatively evaluating the functional heritage of TEs in organisms with faster sequence turnover is more difficult, but there is increasing evidence that TEs have been co-opted repeatedly for cellular and regulatory functions in various eukaryotic lineages. Together, these findings are likely to recapitulate a prominent evolutionary process that has been shaping eukaryotic genomes ever since their origins. Our appreciation of the impact of TEs in regulatory evolution will certainly benefit from a broadening of the diversity of organisms that are under genomic scrutiny. This will necessitate the development of new tools to accelerate the discovery and automate the genomic annotation of TEs. A more rigorous and quantitative examination of the dynamics and the mode of molecular evolution of TEs is also needed. For example, it remains unclear how transposition mechanisms, chromosomal distribution and epigenetic markings of TEs affect their rate of evolution and influence their propensity toward exaptation. Finally, there is a need to continue to develop experimental approaches to further probe the models developed here and elsewhere, which position TEs and their derived proteins as central players in the evolution of eukaryotic gene regulation.

Cédric Feschotte is at the Department of Biology, Life Science Building, BOX 19498, University of Texas, Arlington, Texas 76019, USA.

e-mail: cedric@uta.edu

doi:10.1038/nrg2337
Published online 27 March 2008

1. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**, 529–540 (1968).
2. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* **8**, 973–982 (2007).
3. Brookfield, J. F. The ecology of the genome — mobile DNA elements and their hosts. *Nature Rev. Genet.* **6**, 128–136 (2005).
4. Kidwell, M. G. & Lisch, D. R. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* **55**, 1–24 (2001).
5. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).
6. Deininger, P. L., Moran, J. V., Batzer, M. A. & Kazazian, H. H. Jr. Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* **13**, 651–658 (2003).
7. Gould, S. J. & Vrba, E. S. Exaptation — a missing term in the science of form. *Paleobiology* **8**, 4–15 (1983).
8. Brosius, J. Retroposons — seeds of evolution. *Science* **251**, 753 (1991).
9. Britten, R. J. Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol. Phylogenet. Evol.* **5**, 13–17 (1996).
10. Miller, W. J., McDonald, J. F., Nouaud, D. & Anxolabehere, D. Molecular domestication — more than a sporadic episode in evolution. *Genetica* **107**, 197–207 (1999).
11. Silva, J. C., Shabalina, S. A., Harris, D. G., Spouge, J. L. & Kondrashovi, A. S. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82**, 1–18 (2003).
12. Lowe, C. B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl Acad. Sci. USA* **104**, 8005–8010 (2007).
13. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
14. Bejerano, G., Haussler, D. & Blanchette, M. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* **20** (Suppl. 1), i40–i48 (2004).
15. Xie, X., Kamal, M. & Lander, E. S. A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl Acad. Sci. USA* **103**, 11659–11664 (2006).
16. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
17. Kamal, M., Xie, X. & Lander, E. S. A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl Acad. Sci. USA* **103**, 2740–2745 (2006).
18. Nishihara, H., Smit, A. F. & Okada, N. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.* **16**, 864–874 (2006).
19. Santangelo, A. M. *et al.* Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet.* **3**, 1813–1826 (2007).
20. Maside, X., Bartolome, C. & Charlesworth, B. S-element insertions are associated with the evolution of the HSP70 genes in *Drosophila melanogaster*. *Curr. Biol.* **12**, 1686–1691 (2002).
21. Schlenke, T. A. & Begun, D. J. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc. Natl Acad. Sci. USA* **101**, 1626–1631 (2004).
22. Chung, H. *et al.* *Cis*-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* **175**, 1071–1077 (2007).
23. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
24. Inada, D. C. *et al.* Conserved noncoding sequences in the grasses. *Genome Res.* **13**, 2030–2041 (2003).
25. Brosius, J. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118**, 99–116 (2003).
26. Marino-Ramirez, L., Lewis, K. C., Landsman, D. & Jordan, I. K. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* **110**, 333–341 (2005).
27. Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19**, 68–72 (2003).
28. van de Lagemaat, L. N., Landry, J. R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**, 530–536 (2003).
29. Marino-Ramirez, L. & Jordan, I. K. Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol. Direct* **1**, 20 (2006).
30. Gentles, A. J. *et al.* Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res.* **17**, 992–1004 (2007).
31. Ni, J. Z. *et al.* Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* **21**, 708–718 (2007).
32. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
33. Hambor, J. E., Mennone, J., Coon, M. E., Hanke, J. H. & Kavathas, P. Identification and characterization of an *Alu*-containing, T-cell-specific enhancer located in the last intron of the human *CD8 alpha* gene. *Mol.Cell Biol.* **13**, 7056–7070 (1993).
34. Zhou, Y. H., Zheng, J. B., Gu, X., Saunders, G. F. & Yung, W. K. Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res.* **12**, 1716–1722 (2002).
35. Medstrand, P. *et al.* Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res.* **110**, 342–352 (2005).
36. Thornburg, B. G., Gotea, V. & Makalowski, W. Transposable elements as a significant source of transcription regulating signals. *Gene* **365**, 104–110 (2006).
37. Polak, P. & Domany, E. *Alu* elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**, 133 (2006).

38. Johnson, R. *et al.* Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.* **34**, 3862–3877 (2006).
39. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
40. Wessler, S. R., Bureau, T. E. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**, 814–821 (1995).
41. Ferrigno, O. *et al.* Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nature Genet.* **28**, 77–81 (2001).
42. Romanish, M. T., Lock, W. M., van de Lagemaat, L. N., Dunn, C. A. & Mager, D. L. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet.* **3**, e10 (2007).
43. Borchert, G. M., Lanier, W. & Davidson, B. L. RNA polymerase III transcribes human microRNAs. *Nature Struct. Mol. Biol.* **13**, 1097–1101 (2006).
44. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
45. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46**, 111–138 (1971).
46. Peaston, A. E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
47. Bringaud, F. *et al.* Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. *PLoS Pathog.* **3**, 1291–1307 (2007).
48. Wilkins, A. S. *The Evolution of Developmental Pathways* (Sinauer, Sunderland, Massachusetts, 2002).
49. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic, New York, 2006).
50. Mattick, J. S. A new paradigm for developmental biology. *J. Exp. Biol.* **210**, 1526–1547 (2007).
51. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nature Rev. Genet.* **8**, 93–103 (2007).
52. Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nature Rev. Genet.* **8**, 35–46 (2007).
53. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Rev. Genet.* **8**, 272–285 (2007).
54. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761–764 (2007).
55. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Rev. Genet.* **5**, 522–531 (2004).
56. Smalheiser, N. R. & Torvik, V. I. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21**, 322–326 (2005).
57. Piriyapongsa, J., Marino-Ramirez, L. & Jordan, I. K. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**, 1323–1337 (2007).
58. Smalheiser, N. R. & Torvik, V. I. *Alu* elements within human mRNAs are probable microRNA targets. *Trends Genet.* **22**, 532–536 (2006).
59. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nature Rev. Genet.* **3**, 329–341 (2002).
60. Sijen, T. & Plasterk, R. H. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426**, 310–314 (2003).
61. Piriyapongsa, J. & Jordan, I. K. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* **2**, e203 (2007).
62. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
63. Volff, J. N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913–922 (2006).
64. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
65. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* **17**, 422–432 (2007).
66. Zdobnov, E. M., Campillos, M., Harrington, E. D., Torrents, D. & Bork, P. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res.* **33**, 946–954 (2005).
67. Casola, C., Lawing, A. M., Betran, E. & Feschotte, C. *PIF*-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* **24**, 1872–1888 (2007).
68. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple *gag*-like human proteins. *Trends Genet.* **22**, 585–589 (2006).
69. Muehlbauer, G. J. *et al.* A *hAT* superfamily transposase recruited by the cereal grass genome. *Mol. Genet. Genomics* **275**, 553–563 (2006).
70. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. *Mobile DNA II*, (American Society for Microbiology, Washington, 2002).
71. Makarova, K. S., Aravind, L. & Koonin, E. V. SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. *Trends Biochem. Sci.* **27**, 384–386 (2002).
72. Ros, F. & Kunze, R. Regulation of activator/dissociation transposition by replication and DNA methylation. *Genetics* **157**, 1723–1733 (2001).
73. Aravind, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.* **25**, 421–423 (2000).
74. Siegmund, T. & Lehmann, M. The *Drosophila* Pipsqueak protein defines a new family of helix-turn-helix DNA-binding proteins. *Dev. Genes Evol.* **212**, 152–157 (2002).
75. Roussigne, M. *et al.* The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem. Sci.* **28**, 66–69 (2003).
76. Kapitonov, V. V. & Jurka, J. *Harbinger* transposons and an ancient *HARBI1* gene derived from a transposase. *DNA Cell Biol.* **23**, 311–324 (2004).
77. Babu, M. M., Iyer, L. M., Balaji, S. & Aravind, L. The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res.* **34**, 6505–6520 (2006).
78. Tudor, M., Lobocka, M., Goodwell, M., Pettitt, J. & O'Hare, K. The *pogo* transposable element family of *Drosophila melanogaster*. *Mol. Gen. Genet.* **232**, 126–134 (1992).
79. Franz, G., Loukeris, T. G., Dialektaki, G., Thompson, C. R. & 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* **91**, 4746–4750 (1994).
80. Breitling, R. & Gerber, J. K. Origin of the paired domain. *Dev. Genes Evol.* **210**, 644–650 (2000).
81. Quesneville, H., Nouaud, D. & Anxolabehere, D. Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. *Mol. Biol. Evol.* **22**, 741–746 (2005).
82. Casola, C., Hucks, D. & Feschotte, C. Convergent domestication of *pogo*-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.*, **25**, 29–41 (2008).
83. Piriyapongsa, J., Rutledge, M. T., Patel, S., Borodovsky, M. & Jordan, I. K. Evaluating the protein coding potential of exonized transposable element sequences. *Biol. Direct* **2**, 31 (2007).
84. Cowan, R. K., Hoen, D. R., Schoen, D. J. & Bureau, T. E. *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. *Mol. Biol. Evol.* **22**, 2084–2089 (2005).
85. Lin, R. *et al.* Transposase-derived transcription factors regulate light signalling in *Arabidopsis*. *Science* **318**, 1302–1305 (2007).
86. Cordaux, R., Udit, S., Batzer, M. A. & Feschotte, C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl Acad. Sci. USA* **103**, 8101–8106 (2006).
87. Liu, D. *et al.* The human SETMAR protein preserves most of the activities of the ancestral *Hsmar1* transposase. *Mol. Cell Biol.* **27**, 1125–1132 (2007).
88. Miskey, C. *et al.* The ancient *mariner* sails again: transposition of the human *Hsmar1* element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. *Mol. Cell Biol.* **27**, 4589–4600 (2007).
89. Pathak, R. U., Rangaraj, N., Kallappagoudar, S., Mishra, K. & Mishra, R. K. Boundary element-associated factor 32B connects chromatin domains to the nuclear matrix. *Mol. Cell Biol.* **27**, 4796–4806 (2007).
90. Cam, H. P., Noma, K. I., Ebina, H., Levin, H. L. & Grewal, S. I. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* **451**, 431–436 (2008).
91. Hudson, M. E., Lisch, D. R. & Quail, P. H. The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *Plant J.* **34**, 453–471 (2003).
92. Raizada, M. N., Brewer, K. V. & Walbot, V. A maize *MuDR* transposon promoter shows limited autoregulation. *Mol. Genet. Genomics* **265**, 82–94 (2001).
93. Cui, H. & Fedoroff, N. V. Inducible DNA demethylation mediated by the maize *Suppressor-mutator* transposon-encoded TnpA protein. *Plant Cell* **14**, 2883–2899 (2002).
94. Atkinson, P. W., Warren, W. D. & O'Brochta, D. A. The *hobo* transposable element of *Drosophila* can be cross-mobilized in houseflies and excises like the *Ac* element of maize. *Proc. Natl Acad. Sci. USA* **90**, 9693–9697 (1993).
95. Rezsohazy, R., van Luenen, H. G. A. M., Durbin, R. M. & Plasterk, R. H. A. Tc7, a *Tc1*-hitch hiking transposon in *Caenorhabditis elegans*. *Nucleic Acids Res.* **25**, 4048–4054 (1997).
96. Lampe, D. J., Walden, K. K. & Robertson, H. M. Loss of transposase–DNA interaction may underlie the divergence of *mariner* family transposable elements and the ability of more than one *mariner* to occupy the same genome. *Mol. Biol. Evol.* **18**, 954–961 (2001).
97. Feschotte, C., Osterlund, M. T., Peeler, R. & Wessler, S. R. DNA-binding specificity of rice *mariner*-like transposases and interactions with *Stowaway* MITEs. *Nucleic Acids Res.* **33**, 2153–2165 (2005).
98. Wallace, M. R. *et al.* A *de novo Alu* insertion results in neurofibromatosis type 1. *Nature* **353**, 864–866 (1991).
99. Girard, L. & Freeling, M. Regulatory changes as a consequence of transposon insertion. *Dev. Genet.* **25**, 291–296 (1999).
100. Simons, C., Pheasant, M., Makunin, I. V. & Mattick, J. S. Transposon-free regions in mammalian genomes. *Genome Res.* **16**, 164–172 (2006).
101. Lerman, D. N. & Feder, M. E. Naturally occurring transposable elements disrupt *hsp70* promoter function in *Drosophila melanogaster*. *Mol. Biol. Evol.* **22**, 776–783 (2005).
102. Walser, J. C., Chen, B. & Feder, M. E. Heat-shock promoters: targets for evolution by P transposable elements in *Drosophila*. *PLoS Genet.* **2**, e165 (2006).
103. Ackerman, H., Udalova, I., Hull, J. & Kwiatkowski, D. Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. *Mol. Biol. Evol.* **19**, 884–890 (2002).
104. Martin, C. & Lister, C. Genome juggling by transposons: Tam3-induced rearrangements in *Antirrhinum majus*. *Dev. Genet.* **10**, 438–451 (1989).
105. Koga, A., Iida, A., Hori, H., Shimada, A. & Shima, A. Vertebrate DNA transposon as a natural mutator: the medaka fish *Tol2* element contributes to genetic variation without recognizable traces. *Mol. Biol. Evol.* **23**, 1414–1419 (2006).

Acknowledgements

I owe many thanks to C. Casola, J-M. Deragon, J. Fondon, I. K. Jordan, A. Pires da Silva, E. Pritham and D. Voytas for discussions and comments during the preparation of this article. I also thank the two anonymous reviewers for their constructive comments and useful suggestions. The author apologizes to many colleagues whose relevant work and original articles could not be cited owing to space limitations. Research in the C. F. laboratory is supported by grants from the US National Institutes of Health.

DATABASES

Entrez Conserved Domains:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=Cdd
BED | CENPB | Myb-like | Paired box | PSQ | THAP | WRKY

UniProtKB: http://ca.expasy.org/sprot
Abp1 | FAR1 | FHY3 | p53 | SETMAR

FURTHER INFORMATION

Cédric Feschotte's homepage:
http://www3.uta.edu/faculty/cedric

ALL LINKS ARE ACTIVE IN THE ONLINE PDF