Evolutionary History and Impact of Human DNA Transposons

Cédric Feschotte, *University of Utah School of Medicine, Salt Lake City, Utah, USA*

John McCormick, *University of Utah School of Medicine, Salt Lake City, Utah, USA*

Based in part on the previous version of this eLS article 'Evolutionary History and Impact of Human DNA Transposons' (2008) by Cédric Feschotte.

Introductory article

Article Contents

Online posting date: 18th October 2013

Deoxyribonucleic acid (DNA) transposons are mobile elements that move via a DNA intermediate. The (haploid) human genome harbours more than 300 000 DNA transposon copies, accounting for approximately 3% of the total genomic DNA. Nearly one-third of these elements are specific to the primate lineage, but there is no evidence for transposition activity within the past 40 million years. However, there is growing evidence that DNA transposons have contributed in shaping the current genome architecture of humans and have been a recurrent source of new regulatory and coding DNA throughout mammalian evolution. Notably, more than 50 human genes are currently known to descend from transposase sequences recycled to perform diverse cellular functions.

Introduction

About half of the human genomic deoxyribonucleic acid (DNA) is currently recognisable as being derived from mobile genetic elements (Smit, 1999; Lander *et al*., 2001). These elements are diverse in terms of their origin, mode of amplification and copy numbers. By far the most successful types of transposable elements (TEs) in the human genome are class 1 or retroelements, which are produced by reverse transcription of a ribonucleic acid (RNA) intermediate. Class 2 or DNA transposons, which transpose directly as a DNA intermediate, are also represented in the human genome and they are the focus of this article. This article summarises our current knowledge of the classification, evolutionary history and genomic impact of human DNA transposons. **See also**: Long Interspersed Nuclear Elements (LINEs); Long Interspersed Nuclear Elements (LINEs): Evolution; Retroviral Repeat Sequences; Short Interspersed Elements (SINEs); Transposons: Eukaryotic

All known human DNA transposons belong to the subclass of 'cut-and-paste' TEs. It has been shown using *in-vitro* assays that cut-and-paste transposition generally requires a single element-encoded enzyme called transposase (Craig *et al*., 2002). In a typical DNA transposition reaction, transposase binds in a sequence-specific manner to the terminal inverted repeats (TIRs) located at each end of the transposon and catalyses both the DNA cleavage and strand transfer steps of the transposition reaction (**Figure 1**). Elements that encode active transposase are termed autonomous elements, whereas defective copies unable to encode an active transposase are called nonautonomous. Nonautonomous elements may nevertheless transpose if they contain the *cis*-sequences sufficient for recognition and cleavage by a transposase encoded *in trans* by an autonomous element. For reasons that are not yet fully understood, short nonautonomous elements called MITEs are able to proliferate to a much greater extent than their autonomous partners (Feschotte *et al*., 2002). **See also**: DNA Transposition: Classes and Mechanisms; Transposases and Integrases

Census of Human DNA Transposons

More than 380 000 DNA segments are annotated as DNA transposons in the Hg19 human genome assembly (**Table 1**). These elements fall into 125 families with copy numbers ranging from a hundred to several thousand copies per family. The most abundant is *MER5A*, a MITE family related to the hAT (hobo/Ac/Tam3) superfamily, with more than 30 000 copies per haploid genome. Although only a handful of human DNA transposon families have been subjected to a detailed analysis, it is clear that the diversity of DNA transposons in the human genome is as high or greater than in other eukaryotic species, such as *Drosophila melanogaster*, *Arabidopsis thaliana* or *Fugu rubripes* (Feschotte and Pritham, 2007). Furthermore,

eLS subject area: Evolution & Diversity of Life

How to cite:
Feschotte, Cédric; and McCormick, John (October 2013) Evolutionary History and Impact of Human DNA Transposons. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020996.pub2

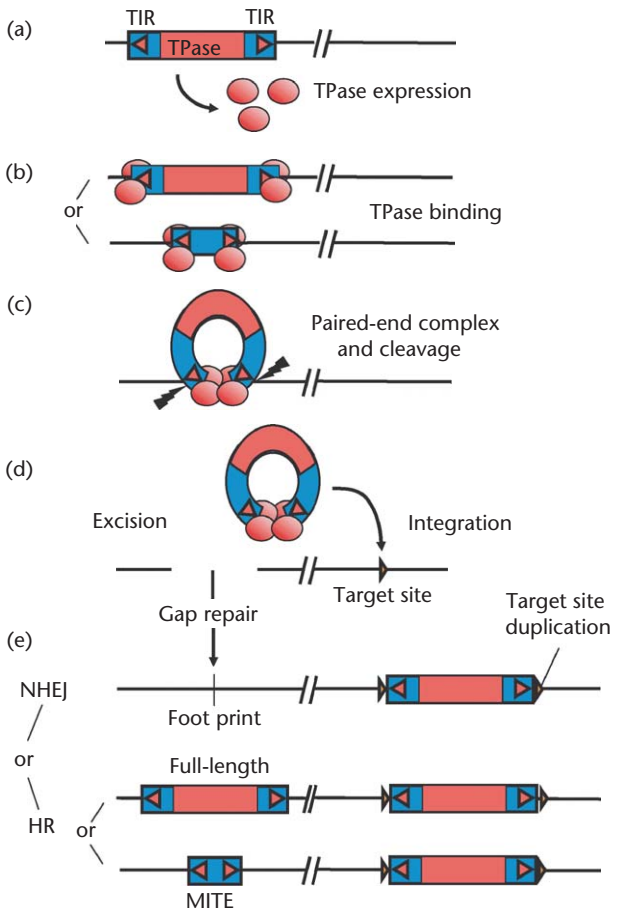

Figure 1 'Cut-and-paste' transposition. (a) An autonomous DNA transposon contains an open reading frame encoding an active source of transposase (TPase) enzyme (circles). TIR: terminal inverted repeats, shown as arrowheads. (b) Transposase molecules return to the nucleus and bind, often as dimers, to the ends of virtually any transposon copy present in the genome (autonomous (top) or nonautonomous (bottom)) that contains intact binding sites for the transposase (usually located within the TIRs). (c) The transposon engages in the formation of a synaptic or paired-end complex and the transposase molecules catalyse cleavage (double-stranded breaks, DSB) at each end of the transposon. (d) The element is now excised out of the chromosome and transposase catalyses its reintegration elsewhere in the genome, either on the same (as shown) or on a different chromosome. (e) Integration results in the duplication of a short host DNA sequence at the target site, called target site duplication. The size of the target site duplication (TSD) varies from 2 to 10 bp and is characteristic of a given transposase superfamily (e.g., usually 8 bp for hAT superfamily). The gap left behind by the excision of the transposon is repaired by the host DNA repair machinery. Two major repair pathways are known to operate in eukaryotic cells. Under the nonhomologous end-joining (NHEJ), the transposon will be essentially lost at the excision site, with short sequences corresponding to the termini of the transposon sometimes remaining, also known as transposon footprint. Under the homologous recombination (HR) pathway the homologous chromosome or sister chromatid may be used as a template to repair the DSB and restore the original insertion at the excision site. This process results in a net increase of one copy of the transposon. If HR is complete, a full-length copy of the excised transposon is restored. However, experiments have shown in *Drosophila* and plants that HR is often incomplete (abortive gap repair) and result in the restoration of an internally deleted copy of the original transposon. These shorter, noncoding elements may still be propagated if they retain the transposase-binding sites, giving rise to homogeneous families of so-called miniature inverted-repeat transposable elements (MITEs).

strictly in terms of their copy numbers, DNA transposons are several orders of magnitude more abundant in the human (approximately 380 000 copies) or mouse (approximately 110 000 copies) genomes than in other eukaryotic species. **See also**: Repetitive Elements: Bioinformatic Identification, Classification and Analysis

Overall, seven out of ten known eukaryotic superfamilies of DNA transposons are represented in the human genome, but the hAT and Tc1/*mariner* superfamilies largely predominate (**Table 1**). hAT elements account for approximately half of the 125 families and two-thirds of all human DNA transposon copies. Human Tc1/*mariner* elements account for approximately one-fourth and can be divided into three evolutionarily distinct lineages: *pogo*-like, *mariner*-like and Tc2-like. The former is the most abundant and diversified lineage, and includes eight families of transposase-encoding *tigger* elements (Smit and Riggs, 1996) and 22 related MITE families. The prevalence of nonautonomous MITEs (74% of the total number of DNA TEs) over transposase-encoding elements (26%) is particularly striking in the human genome and this phenomenon affects all superfamilies (**Table 1**). It is also a characteristic of the DNA transposon population of plants, insects and nematodes (Feschotte *et al*., 2002).

Evolutionary History of Human DNA Transposons

Although the evolutionary history of human *Alu* and L1 retrotransposons has been studied intensively, the history of DNA transposons has been less thoroughly examined. Pace and Feschotte (2007) presented the first systematic assessment of the evolutionary origins of nearly all families of human DNA transposons using a combination of three independent methods (**Figure 2**). **See also**: Evolution of Human Retrosequences: Alu; Transposable Elements: Evolution

Eighty (68%) of the 125 families were found to have originated before the last common ancestor of placental mammals (i.e., eutherian, **Figure 2**). Representative copies of these families are found inserted at orthologous genomic positions in human and at least one of the nonprimate mammalian species for which genome sequences are available (e.g., dog and mouse). Most of the eutherian-wide families belong to the hAT superfamily. A very small subset of these families can be traced back to the split of marsupial and eutherians, as some of their copies can be detected at orthologous genomic positions in humans and opossums.

The remaining families (at least 40 and up to 69, depending on the dating method) appear to result from waves of amplification that are specific to primate genomes. Primate-specific families account for at least approximately 98 000 elements and approximately 38 Mb of DNA in the human genome (**Figure 2**). Seventy-five per cent of these elements (approximately 74 000) were integrated during a period of less than 20 million years (My),

Table 1 Census of human DNA transposon families with copy number > 100

Superfamily	Families		Number of families and subfamilies	Total copy number
hAT	*Autonomous*			
	Blackjack, Charlie1–10, Cheshire, Zaphod1–2		19	46 133
	Nonautonomous			
	Arthur1, FordPrefect, MER102, MER106, MER107, MER112, MER113, MER115, MER117, MER119, MER1, MER20, MER3, MER30, MER33, MER45, MER58, MER5, MER63, MER69, MER81, MER91, MER94, MER96, MER99, ORSL		52	218 059
		Total	71	264 192
Mutator	*Nonautonomous*			
	Ricksha		3	985
		Total	3	985
piggyBac	*Autonomous*			
	Looper		1	521
	Nonautonomous			
	MER75, MER85		3	1569
		Total	4	2090
Tc1/*mariner*	*Autonomous*			
	HSMAR1, HSMAR2, Tigger1–8, Kanga1–2		22	53 320
	Nonautonomous		23	54 718
	MADE1, MARNA, MER104, MER2, MER44, MER46, MER53, MER6, MER8, MER82, MER97			
		Total	45	108 038
Unknown	MER103, MER105		2	7567
		Grand Total	125	382 872

before the emergence of prosimian primates (approximately 63 million years ago (Ma) but after the divergence of a primate ancestor from the closest nonprimates eutherian clades examined (rat, mouse and rabbit; approximately 75–85 Ma). Thus, early primate evolution was a period of high activity for DNA transposons. In comparison, approximately half as many human L1 insertions occurred during the same evolutionary era. This period of intense DNA transposon activity was dominated by Tc1/*mariner* elements, although hAT, *Mutator* and *piggyBac* elements were also active during this era (Pace and Feschotte, 2007).

The activity of DNA transposons continued, albeit with a lesser amplitude, during the next phase of the primate radiation (40–63 Ma), that is, after the split of prosimians, but before the emergence of New World monkeys (**Figure 2**). Approximately 23 000 human DNA elements were integrated during this period, adding at least approximately 5 Mb of DNA to an ancestral anthropoid genome. These elements were from 11 distinct families and three different superfamilies (Tc1/*mariner*, hAT and *piggyBac*). Intriguingly, however, no evidence was found for any DNA transposon families significantly younger than the divergence of New World monkeys, that is approximately 40 My. Consistent with this observation, a systematic survey for the presence/absence of human DNA transposons at orthologous positions in the nearly complete genome of the Rhesus macaque (an Old World monkey) failed to uncover a single instance of a DNA transposon copy present in humans, but precisely missing in macaques (**Figure 2**; Pace and Feschotte, 2007). Thus, to date, there remains no evidence for the activity of any DNA transposons present in the human genome since the emergence of Old World monkeys. The reasons for this apparent cessation in the activity of DNA transposons in the anthropoid primate lineage are unknown. **See also**: Primate Phylogenetics

A dearth of recent DNA transposon activity is also observed in other mammalian lineages, but not in all. Thus far the most glaring exception is the lineage of verspertilionid bats, which has experienced multiple waves of DNA transposition over the past 40 My (Ray *et al.*, 2008; Pritham and Feschotte, 2007). Together, these recently transposed elements account for approximately 10% of the genome

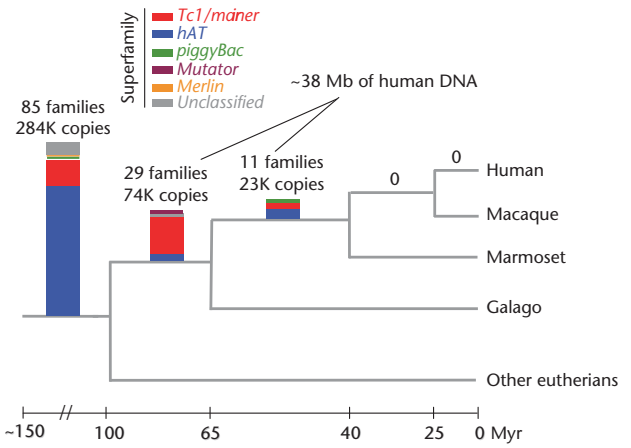

Figure 2 Temporal activity of human DNA transposons throughout the evolution of placental mammals. Figure adapted from Pace and Feschotte (2007). The histograms above the schematic phylogenetic tree show the amount of elements for each superfamily (total copy number for all superfamilies is in thousands) and the total number of families found in the human genome that were inserted at different evolutionary time points (from left to right): eutherian-wide (insertions shared by various placental mammals), primate-specific (insertions shared by all primates, but absent from all nonprimate eutherians examined) and anthropoid-specific (insertions shared by all anthropoid primates examined, but absent from galago). Currently, there is no evidence for the activity of any human DNA transposon families after the emergence of marmoset (New World monkeys), as indicated by '0' above the branches leading to the common ancestor of humans and rhesus macaques (Old World monkeys) and to humans and chimpanzees (great apes). Estimates of the divergence time of the depicted lineages are shown in million years.

sequence of the little brown bat *Myotis lucifugus*. Some of these elements have inserted very recently and at least one family was shown to remain transpositionally active in *M. lucifugus* (Mitra *et al*., 2013). The origin of these recently active DNA transposons remains mysterious. There is evidence that some of the founding members of these transposon families were acquired through horizontal transfer (i.e., the passage of genetic material between species other than by sex) (Pace *et al*., 2008; Gilbert *et al*., 2010). Remarkably, some of these elements were not only transferred to bats, but also independently in a wide range of vertebrate species, including other mammals (Pace *et al*., 2008). The mechanisms underlying such widespread transposon transfers are still obscure, but it is thought that bloodborne parasites have facilitated the spread of some of these transposons across widely diverged animals (Gilbert *et al*., 2010). Thus, the possibility cannot be excluded that active DNA transposons could at some point be reintroduced horizontally in the human population.

Potential Involvement in Genomic Rearrangements and Human Diseases

TEs do not need to be actively transposing to sculpt genomes and have a dramatic effect on phenotype. The potential of TE-mediated genome rearrangements through illegitimate recombination between preintegrated copies is well documented. Notably, DNA transposons in plants and animals have been frequently implicated in large-scale chromosomal rearrangements, such as deletions, inversions, duplications, translocations and chromosome breakage mediated by interelement recombination or aberrant transposition events (Gray, 2000; Feschotte and Pritham, 2007). Thus, despite the cessation of DNA transposition activity in the anthropoid lineage, it is conceivable that DNA transposons have been important contributors to shaping primate genome architecture by promoting chromosomal rearrangements, even in recent evolutionary times. **See also**: Repetitive Elements and Human Disorders; Retrotransposons and Human Disease; Transposons as Natural and Experimental Mutagens

There is an increasingly long list of human diseases collectively known as genomic disorders that result from gross chromosomal rearrangements. In several cases, it has been established that the rearrangements are caused by either nonallelic HR between segmental duplicated blocks (usually >10 kb) or by aberrant events of DNA repair via the NHEJ pathway. In both cases, the initial event triggering the recombination process is a DNA double-strand break (DSB). Whether these breaks are accidental or programmed, hotspots exist in the human genome where the presence of peculiar sequences or features of the DNA greatly stimulate the occurrence of DSBs. The nature of these DSB-enhancing sequences is generally unknown or poorly characterised, but two properties of DNA transposons may qualify them as potential candidates for enhancing *DSB*. **See also**: Segmental Duplications and Genetic Disease

First, some of the *DSB*-enhancing sequences could be the substrate of still catalytically active or partially active transposases. For example, they could be derived from the remnants of transposon sequences carrying binding sites for transposases. Although there remains no evidence for the presence of an active DNA transposon family in the human genome, biochemical studies of the transposase-derived protein SETMAR (see **Table 2**) show that it has preserved some of its DNA-nicking activities (Liu *et al*., 2007; Beck *et al*., 2011). Most recently, Majumdar *et al*. (2013) demonstrated that THAP9, another human transposase-derived protein (**Table 2**), is capable of mobilising the distantly related *P* element from *D. melanogaster* in both human and *Drosophila* cells. Although the biological function and natural substrate (if any) of THAP9 in the human genome is unknown, these results indicate that the protein has retained potent catalytic activities. Finally, it is also established that the transposase-derived protein RAG1, whose normal cellular function is to catalyse V(D)J recombination in immune cells, can induce DNA cleavage at cryptic recombination sites and mediate aberrant recombination events in human cell lines (Reddy *et al*., 2006). These events represent a serious threat to genomic integrity and they may result in oncogenic translocations. **See also**: Immunoglobulin Gene Rearrangements; Translocation Breakpoints in Cancer

Table 2 Examples of human transposase-derived genes

Gene	Origin[a]	Taxonomic distribution	Function	DBD from TPase?[b]
RAG1	*Transib* + *Chapaev*	Jawed vertebrates	V(D)J recombination	Yes
CENP-B	*Pogo*	Mammals	Centromere-binding	Yes
hDREF (ZBED1)	*hAT*	Mammals	Transcription factor	Yes
THAP9	*P element*	Amniotes	Unknown	Yes
ZBED6	*hAT*	Placental mammals	Transcription factor	Yes
PGBD3	*PiggyBac* + CSB	Anthropoid primates	Transcription factor?	Yes
SETMAR	*Mariner* + SET	Anthropoid primates	DNA repair?	Yes

[a]Names in italics refer to the transposase superfamily that gave rise to the gene. *PGBD3-CSB* and *SETMAR* result from the fusion of a transposase to another domain; SET: Su(var)3–9, Enhancer of zeste, Trithorax domain.
[b]Indicates whether the DNA-binding domain of the encoded protein is derived from the ancestral transposase.

Second, many DNA transposons (MITEs) are palindromic in structure. For example, the human *MADE1* MITE consists of two 37-bp TIRs separated by six unique base pairs (Smit and Riggs, 1996; Robertson, 2002). Palindromes and inverted-repeat motifs are a known source of instability and DSBs in both prokaryotic and eukaryotic chromosomes, including those of mammals (Leach, 1994). It is important to note that even repeats relatively divergent in sequence or of very short size (e.g., 20-bp palindrome separated by a short unique spacer) may promote chromosomal rearrangements in bacteria, yeast and mammalian cells (Leach, 1994). Thus, despite the lack of evidence supporting the movement of any DNA transposons in humans, it is conceivable that these elements or the derived transposases could be implicated in human genomic disorders.

An interesting case is Charcot–Marie–Tooth disease type 1 A (CMT1A, OMIM # 118220) and hereditary neuropathy with liability to pressure palsies (HNPP, OMIM # 162500), two genomic disorders caused by unequal recombination events between copies of a large segmental duplication. Two independent studies located a copy of the *mariner*-like family *Hsmar2* as the only peculiar sequence feature near the recombination hotspot (Reiter *et al*., 1996; Kiyosawa and Chance, 1996). Although this copy was apparently unable to encode a functional transposase, it was hypothesised that the presence of this element could promote strand exchange events via cleavage near the 3′ end of the element by a transposase encoded elsewhere in the genome (Reiter *et al*., 1996). This potential source of transposase has yet to be identified. However, there are other transposase-independent mechanisms that might explain the involvement of *Hsmar2* repeat in the recombination events. **See also**: Charcot–Marie–Tooth Disease and Associated Peripheral Neuropathies

Prompted by this discovery, the Lupski group further analysed the chromosomal distribution of members of the *Hsmar2* family using *in-situ* hybridisation techniques (Reiter *et al*., 1999). Although this approach provides somewhat coarse chromosomal coordinates, a significant correlation was found between the location of *Hsmar2* elements and the fragile sites and recombination hotspots involved in human genomic disorders. This supports the view that this particular family may be prone to generate chromosomal rearrangements and may promote genome instability. A refined analysis of the genomic distribution of *Hsmar2* and other families are required to explore further this provocative hypothesis.

Exaptation of Human DNA Transposons

The same properties that make TEs a source of genetic instability and deleterious mutational load, also bestow them with a tremendous potential to create genetic diversity and promote genome structuring. The direct and indirect contributions of TE to micro- and macroevolution have become apparent over the past decade of genomic research, with many studies illustrating TEs as a creative force during evolution (Craig *et al*., 2002; Feschotte and Pritham, 2007). **See also**: Transposons; Transposons: Eukaryotic

One of the most direct evolutionary contributions of TEs is as a source of genetic material recycled as new functional sequences for the host. This process is referred to as ‘exaptation’ or ‘molecular domestication’ of TEs (Miller *et al*., 1999). TE sequences can be exapted into noncoding regulatory sequences, acting either at the DNA or RNA level, or into protein-coding sequences recruited to assemble new genes (Volff, 2006). **See also**: Gene Fusion; Insertion and Deletion of Exons during Human Gene Evolution

One way to identify exapted TEs is through their high level of sequence conservation across large evolutionary distances. Following integration, most TE sequences are under no selective constraint and thus accumulate point mutations at a neutral rate (Robertson, 2002). In contrast, a TE that acquires host function, either immediately after insertion or subsequently in evolution, will become subject to purifying selection. This process results in the selective removal from the population of point mutations in the TE sequence that would affect the proper functionality of the exapted element. The recent availability of large amounts of mammalian genome sequence data in the databases,

combined with their relatively slow mutational rate, has made it possible to align confidently orthologous genomic regions from a wide spectrum of mammalian species. These alignments have provided a unique opportunity to estimate how many of the ancient TEs recognisable in the human genome have evolved under functional constraint. Using fairly stringent criteria, one study detected more than 10 000 human TE fragments that have clearly evolved under purifying selection throughout most of the eutherian radiation and therefore must have acquired a function (Lowe *et al*., 2007). Exapted elements belong to all TE classes, including several hundred ancient DNA transposons.

One pitfall of the comparative phylogenetic approach outlined below is that it can only detect those TEs subject to purifying selection for relatively long periods of time or with extreme intensity. With the current dataset, the method will mostly identify exapted TEs that inserted before the eutherian radiation. However, as pointed out earlier in this article, a large fraction (at least one-fourth) of human DNA transposons is primate specific (Pace and Feschotte, 2007). As more primate species are sequenced, the power to detect exapted DNA transposons and other functional elements in the human genome using comparative sequence analyses will probably increase.

But what are the functions of exapted human TEs? So far, there are only relatively few examples where experimental data have clearly established function of a given human TE. In almost all the cases examined thus far, the experimental data point to the involvement of the elements in regulatory functions, either at the transcriptional (e.g., enhancer) or post-transcriptional (e.g., alternative splicing) levels. So far, the TEs tested for functionality tend to be among the most frequently encountered type of elements in the human genome, such as short interspersed elements (e.g., Bejerano *et al*., 2006). As DNA transposons tend to be numerically less abundant, examples of exapted human DNA transposons with regulatory functions are still scarce in the literature. However, a remarkable example is provided by a copy of *MER113*, a member of the hAT superfamily, located in the distal promoter region of the gene encoding cholesteryl ester transfer protein (*CETP*). The *MER113* element was shown experimentally to contain several *cis*-regulatory sequences driving tissue-specific expression of the *CETP* gene (Jordan *et al*., 2003).

Another intriguing example involves a member of *MER20* (another hAT family), which contains *cis*-regulatory sequences directing the alternative expression of prolactin in extrapituitary tissues in humans, including the endometrium (Gerlo *et al*., 2006). Prolactin plays an essential role for the regulation of lactation in eutherian mammals. As the *MER20* family is known to have amplified before the divergence of eutherians, it is tempting to speculate that the insertion of *MER20* upstream of the prolactin gene was a key step in the regulatory evolution of lactation in mammals. Remarkably, MER20 has also been shown to play a role in coordinating gene regulation during the differentiation of endometrial cells required for pregnancy (Lynch *et al*., 2011). A subset of MER20 elements is evolutionarily conserved across placental mammals, and recruits at hundreds of sites the transcription factors necessary for endometrial differentiation. The MER20 elements can act either as insulators or enhancers to modulate adjacent gene activity in human endometrial cells in a progesterone/cAMP-dependent manner. Thus, the regulatory network apparently assembled from MER20 elements in the common ancestor of placental mammals may have been a key step in the evolution of prolonged pregnancy – a major physiological innovation of this phylum.

Several human TEs have been shown to have given birth to microRNA genes. With their TIRs and frequent palindromic structure, MITEs and other small DNA transposons are good candidates as an evolutionary source of microRNA genes. For example, it was recently established that the *mir-548* family of microRNA genes in humans is directly derived from a subset of *MADE1 mariner*-like MITEs (Piriyapongsa and Jordan, 2007). Interestingly, the *MADE1* family is anthropoid specific, thus *mir-548* and its targets must have emerged relatively recently (less than 50 Ma). This situation presents an excellent opportunity to study in detail the genesis and evolution of microRNA and their targets, issues that remain poorly understood. As more microRNAs are discovered and their precursor characterised in the human genome, the contribution of DNA transposons to the origin and biogenesis of microRNAs will certainly become clearer.

Transposase-Derived Genes in Humans

Another mode of TE exaptation is through the recycling, or domestication, of activities previously encoded by TEs to assemble new genes and evolve novel functions (Volff, 2006). V(D)J recombination, the process by which antigen diversity is generated in the immune system of humans and other jawed vertebrates, offers a spectacular illustration of transposon domestication. Genome sequence analyses (Kapitonov and Jurka, 2005) and biochemical studies (Hencken *et al*., 2012) have provided compelling evidence that RAG1, the enzyme mediating V(D)J recombination, and its associated recombination signal sequences evolved from an ancestral *Transib* DNA transposon. The progenitor element probably integrated in the genome of the common ancestor of all jawed vertebrates approximately 500 Ma. *Transib* elements per se are no longer recognisable in the human genome and seem to have gone extinct in all the vertebrate genomes currently available in the databases, but they have persisted in species that lack V(D)J recombination, such as sea urchin or mosquitoes (Kapitonov and Jurka, 2005). Perhaps, the extinction of *Transib* elements in the vertebrate lineage was concomitant to, and even requisite for, the advent of V(D)J recombination.

At first, it was unclear whether such domestication events would be limited to a few anecdotal yet impressive examples. However, recent studies suggest that exaptations of TE-coding sequences may be a common path for the emergence of new genes, and that DNA transposons in particular seem to represent a frequent source of new coding sequences in mammals (**Table 2**).

A first list of transcribed human genes entirely or largely derived from TE-coding sequences was compiled by Smit (1999) and was further extended to 47 genes in the initial analysis of the human genome sequence (Lander *et al*., 2001). All but four of these genes are derived from DNA transposons, despite the fact that these elements represent a relatively minor fraction of the human repeats (approximately 7%). These include the centromere-binding protein CENP-B (**Table 2**) and the human homologue of the *jerky* gene, which upon ablation in mice induces epileptic seizures (for review, Feschotte and Pritham, 2007). *CENP-B* and *jerky* are distantly related to each other and have no obvious functional connections, but each gene was derived independently from transposases of the *pogo* subgroup of Tc1/*mariner* elements (Smit and Riggs, 1996). *Pogo*-like transposases gave rise to new genes on several additional occasions, but the activities of the corresponding human proteins are generally unknown. Thus, *pogo*-like transposases were a recurrent source of TE-derived proteins, as were hAT transposases (see **Table 2**). Lander *et al*. (2001) stated their surprise regarding the prevalence of transposase-derived genes in the human genome: 'Why there are so many transposase-like genes, many of which contain the critical residues for transposase activity, is a mystery'. Since this seminal publication, additional examples have been described using more stringent criteria for assessing the functionality of the TE-derived genes, such as the presence of intact syntenic orthologues in other vertebrates (see **Table 2** and for an extended list, see Feschotte and Pritham, 2007). Again, the new examples point to the recurrent use of transposase domains as building blocks for the assembly of new proteins.

The transposases encoded by autonomous DNA elements possess several enzymatic activities that could enhance their propensity for domestication. Notably, all transposases studied so far contain an *N*-terminal DNA-binding domain (DBD). This domain has specific affinity for the termini (TIRs) of the cognate transposons to which it binds during the transposition reaction (see **Figure 1**). The transposase is produced by an autonomous element, but it can potentially bind to the TIRs of any related transposons dispersed in the genome. The *trans*-activity of the transposase is therefore largely determined by its DNA-binding specificity and therefore virtually any transposon containing two intact binding sites can be recognised and propagated. This property explains the accumulation of large number of MITEs and other transposase-defective elements that have retained intact TIRs. This characteristic may also explain why transposases appear to be frequently exapted, because not only the DBD of the transposase (**Table 2**) but also a suite of corresponding DNA-binding sites can be potentially recruited at once. Natural selection can then proceed to preserve only those binding sites beneficial to the organism and eliminate from the population those that might be deleterious. In this way, DNA transposon families can be seen as powerful generators of genetic networks poised for exaptation.

To test this model, it would be necessary to study a transposase-derived protein that is recent enough to trace not only the transposon ancestry of its DBD, but also of its binding sites. The human SETMAR protein might be an ideal candidate. Comparative sequence analysis demonstrated that *SETMAR* arose approximately 50 Ma in an anthropoid primate ancestor by fusion of a preexisiting SET histone methyltransferase gene to the transposase gene of an *Hsmar1* transposon inserted downstream of the SET gene (Cordaux *et al*., 2006). The structure and coding sequence of the *SETMAR* gene is highly conserved in all anthropoid primates examined and there is evidence that purifying selection has acted to preserve the transposase domain. The cellular function of SETMAR remains unknown. However, *in-vitro* experiments showed that the transposase region has retained the DNA-binding activity and specificity of the ancestral *Hsmar1* transposase. Notably, SETMAR binds specifically to a 19-bp binding motif derived from the TIRs of the cognate *Hsmar1* and related transposons (Cordaux *et al*., 2006; Miskey *et al*., 2007; Liu *et al*., 2007). These elements create a reservoir of approximately 1500 potential binding sites dispersed on all human chromosomes. This situation is consistent with the hypothesis that the capture of the transposase domain of SETMAR was accompanied by the recruitment of a subset of its DNA-binding sites scattered throughout the genome. It remains to be determined which of these sites are recognised by SETMAR *in vivo* as well as the function of SETMAR once it is bound to genomic DNA.

PGBD3 is another primate-specific transposase-derived gene that has been shown to form a fusion protein (Newman *et al*., 2008). *PGBD3* is located in intron 5 of the chromatin remodeller Cockayne syndrome Group B (*CSB*), and is spliced into the *CSB* transcript just after an acidic *N*-terminal region, replacing the ATPase domain required for CSB's chromatin remodelling activity by a seemingly full-length transposase domain. PGBD3 retains its ability to directly bind hundreds of related non-autonomous element dispersed throughout the human genome (Gray *et al*., 2012), consistent with the above mentioned regulatory network model. Perhaps more surprisingly, the fusion protein is also tethered to many additional genomic sites through protein–protein interactions with other transcription factors possibly influencing their effect on adjacent gene expression (Gray *et al*., 2012). Mutations that disrupt the *CSB* gene cause Cockayne syndrome (CS), a progeria characterised by developmental, and progressive neurological dysfunction (reviewed in Weiner and Gray, 2013). Most mutations that cause CS are after the site of fusion with PGBD3. These mutations presumably do not affect the production of the fusion product, but rather disrupt the chromatin remodelling capacity of

CSB. Intriguingly, expression of the CSB–PGBD3 fusion protein in cells lacking CSB induces an interferon response, which is normally observed in response to a viral infection. The interferon response is possibly due to the production of double-stranded RNAs, but the role of the CSB–PGBD fusion protein in generating the response and its contribution to the aetiology of CS remain unclear (Weiner and Gray, 2013).

Another interesting example of a transposase-derived gene establishing a regulatory network is the eutherian-specific transcription factor *ZBED6*, which is derived from a hAT/Charlie transposon (**Table 2**). In pigs, a naturally occurring point mutation in a ZBED6-binding site located in intron 3 of the *igf2* gene causes upregulation of *igf2* and several developmental defects. Chromatin immunoprecipitation of ZBED6-bound DNA in mouse cells revealed $\sim$2500 binding sites whose consensus sequence perfectly matched the *igf2* binding site in pig. Knocking down ZBED6 in mouse cells also upregulated *igf2* and caused increased cell proliferation (Markljung *et al*., 2009; Butter *et al*., 2010). Thus, ZBED6 possesses the hallmarks of a transcription factor evolved from a domesticated transposase in the common ancestor of placental mammals.

In conclusion, the human genome is host to a sizeable amount and a broad diversity of DNA transposons. These elements have integrated in the genome through multiple waves of transposition that occurred at different time points during mammalian evolution. There was significant activity before and during the primate radiation, but a seemingly general extinction of DNA transposons in the anthropoid lineage, some 40 Ma. Sequences derived from hundreds of human DNA transposon copies have evolved under functional constraint, suggesting that these elements have been a substantial source of genetic material for the emergence of new functional elements, including noncoding regulatory RNAs, such as microRNAs. In addition, more than 40 different human genes have originated by acquisition of coding sequences derived from transposases. We speculate that the capture of DNA-binding domains from transposases has promoted the lineage-specific emergence of new transcription factors and regulatory networks.

References

Beck BD, Lee SS, Williamson E, Hromas RA and Lee SH (2011) Biochemical characterization of metnase's endonuclease activity and its role in NHEJ repair. *Biochemistry* **50**: 4360–4370.

Bejerano G, Lowe CB, Ahituv N *et al*. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87–90.

Butter F, Kappei D, Buchholz F, Vermeulen M and Mann M (2010) A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth. *EMBO Reports* **4**: 305–311.

Cordaux R, Udit S, Batzer MA and Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences of the USA* **103**: 8101–8106.

Craig NL, Craigie R, Gellert M and Lambowitz AM (2002) *Mobile DNA II*. Washington, DC: American Society for Microbiology Press.

Feschotte C and Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* **41**: 331–368.

Feschotte C, Zhang X and Wessler S (2002) Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In: Craig NL, Craigie R, Gellert M and Lambowitz AM (eds) *Mobile DNA II*, pp 1147–1158. Washington, DC: American Society for Microbiology Press.

Gerlo S, Davis JRE, Mager DL and Kooijman (2006) Prolactin in man: a tale of two promoters. *BioEssays* **28**: 1051–1055.

Gilbert C, Schaack S, Pace JK II, Brindley PJ and Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of DNA transposons across animal phyla. *Nature* **464**: 1347–1350.

Gray LT, Fong KK, Pavelitz T and Weiner AM (2012) Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLoS Genetics* **9**: e1002972.

Gray YH (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics* **16**: 461–468.

Hencken CG, Li X and Craig NL (2012) Functional characterization of an acive Rag-like transposase. *Nature Structural and Molecular Biology* **19**: 834–836.

Jordan IK, Rogozin IB, Glazko GV and Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* **19**: 68–72.

Kapitonov VV and Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biology* **3**: e181.

Kiyosawa H and Chance PF (1996) Primate origin of the CMT1A-REP repeat and analysis of a putative transposon-associated recombinational hotspot. *Human Molecular Genetics* **5**: 745–753.

Lander ES, Linton LM, Birren B *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Leach DR (1994) Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays* **16**: 893–900.

Liu D, Bischerour J, Siddique A *et al*. (2007) The human SET-MAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Molecular and Cellular Biology* **27**: 1125–1132.

Lowe CB, Bejerano G and Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the USA* **104**: 8005–8010.

Lynch VJ, Leclerc RD, May G and Wagner GP (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics* **43**: 1154–1159.

Majumdar S, Singh A and Rio DC (2013) The human THAP9 gene encodes an active P-element DNA transposase. *Science* **339**: 446–448.

Markljung E, Jiang L, Jaffe JD *et al*. (2009) ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biology* **12**: e1000256.

Miller WJ, McDonald JF, Nouaud D and Anxolabehere D (1999) Molecular domestication – more than a sporadic episode in evolution. *Genetica* **107**: 197–207.

Miskey C, Papp B, Mates L *et al*. (2007) The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. *Molecular and Cellular Biology* **27**: 4589–4600.

Mitra R, Li X, Kapusta A *et al*. (2013) Functional characterization of piggyBat from the bat *Myotis lucifugus* unveils an active DNA transposon in a mammalian genome. *Proceedings of the National Academy of Sciences of the USA* **110**: 234–239.

Newman JC, Bailey AD, Fan HY, Pavelitz T and Weiner AM (2008) An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genetics* **4**: e1000031.

Pace JK II and Feschotte C (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Research* **17**: 422–432.

Pace JK II, Gilbert C, Clark MS and Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences of the USA* **105**: 17023–17028.

Piriyapongsa J and Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* **2**: e203.

Pritham EJ and Feschotte C (2007) Massive amplification of rolling circle transposons in the lineage of the bat *Myotis lucifugus*. *Proceedings of the National Academy of Sciences of the USA* **104**: 1895–1900.

Ray DA, Feschotte C, Smith JD *et al*. (2008) Multiple waves of recent DNA transposon activity in the bat *Myotis lucifugus*. *Genome Research* **18**: 717–728.

Reddy YV, Perkins EJ and Ramsden DA (2006) Genomic instability due to V(D)J recombination-associated transposition. *Genes and Development* **20**: 1575–1582.

Reiter LT, Liehr T, Rautenstrauss B, Robertson HM and Lupski JR (1999) Localization of mariner DNA transposons in the human genome by PRINS. *Genome Research* **9**: 839–843.

Reiter LT, Murakami T, Koeuth T *et al*. (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nature Genetics* **12**: 288–297.

Robertson HM (2002) Evolution of DNA transposons in eukaryotes. In: Craig NL, Gellert M and Lambowitz AM (eds) *Mobile DNA II*, pp 1093–1110. Washington, DC: ASM Press.

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development* **9**: 657–663.

Smit AF and Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the USA* **93**: 1443–1448.

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* **28**: 913–922.

Weiner AM and Gray LT (2013) What role (if any) does the highly conserved CSB-PGBD3 fusion protein play in Cockayne syndrome? *Mechanisms of Ageing and Development* **134**: 225–233.