The role of Transposable Elements in the Human Genome and their contribution to Evolution

Thomas van Ravesteyn

August 21, 2012

Utrecht University

Cancer Genomics and Developmental Biology

Universiteit Utrecht

cancer genomics & developmental biology

The role of Transposable Elements in the Human Genome and their contribution to Evolution

Author: Thomas van Ravesteyn

University: Utrecht University, The Netherlands

Master programme: Cancer Genomics and Developmental Biology

Study Component: Master thesis

Daily supervisor: Carolien G.F. de Kovel, PhD

Second examiner: Berend Snel, PhD

Table of contents

Introduction

Transposable elements (TEs) were originally discovered in *Zea mays* in the 1950s by Barbara McClintock (MCCLINTOCK 1956). In essence, they are genetic elements that are mobile and can move from one position to another within genomes. This concept fundamentally changed the view on genomes. Instead of rather static entities, it suggested that genomes are actually highly dynamic (Georgiev 1984). The draft of the human genome revealed that nearly half of the human genome is derived from transposable elements (Jurka *et al.* 2005; Lander *et al.* 2001). In fact, this is likely to be an underestimate as many ancient transposable elements probably have diverged beyond recognition. By contrast, it is important to realize that only 1.5% of the human genetic code is protein-coding. The proportions of TE derivatives in eukaryote genomes are highly variable, and each eukaryote has a specific complement of recently active TEs (Kidwell 2002).

Britten and Davidson hypothesized that repetitive elements can act to distribute regulatory sequences throughout the genome, and thereby enriching, possibly even creating, whole pathways (Britten and Davidson 1971). The adoption of a TE to a new function by the genome is called "exaptation" (Gould and Vrba 1982), and would enhance genetic innovation. Moreover, there are examples of TE derived gene products that may have a functional role in human cells. On the one hand, TEs provide many appealing mechanisms that could lead to beneficial variations to the human host. In this sense, they can be viewed as catalysts of evolution because their contribution to variation might have increased the speed of evolution on the human lineage (Britten 2010). On the other hand, these 'parasites of the genome' can lead to deleterious insertions which reduce human fitness. The present activity of TEs can result in *de novo* insertions in essential genes or regulatory regions, and leads to several genetic disorders (Belancio *et al.* 2008b; Callinan and Batzer 2006b).

Here, I will present an overview of the role of TEs in the human genome and their suggested influences on human evolution. First, I will describe TEs based on their classification, manner of transposition, and their distribution within the human genome. Second, I will focus on TEs from a functional perspective. The main question here is: have TEs adopted functional roles within the human genome? I will include analyses related to gene transcription and translation, their contribution to protein sequences, and their impact on genome stability. Third, I will approach TEs from a population genetic and evolutionary view. In this final chapter I will discuss the main findings from population genetic models of TEs, and the implications from elements which are under purifying selection. Using these three different perspectives, I will construct a comprehensive picture that makes it possible to evaluate the significance of TEs to the evolution of the human genome.

1 – Transposable elements from a descriptive perspective

1.1 - Classification of Transposable elements

TEs derived sequences are responsible for at least 45% of the human sequence (see figure 1) (Lander *et al.* 2001). As many elements probably diverged beyond recognition, this number is likely to be an underestimate. By the use of new methods which are able to track smaller elements, is was suggested that even two-thirds of the human genome originates from TEs (de Koning *et al.* 2011). However, TEs show large differences in their structure, copy number and activity (see figure 2). In order to describe TEs properly, they have been classified by their manner of transposition.

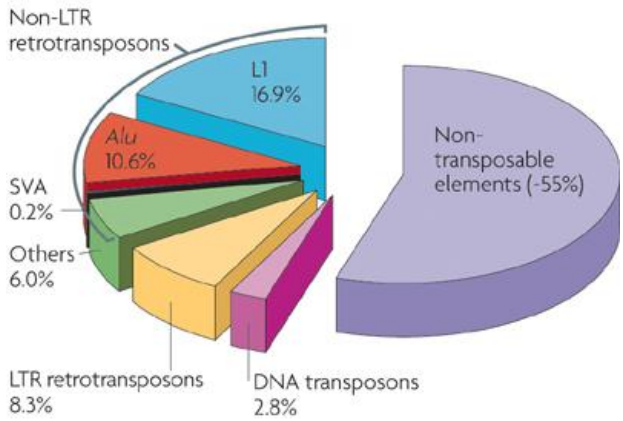

Figure 1: ***Distribution of Transposable elements within the human genome, transposable element derived sequences are estimated to make up at least 45% of the total genome (Cordaux and Batzer 2009).***

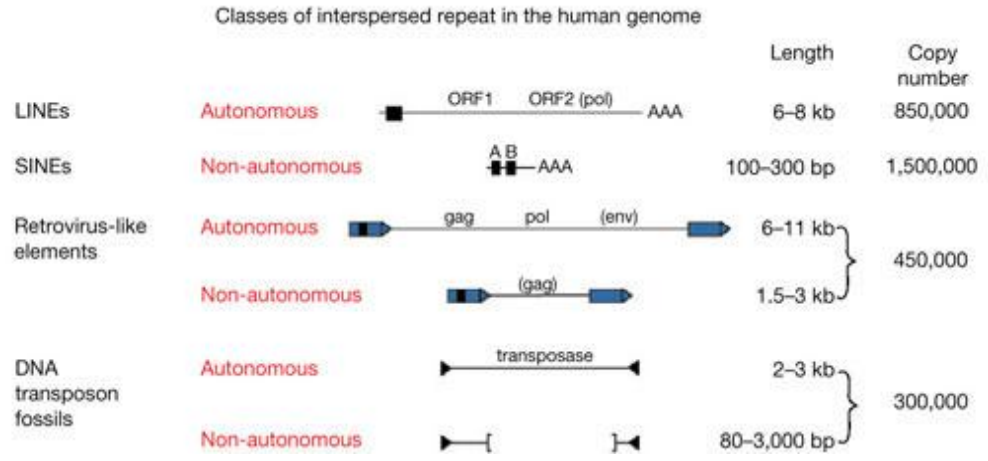

Figure 2: ***Classification, structure and estimated copy number of transposable elements within the human genome (Lander et al. 2001).***

1.1.1 - Class I elements

Class I elements replicate by a copy and paste mechanism. They duplicate through RNA intermediates that are copied into double stranded DNA by reverse transcriptase. After duplication they are inserted into the genomic DNA. The process of duplication and insertion is called retrotransposition. This major class is subdivided into groups which are distinguished by the presence or absence of long terminal repeats (LTRs).

1.1.1.1 - Endogenous retroviruses

Retrovirus-like elements are characterized by their LTRs, which carry all of the necessary transcriptional regulatory sequences. Human endogenous retroviruses (HERVs) are responsible for about 8% of the human genome and consist of three classes of endogenous proviruses, class I (gamma retroviruses), class II (beta retroviruses), and class III (spuma retroviruses). For humans and other mammals all the LTR elements are most likely to be decayed retroviruses that inserted within the cells of the germ line during evolutionary history. This allows for the vertical transmission of retroviral sequences from parent to offspring (Sverdlov 2000). HERVs inserted more than 25 Myr ago into our genome and their activity is presently very limited (Lander *et al.* 2001; Mills *et al.* 2007). Full-length elements contain *gag* and *pol* genes (see figure 1), which code for a protease, reverse transcriptase, RNAse H and integrase. Homologues recombination between the two LTRs results in solitary LTRs, which resemble most of the LTR-derived sequences in the human genome. Before HERVs lost their activity they gave rise to 9 human-specific insertions after the human-chimpanzee split (Chimpanzee Sequencing and Analysis Consortium 2005). Although most HERVs are inert, there might be a single exception. HERV-K elements integrated relatively recently, and might be still active because polymorphic elements still persist in human populations (Moyes *et al.* 2007; Turner *et al.* 2001). Interestingly, it appears that the gain of a cellular envelope gene (*env)* by endogenous retrotransposons resulted in the generation of some contemporary exogenous retroviruses. The acquisition of an envelope-like gene from a viral source allows for the transition of LTR retrotransposons to retroviruses. Phylogenetic evidence was provided for this process for the gypsy and related LTR elements (the insect errantiviruses), the Cer retroviruses in *C. elegans* and Tas element from *Ascaris lumbricoides* (common parasitic roundworm in humans) (Malik *et al.* 2000).

1.1.1.2 - Non-LTR elements

Most of the genomic sequence originating from TEs comes from the activity of non-LTR retrotransposons. These long and short interspersed elements (LINEs and SINEs respectively) form approximately one-third of the human genome. L1 (the most common LINE), Alu, and SVA elements are still active and can therefore lead to deleterious insertions.

1.1.1.2.1 - L1 elements

Over the past 150 Myr, L1 activity resulted in more than 500,000 copies (Lander *et al.* 2001). L1 elements are responsible for 16.9% of the human genome and a full-length L1 is about 6 kb long. A canonical element constitutes of a 5' UTR containing an internal RNA polymerase II (RNAPII) promoter (Swergold 1990), two open reading frames (ORF1 and ORF2), and a 3'UTR containing a poly-adenylation signal with an oligo(dA)-rich tail of variable length (see figure 2) (Babushok and Kazazian 2007). ORF1 encodes a RNA binding protein, and ORF2 encodes the protein with endonuclease and reverse-transcriptase activity which is required for retrotransposition (Babushok and Kazazian 2007). This unique molecular machinery makes the L1 elements the only autonomous TEs in the human genome. Many of the L1 copies are truncated by internal rearrangements and mutations (Szak *et al.* 2002). Therefore only a small subset of approximately 100 copies is functional (Brouha *et al.* 2003).

1.1.1.2.2. - Alu elements

Alu elements are primate specific and are active since ~65 Myr ago (Batzer and Deininger 2002). As a result, there are over 1 million Alu copies in the human genome, constituting 10.6% of the genome. These elements are characterized by their internal Alu endonuclease restriction site (Houck *et al.* 1979). In contrast to L1 elements, Alu elements have no coding capacity which makes them non-autonomous. Instead, they make use of the molecular machinery encoded by L1 elements for retrotransposition. Hence, Alu elements are sometimes referred to as 'a parasite's parasite' (Weiner 2002). The Alu element is in its typical form only 300bp in size and has a dimeric structure. The element contains two monomers derived from the 7SL RNA gene (a component of the signal recognition particle), which are separated by an A-rich linker region (Kriegs *et al.* 2007). The 5'region contains an internal RNA polymerase III (RNAPIII) promoter, the 3'region constitutes an oligo(dA)-rich tail of variable length (Batzer and Deininger 2002). Alus lack internal termination signals, this leads to extended Alu transcripts into the downstream flanking sequence (Comeaux *et al.* 2009; Shaikh *et al.* 1997).

1.1.1.2.3 - SVA elements

Within the human genome approximately 3000 SVA copies have been characterized, these elements have been active during hominoid evolution for ~25 Myr, (Ostertag *et al.* 2003; Wang *et al.* 2005). Like Alu elements, SVA elements are non-autonomous TEs. A full 2 kb SVA element is composed of a hexamer repeat region, an Alu-like region, a variable number of tandem repeats, a HERV K10-like region and a polyadenylation signal ending with an oligo(dA)-rich tail. Whereas Alu elements have an internal promoter, SVA elements might rely on promoter activity in flanking regions and are presumably transcribed by RNAPII (Ostertag *et al.* 2003; Wang *et al.* 2005).

1.1.1.3 - Subfamily structure

Different subfamilies exist for L1, *Alu*, and SVA sequences. Family members share specific insertions, deletions, and nucleotide substitutions. Most subfamilies show a continuous linear sequential evolution pattern in which one subfamily is derived from another. This process is explained by a few 'master' elements that are involved in retrotransposition and are, subsequently, responsible for the formation of all other subfamilies (Deininger *et al.* 1992). For instance, it was shown that all L1 subfamilies in the human genome were derived from a single lineage over the past ~40 Myr (Khan *et al.* 2006). This hypothesis is also supported by reports for Alu and SVA elements (Batzer and Deininger 2002; Wang *et al.* 2005). Furthermore, it was recently proposed that species-specific TE differences are largely determined by the population structure of the host (Jurka *et al.* 2011). It was suggested that the availability of new ecological niches, which led to the formation of sexually isolated human subpopulations, resulted to the generation of new TE families. For example, human AluYa5 and AluYb8 families may have originated from separate proto-human subpopulations. Despite the fact that master elements represent only a tiny fraction of the total human non-LTR retrotransposons, they are responsible for most of the L1, *Alu*, and SVA insertions. Therefore, they are thought to be the ultimate drivers of evolutionary change (Cordaux and Batzer 2009).

1.1.2 - Class II elements

Class II elements, DNA transposons, can excise themselves from a DNA strand, move as DNA, and insert themselves into new genomic locations. They resemble bacterial transposons, having internal inverted repeats and a transposase that binds near the inverted repeats and mediates mobility (see figure 2). In total, DNA transposons form approximately 3% of the human genome (Lander *et al.* 2001). There are at least seven major classes of DNA transposons, which can be subdivided into many families with independent origins (Smit 1996). It was estimated that at least 40 families of DNA transposons were active during the primate radiation, however, they are probably inactive since 37 Myr (Pace and Feschotte 2007). In order to survive, DNA transposon families must move by horizontal transfer to uninfected genomes. Transposition becomes less efficient if inactive copies accumulate in the genome. Transposase, which is produced in the cytoplasm before it enters the nucleus, cannot distinguish active from inactive elements (Clark and Kidwell 1997). As inactive elements increase in allele frequency this decreases the ratio of active over inactive elements. Therefore, transposase increasingly processes inactive instead of active elements and thus reduces the efficiency.

1.2 - Mechanism of retrotransposition

Non-LTR retrotransposons replicate via an RNA-based duplication process called retrotransposition. The process starts with RNAPII-mediated transcription of an L1 locus from an internal promoter at the 5'boundary of the element. Next, the L1 RNA is subsequently exported to the cytoplasm where ORF 1 and ORF2 are translated into an RNA-binding protein and a protein with endonuclease and reverse-transcriptase activity. Both proteins preferentially associate with the L1 RNA transcript that encoded them to produce a ribonucleoprotein (RNP) particle (Wei *et al.* 2001). Eventually, this

particle will be transported back into the nucleus by a poorly understood mechanism (Cordaux and Batzer 2009). Within the nucleus, the integration of the L1 element is proposed to occur through a process named 'target-primed reverse transcriptase' (TPRT, see figure 3) (Cost *et al.* 2002; Feng *et al.* 1996; Moran *et al.* 1996). First, the endonuclease creates a single-strand nick of target DNA, in most cases at 5'-TTTT/AA-3'consensus cleavage sites (Jurka 1997). The L1 RNA transcript anneals by its poly(A) tail, using the free 3'hydroxyl (OH) group generated by the DNA nick to the cleavage site and primes transcription. After synthesis of the complementary DNA copy, second strand synthesis is carried out by using the first strand as a template. Next, the second strand of the target DNA is cleaved and used to prime second-strand synthesis. The single stranded regions remaining in the genomic DNA at both ends are filled in, this creates target site duplications (TSDs) of 2-20 base pairs in length. Furthermore, the integration process generates distinct signatures like 5' truncations and an oligo(dA)-rich tail at the 3'end (Lander *et al.* 2001; Szak *et al.* 2002). If L1 integration causes partial deletions of target DNA, TSDs may not be formed (Gilbert *et al.* 2005; Morrish *et al.* 2002).

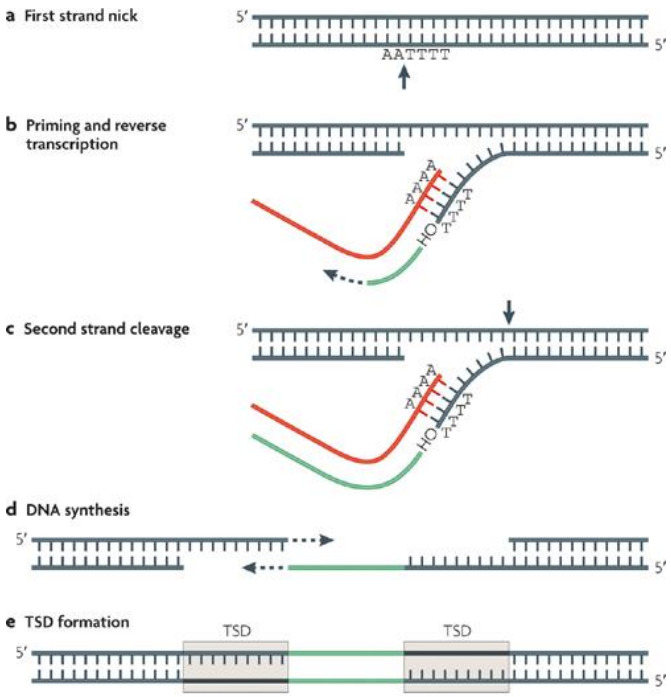

Figure 3: ***Integration of L1 elements, named Target-primed reverse transcription. At first, the endonuclease cleaves the first strand of target DNA (a). Second, L1 reverse transcriptase primes reverse transcription of L1 RNA (red) at free 3' hydroxyl group (b). Third, formation of a double strand DNA break (c). Fourth, second strand synthesis and recovery of double strand break (green, d). 3' Poly(A)-rich sequences and target site duplications at 3' and 5' ends resemble the hallmarks of the integration process (Cordaux and Batzer 2009).***

The precise mechanism of Alu and SVA trans-mobilization by L1 proteins still needs to be deciphered (Cordaux and Batzer 2009). Alu transcription is mediated by RNAPIII, and transcripts are exported to the cytoplasm and bound to signal recognition particle protein SRP9 or SRP14 to form stable RNPs (Chang *et al.* 1996; Sinnett *et al.* 1991). It remains unclear whether Alu RNPs gain access to the L1 machinery in the nucleus or the cytoplasm (Bennett *et al.* 2008; Kroutter *et al.* 2009).

Because the spread of TEs can be deleterious to the host cell, the cell has developed various processes to control TE activity. L1 elements have been regulated at the transcriptional level by the recruitment of new regulatory regions (Khan *et al.* 2006). Methylation of promoter DNA is known to repress L1 expression (Bourc'his and Bestor 2004; Hata and Sakaki 1997). Furthermore, L1 elements might be silenced by RNA interference (Soifer *et al.* 2005; Yang and Kazazian 2006). Another form of post-transcriptional regulation is the truncation of full length transcripts by premature polyadenylation (Perepelitsa Belancio and Deininger 2003). In addition, cells may produce proteins that inhibit L1, and therefore also Alu activity (Muckenfuss *et al.* 2006). Other factors which can reduce TE activity include: alteration of important motifs (Bennett *et al.* 2008; Comeaux *et al.* 2009), the genomic environment where Alu elements become inserted (Aleman *et al.* 2000; Chesnokov and Schmid 1996)(Aleman *et al.* 2000; Aleman *et al.* 2000; Chesnokov and Schmid 1996), and the overall length of Alu transcripts (Comeaux *et al.* 2009). The use of super-active versions of human and mouse L1 provide insight into the mechanisms of transposition and host control. These versions, with up to a 200-fold enhanced activity in retrotransposition assays, have been genetically engineered by recoding the open reading frames (An *et al.* 2006; Han and Boeke 2004).

1.3 - Insertion of Transposable elements

Analysis of *de novo* TE insertions can give insight in their behaviour and thereby the relevance of TEs to the human genome. Recent insertions generate polymorphic alleles which are hard to identify using conventional methods. Genomic repeats with extreme high copy numbers overwhelm hybridization-based assays, introduce artefacts in PCR amplifications, and generate unmappable reads. Therefore they are relatively hard to characterize in comparison with unique DNA elements. For TE unrelated studies such data is often left out from further processing. As a result, the appreciation of the importance of polymorphic repeats has lagged behind other areas of genomics (Burns and Boeke 2012). In order to detect *de novo* insertions new sequencing and data analysis techniques are in development which enable the identification of recent, family or even cell specific, insertions.

1.3.1 - Methods to identify insertions

Approaches for targeted recovery of insertions have been PCR-based methods. These methods amplify a known TE sequence along with their neighbouring, unique sequences. By exploiting internal characteristic repeat sequences, reactions gain specificity (Skowronski *et al.* 1988). For humans, this allows selective amplification of partially polymorphic L1 insertions which were inserted about 2 Myr ago (Myers *et al.* 2002). On the one hand, recent developments in genomic technology have led to more comprehensive methods for the discovery of retrotransposons insertion polymorphisms (RIPs). Complex mixtures of PCR-products are resolved by either genomic tilling microarrays (TIP-chip) or next-generation sequencing (TIP-seq) (Huang *et al.* 2010; Wheelan *et al.* 2006). On the other hand, computational methods outpace these advances as their power increases to identify TE insertion events *in silico* (Burns and Boeke 2012). The comparison of genomes is also useful for the analysis of rare archaic hominid DNA. By analysis of the individual sequence reads used to assemble the published Neanderthal and Denisovan genomes, new

insertions of HERV-K sequences were identified (Agoni *et al.* 2012). In total, 14 new integration sites were defined where modern humans contain the corresponding, empty, preintegration sites.

1.3.2 - Insertion rates of Transposable elements

The use of new techniques for transposon identification provides a more complete catalogue of common polymorphisms. Additionally, this made it possible to estimate the level of activity of these elements in present day humans. By comparing the reference genome with newly discovered insertions, estimations were made for insertion rates. It was expected that 1 in 21 individuals would have a new Alu, 1 in 211 would have a new L1, and 1 in 916 would have a new SVA element (Xing *et al.* 2009).

Another study based on TIP-chip, estimated new L1 insertions for 1 in 108 births (Huang *et al.* 2010). Using a distinct prediction method, these numbers were confirmed, the rate of *de novo* L1 transposition was estimated between 1 in 95 and 1 in 230 births (Ewing and Kazazian 2010). These high transposition rates imply the presence of highly active mobile elements within the human DNA. These rates in combination with population size, enables the calculation of overall numbers of human insertion alleles. For L1, this may be as many as 12,000 segregating insertions with allelic frequencies greater than 0.05. This suggests that there will be a growing list available of common variants for transposable elements in the near future (Burns and Boeke 2012).

L1 retrotransposition kinetics can be analyzed using *in vitro* retrotransposition assays. For example, the use of an enhanced green fluorescent protein (EGFP) retrotransposition cassette would allow the detection of acquisition of a retrotransposition event in single cells (see figure 4) (Ostertag *et al.* 2000). In this case, HeLa cells are transfected with a L1-EGFP construct. Cells only express EGFP when an L1 transcript, that contains an antisense EGFP marker, undergoes the process of splicing, reverse transcription and integration into genomic sequences. Then, EGFP is expressed from a pCMV promoter and cells can be analyzed by FACS-sorting.

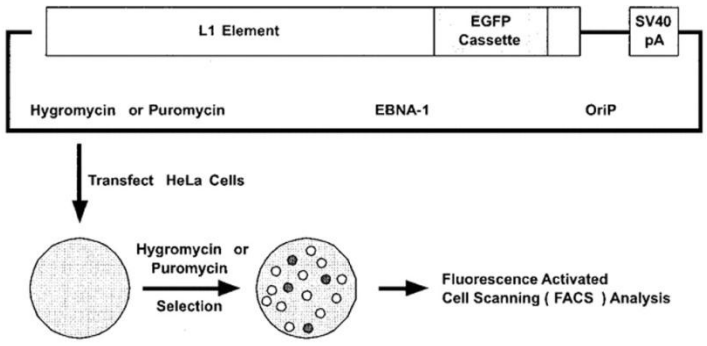

Figure 4: ***L1 elements containing the EGFP cassette are cloned into mammalian expression vectors, which places a SV40 poly(A) signal downstream of the L1 element. These vectors can replicate in HeLa cells by using an eukaryotic origin of replication. Vectors contain antibiotic resistance genes which allow for selection. Finally, cells are analyzed by FACS under UV light (Ostertag et al. 2000).***

The use of such an assay demonstrated that a small proportion of the potentially active L1 elements in humans dominate the transpositional potential (Brouha *et al.* 2003). It was estimated that on average each individual contains 80-100 potentially active L1's, and 6 L1s per haploid genome that are highly active. More recently, the activity of novel L1 insertions within the human population was examined in detail. Low frequency alleles were recovered using a fosmid paired-end DNA sequencing strategy to identify indels. Although these "hot-L1s" are relatively uncommon, the majority of 68 full-length L1s found in fosmid libraries from six individuals were shown to be 'hot' (Beck *et al.* 2010). This provides evidence that each genome harbours about hundred competent L1s with some highly active variants. These variants are about 1 Myr old and segregate in low frequency through the human population. Considering the small number of hot-L1s and the potential loss of their activity suggest that there is substantial individual variation in retrotransposition capability (Seleme *et al.* 2006). Overall, these results strongly support the 'master gene' hypothesis as suggested by Deininger.

Human non-LTR TEs have been active over tens of millions of years and lead to differences in copy number among primates. For instance, more than 7500 copies were accumulated in our lineage since the split from chimpanzees (Mills *et al.* 2006). However, the rate of amplification was variable over time. For instance, most L1 subfamilies were extensively amplified 12-40 Myr ago (Khan *et al.* 2006). Most Alu elements inserted about 40 Myr ago, during this amplification peak there was approximately one new Alu integration in every new birth (Shen *et al.* 1991).

Furthermore, is was found that the Alu Yb lineage, which was found in all hominoid genomes examined, expanded to about 2000 copies specifically within the human genome (Han *et al.* 2005a). By comparing the human genome with the sequence data from non-human hominoid primates, it was shown that these non-human primates carry only a handful of Alu Yb elements (Carter *et al.* 2004; Gibbons *et al.* 2004). This element probably inserted in hominoid genomes between 18 and 25 Myr ago. This means that these TEs survived within the human genome with low retrotranspositional activity for a period of approximately 20 Myr. As the Alu Yb lineage underwent a remarkable expansion in the human lineage during the past few million years, a 'stealth driver' model was proposed (Han *et al.* 2005a). This model suggests that long-lived, and low activity master elements occasionally produce short-lived hyperactive copies that are responsible for the expansion of Alu elements in the genome. In other words, these elements are quiescent enough to escape negative selection but are still capable of producing progeny and thereby increase in allelic frequency. The reason for the low level of activity within these Alu subfamilies is unknown, multiple reasons have been proposed for reduction of Alu retrotransposition, including altered transcription, Alu RNA secondary structure, or reduced TPRT ability (Deininger and Batzer 1999). Reducing the retrotranspositional activity might be a common evolutionary strategy of various retrotransposons (Han and Boeke 2004; Li and Schmid 2004). In fact, Alu Yb elements escaped from negative selection successively and were able to persist in the long term.

1.4 - Distribution of TEs in the human genome

Some genomic regions are extremely dense in TEs, for example, up to 89% overall density in a 525kb region (Lander *et al.* 2001). While other regions of the genome are nearly devoid of these elements, these regions include the four homeobox gene clusters, HOXA, HOXB, HOXC and HOXD. They accommodate less than 2% of interspersed elements over a region of 100kb. Probably, these regions are under purifying selection because they contain many *cis*-acting regulatory regions which do not tolerate any insertions. This example shows that evaluation of the distribution of TEs may help to elucidate which regions are under purifying selection in the human genome. Moreover, it might be useful to examine TE distribution in order to gain understanding about the possible regulatory functions of TEs and their derivates. In addition, this may give insight in the mechanism of transposition and the consequences of integration events. Hence, many studies have been performed to analyze the distribution of TEs in genomes more precisely. Correlations have been reported for GC-content, imprinted regions, expression patterns and transcription factor binding sites.

1.4.1 - GC content

GC content is the percentage of guanine and cytosine bases in a region of the genome. Interestingly, genes are often found to have a higher GC content in contrast to the background level of the genome. On the whole, LINE sequences are roughly fourfold enriched in AT-rich regions, by contrast SINEs show an opposite trend (Lander *et al.* 2001; Smit 1999; Soriano *et al.* 1983). LTR elements and DNA transposons show a more uniform distribution, although they generally are less prevalent in GC-rich regions. LINE sequences preferentially insert at locally AT-rich regions (Jurka 1997), this explains why they are not distributed randomly. This strategy seems reasonable because insertions within gene poor regions impose a lower negative effect, hence they are tolerated. For example, relatively young LTR elements (class II ERV) are less prevalent than expected within genes based on the GC content of the evaluated segment and the whole genome GC (Medstrand *et al.* 2005). These patterns reflect an true integration preference for gene-poor regions (Kurdyukov *et al.* 2001).

However, the inverse correlation between SINEs and GC content is less obvious. Is has been suggested that the insertion of Alu elements in gene-rich, or in high GC-content, regions is beneficial (Schmid 1998). Proposed was that SINE RNAs may regulate protein synthesis in response to cell stress by signalling protein kinase R (PKR), which eventually inhibits translation. This regulatory mechanism would imply a selective advantage for the maintenance of TEs within the host genome. In accordance, Alu sequences are not distributed randomly, they are found in high densities in gene-rich regions (Medstrand *et al.* 2002). Furthermore, three-quarters of all genes have Alu sequences in their flanking regions (Grover *et al.* 2004). These findings highlight the possibility of a relevant role for transposable elements in gene regulation.

Nonetheless, alternative explanations for high Alu numbers in gene-rich regions have been put forward. Young TEs have high GC content in comparison with their surroundings. Thus, a constant influx of TEs tends to increase overall GC content. Therefore, the accumulation and fixation of Alus in GC-poor regions would be passively avoided because it would severely change the local composition and thereby affect gene transcription. This would be the effect of evolutionary dynamics rather than insertion preferences. The current pattern of Alu and LINE distribution would be the result of

genomic stability, and a major part of Alu elements in GC poor regions would have been lost (Pavlcek *et al.* 2001; Rynditch *et al.* 1998). Others suggest that insertions are more likely to be retained in GC-rich regions, because a deletion event would simultaneously remove functionally important sequences (Brookfield 2001). Thus, once a TE is inserted in an important region and is not directly harmful, it is unlikely to be deleted because this has a high chance to destroy functionally important regions at the same time.

1.4.2 - Gene expression

In order to gain insight into the often suggested role of TEs on gene expression, studies have been conducted to analyse the distribution of TEs among promoters, regulatory elements, and UTRs. Here, I will briefly go over the most basic findings, the next chapter continues with a more functional evaluation of the relation between TEs and gene expression.

In short, promoters are the genomic elements which bind RNA polymerase directly upstream from the transcription start site of a gene. They regulate and initiate gene transcription. By analysis of two thousand promoters Jordan *et al.* (2003) found that 24% contained TE-derived sequences from all common human TEs. This suggests that TEs potentially have a role in the regulation of gene transcription. Within promoter sequences it was shown that SINEs are overrepresented. This observation is in agreement with the notion that SINEs are mostly found in GC-rich regions and LINEs within AT-rich regions. There is consistent decrease in the contribution of TEs to promoters from distal regions to the transcription start site. Probably, proximal insertions have on average more negative effects with respect to gene regulation in comparison with insertions further upstream.

In order to prove functionality of TEs as regulatory site, it is necessary to relate TE insertions to experimentally confirmed transcription binding sites. However, Jordan *et al.* (2003) found only 21 *cis*-regulatory sequence elements overlapping with TEs. Nonetheless, some interesting individual examples were identified. For example, several TE derived regulatory elements were found within the β-globin locus. Furthermore, a relatively high abundance of TEs was found in 3'UTRs of mRNA in comparison with 5'UTRs and coding sequences. Both UTRs often harbour relevant regulatory elements that act either on the level of transcription or translation. The enrichment of TEs in 3'UTRs may reflect a lack of selection against fixation of TEs within these relatively long segments. Nevertheless, there have been examples reported in which TEs contributed to regulatory elements on 3' UTRs (Brosius 1999b).

TEs have also been related to another form of transcriptional regulation, this involves the formation of chromatin loops by the attachment of specific stretches of DNA to the nuclear scaffold or matrix (Bode *et al.* 1996). It was found that described nuclear scaffold/matrix attachment regions (S/MAR sequences) are enriched in TE-derived sequences (Jordan *et al.* 2003). Especially LINE elements are abundant, in total 98 consensus sequences were found to contain 14 different S/MAR recognition signatures (van Drunen *et al.* 1999). As a result, TEs appear to have a role in gene regulation by promoting the partitioning of the human genome into distinct transcriptional regions. The isolation of specific genetic regions enables the regulation of the genes in a more collective manner. A genomic region that contains multiple genes can be actively rearranged which leads to a general up or down regulation of gene expression. This may involve histone modifications or the movement of

genetic regions towards local high concentrations of transcriptional and mRNA-processing machinery within the nucleus. Such mechanism may simplify the regulation of multiple genes in a tissue specific manner or in response to external stimuli (Fraser and Bickmore 2007).

Another study specifically evaluated the role of Alus on gene expression. As Alu sequences are enriched for CpG dinucleotides (Jurka 2004), they could contribute to an increase in expression breadth by introducing CpG islands upon insertion. CpG islands are rich in CpG dinucleotides, and these stretches of DNA are commonly found upstream of genes which are expressed over a wide range of tissues (Larsen *et al.* 1992). This hypothesis has recently been tested, however it was found that genes which have always had broad expression are richest in Alus, whereas those that are more likely to have become more broadly expressed have lower enrichment (Urrutia *et al.* 2008). This enrichment is not explained by the relation of both expression breadth and Alu density, to regional GC content. This finding is consistent with a model in which Alus accumulate near broadly expressed genes, but do not affect their expression breadth. The abundance of Alu near broadly expressed genes is better explained by their preferential preservation near to housekeeping genes rather than by a modifying effect on expression of genes (Urrutia *et al.* 2008). These results provide no evidence for a functional role of TEs on the regulation of housekeeping genes across the genome; instead it suggests that Alus just tend to accumulate in the vicinity of housekeeping genes. This finding is in agreement with the earlier suggested conservation of Alu sequences in gene-rich regions.

1.4.3 - Genomic imprinting

TEs have also been related to genomic imprinting. Genomic imprinting is an epigenetic process that ensures monoallelic gene expression of a pair of genes, present on homologous chromosomes. The process involves gene silencing by DNA methylation and histone modifications, which are established in the germ line and are maintained throughout the somatic cells. The gene that is imprinted and thus inactive, is always the same member of a pair. For some genes this affects the maternal allele, for others the paternal allele. Imprinted genes are distributed around the genome, but tend to cluster.

It has been observed that imprinted regions lack SINEs significantly, both primate specific Alu and more ancient mammalian SINEs, are less frequently associated with imprinted than non-imprinted genomic regions (Greally 2002). The sharp reduction of SINE contribution to imprinted regions may help to predict the presence and extent of these characteristic regions. A disturbed insertion pattern at imprinted regions, was not found for L1 elements. Apparently, L1s continuously inserted into imprinted regions during mammalian evolution. This shows that the low abundance of SINEs in comparison with L1s, cannot be easily explained by the fact that insertions in general can be deleterious. Moreover, SINEs are probably mobilized by the same L1 machinery (Jurka 1997), the low frequency of SINEs is therefore unlikely to be caused by a primary failure of retrotransposition. Therefore, another explanation has been put forward. At loci where the paternal or maternal allele will be imprinted, there is a higher chance of negative effects if SINEs would infer with gene regulation. As SINEs may have a functional role in gene regulation by attracting and spreading methylation *in cis* to flanking sequences (Hasse and Schulz 1994; Yates *et al.* 1999), SINE insertions at imprinted regions could lead to severe effects. In contrast to non-imprinted regions, there is no back-up available for alleles of which the counterpart on the homologous chromosome is already

silenced by imprinting. As a result, SINE integrations at imprinted regions will be less tolerated in contrast to L1 insertions, or SINE insertions at other genomic regions. Negative selection against SINE integrations at imprinted regions may explain the observed SINE depletion.

1.4.4 - Presence of Transposable elements near conserved genes

Shown was that highly conserved genes, e.g., genes with essential functions in metabolism, development or cell structure, have a low prevalence of TEs in their mRNAs (van de Lagemaat *et al.* 2003). This intuitively means that TEs insertions which change the expression of fundamental genes, are not tolerated and hence are strongly selected against. By contrast, it was observed that Alu and L1 sequences are more common in flanking regions of highly expressed and housekeeping genes (Kim *et al.* 2004). Another study, that also took into account the isochore type, long DNA segments which are relatively homogeneous in GC content, found that TE enrichment near, not in, housekeeping genes is not a by-product of variable insertion rates among different genomic compartments (Eller *et al.* 2007). Interestingly, it was shown that repetitive sequence environment actually distinguishes housekeeping genes from tissue-specific genes in every isochore. This implies that TEs have a role in the separation of gene regulation programs of essential and tissue-specific genes.

However, younger and mammalian-specific genes, such as those involved in immunity and those that have expanded during mammalian evolution, are enriched with TEs in their mRNAs (van de Lagemaat *et al.* 2003). A possible explanation for this finding is that these relatively young genes are initially more tolerant to insertion as they have more 'freedom' in fulfilling their tasks. Over time, inserted TEs might evolve to more functional roles in transcription regulation. TE insertions might constrain and specify the functionality of specific genes that are partially under control of TE derived sequences. Therefore, the donation of regulatory elements by TEs would enhance the diversification and evolution of young mammalian genes (van de Lagemaat *et al.* 2003).

1.5 - Conservation of Transposable elements

When genes are highly conserved, meaning that they are highly similar over long evolutionary time scales and hence between orthologs in related species, they often are essential to the organism. For example, the crucial genes that code for ribosomal RNA change relatively slowly. Therefore, the comparison of orthologous genes can help to identify genes or regulatory elements that have a functional role. The same is true for TEs, although TEs are often used to calibrate the molecular clock, highly conserved TEs imply some benefit to the host.

At first, it was thought that specific SINE families were restricted to only a few species or a single genus (Shedlock and Okada 2000). However, examples are now provided for more widely distributed SINE families, of which their members share the same conserved sequence. For example, members of the CORE-SINE superfamily share a 65 bp of "core"-sequence in their central regions (Gilbert and Labuda 1999). They are referred to as mammalian-wide interspersed repeats that proliferated before the radiation of placental mammals. The CORE element is supposed to identify an ancient tRNA-like SINE element which survived presumably for more than 550 Myr in different lineages. The

authors relate the successful longevity of this superfamily to the recruitment of an internal promoter from highly transcribed host RNA, and its presumed capacity to exchange sequences with active LINEs.

Another example is presented by Nishihara *et al.* (2006), they reported high conservation for a new SINE family which was presumably active during the Carboniferous period, at least 310 Myr ago. At this time the amniotes evolved, they have an adapted egg that allowed egg survival without water. This tetrapod vertebrate group includes the mammals, reptiles and birds. This newly identified amniota-wide SINE family (AmnSINE1) is characterized by a central domain, in total 105 human AmnSINE1 copies were described that are phylogenetically conserved among mammalian orthologs. The chimeric structure of AmnSINE1 consists of a 5S rRNA and a tRNA-derived SINE. Moreover, this study reported related SINE families that belong to the DeuSINE superfamily (deuterostomia SINE, see figure 5). These newly described SINE families share a common central domain that is also found in zebrafish, namely SINE3. Because of the high conservation of the central Deu-domain, it has been suggested that these TEs have been exaptated within mammalian genomes (Nishihara *et al.* 2006). A search within the human expressed sequence tag database, resulted in three positive hits. These hits code for mRNA of Protein phosphatase 2 regulatory subunit B α isoform, Epsin 2, and cGMP-inhibited 3',5'-cyclic phosphodiesterase A. Possibly, AmnSINE1 elements function as part of each mRNA. In addition, these SINEs are present in introns, they might be transcribed as pre-mRNA and have some role in mRNA processing. However, it is clear that additional research is required to elucidate their possible functionality. Eventually, this may lead to better understanding of their contribution to the evolution of mammals.

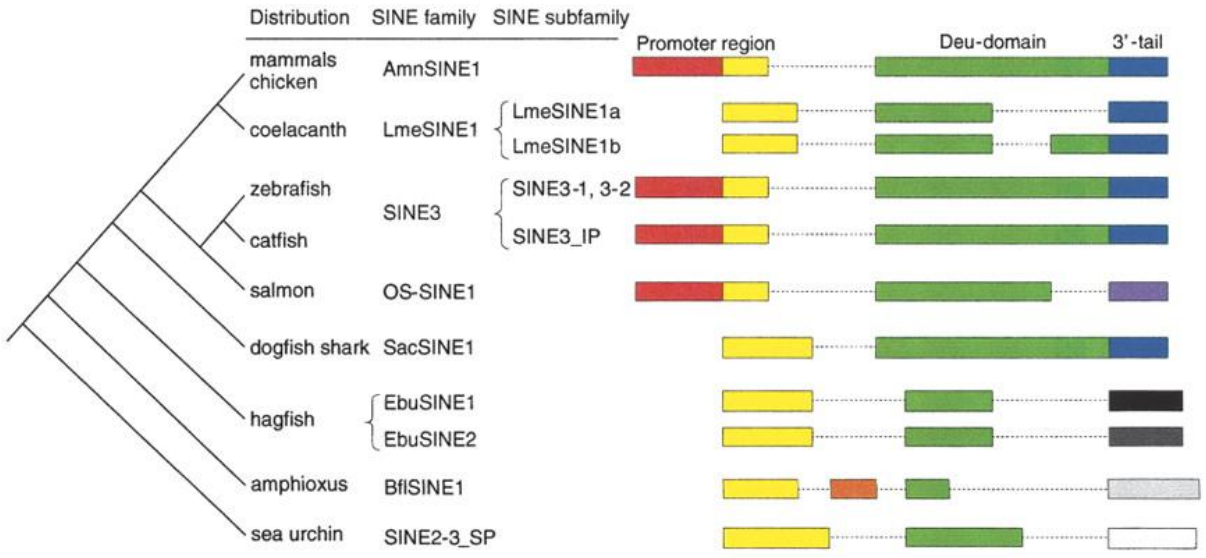

Figure 5: ***Phylogenetic reconstruction, structure and distribution of the DeuSINE superfamily. Shown are common DeuSINE sequences (green boxes), promoter regions derived from tRNA (yellow boxes) and 5S rRNA (red boxes), 3'tails similar to that of zebrafish (blue boxes) and of rainbow trout (purple). Grey and white boxes are distinct and of unknown origin (Nishihara et al. 2006).***

Last, an interesting example is provided by the MER121 repeat family. Whereas the family members show considerable variation, the individual copies within families are highly conserved in orthologous locations across the human, dog, mouse and rat genomes (Kamal *et al.* 2006). Most copies retained a ~ 150 bp central region, although the flanking regions are not always present, they reveal a similar conservation rate. Although these elements are non-coding and lack transcripts, MER121 might encode cis-acting regulatory or structural elements. The conservation of some 6-mers within the elements may reflect protein binding sites. For example, the most highly conserved 6-mer, is a consensus for E-box motif bound by transcription factor USF. The authors speculate that this repeat element may have been picked up by a TE around 200 Myr ago. The element was able survive if insertions were advantageous, and were locally fine-tuned to result into functional elements. The distribution of a Drosophila insulator by the gypsy LTR retrovirus was achieved by a similar mechanism (Gdula *et al.* 1996).

2 – Transposable elements from a functional perspective

If TEs lose their ability of transposition they may get 'domesticated' by the genome. Since the discovery of TEs, speculations have been published about their potential functional roles in modulating gene expression and their contribution to protein coding sequences (Flavell 1995; Georgiev 1984; Jacob 1977; McDonald 1993). The distribution of TEs in the genome highlighted already some of the possible functional roles. Although there are many examples of recruited neogenes in a variety of organisms (Pinsker *et al.* 2001; Sarkar *et al.* 2003), their function is often unknown.

However, it possible that the genetic code of some functional proteins to date, are actually derived from the bases that were provided originally by TEs. For example, the RAG proteins, which play an important role within the somatic VDJ recombination system, share important similarities with DNA transposons. Both involve recombinase activity and recognize DNA which is enclosed by recombinase binding sites. It was shown that RAG1 and RAG2 together form a transposase that can mediate a complete transposition in vitro (Agrawal *et al.* 1998; Hiom *et al.* 1998). This process gives rise to the hallmarks of TE insertion, the final DNA product contains a short duplication of target sequence that immediately flanks the transposed fragment. These findings support the hypothesis that RAG proteins were once components of a TE.

Another ancient protein that might have arisen from TEs is telomerase. Telomerase is the enzyme that maintains the chromosome ends of eukaryotes by synthesizing telomeric repeat sequences. This essential enzyme might be derived from a reverse transcriptase that originates from non-LTR retrotransposons. Actually, it is not precisely known who gave rise to who (Eickbush 1997). Interestingly, the telomeres in Drosophila are extended by retrotransposition (Levis *et al.* 1993). Nevertheless, a comprehensive phylogenetic study based on the reverse transcriptase domain was conducted to shed light on the origin of non-LTR elements (Malik *et al.* 1999). This study suggested that non-LTR elements are as old as eukaryotes, with eleven clades dating back to the Precambrian era.

The Mart gene family, which is related to the gag gene of Sushi-like terminal repeat transposons from fish and amphibians, is thought to be an ancient TE that is exaptated by primates and mammals (Brandt *et al.* 2005). The Mart gene family is present in human (11 copies) and other primates, Mart expression was confirmed for six genes during mouse embryonic development. Gene expression was also observed for adult mice, with variable tissue-specific gene expression patterns. Next, it was found that Mart2, Mart3, and Mart4 contain a zinc finger domain which suggests a function that involves the binding of DNA. In addition, two autosomal *Mart* genes are subject to imprinting. Last, some mart genes might be involved in the regulation of cell proliferation and apoptosis (Nagasaki *et al.* 2003).

In conclusion, exaptation could explain why some TEs have been maintained over evolutionary time scales. However, only reports of functional TEs are able to support this model. Here I will give an overview of studies that suggest, or contradict a functional role for TEs.

2.1 – Effect of Transposable elements on gene regulation

During evolution TEs have provided the base pairs for promoters and enhancers for numerous examples, and thereby likely influenced gene expression in more or less significant ways (Bejerano *et al.* 2006; Bourque *et al.* 2008; Jordan *et al.* 2003; van de Lagemaat *et al.* 2005). Moreover, TEs can provide alternative splice sites, poly(A)-signals, silencers, binding sites, response elements and increase mRNA stability (Brosius 1999a). Taken together, there have been many suggestions for the influence of TEs on different levels of gene regulation, including gene transcription and translation.

2.1.1 - Promoters

In order to express a gene, it is required that a protein complex is formed 5' of the coding sequence. RNA polymerase is a multi subunit protein that eventually synthesizes the primary transcript while it travels along the DNA strand. The core promoter is the genomic region where the initiation complex is formed. The assembly of this complex can occur in the absence of enhancer elements, but then it is rather inefficient. The composition of the promoter and distal enhancer elements is an essential feature in the recruitment of the initiation complex, and both activating and repressing transcription factors. As a consequence, the promoter is an essential determinant in general and tissue-specific gene regulation.

TEs can integrate near promoter elements, 25% of human promoter regions that have been analysed have been shown to contain DNA sequences that are derived from mobile DNA (Jordan *et al.* 2003). TE integrations upstream of silent genes can occasionally lead to the activation of these genes, this is often accompanied by the alteration of the gene's the tissue specificity. SINEs can act as migrant carriers of promoters/enhancers by integrating near or even into genes and thereby changing their temporal and spatial expression patterns (Brosius 1999a). Even dormant pseudogenes may be reactivated by a neighbouring SINE. Resident Alu elements have the potential to turn into a regulatory element by changes in its sequence, in the associated gene or by juxtaposition to a gene by recombination at the locus of the SINE (Brosius and Gould 1992)

Analysis of approximately 250,000 TE-derived transcription start sites revealed that their transcripts are generally tissue specific (Faulkner *et al.* 2009). Furthermore, this study identified 2000 TE derived bidirectional promoters genome-wide, bidirectional promoters can initiate transcription on both complementary DNA strands in opposite direction. In addition, the same study showed that TEs located directly 5' of protein-coding loci often have a role as alternative promoter and/or express non-coding RNAs. TE derived promoters can lead to read-through transcription by RNA polymerase II and thereby interfere with regular gene expression (Speek 2001). In order to prevent read-through transcription, termination signal must be placed upstream from inserted promoter sequences.

Promoters that are enriched for TEs are both more highly and broadly expressed, on average, than promoters that lack TEs (Huda *et al.* 2009). In addition, promoters that have similar repetitive DNA profiles regulate genes that have more similar expression patterns and encode proteins with more similar functions than promoters that differ with respect to their repetitive DNA. Moreover, distinct repetitive DNA promoter profiles are correlated with tissue-specific patterns of expression.

2.1.2 - Transcription factor binding sites

The binding of transcription factors to specific recognition sites influences the processes involved in transcription initiation and RNA synthesis. Stimulating and inhibitory molecules can bind to DNA sequences that regulate gene expression in tissue-specific manner, or in response to signals from outside the cell or during specific developmental stages.

Numerous examples have been published for cases of TE derived sequences which bind transcription factors and thereby functionally regulate gene expression (Bejerano *et al.* 2006; Brosius 2003; Jordan *et al.* 2003; van de Lagemaat *et al.* 2003). For example, it was demonstrated, using a mouse enhancer assay, that two SINEs function as distal transcriptional enhancers in developing mouse embryos (Sasaki *et al.* 2008). One SINE locus is an enhancer which regulates FGF8 expression in two regions of the developing forebrain. Both reported enhancers function specifically within the developing forebrain, possibly they contributed to mammalian-specific brain formation.

Many promoters and enhancers that have been derived from LTR sequences are primarily active in the placenta. Examples include the endothelin-B receptor (Medstrand *et al.* 2001), midline 1 (Landry *et al.* 2002), pleiotrophin (Schulte *et al.* 1996) and aromatase CYP19 (van de Lagemaat *et al.* 2003). These examples show that TEs might function as a transcriptional linker. By TE transposition it is possible to spread specific transcription binding sites throughout the genome that influence gene expression similarly. Binding sites could either be formed by specific mutations within TE sequences that don't interfere with the ability of transposition, or by the acquirement of binding motifs by active TEs due to small chromosomal rearrangements. This enables the transcriptional linkage of genes that are expressed in the same tissue (Lercher *et al.* 2002). This mechanism can produce new regulatory elements rather instantaneously, while the formation of new transcription factor binding sites would otherwise need multiple mutations.

Based on the enrichment of rather general GO terms, it was suggested that Alu associated binding sites have mainly a role in developmental processes (Polak and Domany 2006). It was observed that there is an enrichment of biosynthesis genes with Alu associated binding sites for transcription factors that are active during stages of development. The expansion of complex binding sites for transcription factors via TEs is supported by several studies (Johnson *et al.* 2006; Mortazavi *et al.* 2006). For instance, binding motifs of the neuronal repressor NRSF/REST would have been generated by lineage-specific TEs. In addition, a recent study found that five transcription factor binding sites (ESR1, TP53, POU5F1, SOX2, and CTCF) are embedded in distinctive families of transposable elements (Bourque *et al.* 2008). These repeat-associated binding sites (RABS) are associated with major regulatory expansions throughout the mammalian lineage and were subjected to evolutionary selection toward good binding motifs. These results were supported by the trend that older repeats show an enrichment for binding motifs. Furthermore, there is recent evidence that TEs contain sequences for regulatory assemblies that restructure tissue-specific transcriptomes (Kunarso *et al.* 2010; Lynch *et al.* 2011). The role of TEs in gene regulation is additionally supported by another recent report in which the binding the mammalian insulator CTCF was found to be related to a transposable element (Schmidt *et al.* 2012).

In short, once a binding site has been acquired by a TE, it may spread through the genome generating copies of the original motif sequence. These studies indicate that transcriptional networks are highly dynamic in eukaryote genomes and that transposable elements might play an important role in expanding binding motifs. In conclusion, these findings argue for a transcriptional linkage model of transcription binding sites that were amplified by TE mobility.

2.1.3 - Effect of TEs on Transcriptional elongation

TE insertions can lead to changes in the composition and characteristics of the genomic DNA. This could affect processes in which the DNA serves as a substrate. For instance, it was found that L1 ORF sequences in the sense orientation serve as a poor substrate for gene transcription by RNAPII (Han *et al.* 2004). A nuclear run-on assay was performed to evaluate RNAPII activity across ORF2 containing sequences. Although ORF2 does not inhibit transcription initiation, these experiments revealed that RNAPII gradually is reduced as transcription runs over the ORF2 element. If RNAPII stalls, or even dissociates from the DNA strand, as it runs across long L1 elements, it is expected that L1 insertions weaken endogenous gene expression. Therefore the insertion of L1 sequences on a transcript decreases RNA expression and therefore protein production. Although such L1 insertion seems to have rather adverse effects, the authors argue for a model in which L1 elements affect gene expression genome-wide by acting as a "molecular rheostat" of target genes. In this view, L1 insertions are thought to provide variation to gene expression levels during evolution. Thus, the cumulative effect of L1 insertions would help to fine-tune the human transcriptome. However, one could discuss the real beneficial value of such a mechanism. Possibly, negative selection had simply not enough power to eliminate these L1 insertions and other solutions were brought upon.

2.1.4 - Alternative splicing mediated by Transposable elements

Upon the discovery of introns, it was thought that each single gene always produces the same mRNA. However, we know nowadays that the primary transcript can follow several alternative splicing pathways. Each transcript can therefore lead to the synthesis of a range of proteins. Thus, the central dogma of 'one gene, one protein' has been completely overthrown. In conclusion, alternative splicing is an important way of the genome to regulate and enhance the number of possible proteins available.

For humans, at least 5% of all alternatively spliced exons within protein coding regions contain sequences from Alu elements (Sorek *et al.* 2002). It has been shown that Alu consensus sequences contain up to ten potential 5′ donor splice sites and 13 potential 3′ acceptor splice sites. Alternative splicing is often regulated in a tissue-specific manner, and hence leads to different protein products (Yeo *et al.* 2004). A mechanism was proposed that governs 3′ splice-site selection during alternative splicing in gene exons with Alu insertions (Lev Maor *et al.* 2003). Two positions on antisense orientated *Alu* sequences are mostly used as 3′ splice sites in *Alu* exonizations.

Besides Alu elements, L1 elements also contain several functional splice donor and acceptor sites, although the largest part is predicted to be weak (Belancio *et al.* 2006). Evidence has been provided that the splicing of primary transcripts which contain inserted L1 elements, is delayed by

endogenous protein that influences the splicing process (Belancio *et al.* 2008a). Only if gene introns are defined by weak splice sites, L1 splice sites might be able to compete and interfere with normal splicing. Nevertheless, intronic L1s could result in the production of aberrantly spliced mRNAs. As a consequence, L1 sequences could have a tissue-specific effect on gene expression. This may be beneficial for the regulation of the proteome within tissues. On the other hand, if L1 sequences delay the process of splicing, this might hint for a defence mechanism of the host which reduces the amount of functional L1 proteins produced. In this case, the alternative splicing pathway is adapted to repress the transpositional activity of L1 elements.

2.1.5 - The initiation of polyadenylation by Transposable element derived signals

The synthesis of most mRNAs is combined with a process called polyadenylation. At the 3' end of eukaryotic transcripts, a template-independent RNA polymerase adds a series of up to 250 adenosines. Polyadenylation is directed by a signal encoded in the RNA, often 5'AAUAAA-3', and is located upstream of the polyadenylation site. This internal site is cleaved to create a new 3' end to which the poly(A) tail is subsequently added. Many genes have more than a single poly(A) signal, this means that termination can take place at several positions, and thus leads to transcripts with different 3' UTRs. This enables tissue-specific regulation of the processing of gene transcripts. Again, this is a mechanism that has the potential to regulate protein production (Zhang *et al.* 2005).

L1 elements contain multiple functional poly(A) signals. Intronic insertion of a L1 can therefore result in the truncation of full-length transcripts by premature polyadenylation (Perepelitsa Belancio and Deininger 2003). To investigate the direct effect of L1 ORF2 elements on expression, Jeffrey *et al.* (2004) fused ORF2 coding sequences downstream of the green fluorescent protein ORF. This showed that the anti-sense insertion of a L1 inhibits full-length transcript synthesis primarily because of premature polyadenylation. Besides L1 features of abortive polyadenylation, a tiny fraction (1%) of the approximately 10,000 *Alu* sequences in human 3' UTRs is functionally active as poly(A) signal (Chen *et al.* 2009). Interestingly, it appears that just a few point mutations within Alu hexamers that resemble poly(A) signals and/or within flanking GT-rich regions, can produce efficient poly(A) signals. The authors claim that Alu inserts not necessarily represent weak poly(A)-signals, instead they would often represent the major or even the only poly(A)-signal in a gene.

In the end, these examples show that the insertion of TEs, whether or not in combination with a few mutations, can lead to premature polyadenylation of primary transcripts. Possibly, the TE derived alternative poly(A)-signals had a functional role in changing the expression patterns of some proteins during evolution.

2.1.6 - The effect of Transposable elements on nucleosome binding

The modification or remodelling of nucleosomes, the DNA-histone complex, can change the accessibility of the DNA for transcription factors and can thus influence gene expression (see also section 1.4.2). In some cases nucleosome remodelling within a local genomic region is required for gene activation. Moreover, dense packaging of DNA is related to gene inactivation, these

heterochromatic regions are so compact that the accessibility for proteins in severely affected. In contrast, euchromatin has a more open conformation and harbour actively transcribed genes.

It has been observed that L1 insertions are drawn towards open chromatin in transcribed regions (Cost *et al.* 2001). In addition, human genes are more likely to be highly expressed, and in broad patterns if their promoters are TE rich (Huda *et al.* 2009). Although TEs are enriched distal from TSSs, they are in general excluded from core promoters (Jordan *et al.* 2003). These findings indicate that TEs are related to the chromatin structure upstream of the transcription start site (Huda *et al.* 2009). In fact, it has been shown that different classes of repetitive elements mediate nucleosome binding in different ways (Huda *et al.* 2009). TEs bind nucleosomes tightly whereas microsatellites, repeating DNA sequences of 2-6 base pairs, have low nucleosome affinity and are enriched upstream of transcription start sites.

If purifying selection or the specific inhibition of a TE was not strong enough, other mechanisms might have been evolved that limit the negative effects of TEs. Therefore, it has been suggested that heterochromatin evolved as a control mechanism to silence TEs (Henikoff and Matzke 1997). Earlier, it was shown that compact heterochromatin is enriched for both TEs and microsatellites in several organisms (Dimitri and Junakovic 1999). Actually, heterochromatin does reduce the deleterious effects from TEs by repressing transcription and ectopic recombination between dispersed element sequences (Grewal and Jia 2007). Although interestingly and initially appealing, the suggestion of heterochromatin as a mechanism to silence TEs is highly speculative.

2.2 - The adoption of Transposable elements in coding sequences

Because TEs contain several splice sites, they can contribute to gene diversity and versatility (Belancio *et al.* 2006; Yeo *et al.* 2004). At first it was estimated that up to 4% of the human protein coding sequences harbours TEs (Nekrutenko and Li 2001), however, a more recent analysis on protein level suggested that only about 0.1% of all protein coding genes includes sequences derived from TEs (Gotea and Makalowski 2006). For only three (*CAPN1, GZMA and PTPN1*) out of the 3764 Protein Databank (PDB) entries a TE cassette could be convincingly identified. It should be noted that this percentage is likely to be an underestimate, because the used PDB-collection contains only well characterized proteins.

Older Alu subfamilies are significantly overrepresented in Alu-containing exons (Sorek *et al.* 2002). An intuitive explanation is that older elements had more time after insertion to diverge and acquire the potential to serve as splice sites. By investigation of 152 human chromosomal loci where Alu elements were exonized based on expressed sequence tags, and detailed phylogenetic reconstructions of four specific examples (*RPE2, C-rel, MTO1 and PKP2b*), it was suggested that Alu exonization took place at various evolutionary time points within primate lineages (Krull *et al.* 2005). Predictions of inclusion or exclusion of a sequence to produce mature RNA probably remain difficult, besides the requirement to acquire prominent motifs, the local environment, secondary structures and the presence of additional sequence motifs need to be considered. A protein-coding function may therefore be acquired relatively soon or could instead take millions of years.

Singer *et al.* (2004) reconstructed the key events which lead to the formation of a novel receptor isoform. In fact, a 5'exon was generated from an alternative transcript in the human tumour necrosis factor receptor gene (*p75TNFR*) that contains an ancient Alu element. Insertion of the Alu element and the formation of a new, alternative transcription start site took place around 59 Myr ago in the common ancestor of the higher primates. After the creation of an alternative start codon and splice site, an open reading frame was introduced between 40 and 25 Myr ago on the catarrhine lineage (Old World monkeys including apes) (Singer *et al.* 2004). This reconstruction illustrates that the exonization of an Alu sequence was based on multiple key mutations that were created just by chance. In short, although Alu exonization is possible, it is highly unlikely that this happens on short evolutionary time scales.

Only a few reports are available for functional proteins containing TE Insertions that encode amino acids. For example, it was shown that one expressed splice form of the human *RED1* in brain and heart tissues contains an Alu element (Gerber *et al.* 1997). The insertion of this TE is located within the centre of the deaminase motif, the catalytic core the protein. Characterization of two alternative spliced isoforms showed that they have the same substrate specificity, but differ in their catalytic activity. Another study observed that an isoform of Casein kinase 2 (CK2), a highly conserved and ubiquitously expressed tetrameric enzyme, has a liver-specific subcellular localization (Hilgard *et al.* 2002). The unique CK2α" isoform contains an Alu sequence and is either a CK2α-derived retrotransposon or is the result of alternative splicing.

Last, it was demonstrated that an Alu-derived motif interacts *in vitro* with tau, a microtubule associated protein that has been implicated in several neurodegenerative diseases (Hoenicka *et al.* 2002). This interaction might be involved in the regulation of tau phosphorylation, and may therefore play a role in cellular localization of tau.

Although there are many TEs found within the coding regions of many genes (Nekrutenko and Li 2001), there are only a handful of reports of potentially functional proteins containing TE derived sequences. The fact that TEs are found within coding sequences and their transcripts does not necessarily have to lead to translation of these elements. Several mechanisms within cells can initiate degradation of the transcript, or destruction of the protein product directly after translation (Wagner and Lykke Andersen 2002). Moreover, the TE containing protein might be non-functional or possibly deleterious (Lovell 2003). As a result, it is hard to predict whether TE sequences are really incorporated as protein coding elements. Although the present count of functional protein proteins containing TEs is rather low, it is expected that this number will increase slightly with ongoing research in the future. Nevertheless, the role of TEs in regulatory elements is more profound.

2.2.1 - Gene breaking – a mechanism by which Transposable elements give rise to new genes

Several mechanisms can lead to the formation of new protein coding genes, the most important step might be the formation of an open reading frame and the arrangement of a promoter. After all, only if the assembly of a transcription initiation complex is successful, gene expression and subsequent processing can lead to functional RNAs or proteins.

With this in mind, it was found that the antisense promoter and poly(A)-signals of full-length L1 elements can lead to the formation of new genes by division of pre-existing genes *in vitro* (Wheelan *et al.* 2005). If an L1 element inserts in an intron this may split the gene's transcript into two smaller transcripts (see figure 6). In this scenario, one transcript starts normally from the native gene promoter and includes the 5' exons, however, it terminates at the major antisense poly(A)-site of the inserted L1. This results in a shortened transcript with a premature abortion as consequence from the inserted poly(A) signal. Transcription of the second transcript starts at the L1 antisense promoter and contains the remaining 3' exons. The endogenous poly(A) signal is used and the transcript contains the original 3'UTR. Furthermore, Wheelan *et al.* (2005) identified three human genes and 12 candidate genes that were divided by L1 elements. These new transcripts might encode potentially interacting (e.g. MET*)* or novel proteins (e.g. BCAS3*)*. The relative abundance of transcripts from broken genes is expected to be influenced by numerous factors. These include: the strength of the endogenous promoter, RNA stability, the activity of the inserted anti-sense promoter, and the possible effect of RNA interference on L1 containing transcripts. The majority of the three identified and 12 candidate L1 elements leading to gene-breaking integrated probably before the split between the human and chimpanzee lineages (Wheelan *et al.* 2005).

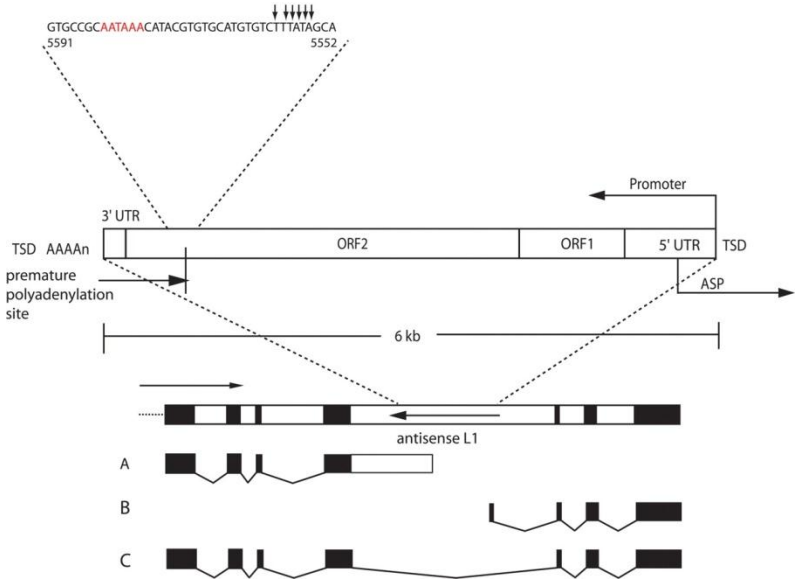

Figure 6: ***Gene-breaking model. A generic gene is shown with an intronic antisense L1 integration. Three transcripts may result from this arrangement: transcript A, which terminates premature. This transcript contains the 5' exons, part of the intron, and part of L1. Transcript B, transcription initiates at the antisense promoter in L1 and contains part of L1 sequence and the downstream exon. Transcript C shows the native (expected) transcript. Arrows indicate direction of transcription; the arrow for the antisense L1 indicates transcription from the native L1 promoter. Poly(A) signal (red); small arrows, polyadenylation sites; ASP, antisense promoter (Wheelan et al. 2005).***

2.2.2 - Retrotransposon-mediated transduction

In the case of gene breaking, new genes arise by the destruction of original or duplicated genes. Another mechanism was proposed in which coding sequences are duplicated by the insertion of TEs. In turn, when L1 and SVA elements duplicate within the genome, they can take flanking sequences along. This process of retrotransposon-mediated transduction (RMT) can duplicate either 3' or 5' sequences (Damert *et al.* 2009; Hancks *et al.* 2009). In the case of 3' transduction, transcription is initiated at the TEs promoter, however, during the transcription process the RNA polymerase machinery skips the poly(A) signal and continues through the flanking sequences. This leads to an extended transcript with downstream flanking sequences incorporated. Next, the transcriptional start from an alternative upstream promoter leads to 5' transduction in which the transcript is lengthened by upstream elements. The resulting TE transcript and additional flanking sequences can integrate at new genomic locations and thereby generates duplications. Thus, RMT can lead to the duplication of coding sequences. For example, the AMAC1 gene was duplicated multiple times through SVA mediated transduction and generated three transduced copies (see figure 7) (Xing *et al.* 2006). These events took place during primate evolution approximately 7-14 Myr ago and duplicated promoter sequences along. The combined duplication of both coding and promoter elements, suggests that duplicated genes can retain their functional potential within the target genomic environment. Hence, RMT might lead to the rapid expansion of functional gene families.

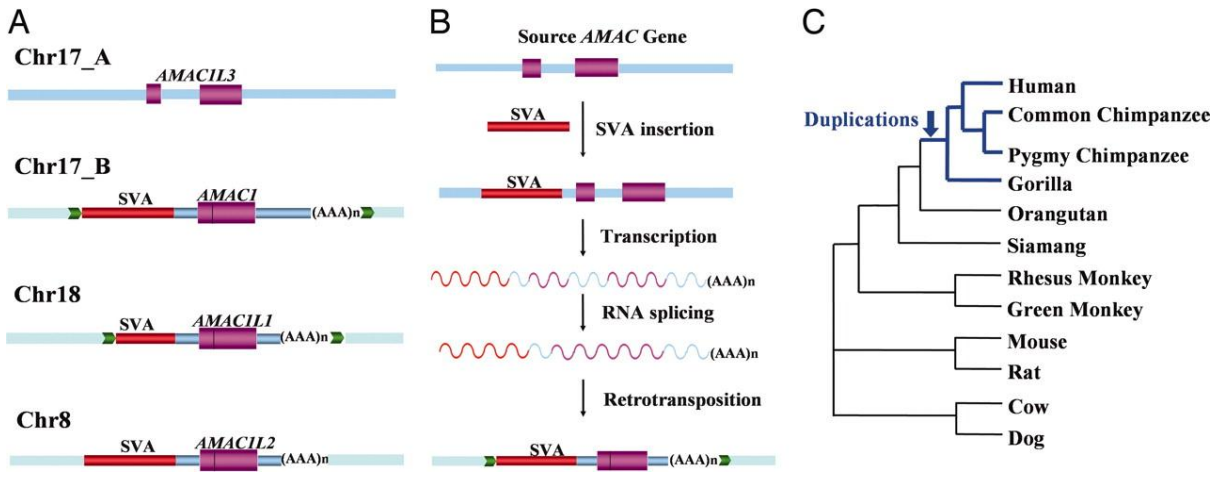

Figure 7: ***SVA transduction-mediated gene duplication for the AMAC genes. (A) Schematic diagram, showing SVA elements (red), coding sequences (purple) , transduced sequences (blue), flanking sequences of transduced loci (light blue) and TSDs (green). (B) Schematic diagrams for putative evolutionary scenarios of the SVA transduction-mediated gene duplications. First, insertion of SVA element upstream of AMAC gene locus. Next, transduction of full-length AMAC gene by transcription of active SVA. Removal of intron during RNA processing, and integration into new genomic regions (Xing et al. 2006).***

2.3 - The induction of structural variation by Transposable elements

Structural variation refers to changes in chromosome build. Examples include: deletions, duplications, copy-number variants, insertions, inversions and translocations. Large variations are often not tolerated, because rearrangements of genes and regulatory elements often leads to disrupted gene expression.

2.3.1 - Recombination mediated deletions

Because of the extremely high copy numbers for TEs, e.g. more than 1 million Alus within the human genome, they can create structural variation between non-allelic homologous elements. Hence, genome rearrangement by TEs is not related to the activity of TEs. Ectopic rearrangement can result in deletions, duplications and inversions.

Since the human-chimpanzee split there have been 492 Alu recombination mediated deletions (RMDs) and 73 L1 RMD events, as identified by genome-wide comparisons (Han *et al.* 2008; Sen *et al.* 2006). Together these events might have removed nearly 1Mb of genomic sequence from the human genome. In comparison with polymorphisms found in present-day humans this number is rather small. As there are more than 70 reported cases of Alu RMDs and 3 cases of L1 RMDs related to cancer and genetic disorders (Callinan and Batzer 2006a; Han *et al.* 2008), there is probably a strong negative selection against RMDs. This shows that TEs can have a direct negative effect on human fitness.

Retroelements can, in rare cases, be precisely deleted from primate genomes, most likely via recombination between 10- to 20-bp target site duplications flanking the retroelement (van de Lagemaat *et al.* 2005). Recombination between the target site duplication does not involve the TE sequence, and hence can be removed completely. Therefore, the deleted loci are indistinguishable from pre-integration sites, which effectively reverses the insertion event. Through human-chimpanzee-Rhesus monkey genomic comparisons it was estimated that 0.5%-1% of the apparent retroelement insertions distinguishing humans and chimpanzees actually represent deletions. This means that some apparent lineage specific insertions were already present in the common ancestor. As a result, this study challenges the idea of the unidirectionality of retrotransposons. Moreover, these mechanisms reveal that the most parsimonious explanations are not always true. This is of major importance when inferring functional novelties from the integration of TEs.

2.3.2 - Insertion-mediated deletions

The insertion of TEs at target DNA can lead to the deletion of adjacent genomic DNA. In short, small deletions are likely to be caused by the formation of double stranded cleavages by L1 endonuclease activity that are inexactly opposed, and is followed by 5'–3' exonuclease activity on both exposed 5' ends. Large deletions are explained by the invasion of L1 cDNA at a double strand DNA break, in combination with processes involved in gap repair that delete intervening single stranded DNA. The deletion of target DNA was first observed by Gilbert et al. (2002) and Symer et al. (2002) through analysis of L1 integrations in vitro. Whereas 16–25% of L1 insertions identified *in vitro* cause

deletions at the target site (Gilbert *et al.* 2002; Gilbert *et al.* 2005; Symer *et al.* 2002), only about 2.2% of existing human-specific L1 insertions seem to be directly linked to genomic deletions (Han *et al.* 2005b). The large difference between *in vivo* and *in vitro* L1 related deletions can't be fully explained by a slight underestimation for the *in vivo* rates due to the different levels of completion of human and chimpanzee genome projects. Instead, the authors suggest that it reflects natural selection that limits the number of deletions after L1 insertion that can persist.

By evolutionary analysis of the human and chimpanzee genome it was found that 50 deletions were directly linked to the insertion of L1 elements (Han *et al.* 2005b). Over the last 4-6 Myr approximately 18 kb of human genomic sequence and about 15 kb of chimpanzee genomic sequence was lost. In total, L1 insertions may have lead to more than 11,000 deletion events - up to 7.5 MB of target sequences – during the 60 Myr of primate radiation. These numbers are rather low, especially in comparison with the amount of base pairs that is added to the genome by L1 insertion in the same period. It is obvious that the large L1 mediated deletions such as those indentified in cell culture assays do not persist over longer time scales. These aberrations result in the loss of genes, which cannot be tolerated. Therefore, the authors proposed new mechanisms for the creation of some specific L1 structures (Han *et al.* 2005b). A similar study identified that a single insertion-mediated deletion caused a coding difference between humans and chimpanzees in the past 5 Myr (Callinan *et al.* 2005). Specifically, the gene *C-rel* was lost within the human lineage. This gene might have a functional role in regulating cell proliferation and differentiation (Bishop and Varmus 1992).

2.3.3 - Duplications

Gene duplication is a very important mechanism by which new genes can be generated. After gene duplication, both genes will be identical, although distal regulatory elements might differ. Selective constraints will ensure that one of both copies remains identical. It is likely that this copy continues to supply the coding sequence for its function. However, the duplicated copy is free from evolutionary constraints and will accumulate mutations at random. For some instances this might lead to changes in protein composition and structure which enables the gene to acquire new functions.

Within the human genome there are many duplicated regions which are highly similar. Alu elements are found within approximately 24% of the boundaries of these recent segmental duplications (Bailey *et al.* 2003). It was shown that in particular the young Alu elements were responsible for the enrichment at the junctions of duplicated regions. This observation might indicate that Alu elements were involved in the expansion of 5% of the human genome over the last 40 Myr. The authors proposed a model in which the primate-specific burst of Alu transposition 'sensitized' the human genome for Alu recombination-mediated duplications and thereby formed the basis of gene-rich segmental duplications. The model is in agreement with the expansion of interchromosomal duplications and the primate burst of Alu retrotransposition about 35 Myr ago. This shows that the insertion of TEs may catalyze the speed by which genomes evolve. For example, the recombination between a L1 pair about 35 Myr ago, is suggested to be responsible for the duplication of the ß-globin gene that eventually generated the Gγ and Aγ members of this gene family (Brown 2007).

2.3.4 - Inversions

In the case of inversions, another type of structural variation, recombination between homologues recombination can result the reversion of genomic regions. By another human and chimpanzee genome comparison it was suggested that TEs can cause chromosomal inversions (Lee *et al.* 2008). The proposed mechanism, retrotransposon recombination-mediated inversion (RRMI), involves the formation of a secondary structure by TEs or an increased probability of double stranded breaks. From a total of 49 observed RRMI loci, 28 human specific inversions were identified. Whereas RMD leads to genomic deletions which can alter or disrupt gene function, RRMI does not change genome size. By contrast, it could invert sequences within genes, as a result alternative splice sites may be introduced. Three RRMI events lead to changes in the exonic regions of known genes and ten RRMI events were found polymorphic within a species. As a result, the authors suggest that RMMI generates variation between and within species (Lee *et al.* 2008).

In conclusion, TEs are related to a large number of structural variations. Based on their high copy number they increase the chance of recombination events that lead to genomic deletions, duplications and inversions. As recombination is one of the driving forces that can generate diversity within genomes, this indicates that TEs probably had a profound effect on chromosome structure. However, it is hard to really prove that specific recombination events, and for example gene duplications, are indeed the result of the presence of TEs.

3 - Transposable elements from a population genetic and evolutionary perspective

The persistence of TEs in the face of negative selection was originally explained by their success as replicating units (Doolittle and SAPIENZA 1980; Orgel and CRICK 1980). Because of their high amplification rates they were able to maintain their position within the vertebrate lineage. This concept resolved the C-value paradox, which refers to the immense, counterintuitive and seemingly arbitrary differences in genome size observed among organisms (Hartl 2000). It has been argued that the transition from prokaryotes to multi cellular eukaryotes is associated with immense reductions in population size (Lynch and Conery 2003). This hypothesis is strongly supported by the observed correlation between effective population size (N_e) and total genome size. N_e determines the degree to which gene frequencies are faithfully transmitted across generations. Reductions in N_e enhance the power of genetic drift, the power of random genetic drift appears to vary by multiple orders of magnitude between prokaryotes and vertebrates. Increased random genetic drift causes wilder fluctuations of allele frequencies and faster fixation of neutral elements. Reduced N_e lowers the power of natural selection which enables the fixation of mildly deleterious elements. Thus, decreased N_e leads to more random changes in allele frequency, and thereby enabled the extensive proliferation of various slightly deleterious genomic features that would otherwise be eliminated by purifying selection (Lynch and Conery 2003).

N_e is influenced by the number of individuals that contribute to reproduction, the level of random mating, and is related to the actual population size but cannot exceed this number. N_e is an important population parameter in population genetics, it helps to evaluate how genetic architecture and human populations evolved during history. N_e can be estimated by genome comparisons or from linkage disequilibrium (LD) data. In general, human N_e is quoted as 10,000, however, estimates appear to be much lower (Takahata 1993). Interestingly, N_e was also estimated from Alu evolution, this resulted in a N_e of approximately 18,000 during the last one to two million years (Sherry *et al.* 1997). By contrast, a more recent estimate from LD data suggest an effective population size of less than 3000 from European ancestry samples during recent evolution (Tenesa *et al.* 2007). However, it is important to notice that it is not really meaningful to discuss N_e without a reference in time. For instance, population bottlenecks and expansions can have large effects on N_e and therefore purifying selection (Hayes *et al.* 2003).

If TE insertions have a negative effect on fitness, then, the number of insertions capable of drifting to fixation decreases with increasing N_e. Moreover, TEs acquire deleterious mutations suggesting that individual members of TE families must generate new insertions in order to survive on evolutionary time scales within a host (Le Rouzic and Deceliere 2005). It was observed that TEs appear to have a genome size threshold (see figure 8). Lynch and Conery (2003) identified that TEs are unable to establish themselves below the genome size threshold, and are present in all genomes above about 100 Mb. In short, they suggested an overall correlation between the N_e of contemporary populations and the total genome size. If N_e is sufficiently small, then, even slightly deleterious TEs might drift to fixation because they are able to move under the radar of purifying selection.

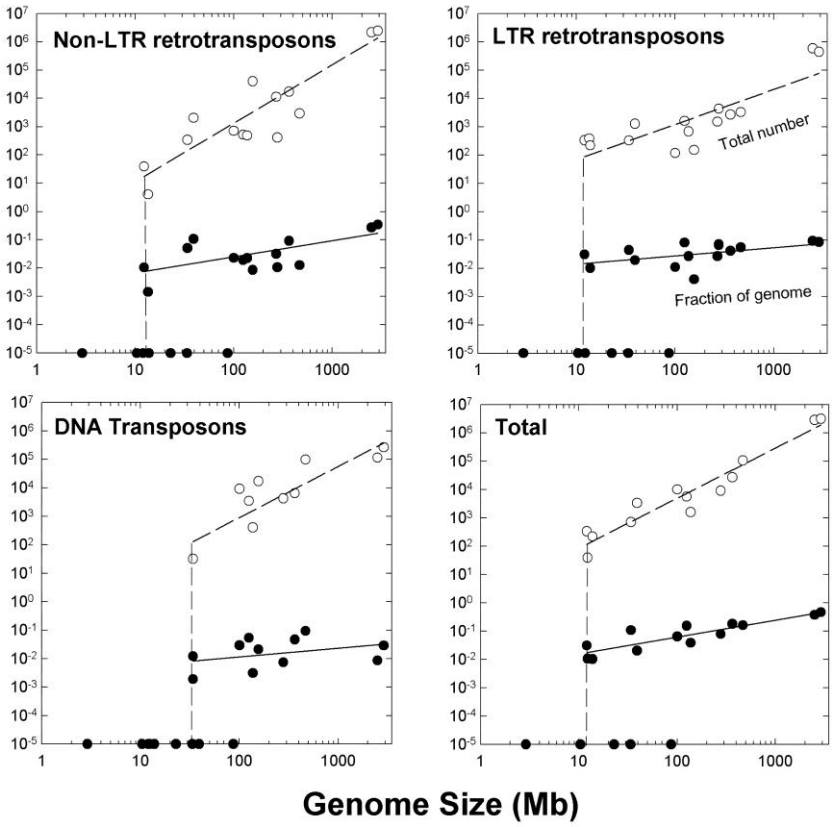

Figure 8: ***The expansion and observed genome size threshold for the three major classes of transposable elements. Species without transposable elements are plotted on the x-axis, but are no part of the reported regression (Lynch and Conery 2003).***

By contrast, the CASP (carrier subpopulation) hypothesis suggests that the division of a population triggers fixation of TE families by genetic drift (Jurka *et al.* 2011). The number of different fixed TE families within the individual genome of an organism should positively correlate with the number of division events from the ancestral metapopulation rather than the overall correlation between the N_e of contemporary populations and the total genome size. Therefore, the fixation of TEs by genetic drift cannot be separated from their phylogenetic history and must be analyzed in the context of historical populations. Moreover, fluctuations in amplification rates over short evolutionary time scales suggests there were important influences at the host population level that affected TE mobility (Hedges *et al.* 2004; Seleme *et al.* 2006).

3.1 - Models of population genetics of Transposable elements

Population genetics refers to the inference of population genetic and evolutionary parameters from genome-wide data sets (Black *et al.* 2001). The main goal of models of population genetics for TEs is to define the conditions of TE maintenance, meaning that TE copy number is in equilibrium within a species genome (reviewed by Rouzic and Deceliere, 2005). Without any mechanism except transposition, TE numbers are expected to grow exponentially. Therefore, basically two general models have been proposed that limit TE expansion, a 'neutral model' and a 'selection model'

(Charlesworth and Charlesworth 1983). In the first model, TE colonization is supposed to be limited by a decrease in the transposition rate if TE copy number increases. The reduced activity of Alu Yb over millions of years within the human lineage could be an example of this model. The regulation of TE activity could emerge from the host, or from the element itself (Badge and Brookfield 1997). However, self regulation may only evolve if TE integrations lead in a substantial number of instances to infertility (Brookfield 1991). For the second model it is assumed that the increase in TE numbers is restrained by selection against the negative effects of insertion to the host. Interestingly, it is unlikely that the deleterious insertion model is true (Le Rouzic and Deceliere 2005). According to this model, the selection coefficient associated with the deleterious effect of a single insertion must be of the same order of magnitude as the transposition rate to reach a realistic non-null copy number equilibrium (Charlesworth 1991).

Both general models, including their refinements, do not take into account specific features of either the host or TE. However, specific interactions between TEs and host genomes exist (Engels *et al.* 1990; Kidwell 1985). Therefore, some specific models have been developed to model these interactions (see Rouzic and Deceliere, 2005). Although they can improve our insight in these matters, they cannot be really validated because their complexity requires very specific parameters which are hard to estimate.

The general picture that arises from these population genetic models is that there is a 'robust theoretical base' for the selfish gene hypothesis (Le Rouzic and Deceliere 2005). Notably, this is also true for TEs with some negative effects on the fitness of the host. Regulation and selection are the two main proposed evolutionary forces that could explain the restriction of TE copy number in populations. However, observations indicate that TEs within human population are not in equilibrium, i.e., the high insertion rates and recently expanded Alu elements (Burns and Boeke 2012; Han *et al.* 2005b). TE mobilization and demographic events may push populations out of equilibrium (Tsitrone *et al.* 1999). Changes in equilibrium state may lead to abrupt changes in TE activity, and possibly explains rapid expansion or removal of TE families in the past (Le Rouzic and Deceliere 2005). Therefore, it is also possible that a subset of populations within a species is in equilibrium while the majority of populations is not.

As an example, mouse TEs are younger in comparison with human TEs and the spectrum of active elements is different (Mouse Genome Sequencing Consortium 2002). Here, L1 and LTR elements are the main active elements that find new insertion sites. The active pool may even be 5-6 times larger than in humans. Explanations for differences in the fraction of active TEs may originate from variations in population size, generation time, population bottlenecks, and factors that influence TE-host interactions (Brookfield 2005).

Many of the complex interactions between TEs and their host are argued to resemble issues found in community ecology (Brookfield 2005; Kidwell and Lisch 1997). Analogue to ecosystems, the genome can be seen as an ecological community in which genes and TEs survived over hundreds of millions of years. Questions about TE diversity and copy number are similar to those asked in the field of community ecology. Making ecological parallels with TE biology might give insight in TE evolution and survival strategies, and host defence mechanisms (Brookfield 2005).

3.2 - Evolutionary selection

The comparisons of human, mouse and rat genomes suggested that about 5-6% of all bases in mammalian genome show evidence of past purifying selection (Cooper *et al.* 2004; Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004). In fact, the majority of conserved sequences appear to be outside coding regions because coding regions represent only 1.5% of the human genome. Mutations within functional elements are more prone to purifying selection because of their negative effects. This means that mutations in functional sequences are less likely to become fixed in population (Kimura 1983). Past purifying selection within functional elements can be observed as a relative lack of substitutions in comparison with neutral sequences. In turn, the magnitude of the deficit can be correlated to the strength of selection.

3.2.1 - General method used to identify potential loci under selection

Population genetic studies start with sampling loci by SNPs, microsatellites, TEs or sequence data throughout the genome. Using this data an overall statistic is calculated that quantifies an aspect of genetic variation (Akey 2009). The null-hypothesis argues for evolution under neutrality. Using this quantification an empirical distribution is constructed over all loci. Next, supposed targets of selection are based on outliers within the extreme tail of the empirical distribution. For these outliers the null-hypothesis is rejected. Implicit assumptions involve that loci are independent from each other, genetic drift influences all loci equally, and selection is strong enough to push individual loci into the tail of the empirical distribution. In addition, it is unavoidable that some selected loci will not appear as outliers (false negatives), and that some neutral loci end up within the tail (false positives). As the empirical distribution is continuous, it is required to set up an arbitrary boundary in order to select outliers. Simulations of neutral evolution and increasingly realistic models for human demographic history, recombination, gene conversion, and mutation rate heterogeneity allow for more robust outlier criteria (Schaffner *et al.* 2005).

One of the major challenges is to distinguish indentified regions between real positive selection and effects of demographic history (Akey 2009). Moreover, if a large selective sweep is found, it is complicated to pinpoint the allele that was under positive selection. The identified region might contain many genes with low variation in the population. More specific follow-up experiments are necessary to identify the causal gene.

In summary, statistics can only reveal that specific loci have a pattern of genetic variation that is unusual in comparison to the rest of the genome. Taken together statistics cannot prove that a locus has been influenced by selection. Important to notice is that many instances of selection are probably not detectable at all. For example if selection is too small or when selection starts to act when an allele is already at an appreciable frequency within the population (Teshima *et al.* 2006). In addition, loci can also acquire non-neutral patterns of genetic variation by the confounding effects of population demographic history.

3.2.2 - Evolutionary constrained Transposable elements

By sequence alignment and comparison of orthologous sequences from 29 mammalian sequences it was found that constrained elements, regions that have been subject to purifying selection, tend to cluster and are inversely correlated with TE density (Cooper *et al.* 2005). Therefore, the authors state that TEs remain a proper model for neutral evolution in the human genome. Nevertheless, they show the presence of some constrained elements that overlap with TEs. A small group of elements (i.e. class L3, L2 and MER121) were claimed to have been under 'intense' purifying selection. In some of these cases the level of purifying selection was comparable, or even greater, as experienced by protein coding exons. In fact, similar elements were identified earlier in human, mouse, and rat alignments (Bejerano *et al.* 2004). Another example involves the previously discussed ancient SINE superfamily that showed strong conservation over the central domain in mammals and birds (Nishihara *et al.* 2006).

Another genome-wide study found that TEs contributed to at least 5.5% of all constrained nonexonic elements unique to mammals (Lowe *et al.* 2007). Although all four main TE classes were represented within the constrained segments, LINEs and SINEs contributed to the majority. Moreover, constrained TEs show a strong preference for genes involved in development and transcription regulation. It was suggested that these elements have been under purifying selection since at least 100 Myr. These results were more recently expanded by the identification of 284,857 conserved non-exonic TEs, totalling almost 7 Mb of genomic DNA (Lindblad Toh *et al.* 2011a). This number accounts for at least 19% of the approximately 1.1 million constrained elements that arose during the 90 Myr between the divergence from marsupials and the eutherian radiation. Again, suggested conserved elements are significantly enriched near regulatory sequences. Although gene poor regions are abundant in TEs, they reveal only a few instances of suggested exaptations.

In order to evaluate the genome-wide effect of TE activity on human gene expression, the expression divergence (ED) was calculated (Warnefors *et al.* 2010). Here, ED was used to measure the difference in gene expression levels between humans and chimpanzees. No increase in ED was detected due to new TE insertions; as a result the authors concluded that TE activity has not contributed to the genome-wide evolution of gene expression levels in humans and chimpanzees (Warnefors *et al.* 2010). This result is in agreement with the finding that Alu sequences did not boosted the expression breadth of neighbouring genes during primate evolution (Urrutia *et al.* 2008). Possibly, the short time scale of both studies, only 6 Myr, does not allow to fully measure the impact of TEs on human gene expression, and thus explains the reported outcome. TEs may initially show only a weak impact on gene expression and regulatory functions, but are refined over longer time scales by selection (Faulkner *et al.* 2009).

In conclusion, the results from comparative studies support the model of extensive exaptation of TEs by a diversity of mammalian genomes. Furthermore, TEs may have played a large role in the formation of gene regulation networks during mammalian evolution (see section 2.1.2). Nevertheless, for many instances this may take more time than sometimes expected, it is therefore very hard to relate specific TE insertions to the recent evolution of humans.

3.2.3 - Ta1 elements have been subject to negative selection

Besides a positive role for TEs, it is clear that TE insertions can also lead to severe abnormalities which alter gene function negatively. Although selection against individual elements is rather weak, cells have developed several mechanisms that limit the activity of TEs collectively (see section 1.2). In a more specific case it was found that full-length Ta1 elements, an active L1 family, were subject to negative selection (Boissinot *et al.* 2006). The authors determined the selective constraints on loci with full-length and truncated elements by using a maximum likelihood method. Interestingly, it was found that this is not the case for truncated L1 elements and Alu inserts, they were apparently not under negative selection. Of course, insertions within essential genes could definitely be very deleterious. However, these alleles would be lost within the population and were thus no part of this study. This means that there is an incredible bias for non-deleterious insertions, which makes it hard to identify slightly deleterious examples. The fact that full-length elements contain regulatory sequences (Jordan *et al.* 2003), and therefore may perturb gene expression, might explain the observed selection against these elements. In addition, L1 products may bind essential host factors or impose deleterious effects by enzymatic activity (Feng *et al.* 1996). Most L1 elements probably inserted before humans spread over the world. Although population expansions and migrations reduced selection, full-length L1 frequencies remained consistent in major subpopulations. This suggests that natural selection constantly acted against these elements (Boissinot *et al.* 2006).

3.2.4 - Alu elements

Alu elements are primate specific and have the highest copy number of all TEs in the human genome. An interesting question is whether these elements had a special role in human evolution. Therefore Cordaux *et al.* (2006) analyzed the genomic distribution and the insertion polymorphisms of three youngest human Alu subfamilies, namely Ya5a2, Ya8 and Yb9. These three subfamilies are estimated to be 0.6-1.8 Myr old (Bandelt *et al.* 1999; Cordaux *et al.* 2004) Recently integrated, polymorphic TEs are expected to show an even distribution, which would reflect their initial insertion pattern (Salem *et al.* 2003; Watkins *et al.* 2003). It was observed that both polymorphic and fixed Alu elements reside in genomic regions that are indistinguishable with respect to their GC content. Moreover, it was found that recently integrated Alu elements are inserted randomly, regardless of the GC content of the surrounding DNA. Overall, these results suggest that young Alu elements are on the whole not under the influence of natural selection (Cordaux *et al.* 2006). This finding supports the neutral and TE marker model. Nevertheless, it is possible that some specific insertions did have a more specific role in human genome evolution.

Overall, there are many cases which show a positive role for TEs during human evolution. However, using population genetic studies it is hard to provide evidence for, and to pinpoint the elements that are positively selected for. It will be hard, but necessary, to prove the functional benefit of many more individual TE insertions in order to claim their assumed contribution to evolution. Probably, this will be very hard for recent integrations, especially in contrast to integrations that took place tens of millions of years ago which have a much higher chance of exaptation.

Conclusion

It was already observed in the early days of molecular biology that genome size does not correlate well with the complexity of organisms. The C-value paradox was elucidated by the finding that a substantial part of genomes can contain a large amount of repetitive sequences (Hartl 2000). As already indicated, almost half of the sequences within the human genome are derived from TEs, and even this high number is likely to be an underestimate (de Koning *et al.* 2011; Lander *et al.* 2001). It remains unclear whether the particular high conservation of SINEs indicates endogenous functional properties, is a by-products of their high copy numbers or results from their distinctive sequence architecture which makes them more easy to detect in comparison to old retrotransposons (Cordaux and Batzer 2009; Feschotte 2008). The confirmation that some TEs are present at the protein level is not really surprising (Gotea and Makalowski 2006). By experimental evolution of arbitrary sequences in a bacteriophage it was shown that random base pairs could acquire biological functions if it had sufficient time to evolve (Hayashi *et al.* 2003). Nevertheless, because of the extremely high copy numbers and the activity of TEs over tens of millions of years, it is likely they played at least some role during human evolutionary history (Cordaux and Batzer 2009).

It is clear that due to the large number of TE copies, they promote to all sorts of structural variation through unequal crossing-over. Although it is hard to prove specific TE related recombination events, they likely played an important role in shaping the genome. The acquisition of new genes is directly related to, for example, genome duplications. For instance, whereas Drosophila has one Hox cluster, vertebrates have four which is the result of two duplication events during evolutionary history. Is has been shown that TEs probably have led to a fair number of duplications (Bailey *et al.* 2003). As a result, Alu sequences are found within 24% of the boundaries of segmental duplications. This could explain a large number of interchromosomal duplications during the last 40 Myr. This shows that TEs likely contributed to the duplication of many genomic regions and thereby influenced genome evolution.

In addition, several mechanisms have been proposed that explain duplications, deletions and the formation of new genes directly by the activity of TEs (Damert *et al.* 2009; Gilbert *et al.* 2002; Wheelan *et al.* 2005). Although TEs are still active, these processes are not expected to influence genome structure as much as the structural variation caused by already integrated TEs. These proposed mechanisms only act on relatively small stretches of DNA. For example, the transcription machinery that skips a termination signal and leads subsequently to 3' transduction, cannot duplicate 1 Mb of DNA. This is in large contrast to recombination events, which can lead to the duplication of whole chromosome arms.

To date, it seems rather clear that TEs can provide regulatory sequences and hence affect gene expression levels in different tissues. For example, is has been shown that a HERV LTR element functions as a parotid-specific enhancer (Samuelson *et al.* 1990). The integration in the amylase loci took place between the split of the New and Old world monkey and the split of the human-ape lineage (Samuelson *et al.* 1996). It has been shown that transgenic mice for this element - mice normally lack salivary amylase - could direct amylase expression to their salivary glands (Ting *et al.* 1992). For humans it is likely that this enhancer replaced an ancestral enhancer because Old world

monkeys express amylase in their saliva by another mechanism (Samuelson *et al.* 1996). Although TEs can provide important regulatory elements, this example also indicates that the genome probably already found different ways to express genes correctly. In this case, TE insertion was likely tolerated because it did not functionally change the expression pattern significantly. Such examples really question the direct relevance of TEs with regulatory functions.

Nonetheless, there is rather convincing evidence that TEs supported the formation of transcriptional networks of tissue specific genes (Bourque *et al.* 2008; Kunarso *et al.* 2010; Lynch *et al.* 2011). The spread of tissue specific transcription binding sites within TE sequences could lead to transcriptional linkage, which enables the collective regulation of multiple tissue specific genes. For instance, the finding that MER20 directly binds transcription factors that are essential for specific gene regulation pathways supports this view. Here, gene regulation dedicated to pregnancy in placental mammals is suggested to be regulated in response to progesterone and cAMP (Lynch *et al.* 2011). Nevertheless, there are still findings that appear to contradict each other. For instance, although SINEs are enriched near housekeeping genes, it was argued that they did not significantly contribute to the expression breadth of these genes during primate evolution (Urrutia *et al.* 2008). Moreover, it could not be proven that TEs supported the expression divergence between humans and chimpanzees (Warnefors *et al.* 2010). Such findings are not in agreement with a general suggested role of TEs as transcriptional linkers that may simplify regulatory networks. However, this does not excludes specific events that involve specific TE families or gene networks. In order to understand how transcriptional networks evolved and function, it is important to clarify these issues. TEs should therefore be an integral part of studies that analyse the transcriptome in a systems biology approach.

As expected, TEs are in general negatively correlated with elements that were constrained by purifying selection during long periods of evolution (Cooper *et al.* 2005). As the interpretation of maps of positive selection is not that simple, many considerations should be taken into account and examined before one can relate a constrained element to a functional role in cells (Akey 2009). Therefore it is worthwhile to invest in studies that attempt to provide evidence for functionally adopted TEs. In fact, clear and established examples are required to put TEs convincingly on a higher pedestal of evolutionary relevant determinants. This exercise should include the role of TE derived sequences on both regulatory and coding elements. The latter could end up rather disappointing because amino acids encoded by TE derived sequences are only found in a few expressed proteins (Gotea and Makalowski 2006).

Interestingly, recent discoveries suggested a positive role for TE activity in somatic cells. Somatic retrotransposition in neurons might endow specific populations of cells with beneficial genetic diversity which allows for selection of phenotypes on a cellular scale (Muotri *et al.* 2005). As identified genomic evidence is extremely low in abundance, the suggestion is that these somatic insertions belong to tiny clonal lineages or even individual cells within the human brain (Baillie *et al.* 2011). Somatic events are present for only a single generation and may affect protein coding genes in a specific environmental context.

In summary, it is obvious that TEs are really an integrated part of our genomes. It seems that the reduction of N_e during evolution enabled the accumulation of TEs in most of the eukaryotic organisms. TEs provided the necessary base pairs which were needed to generate new genes and to

acquire regulatory functions. They had probably the most profound influences on genome architecture in the early times of evolution and during moments of rapid population expansions. During these periods many insertions were tolerated because of decreased selection. This implies that TEs were also able to invade near or in functionally important regions. During following evolutionary periods TE derived fragments may have evolved slowly into functional elements involved in for instance gene regulation. Even with modern sequencing and data analysis techniques it is hard to pinpoint the true elements which were positively selected in the human lineage. Most of the reported exaptations are based on the identification of highly conserved TE derived elements, functional studies clearly lag behind. It appears that the amount of TE derived sequences that code for amino acids in functional proteins is rather low. In addition to effects caused by direct insertion, it seems that TEs had a profound impact on genome evolution by recombination events. The immense TE copy number within the human genome increases the chance of recombination events which can lead to structural variation such as genome duplications, deletions and inversions. Thus, TEs may also affect genome structure after insertion by TE mediated recombination events. Furthermore, it would be interesting to evaluate the role of TEs in the recent evolution of humans. It is possible that exaptated TEs lead to important functional changes that contributed to the evolution towards Modern humans. However, because of the short evolutionary distance between humans and our closest relatives, i.e., chimpanzees and gorilla, is seems difficult to find TEs that can be related to recent human evolution. Nevertheless, it is tempting to evaluate recently sequenced archaic genomes (e.g. Neanderthal and Denisovian) in the light of TEs and evolution (Green *et al.* 2010; Rasmussen *et al.* 2010; Reich *et al.* 2010). These closest relatives might learn us more about the activity of TEs during the most recent periods of human evolution. Archaic genomes may be useful as a TE transposition control or as an independent model alongside of the transposition assays used today. The identification of recent integrations might learn us more about the mechanism of transposition, evolutionary selection in the human lineage and the current effect of TE activity on human health.

Abbreviations

bp	base pair
CASP	carrier subpopulation hypothesis
DNA	deoxyribonucleic acid
HERV	human endogenous retrovirus
L1	long interspersed element 1
LD	linkage disequilibrium
LINE	long interspersed element
LTR	long terminal repeat
Kb	kilo base pairs
Myr	million years
N_e	effective population size
ORF	open reading frame
PCR	polymerase chain reaction
PDB	protein databank
PKR	protein kinase R
RAB	repeat associated binding site
RIP	retrotransposon insertion polymorphism
RMD	recombination mediated deletion
RMT	retrotransposon mediated transduction
RNA	ribonucleic acid
RNAPII	RNA polymerase II
RNAPIII	RNA polymerase III
RNP	ribonucleoprotein
RRMI	retrotransposon recombination-mediated inversion
S/MAR	scaffold/matrix attachment regions
SSR	single sequence repeats
TE	transposable element
TF	transcription factor
TSD	target site duplications
TSS	transcription start site
TPRT	target-primed reverse transcriptase
SINE	short interspersed element
UTR	untranslated region

References

Agoni L, Golden A, Guha C, Lenz J. 2012. Neandertal and denisovan retroviruses. Current Biology 22(11):R437-8.

Agrawal A, Eastman QM, Schatz DG. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394(6695):744-51.

Akey J. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res 19(5):711-22.

Aleman C, Roy Engel AM, Shaikh TH, Deininger PL. 2000. Cis-acting influences on alu RNA levels. Nucleic Acids Res 28(23):4755-61.

An W, Han J, Wheelan S, Davis E, Coombes C, Ye P, Triplett C, Boeke J. 2006. Active retrotransposition by a synthetic L1 element in mice. Proc Natl Acad Sci U S A 103(49):18662-7.

Babushok D and Kazazian H. 2007. Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat 28(6):527-39.

Badge RM and Brookfield J. 1997. The role of host factors in the population dynamics of selfish transposable elements. J Theor Biol 187(2):261-71.

Bailey J, Liu G, Eichler E. 2003. An alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 73(4):823-34.

Baillie JK, Barnett M, Upton K, Gerhardt D, Richmond T, De Sapio F, Brennan P, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479(7374):534-7.

Bandelt HJ, Forster P, Rohl A, Rhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37-48.

Batzer M and Deininger P. 2002. Alu repeats and human genomic diversity. Nature Reviews.Genetics 3(5):370-9.

Beck C, Collier P, Macfarlane C, Malig M, Kidd J, Eichler E, Badge R, Moran J. 2010. LINE-1 retrotransposition activity in human genomes. Cell 141(7):1159-70.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick J, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304(5675):1321-5.

Bejerano G, Lowe C, Ahituv N, King B, Siepel A, Salama S, Rubin E, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441(7089):87-90.

Belancio VP, Roy-Engel AM, Deininger P, Roy Engel AM. 2008a. The impact of multiple splice sites in human L1 elements. Gene 411(1-2):38-45.

Belancio V, Hedges D, Deininger P. 2008b. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. Genome Res 18(3):343-58.

Belancio V, Hedges D, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34(5):1512-21.

Bennett EA, Keller H, Mills R, Schmidt S, Moran J, Weichenrieder O, Devine S. 2008. Active alu retrotransposons in the human genome. Genome Res 18(12):1875-83.

Bishop JM and Varmus H. 1992. **Functions and origins of retroviral oncogenes**. In: Molecular biology of the tumor viruses: RNA tumor viruses. Weiss R, Teich N, Varmus H, and others, editors. Cold Spring Harbor, NY (1982): Cold Spring Harbor Laboratory Press. 999 p.

Black WC, Baer CF, Antolin MF, DuTeau NM. 2001. Population genomics: Genome-wide sampling of insect populations. Annu Rev Entomol 46:441-69.

Bode J, StengertIber M, Kay V, Schlake T, DietzPfeilstetter A, Stengert Iber M, Dietz Pfeilstetter A. 1996. Scaffold/matrix-attached regions: Topological switches with multiple regulatory functions. Crit Rev Eukaryot Gene Expr 6(2-3):115-38.

Boissinot S, Davis J, Entezam A, Petrov D, Furano A. 2006. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A 103(25):9590-4.

Bourc'his D and Bestor T. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431(7004):96-9.

Bourque G, Leong B, Vega V, Chen X, Lee Y, Srinivasan K, Chew J, Ruan Y, Wei C, Ng H, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18(11):1752-62.

Brandt J, Schrauth S, Veith A, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff J. 2005. Transposable elements as a source of genetic innovation: Expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345(1):101-11.

Britten RJ and Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46(2):111-38.

Britten R. 2010. Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A 107(46):19945-8.

Brookfield JF. 2001. Selection on alu sequences? Current Biology 11(22):R900-1.

Brookfield JF. 1991. Models of repression of transposition in P-M hybrid dysgenesis by P cytotype and by zygotically encoded repressor proteins. Genetics 128(2):471-86.

Brookfield J. 2005. The ecology of the genome - mobile DNA elements and their hosts. Nature Reviews.Genetics 6(2):128-36.

Brosius J. 1999a. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238(1):115-34.

Brosius J. 1999b. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107(1-3):209-38.

Brosius J and Gould SJ. 1992. On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci U S A 89(22):10706-10.

Brosius J. 2003. The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118(2-3):99-116.

Brouha B, Lutz Prigget S, Schustak J, Badge R, Lutz Prigge S, Farley A, Moran J, Kazazian H. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100(9):5280-5.

Brown. 2007. Genomes 3. Garland Science Publishing, New York

Burns K and Boeke J. 2012. Human transposon tectonics. Cell 149(4):740-52.

Callinan PA and Batzer MA. 2006a. Retrotransposable elements and human disease. Genome Dynamics 1:104-15.

Callinan PA and Batzer MA. 2006b. Retrotransposable elements and human disease. Genome Dynamics 1:104-15.

Callinan P, Wang J, Herke S, Garber R, Liang P, Batzer M. 2005. Alu retrotransposition-mediated deletion. J Mol Biol 348(4):791-800.

Carter A, Salem A, Hedges D, Keegan C, Kimball B, Walker J, Watkins WS, Jorde L, Batzer M. 2004. Genome-wide analysis of the human alu yb-lineage. Human Genomics 1(3):167-78.

Chang DY, Hsu K, Maraia RJ. 1996. Monomeric scAlu and nascent dimeric alu RNAs induced by adenovirus are assembled into SRP9/14-containing RNPs in HeLa cells. Nucleic Acids Res 24(21):4165-70.

Charlesworth B. 1991. Transposable elements in natural populations with a mixture of selected and neutral insertion sites. Genet Res 57(2):127-34.

Charlesworth B and Charlesworth D. 1983. The population-dynamics of transposable elements. Genet Res 42(1):1-27.

Chen C, Ara T, Gautheret D. 2009. Using alu elements as polyadenylation sites: A case of retroposon exaptation. Mol Biol Evol 26(2):327-34.

Chesnokov I and Schmid CW. 1996. Flanking sequences of an alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. J Mol Evol 42(1):30-6.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69-156.

Clark JB and Kidwell MG. 1997. A phylogenetic perspective on P transposable element evolution in drosophila. Proc Natl Acad Sci U S A 94(21):11428-33.

Comeaux M, Roy Engel A, Hedges D, Deininger P. 2009. Diverse cis factors controlling alu retrotransposition: What causes alu elements to die? Genome Res 19(4):545-55.

Cooper G, Stone E, Asimenos G, Green E, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15(7):901-13.

Cooper G, Brudno M, Stone E, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res 14(4):539-48.

Cordaux R and Batzer M. 2009. The impact of retrotransposons on human genome evolution. Nature Reviews Genetics 10(10):691-1394.

Cordaux R, Hedges D, Batzer M. 2004. Retrotransposition of alu elements: How many sources? Trends in Genetics 20(10):464-7.

Cordaux R, Lee J, Dinoso L, Batzer M. 2006. Recently integrated alu retrotransposons are essentially neutral residents of the human genome. Gene 373:138-44.

Cost GJ, Golding A, Schlissel MS, Boeke JD. 2001. Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29(2):573-7.

Cost G, Feng Q, Jacquier A, Boeke J. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J 21(21):5899-910.

Damert A, Raiz J, Horn A, Loewer J, Wang H, Lwer J, Xing J, Batzer M, Lwer R, Schumann G. 2009. 5'-transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res 19(11):1992-2008.

de Koning APJ, Gu W, Castoe T, Batzer M, Pollock D. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLOS Genetics 7(12).

Deininger PL and Batzer MA. 1999. Alu repeats and human disease. Mol Genet Metab 67(3):183-93.

Deininger PL, Batzer MA, Hutchison CA, Edgell MH. 1992. Master genes in mammalian repetitive DNA amplification. Trends in Genetics 8(9):307-11.

Dimitri P and Junakovic N. 1999. Revising the selfish DNA hypothesis: New evidence on accumulation of transposable elements in heterochromatin. Trends in Genetics 15(4):123-4.

Doolittle WF and SAPIENZA C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601-3.

Eickbush TH. 1997. Telomerase and retrotransposons: Which came first? Science 277(5328):911-2.

Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, Marahrens Y. 2007. Repetitive sequence environment distinguishes housekeeping genes. Gene 390(1-2):153-65.

Engels WR, Sved J, Johnson Schlitz DM, Eggleston WB. 1990. High-frequency P element loss in drosophila is homolog dependent. Cell 62(3):515-25.

Ewing A and Kazazian H. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20(9):1262-70.

Faulkner G, Kimura Y, Daub C, Wani S, Plessy C, Irvine K, Schroder K, Cloonan N, Steptoe A, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41(5):563-71.

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87(5):905-16.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nature Reviews.Genetics 9(5):397-405.

Flavell AJ. 1995. Retroelements, reverse transcriptase and evolution. Comparative Biochemistry and Physiology.Part B, Biochemistry Molecular Biology 110(1):3-15.

Fraser P and Bickmore W. 2007. Nuclear organization of the genome and the potential for gene regulation. Nature 447(7143):413-7.

Gdula DA, Gerasimova TI, Corces VG. 1996. Genetic and molecular analysis of the gypsy chromatin insulator of drosophila. Proc Natl Acad Sci U S A 93(18):9378-83.

Georgiev GP. 1984. Mobile genetic elements in animal cells and their biological significance. European Journal of Biochemistry 145(2):203-20.

Gerber A, OConnell M, Keller W, O'Connell MA. 1997. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an alu cassette. RNA 3(5):453-63.

Gibbons R, Dugaiczyk L, Girke T, Duistermars B, Zielinski R, Dugaiczyk A. 2004. Distinguishing humans from great apes with AluYb8 repeats. J Mol Biol 339(4):721-9.

Gilbert N and Labuda D. 1999. CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. Proc Natl Acad Sci U S A 96(6):2869-74.

Gilbert N, Lutz Prigge S, Moran J. 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110(3):315-25.

Gilbert N, Lutz S, Morrish T, Moran J. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 25(17):7780-95.

Gotea V and Makalowski W. 2006. Do transposable elements really contribute to proteomes? Trends in Genetics 22(5):260-7.

GOULD S and VRBA E. 1982. EXAPTATION - A MISSING TERM IN THE SCIENCE OF FORM. Paleobiology 8(1):4-15.

Greally J. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci U S A 99(1):327-32.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. 2010. A draft sequence of the neandertal genome. Science 328(5979):710-22.

Grewal SIS and Jia S. 2007. Heterochromatin revisited. Nature Reviews.Genetics 8(1):35-46.

Grover D, Kannan K, Mukerji M, Bhatnagar P, Brahmachari S. 2004. Alu repeat analysis in the complete human genome: Trends and variations with respect to genomic composition. Bioinformatics 20(6):813-7.

Han J and Boeke J. 2004. A highly active synthetic mammalian retrotransposon. Nature 429(6989):314-8.

Han J, Szak S, Boeke J. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429(6989):268-74.

Han K, Lee J, Meyer T, Remedios P, Goodwin L, Batzer M. 2008. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A 105(49):19366-71.

Han K, Xing J, Wang H, Hedges D, Garber R, Cordaux R, Batzer M. 2005a. Under the genomic radar: The stealth model of alu amplification. Genome Res 15(5):655-64.

Han K, Sen S, Wang J, Callinan P, Lee J, Cordaux R, Liang P, Batzer M. 2005b. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res 33(13):4040-52.

Hancks D, Ewing A, Chen J, Tokunaga K, Kazazian H. 2009. Exon-trapping mediated by the human retrotransposon SVA. Genome Res 19(11):1983-91.

Hartl DL. 2000. Molecular melodies in high and low C. Nature Reviews.Genetics 1(2):145-9.

Hasse A and Schulz WA. 1994. Enhancement of reporter gene de novo methylation by DNA fragments from the alpha-fetoprotein control region. J Biol Chem 269(3):1821-6.

Hata K and Sakaki Y. 1997. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene 189(2):227-34.

Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T. 2003. Can an arbitrary sequence evolve towards acquiring a biological function? J Mol Evol 56(2):162-8.

Hayes B, Visscher P, McPartlan H, Goddard M. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res 13(4):635-43.

Hedges D, Callinan P, Cordaux R, Xing J, Barnes E, Batzer M. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res 14(6):1068-75.

Henikoff S and Matzke MA. 1997. Exploring and explaining epigenetic effects. Trends in Genetics 13(8):293-5.

Hilgard P, Huang T, Wolkoff A, Stockert R. 2002. Translated alu sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. American Journal of Physiology.Cell Physiology 283(2):C472-83.

Hiom K, Melek M, Gellert M. 1998. DNA transposition by the RAG1 and RAG2 proteins: A possible source of oncogenic translocations. Cell 94(4):463-70.

Hoenicka J, Arrasate M, de Yebenes J, Avila J. 2002. A two-hybrid screening of human tau protein: Interactions with alu-derived domain. Neuroreport 13(3):343-9.

Houck CM, Rinehart FP, Schmid CW. 1979. A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol 132(3):289-306.

Huang CRL, Schneider A, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka J, Valle D, Civin C, Wang T, et al. 2010. Mobile interspersed repeats are major structural variants in the human genome. Cell 141(7):1171-82.

Huda A, Marino Ramirez L, Landsman D, Jordan IK, Mario-Ramrez L. 2009. Repetitive DNA elements, nucleosome binding and human gene expression. Gene 436(1-2):12-22.

Jacob F. 1977. Evolution and tinkering. Science 196(4295):1161-6.

Johnson R, Gamblin R, Ooi L, Bruce A, Donaldson I, Westhead D, Wood I, Jackson R, Buckley N. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. Nucleic Acids Res 34(14):3862-77.

Jordan IK, Rogozin I, Glazko G, Koonin E. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19(2):68-72.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94(5):1872-7.

Jurka J, Pavlicek A, Klonowski P, Kohany O, Kapitonov VV, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110(1-4):462-7.

Jurka J. 2004. Evolutionary impact of human alu repetitive elements. Current Opinion in Genetics Development 14(6):603-8.

Jurka J, Bao W, Kojima K. 2011. Families of transposable elements, population structure and the origin of species. Biology Direct 6:44.

Kamal M, Xie X, Lander E. 2006. A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A 103(8):2740-5.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16(1):78-87.

Kidwell MG. 1985. Hybrid dysgenesis in drosophila melanogaster: Nature and inheritance of P element regulation. Genetics 111(2):337-50.

Kidwell MG and Lisch D. 1997. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94(15):7704-11.

Kidwell M. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115(1):49-63.

Kim T, Jung Y, Rhyu M. 2004. Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. J Korean Med Sci 19(6):783-92.

Kimura. 1983. The neutral theory of molecular evolution. .

Kriegs J, Churakov G, Jurka J, Brosius J, Schmitz J. 2007. Evolutionary history of 7SL RNA-derived SINEs in supraprimates. Trends in Genetics 23(4):158-61.

Kroutter E, Belancio V, Wagstaff B, Roy Engel A. 2009. The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. PLOS Genetics 5(4).

Krull M, Brosius J, Schmitz J. 2005. Alu-SINE exonization: En route to protein-coding function. Mol Biol Evol 22(8):1702-11.

Kunarso G, Chia N, Jeyakani J, Hwang C, Lu X, Chan Y, Ng H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42(7):631-4.

Kurdyukov SG, Lebedev YB, Artamonova II, Gorodentseva TN, Batrak AV, Mamedov IZ, Azhikina TL, Legchilina SP, Efimenko IG, Gardiner K, et al. 2001. Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: Map locations and evolutionary history. Gene 273(1):51-61.

Lander ES, Linton LM, Birren B. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):p860-62p.

Landry J, Rouhi A, Medstrand P, Mager D. 2002. The opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol 19(11):1934-42.

Larsen F, GUNDERSEN G, LOPEZ R, PRYDZ H. 1992. CpG islands as gene markers in the human genome. Genomics 13(4):1095-107.

Le Rouzic A and Deceliere G. 2005. Models of the population genetics of transposable elements. Genet Res 85(3):171-81.

Lee J, Han K, Meyer T, Kim H, Batzer M. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS ONE 3(12):e4047.

Lercher M, Urrutia A, Hurst L. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31(2):180-3.

Lev Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in alu exons. Science 300(5623):1288-91.

Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM. 1993. Transposons in place of telomeric repeats at a drosophila telomere. Cell 75(6):1083-93.

Li T and Schmid C. 2004. Alu's dimeric consensus sequence destabilizes its transcripts. Gene 324:191-200.

Lindblad Toh K, Garber M, Zuk O, Lin M, Parker B, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011a. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478(7370):476-82.

Lindblad Toh K, Garber M, Zuk O, Lin M, Parker B, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011b. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478(7370):476-82.

Lovell S. 2003. Are non-functional, unfolded proteins ('junk proteins') common in the genome? FEBS Lett 554(3):237-9.

Lowe C, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A 104(19):8005-10.

Lynch M and Conery JS. 2003. The origins of genome complexity. Science 302(5649):1401-4.

Lynch V, Leclerc R, May G, Wagner G. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 43(11):1154-9.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10(9):1307-18.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16(6):793-805.

McClintock B. 1956. Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21:197-216.

McDonald JF. 1993. Evolution and consequences of transposable elements. Current Opinion in Genetics Development 3(6):855-64.

Medstrand P, Landry JR, Mager DL. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276(3):1896-903.

Medstrand P, van de Lagemaat LN, Landry J, Svenback D, Dunn CA, Mager DL. 2005. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenetic and Genome Research 110(1-4):342-52.

Medstrand P, van de Lagemaat LN, Mager D. 2002. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. Genome Res 12(10):1483-95.

Mills R, Bennett EA, Iskow R, Devine S. 2007. Which transposable elements are active in the human genome? Trends in Genetics 23(4):183-91.

Mills R, Bennett EA, Iskow R, Luttig C, Tsui C, Pittard WS, Devine S. 2006. Recently mobilized transposons in the human and chimpanzee genomes. Am J Hum Genet 78(4):671-9.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87(5):917-27.

Morrish T, Gilbert N, Myers J, Vincent B, Stamato T, Taccioli G, Batzer M, Moran J. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31(2):159-65.

Mortazavi A, Leeper Thompson EC, Garcia S, Myers R, Wold B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire. Genome Res 16(10):1208-21.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520-62.

Moyes D, Griffiths DJ, Venables PJ. 2007. Insertional polymorphisms: A new lease of life for endogenous retroviruses in human disease. Trends in Genetics 23(7):326-33.

Muckenfuss H, Hamdorf M, Held U, Perkovic M, Loewer J, Lwer J, Cichutek K, Flory E, Schumann G, Mnk C. 2006. APOBEC3 proteins inhibit human LINE-1 retrotransposition. J Biol Chem 281(31):22161-72.

Muotri A, Marchetto M, Chu V, Deng W, Moran J, Gage F. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435(7044):903-10.

Myers J, Vincent B, Udall H, Watkins WS, Morrish T, Kilroy G, Swergold G, Henke J, Henke L, Moran J, et al. 2002. A comprehensive analysis of recently integrated human ta L1 elements. Am J Hum Genet 71(2):312-26.

Nagasaki K, Rosel F, Schem C, von Kaisenberg C, Biallek M, Rsel F, Jonat W, Maass N. 2003. Leucine-zipper protein, LDOC1, inhibits NF-kappaB activation and sensitizes pancreatic cancer cells to apoptosis. International Journal of Cancer 105(4):454-8.

Nekrutenko A and Li W. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends in Genetics 17(11):619-21.

Nishihara H, Smit AFA, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res 16(7):864-74.

Orgel LE and CRICK F. 1980. Selfish DNA: The ultimate parasite. Nature 284(5757):604-7.

Ostertag EM, Prak E, DeBerardinis RJ, Moran JV, Kazazian HH. 2000. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 28(6):1418-23.

Ostertag E, Goodier J, Zhang Y, Kazazian H. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73(6):1444-51.

Pace J and Feschotte C. 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res 17(4):422-32.

Pavlcek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G. 2001. Similar integration but different stability of alus and LINEs in the human genome. Gene 276(1-2):39-45.

Perepelitsa Belancio V and Deininger P. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 35(4):363-6.

Pinsker W, Haring E, Hagemann S, Miller WJ. 2001. The evolutionary life history of P transposons: From horizontal invaders to domesticated neogenes. Chromosoma 110(3):148-58.

Polak P and Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7:133.

Rasmussen M, Li Y, Lindgreen S, Pedersen J, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct palaeo-eskimo. Nature 463(7282):757-62.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature 428(6982):493-521.

Reich D, Green R, Kircher M, Krause J, Patterson N, Durand E, Viola B, Briggs A, Stenzel U, Fu Q, et al. 2010. Genetic history of an archaic hominin group from denisova cave in siberia. Nature 468(7327):1053-60.

Rynditch AV, Zoubak S, Tsyba L, Tryapitsina Guley N, Bernardi G. 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222(1):1-16.

Salem A, Kilroy G, Watkins WS, Jorde L, Batzer M. 2003. Recently integrated alu elements and human genomic diversity. Mol Biol Evol 20(8):1349-61.

Samuelson LC, Phillips RS, Swanberg LJ. 1996. Amylase gene structures in primates: Retroposon insertions and promoter evolution. Mol Biol Evol 13(6):767-79.

Samuelson LC, Wiebauer K, Snow CM, Meisler MH. 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol Cell Biol 10(6):2513-20.

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH. 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Molecular Genetics and Genomics 270(2):173-80.

Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A 105(11):4220-5.

Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576-659.

Schmid CW. 1998. Does SINE evolution preclude alu function? Nucleic Acids Res 26(20):4541-50.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages (vol 148, pg 335, 2012). Cell 148(4):832-.

Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A. 1996. Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci U S A 93(25):14759-64.

Seleme MdC, Kazazian HH, Vetter M, Cordaux R, Bastone L, Batzer M. 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A 103(17):6611-6.

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. 2006. Human genomic deletions mediated by recombination between alu elements. Am J Hum Genet 79(1):41-53.

Shaikh TH, Kim J, Roy AM, Batzer MA, Deininger PL. 1997. cDNAs derived from primary and small cytoplasmic alu (scAlu) transcripts. J Mol Biol 271(2):222-34.

Shedlock AM and Okada N. 2000. SINE insertions: Powerful tools for molecular systematics. Bioessays 22(2):148-60.

Shen MR, Batzer MA, Deininger PL. 1991. Evolution of the master alu gene(s). J Mol Evol 33(4):311-20.

Sherry ST, Stoneking M, Harpending HC, Batzer MA. 1997. Alu evolution in human populations: Using the coalescent to estimate effective population size. Genetics 147(4):1977-82.

Singer S, Mannel D, Mnnel D, Hehlgans T, Brosius J, Schmitz J. 2004. From "junk" to gene: Curriculum vitae of a primate receptor isoform gene. J Mol Biol 341(4):883-6.

Sinnett D, Richer C, Labuda D, Deragon JM. 1991. Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. J Biol Chem 266(14):8675-8.

Skowronski J, Fanning TG, Singer MF. 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol 8(4):1385-97.

Smit AF. 1996. The origin of interspersed repeats in the human genome. Current Opinion in Genetics Development 6(6):743-8.

Smit A. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics Development 9(6):657-63.

Soifer H, Zaragoza A, Peyvan M, Behlke M, Rossi J. 2005. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. Nucleic Acids Res 33(3):846-56.

Sorek R, Ast G, Graur D. 2002. Alu-containing exons are alternatively spliced. Genome Res 12(7):1060-7.

Soriano P, Meunierrotival M, Bernardi G, Meunier Rotival M. 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc Natl Acad Sci U S A 80(7):1816-20.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21(6):1973-85.

Sverdlov ED. 2000. Retroviruses and primate evolution. Bioessays 22(2):161-71.

Swergold GD. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol 10(12):6718-29.

Symer D, Connelly C, Szak S, Caputo E, Cost G, Parmigiani G, Boeke J. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110(3):327-38.

Szak S, Pickeral O, Makalowski W, Boguski M, Landsman D. 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol 3(10).

Takahata N. 1993. Allelic genealogy and human evolution. Mol Biol Evol 10(1):2-22.

Tenesa A, Navarro P, Hayes B, Duffy D, Clarke G, Goddard M, Visscher P. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res 17(4):520-6.

Teshima K, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res 16(6):702-14.

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Development 6(8):1457-65.

Tsitrone A, Charles S, Biemont C. 1999. Dynamics of transposable elements under the selection model. Genet Res 74(2):159-64.

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. Current Biology 11(19):1531-5.

Urrutia A, Balladares Ocana L, Hurst L, Ocaa L. 2008. Do alu repeats drive the evolution of the primate transcriptome? GenomeBiology.Com 9(2):R25.

van de Lagemaat LN, Gagnier L, Medstrand P, Mager D. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res 15(9):1243-9.

van de Lagemaat LN, Landry J, Mager D, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends in Genetics 19(10):530-6.

van Drunen CM, Sewalt R, Smeekens S, Oosterling RW, Weisbeek PJ, van Driel R. 1999. A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res 27(14):2924-30.

Wagner E and Lykke Andersen J. 2002. mRNA surveillance: The perfect persist. J Cell Sci 115(Pt 15):3033-8.

Wang H, Xing J, Grover D, Hedges D, Han K, Walker J, Batzer M. 2005. SVA elements: A hominid-specific retroposon family. J Mol Biol 354(4):994-1007.

Warnefors M, Pereira V, Eyre Walker A. 2010. Transposable elements: Insertion pattern and impact on gene expression evolution in hominids. Mol Biol Evol 27(8):1955-62.

Watkins WS, Rogers A, Ostler C, Wooding S, Bamshad M, Brassington A, Carroll M, Nguyen S, Walker J, Reddy PG, et al. 2003. Genetic variation among world populations: Inferences from 100 alu insertion polymorphisms. Genome Res 13(7):1607-18.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: Cis preference versus trans complementation. Mol Cell Biol 21(4):1429-39.

Weiner A. 2002. SINEs and LINEs: The art of biting the hand that feeds you. Curr Opin Cell Biol 14(3):343-50.

Wheelan S, Aizawa Y, Han J, Boeke J. 2005. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. Genome Res 15(8):1073-8.

Wheelan S, Scheifele L, Martinez Murillo F, Irizarry R, Boeke J, Martnez-Murillo F. 2006. Transposon insertion site profiling chip (TIP-chip). Proc Natl Acad Sci U S A 103(47):17632-7.

Xing J, Wang H, Belancio V, Cordaux R, Deininger P, Batzer M. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A 103(47):17608-13.

Xing J, Zhang Y, Han K, Salem A, Sen S, Huff C, Zhou Q, Kirkness E, Levy S, Batzer M, et al. 2009. Mobile elements create structural variation: Analysis of a complete human genome. Genome Res 19(9):1516-26.

Yang N and Kazazian H. 2006. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature Structural Molecular Biology 13(9):763-71.

Yates PA, Mummaneni P, Krussel S, Burman RW, Turker MS. 1999. Tandem B1 elements located in a mouse methylation center provide a target for de novo DNA methylation. J Biol Chem 274(51):36357-61.

Yeo G, Holste D, Kreiman G, Burge C. 2004. Variation in alternative splicing across human tissues. GenomeBiology.Com 5(10):R74.

Zhang H, Lee J, Tian B. 2005. Biased alternative polyadenylation in human tissues. GenomeBiology.Com 6(12):R100.