MECHANISMS INVOLVED IN TARGET SEQUENCE RECOGNITION AND INTEGRATION OF HUMAN LINE-1 RETROTRANSPOSONS

Dissertation
zur Erlangung des Doktorgrades im Fachbereich Chemie
der Universität Hamburg

vorgelegt von
Nora Zingler
aus München

Hamburg 2004

Die vorliegende Arbeit wurde in der Zeit von August 1999 bis Juli 2004 unter der Leitung von Prof. Dr. Wolfram Ostertag und PD Dr. Gerald G. Schumann am Heinrich-Pette-Institut für Experimentelle Virologie und Immunologie (Hamburg), am Universitätsklinikum Eppendorf (Hamburg) und am Paul-Ehrlich-Institut, Bundesamt für Sera und Impfstoffe (Langen) angefertigt.

1. Gutachter: Herr Prof. Dr. W. Ostertag

2. Gutachter: Herr Prof. Dr. H. Marquardt

Tag der Disputation: 9.7.04

INDEX

Appendices

TABLE OF FIGURES AND TABLES

TABLE OF ABBREVIATIONS

aa	amino acid
AP	apurinic/apyrimidinic
ATCC	American Type Culture Collection
bp	base pair
cDNA	complementary DNA
Ci	Curie
CMV	cytomegalovirus
cpm	counts per minute
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynucleosid triphosphate
EDTA	ethylendiamine tetraacetate
EN	endonuclease
EtOH	ethanol
FCS	fetal calf serum
G418	geniticin sulfate
gag	group-specific antigen
GFP	green fluorescent protein
GMCSF	granulocyte-macrophage colony stimulating factor
HGWD	human genome working draft
kb	kilobasepairs
kDa	kilodalton
L1	the human LINE-1 element
LINE	long interspersed nuclear element (autonomous non-LTR retrotransposon)
LTR	long terminal repeat
MBD	methyl-CpG-binding domain
MBq	Megabequerel
mRNA	messenger RNA
Myr	million years
neo^R	neomycin/geniticin sulfate resistant
OD	optical density
ORF	open reading frame
p.	page
PBS	phosphate buffered saline
PCR	polymerase chain reaction
RNA	ribonucleic acid
rpm	revolutions per minute
RT	reverse transcriptase
SA	splice acceptor
SCID	severe combined immunodeficiency disease
SD	splice donor
SDS	sodium dodecyl sulphate
SV40	simian virus 40
TE	transposable element
TPRT	target-primed reverse transcription
Tris	Tris(hydroxymethyl)aminomethane
U	units

SUMMARY

The human LINE-1 (Long Interspersed Nuclear Element-1, L1) retrotransposon is a member of the group of autonomous non-LTR retrotransposons found in the genome of almost every eukaryotic organism. As mobile genetic element, it generates copies of its genetic information by reverse transcription of an RNA intermediate and integrates them into the host genome. Due to its endogenous and basically non-pathogenic nature, L1 is a promising candidate as vector for gene delivery in somatic gene therapy. However, since many details of L1 biology are still insufficiently characterised, the present study focussed on several aspects of L1 replication to better assess the potential of L1 as safe and efficient gene shuttle.

- In order to use L1 as vector for gene therapy, it is indispensable to understand the regulation of its expression. In collaboration with the laboratory of Prof. Strätling (Universitätsklinikum Hamburg-Eppendorf) we characterised the roles of several methyl-CpG-binding proteins in the regulation of methylated L1 elements and showed that methyl-CpG-binding protein 2 (MeCP2) is a major repressor of L1 transcription and retrotransposition. The gene delivery efficiency of (unmethylated) retrotransposon-based vectors should not be affected by this mechanism since *de novo* methylation is a slow process that takes place after the initial replication and integration phase of L1.
- When compared to conventional gene therapy vectors like retroviral vectors, one of the major advantages of non-LTR retrotransposons is the ability of a subset of these elements to insert into the host genome without harming their host, by specifically integrating into defined DNA sequences. One aim of this study was to answer the question which protein moieties or structural motifs are responsible for this target site specificity. The endonuclease (EN) domain of the semi-site-specific human L1 was replaced with ENs borrowed from the closely related site-specific non-LTR retrotransposons Tx1L from *X. laevis* and R1Bm from *B. mori*. Various swapping experiments led to the identification of a hairpin loop region in L1 EN that influences target site recognition and at the same time tolerates amino acid substitutions without severe adverse effects on retrotransposition efficiency.
- The predisposition toward frequent and variable truncations at the 5' end of newly integrated L1 copies is an ambivalent property with respect to gene therapy: While deletion of the promoter region provides a built-in safety mechanism that prevents subsequent transposition of a possible transgene-containing L1 insertion, more extensive truncations might damage or delete the therapeutic gene. Analyses of 55 *de novo* L1 integrants and 10,034 endogenous

L1 copies suggest that the cellular DNA repair pathway of ‘microhomology-driven single strand annealing’ is involved in the process of 5’ truncation. In contrast, full-length insertions seem to derive from a different mechanism not involving microhomologies. Based on these results a model was developed explaining both the occurrence of 5’ truncated L1 elements and the mechanism of second-strand synthesis.

This study provides novel, promising insights into the regulation, target site recognition and integration mechanism of L1 elements. It thus lays the foundation for subsequent investigations that could lead to the utilisation of retrotransposons in gene therapy.

ZUSAMMENFASSUNG

Das menschliche Retrotransposon LINE-1 (Long Interspersed Nuclear Element-1, L1) gehört zur Gruppe der autonomen Non-LTR-Retrotransposons, die ein Bestandteil des Genoms der meisten Eukaryonten sind. Als mobiles genetisches Element ist es in der Lage, über den Mechanismus der reversen Transkription eines RNA-Intermediates neue Kopien seiner genetischen Information anzufertigen und in das Wirtsgenom zu integrieren. Die grundsätzlich nicht pathogenen Eigenschaften von L1 legen dessen Verwendung als Vektor für die somatische Gentherapie nahe. Da der L1-Lebenszyklus noch unzureichend charakterisiert ist, wurden in der vorliegenden Arbeit einige Aspekte der L1-Replikation genauer untersucht.

- Beabsichtigt man, L1-Elemente als Vektoren zur Einführung von Fremdgenen ins menschliche Genom zu verwenden, so muß vorher die Regulation der Expression dieser Elemente verstanden werden. In Zusammenarbeit mit der Arbeitsgruppe von Herrn Prof. Strätling (Universitätsklinikum Hamburg-Eppendorf) untersuchten wir den Einfluß von verschiedenen Methyl-CpG-bindenden Proteinen auf die Regulation methylierter L1-Elemente und zeigten, daß das Methyl-CpG-bindende Protein 2 (MeCP2) eine wesentliche Rolle bei der Repression von Transkription und Retrotransposition von L1-Elementen spielt. Dieser Prozeß sollte jedoch die Effizienz eines (unmethylierten) retrotransposon-basierten Vektors nicht beeinflussen, da *de-novo*-Methylierung ein langsamer Prozeß ist, der erst nach der anfänglichen Replikations- und Integrationsphase von L1 zum Tragen kommt.
- Ein großer Vorteil von Non-LTR-Retrotransposons im Vergleich zu konventionellen Gentherapievektoren wie z. B. retroviralen Vektoren ist die Fähigkeit einiger dieser Elemente, das Wirtsgenom zu besiedeln, ohne dem Wirt zu schaden. Dies geschieht durch gerichtete Integration in definierte DNA-Sequenzen. Ein Ziel dieser Arbeit war die Beantwortung der Frage, welche Proteinbereiche oder Strukturmotive diese sequenzspezifische Integration vermitteln. Die Endonuklease (EN)-Domäne des nur schwach zielgerichteten L1-Elements wurde durch EN-Domänen der eng verwandten, spezifisch integrierenden Non-LTR-Retrotransposons Tx1L aus *X. laevis* und R1Bm aus *B. mori* ersetzt. Zahlreiche Austauschexperimente führten zur Identifizierung einer

Haarnadelschleife in der L1-Endonukleasedomäne, die die Zielsequenzerkennung von L1 beeinflußt und deren Austausch nicht zu einem drastischen Aktivitätsverlust führt.

- Ca. 95% aller integrierter L1-Kopien sind an ihrem 5'-Ende nicht vollständig. Diese Eigenschaft kann in der Gentherapie sowohl von Vor- als auch von Nachteil sein: während das Fehlen der Promotorregion nach Retrotransposition einen inhärenten Sicherheitsmechanismus darstellen könnte, der verhindert, daß sich ein Transgen-tragendes integriertes L1-Element unkontrolliert weitervermehrt, könnte ein näher am 3'-Ende stattfindender Abbruch zur Beschädigung oder Entfernung des therapeutischen Gens führen. Die Analyse von 55 *de-novo*-L1-Insertionen und von 10.034 endogenen L1-Kopien deutet darauf hin, daß die 5'-Abbrüche mit dem zellulären DNA-Reparaturmechanismus des „microhomology-driven single strand annealing“ zusammenhängen. Vollständige Insertionen hingegen scheinen unseren Daten zufolge durch einen anderen Mechanismus vermittelt zu werden, der nicht auf der Nutzung von Mikrohomologien beruht. Aufgrund dieser Ergebnisse wurde von mir ein Modell entwickelt, das sowohl die häufigen 5'-Abbrüche von L1-Elementen als auch den Mechanismus der Zweitstrangsynthese erklärt.

Die vorliegende Arbeit liefert neue vielversprechende Erkenntnisse über Regulation, Zielsequenzerkennung und den Integrationsmechanismus von L1-Retrotransposons. Dadurch legt sie die Grundlage für weiterführende Untersuchungen, die die Nutzung von Retrotransposons in der Gentherapie ermöglichen könnten.

1. INTRODUCTION

The first concepts of genome organisation assumed that a genome is an assembly of genes interrupted by regulatory elements. But soon it was recognised that the complexity of an organism does not necessarily directly correlate with its genome size (Thomas, 1971; Gregory and Hebert, 1999). For example, the genome of the yeast *S. cerevisiae* with its 6,200 genes has one fifth the number of genes of the human genome containing 30,000 – 40,000 genes. However, the overall size of the human genome, which comprises approximately $3 \cdot 10^9$ basepairs (bp) per haploid chromosome set, is 200 times bigger than *S. cerevisiae*'s 14 Mb (Lander *et al.*, 2001).

With progressive success in sequencing, which culminated in the elucidation of the almost complete human genome (Lander *et al.*, 2001), it became clear that genomes can contain a substantial amount of repetitive sequences. These sequences were often described as "junk DNA" because they had no evident beneficial function for the host (Ohno, 1972; Pagel and Johnstone, 1992). A small percentage of repetitive sequences is comprised of simple repeats like microsatellites (di-, tri-, and tetranucleotide repeats) or telomeric repeats, but the vast majority derives from transposable elements (TEs). Interestingly, transposable elements were first identified in the late 1940s by Barbara McClintock (McClintock, 1950), even before the structure of DNA had been elucidated. McClintock discovered that genetic elements could be mobile and thus challenged the conservative view of a static genome. Today, many different kinds of mobile DNA have been identified in virtually all species ranging from bacteria and yeast to plants and mammals.

The question why TEs have been so successful in evolution is the subject of ongoing discussion. The notion of “selfish genes” (Dawkins, 1976) or “genomic parasites” (Yoder *et al.*, 1997) implies a purely detrimental effect on the host. However, evidence accumulated over the last several years demonstrating that despite their disease causing potential (Kazazian, 1998), TEs might have an overall beneficial effect, for example by increasing the genomic diversity within a species (Boeke and Pickeral, 1999; Nekrutenko and Li, 2001), playing a role in stress response of the host cell (Li and Schmid, 2001) or taking over vital cell functions (Pardue *et al.*, 1996).

1.1 Distribution of Transposable Elements in the Host Genome

Transposition events within the genome can generate deleterious effects by inserting into coding or regulatory regions (Charlesworth and Langley, 1989; Biemont *et al.*, 1997). Therefore, the mobility of all transposable elements is believed to be regulated at some level (Labrador and Corces, 1997). It has been suggested that transcriptional repression of transposable elements by hypermethylation is a major defence mechanism against TEs in eukaryotes (Yoder *et al.*, 1997), a hypothesis corroborated by our results (see 3.2, p.59 and Yu *et al.*, 2001). Nevertheless, many TEs have found ways to circumvent or break down this defence and have been highly successful in colonising their host's genome.

The distribution of transposable elements in the genome is usually not random. Several studies found an accumulation of retrotransposons in regions with low recombination rates (Rizzon *et al.*, 2002 and references therein). Heterochromatic regions are usually gene poor and consist largely of TEs and satellite DNA (Tilford *et al.*, 2001; Hilliker *et al.*, 1980; Cold Spring Harbor Laboratory, 2000). It has been proposed that this accumulation can be explained either by selection against TE-induced mutations (insertion model) or selection against rearrangements caused by ectopic recombination between copies of these elements (ectopic exchange model) (Charlesworth and Langley, 1989; Rizzon *et al.*, 2002).

However, it is becoming increasingly evident that the biased distribution of TEs is not merely the result of passive accumulation caused by the absence of strong forces to eliminate them. The TEs themselves often contribute to their localisation in the genome by coding for proteins that target their integration into preferred sites. They have developed strategies by which they can establish and maintain active populations without causing mutations, i.e. strategies to integrate at positions not occupied by functional host genes or their regulatory elements. Protein-protein interactions can mediate specific targeting, for example through interaction of TE-encoded proteins with chromatin-associated proteins. In fact, the yeast retrotransposon Ty5 has a strong bias to integrate into silent chromatin which is mediated by direct interaction of Ty5 integrase with Sir4p bound to heterochromatic regions (Zou and Voytas, 1997; Xie *et al.*, 2001; Zhu *et al.*, 2003). Targeting of the telosome, a complex of telomere-associated proteins, appears to be an alternative successful strategy for genome colonisation by TEs, which is used by mobile elements from various species (Danilevskaya *et al.*, 1994; Okazaki *et al.*, 1995; Takahashi *et al.*, 1997; Arkhipova and Morrison, 2001; Goodwin *et al.*, 2001). In *D. melanogaster*, the retrotransposons HeT-A and TART have completely taken over the function of telomeres (Pardue *et al.*, 1996), which is a spectacular example of domestication of functions from selfish mobile elements into vital cell functions.

However, specific targeting of certain chromosomal loci is most commonly achieved by recognition of genomic DNA sequences by the TE's integration machinery. This strategy has the principle disadvantage that potential target sequences within a genome are limited. Ideal sites for targeted integration enable exclusive recognition by the TE and are found frequently in the genome. Additionally, insertion into these sites should not be selected against. Therefore, most mobile elements that show significant target site specificity selected reiterated sequences in the genome. These include ribosomal RNA gene clusters, tRNA genes, snRNA genes, transposable elements, telomeric repeats and microsatellites (reviewed in Zingler *et al.*, in press).

1.2 Transposable Elements in the Human Genome

In *Homo sapiens*, transposable elements were responsible for the formation of at least 45% of the genome (Lander *et al.*, 2001). Fig. 1 shows an overview of the different types of mobile elements that caused human genome expansion.

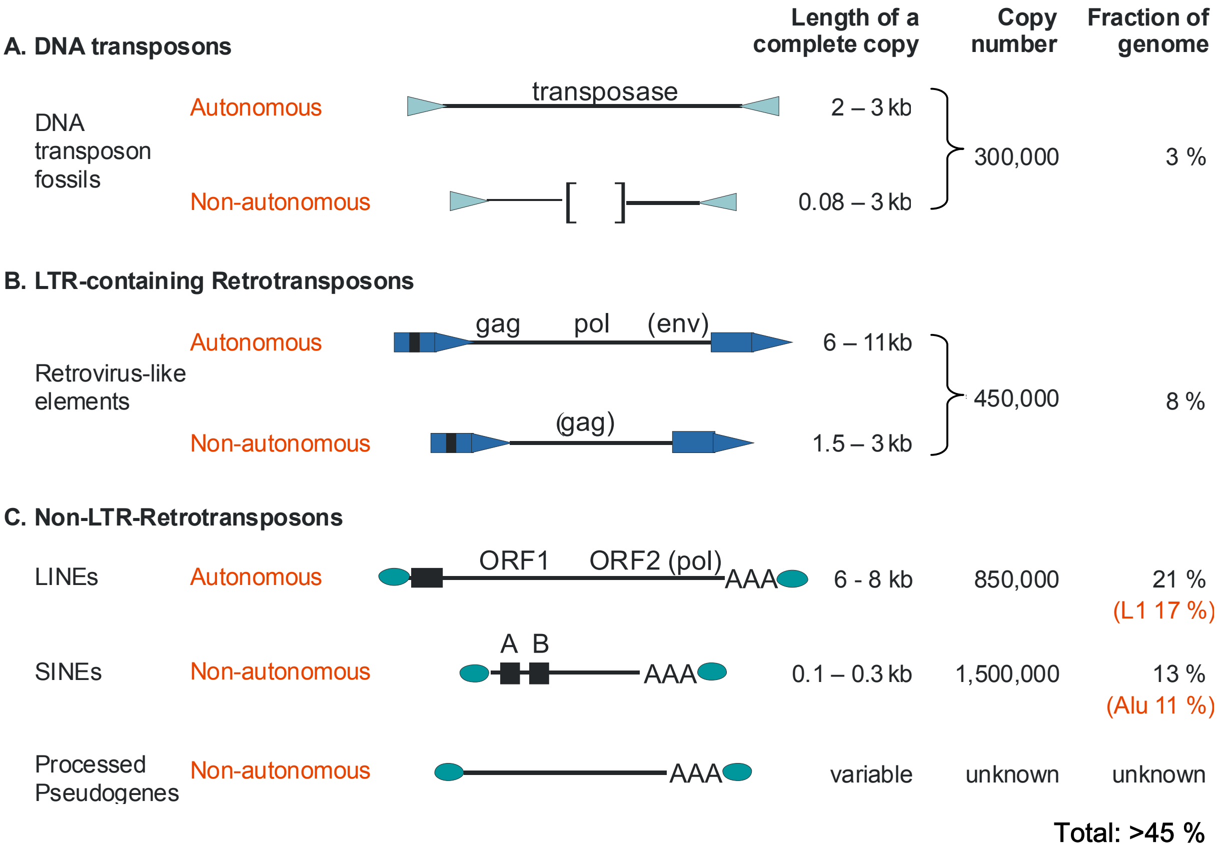


Fig. 1: Schematic representation of the mobile elements present in the human genome (adapted from Lander *et al.*, 2001). Light blue triangles, inverted repeats; dark blue rectangles and triangles, long terminal repeats; green ovals, target site duplications; black boxes, promoter regions; AAA, poly(A) tails.

Only about 3% of the human genome is derived from DNA transposons (Fig. 1A). They move via a DNA intermediate by a cut-and-paste mechanism mediated by a transposase, but all ~300,000 DNA transposons identified in the human genome are genomic fossils that have been inactive for at least 50 Myr (Lander *et al.*, 2001; Smit and Riggs, 1996).

By far the largest portion of mobile DNA in humans originates from retrotransposons. They replicate via a copy-and-paste mechanism involving transcription of the complete element, reverse transcription of the RNA to cDNA, and integration of the cDNA into a new locus in the genome. Thus, one functional retrotransposon can generate multiple copies of itself. This circumstance and the fact that there is at least one family of retrotransposons still active in humans account for the excess of retroelements in the genome.

Retrotransposons can be devided into two major classes that are phylogenetically and structurally unrelated. The LTR-containing retrotransposons, accounting for 8% of the human genome, are characterised by long terminal repeats (LTRs) flanking the elements' coding regions (Fig. 1B). They are also called 'retrovirus-like elements' or 'endogenous retroviruses' because their structure and replication pathway is highly similar to retroviruses. They are thought to originate from retroviruses that have lost a functional *env*-gene. Therefore, endogenous retroviruses cannot infect other cells, and are forced to go through their replicative cycle within a single cell. With the possible exception of HERV-K, which is a putatively active human endogenous retrovirus, all known human LTR-retrotransposons are genomic fossils that have not been active for the last 40 Myr (Lander *et al.*, 2001; Turner *et al.*, 2001).

Non-LTR retrotransposons, the dominating class of retrotransposons, are evolutionary more ancient. Sequence comparisons indicate that they share a common origin with RT-bearing group II introns of bacteria and mitochondria (Xiong and Eickbush, 1990; Yang *et al.*, 1999). Comprising more than one third of the human DNA (34%), non-LTR retrotransposons have had -and continue to have- the greatest impact on our genome (Fig. 1C). 'Long interspersed nuclear elements' (LINEs) are autonomous non-LTR retrotransposons that encode the proteins required for their own retrotransposition. In the human genome, three LINE-families exist, called L1, L2 and L3, or LINE-1 to 3. In order to avoid confusion, in this text LINE will be used as a general term for autonomous non-LTR retrotransposons only, and L as abbreviation for human LINE elements. The diploid human genome contains 92 active L1 elements with ORFs coding for functional proteins (Brouha *et al.*, 2003). In contrast, L2 and L3 have accumulated numerous mutations over time which led to the loss of their ability to transpose autonomously 80-100 Myr ago (Lander *et al.*, 2001).

LINE elements display a marked *cis* preference, i. e. they preferentially copy their own RNA, thus assuring that only functional copies are propagated (Boeke, 1997; Esnault *et al.*, 2000; Wei *et al.*, 2001). However, it has long been proposed that some RNAs can interfere with this *cis* preference and recruit LINE proteins for their own proliferation. The most prominent example of such “parasitic” RNA is the RNA transcribed from the non-autonomous *Alu* element, a 300 nucleotide DNA sequence that is derived from the 7SL RNA gene (Ullu and Tschudi, 1984). It is a member of the class of ‘short interspersed nuclear elements’ (SINEs) that are between 100-300 bp long and are characterised by an internal PolIII-promoter. As *Alus* have no protein coding capacity, they only ensure that their RNA is transcribed. For reverse transcription and integration they rely on L1 elements (Smit, 1996; Boeke, 1997). This relationship between LINEs and SINEs has recently been proven experimentally by the Heidmann laboratory (Dewannieux *et al.*, 2003).

The *Alu* elements’ extraordinary success - more than 1.5 million copies of *Alu* exist in the human genome (Fig. 1C) - is thought to arise from its structure: *Alus* are derived from 7SL RNA, the RNA scaffold of the signal recognition particle (SRP) that binds to nascent signal peptide sequences and transiently arrests translation (Siegel and Walter, 1988). As the secondary structure of *Alu* RNA resembles this ribosomal RNA, *Alu* RNA may be able to associate with ribosomes, get in close physical proximity to nascent LINE proteins and misappropriate them for its own replication (Boeke, 1997; Weichenrieder *et al.*, 2000; Dewannieux *et al.*, 2003).

In rare cases, the *cis* preference of LINEs is also circumvented by spliced mRNAs of cellular genes. This results in an intronless and promoterless copy of the original gene, followed by a polyA tail and flanked by target site duplications (Vanin, 1985). Therefore, these so-called processed pseudogenes (Fig. 1C) are also a direct result of L1 activity (Esnault *et al.*, 2000).

1.3 Non-LTR Retrotransposons

1.3.1 Classification

There are three indispensable constituents of autonomous retrotransposons: (1) a promoter to ensure transcription of a full-length RNA, (2) a reverse transcriptase (RT) to produce a cDNA copy of this RNA and (3) a protein machinery that mediates integration of the cDNA into a new genomic site. While in LTR retrotransposons, the latter function is taken over by an element-encoded classical integrase (Curcio and Derbyshire, 2003), in non-LTR retrotransposons the integration process is initiated by an element-encoded endonuclease (EN).

Based on structural differences and the kind of EN they encode, non-LTR retrotransposons can be classified into two subtypes (Yang *et al.*, 1999) (Fig. 2).

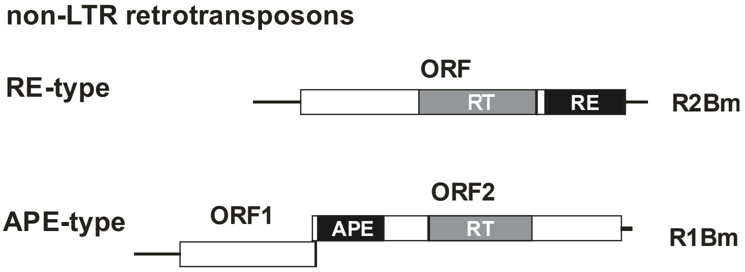


Fig. 2: RE-type and APE-type non-LTR retrotransposons differ in their structural organisation and in their coding capacity. The organisation of R2Bm and R1Bm is depicted with each representing another subtype of non-LTR retrotransposons. RT, reverse transcriptase; RE, restriction enyzme-like endonuclease; APE, apurinic/apyrimidinic endonuclease-like endonuclease. Open bars represent ORFs, thin lines the 5' and 3' UTRs.

RE-type non-LTR retrotransposons are characterised by a single open reading frame (ORF) with a restriction enzyme (RE)-like EN domain following the C-terminal end of the RT domain. This EN domain is similar to type-IIS restriction endonucleases with separate DNA-cleavage and DNA-binding domains (Yang *et al.*, 1999) and is usually sequence-specific (Eickbush, 2002). RE-type elements represent the oldest lineage of non-LTR retrotransposons (Eickbush and Malik, 2002), but as the human genome does not harbour members of this lineage (Lander *et al.*, 2001), they will not be discussed here in detail.

Most retrotransposons discovered so far belong to the second subtype, the class of APE-type non-LTR retrotransposons. They are hallmarked by two ORFs and the existence of an EN domain that is distantly related in sequence to the apurinic/apyrimidinic (AP) endonucleases (Martín *et al.*, 1995; Feng *et al.*, 1996) (see 1.4). The EN domain is localised at the N-terminal end of ORF2p, upstream of the RT domain (Fig. 2). Based on the elements' structures and on phylogenetic analyses of their RT domains, we can currently distinguish four groups of APE-type non-LTR retrotransposons, which can further be subdivided into 11 clades (Burke *et al.*, 1999; Malik *et al.*, 1999; Eickbush and Malik, 2002; Lovsin *et al.*, 2001). Structural and organisational features of members of the 11 clades of APE-type elements are listed in Fig. 3, with the three elements used in this study (see 1.5) highlighted in red.

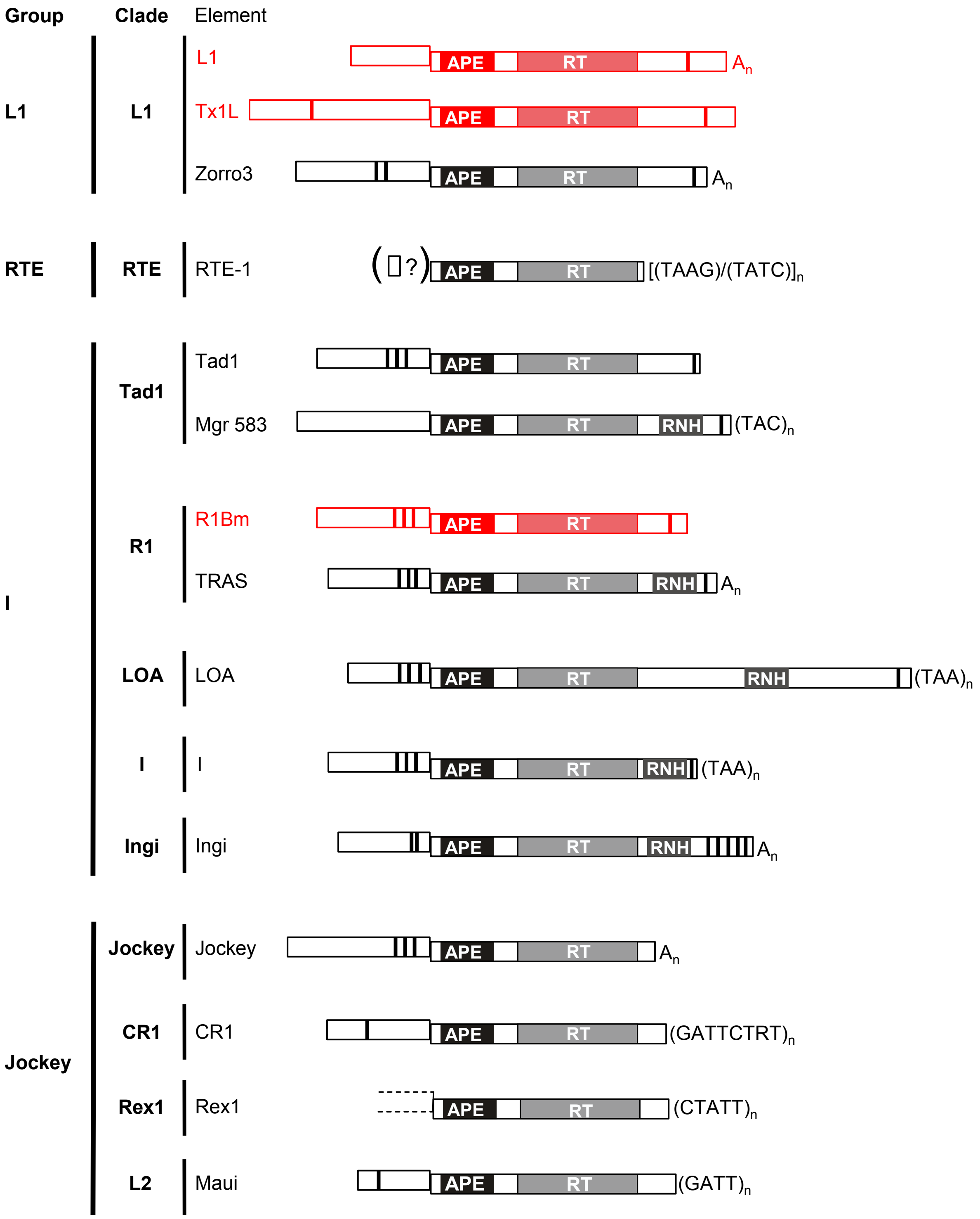


Fig. 3: Schematic ORF structures of representative members of each APE-type non-LTR retrotransposon clade. Open boxes indicate ORFs, shaded boxes represent the enzymatic domains encoded on each element. The stippled box in Rex1 indicates that the 5' end of this retrotransposon has not been identified yet. APE, apurinic/apyrimidinic endonuclease; RT, reverse transcriptase; RNH, RNase H; While ORF lenghts are approximately to scale, enzymatic domains are indicated by rectangles of fixed size. Vertical bars represent cysteine-rich motifs. The three elements used in this study are highlighted in red.

1.3.2 Mechanism of retrotransposition

The mechanism of retrotransposition of non-LTR retrotransposons is not entirely understood. However, the first steps of integration of these elements have been elucidated by biochemical work on the site-specific RE-type retrotransposon R2 from *B. mori* (Luan *et al.*, 1993), which led to a model called 'target primed reverse transcription' (TPRT) (Fig. 4).

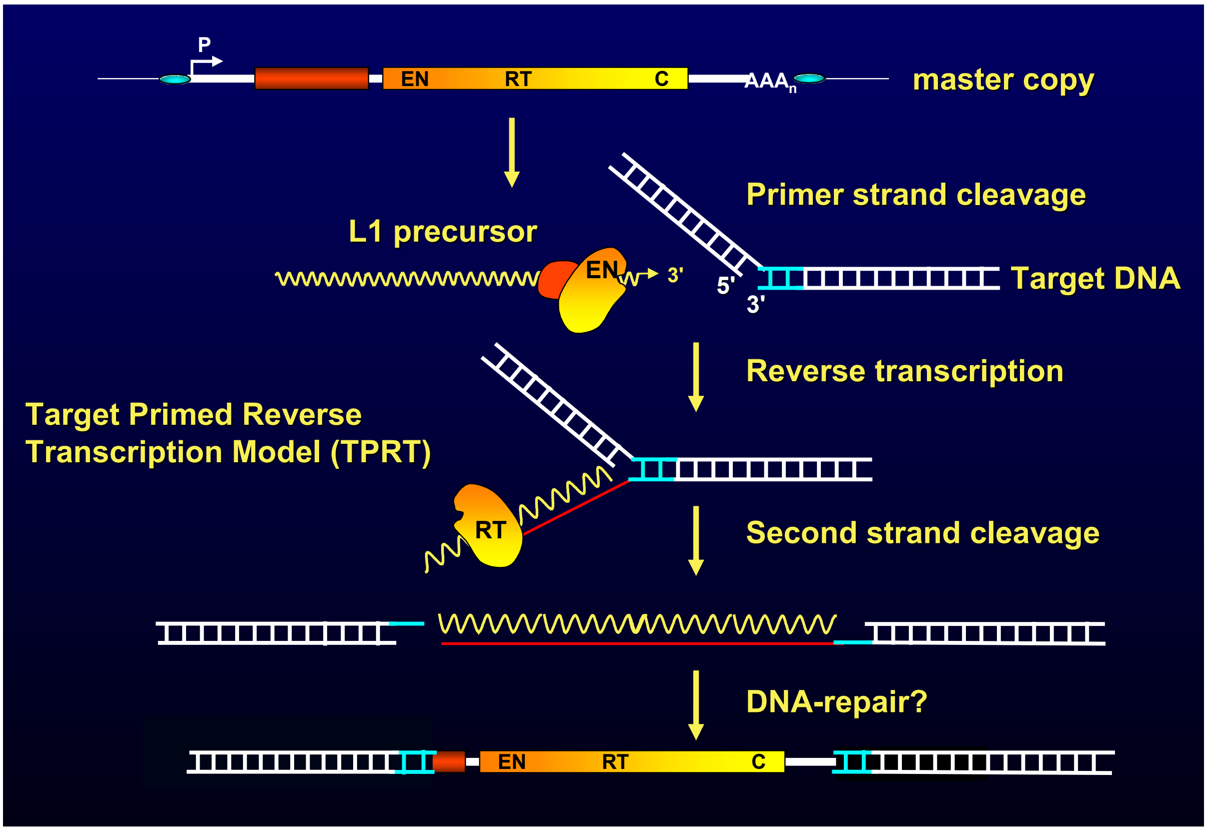


Fig. 4: Schematic representation of the 'target primed reverse transcription' (TPRT) mechanism. After L1 transcription and translation, ORF1 and ORF2 proteins associate with their own mRNA transcript. The EN domain of ORF2p initiates integration by generating a nick in the lower strand of the genomic target DNA. Then the RT domain uses the exposed 3' hydroxyl end to prime reverse transcription. After reverse transcription, cleavage of the upper DNA strand occurs, creating a staggered cut. Second-strand synthesis and ligation of the newly synthesised strands may be brought about either by L1-encoded enzymatic activities or by cellular DNA repair mechanisms. The genomic target DNA is represented as white ladder, with the sequence duplicated during retrotransposition (TSD) coloured blue. L1 RNA is depicted as yellow wavy line. EN, endonuclease; RT, reverse transcriptase.

Although RE-type and APE-type elements belong to different families of non-LTR retrotransposons that share only little structural similarities, the basic mechanism of transposition initiation by TPRT seems to be conserved. This was demonstrated by Cost and co-workers, who reconstituted the initial steps of L1 element transposition *in vitro*, requiring only the complete L1 ORF2 protein, L1 RNA, and a target DNA (Cost *et al.*, 2002). Their work provided the first direct, experimental evidence that the human L1 element, a member of the group of APE-type elements, uses TPRT for retrotransposition. Both studies showed that the EN domains of the two retrotransposons initiate the integration process by nicking the target DNA. The generated 3' hydroxyl group serves as primer for reverse transcription of the elements' RNA. It was demonstrated for L1 that TPRT can also occur at pre-formed nicks and double strand breaks in the target DNA. Therefore, it was concluded that nicking and reverse transcription are two independent steps in TPRT that can be uncoupled (Cost *et al.*, 2002).

The second strand of the target DNA can also be cleaved by the EN domain, though at a much slower rate than the rapidly nicked first strand. Depending on the position of the second nicking site relative to the initial one, TPRT can generate a target site deletion (as for R2 integration), a simple "blunt" integration, or a perfect target site duplication (TSD) flanking the newly inserted element.

A major unresolved issue regarding the mechanism of LINE retrotransposition is what occurs after second-strand cleavage. Despite extensive efforts, *in vitro* experiments with R2 protein did not lead to detection of intermediates expected for second-strand synthesis (Luan *et al.*, 1993). In contrast, *in vitro* TPRT of L1 yielded 5' junctions between L1 sequence and the target DNA. This result indicates that the RT is able to accept cDNA as template for second-strand synthesis, probably by a second round of TPRT (Cost *et al.*, 2002). However, this *in vitro* process is very inefficient. It does not necessarily reflect the natural mode of retrotransposon integration and still leaves open the question how the damaged genomic DNA is repaired. It is generally assumed that cellular DNA repair pathways are involved in these last steps of integration.

1.4 The Family of AP-like Endonucleases

The TPRT model implies that the EN domain is the prime determinant of target site specificity, as the nicking site is identical to the site of integration. However, when this model was developed by Luan *et al.* in 1993, identification of an EN was impossible in many RT-bearing repetitive elements. The breakthrough came in 1995, when Martín *et al*. recognised a sequence homology between the N-terminal part of ORF2p of the retrotransposon L1Tc from *Trypanosoma cruzi* and the AP class II endonuclease family (Martín *et al.*, 1995).

AP class II endonucleases constitute a family of highly conserved, multifunctional DNA repair enzymes with representatives identified in bacteria, plants, insects, and mammals (Barzilay and Hickson, 1995 and references cited therein). They are versatile proteins which, in addition to their endonuclease activity, possess 3‘-phosphatase, 3‘-phosphodiesterase, RNase H, and 3'→5'-exonuclease activities (Demple and Harrison, 1994; Barzilay and Hickson, 1995; Evans *et al.*, 2000). They are involved in the predominant pathway for the repair of oxidative DNA damage and the resulting apurinic/apyrimidinic (AP) sites *in vivo* (Barzilay and Hickson, 1995; Demple and Harrison, 1994).

The existence of a conserved AP-like EN domain in non-LTR retrotransposons raised questions about its function. Any of the activities of AP ENs could potentially play a role in retrotransposition, but in the last decade evidence accumulated indicating that the endonucleolytic cleavage activity is the crucial function of the APE-like domain of retrotransposon ORF2 proteins (Feng *et al.*, 1996; Feng *et al.*, 1998; Christensen *et al.*, 2001; Takahashi and Fujiwara, 2002).

When this project was started, only three members of the family of AP-like ENs had been structurally characterised: bovine pancreatic deoxyribonuclease I (DNase I) (Lahm and Suck, 1991), *E. coli* exonuclease III (ExoIII) (Saporito *et al.*, 1988; Mol *et al.*, 1995) and human AP endonuclease 1 (APE1, HAP1) (Gorman *et al.*, 1997; Mol *et al.*, 2000). Comparison of their structures showed a similar tertiary structure: the core consists of two parallel β-sheets surrounded by several α-helical structures. Flexible loops, especially on the DNA binding surface, connect these structural elements (see Fig. 5).

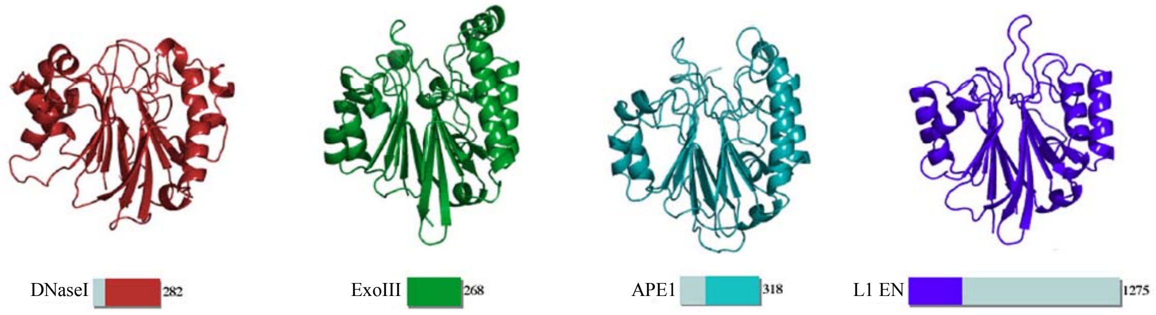


Fig. 5: Crystal structures of four members of the family of AP-like ENs, DNase I, ExoIII, APE1 and L1 EN (kindly supplied by O. Weichenrieder, The Netherlands Cancer Institute). The four enzymes are depicted in the same relative orientation, with the putative DNA binding surface on top. The bars indicate the lengths of the ORFs coding for the respective enzymes and the relative position of the EN domains within.

No information was available on the three-dimensional structure of any retrotransposon-encoded AP-like EN. It was just assumed from the alignment of amino acid sequences and predicted secondary structures that the overall fold of APE1 is maintained. Only very recently, the laboratory of A. Perrakis at the Netherlands Cancer Institute succeeded in elucidating the crystal structure of the human L1 EN (Weichenrieder *et al.*, in press). Their results confirmed that the structures of L1 EN and APE1 are indeed largely superimposable (Fig. 5). The active site residues and the supporting structural elements that place them into their respective positions are highly conserved. This suggests that the DNA cleavage mechanism that has been proposed for APE1 (Mol *et al.*, 2000) applies for human L1 EN as well.

Modulation of cleavage specificity of AP-like ENs is thought to be achieved mainly via variations in the surface loops that contact the DNA. Transplant experiments with ExoIII and DNaseI supported this notion: by grafting a prominent α-helix from the AP-site-specific nuclease ExoIII onto the DNA binding surface of DNaseI, the latter enzyme could be converted from an unspecific endonuclease to a nicking enzyme with high selectivity for abasic sites (Cal *et al.*, 1998).

1.5 Retrotransposons Used in this Study

One focus of my work was to elucidate which regions of the EN domain determine the target site specificity of APE-type retrotransposons. For this purpose I worked with three APE-type retrotransposons: the human L1 element, which prefers to integrate into a short consensus sequence (5’-T/AAAA-3’, where / designates the integration site), and two highly specifically integrating retrotransposons, Tx1L and R1Bm. The latter two elements were selected due to the thorough biochemical characterisation of their sequence-specific EN domains (Feng *et al.*, 1998; Christensen *et al.*, 2000; Christensen *et al.*, 2001). Since Tx1L is phylogenetically closely related to L1, it was grouped in the same clade as L1 (Fig. 3). In contrast, R1Bm belongs to the I group and is the founder member of the R1 clade (Malik *et al.*, 1999). This relationship is also reflected in the phylogeny of the element's host species: the vertebrate species *H. sapiens* and *X. laevis* (African clawed frog) harbour the two closely related elements L1 and Tx1L, while the arthropod species *B. mori* (Mulberry silkworm) is the host of the more distantly related R1Bm.

Organisation and structure of these three elements are very similar. They all display a bicistronic structure and encode APE-type endonucleases. Their structures and integration sites are shown in Fig. 6 and described in detail in the following paragraphs.

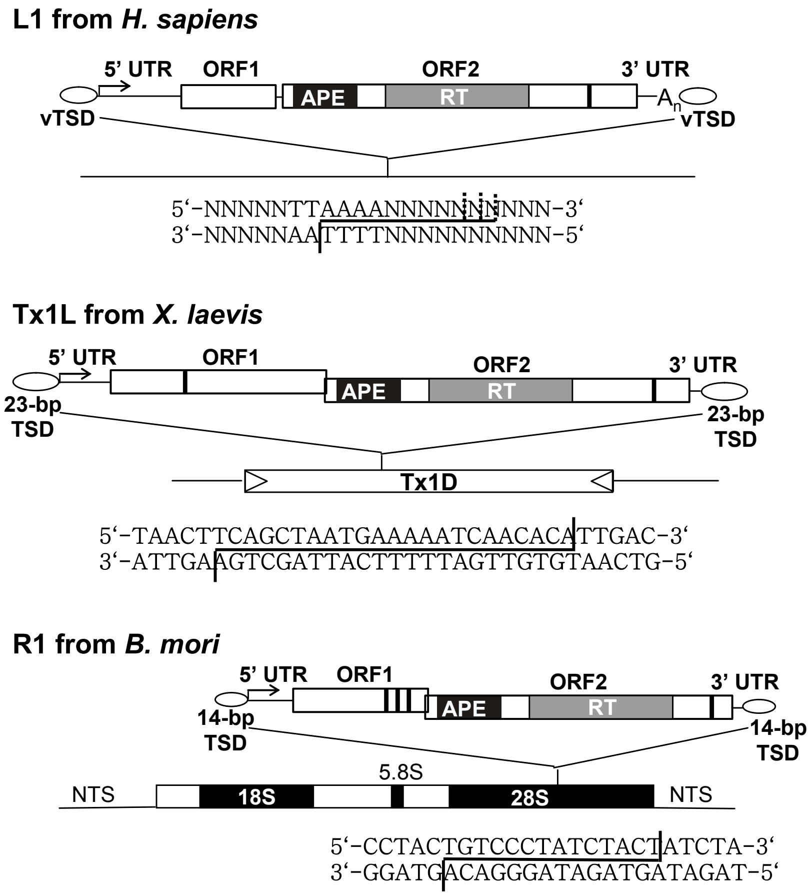


Fig. 6: Structures of L1, R1Bm, and Tx1L with schematic integration sites. Retrotransposons are depicted as in Fig. 2 and 3. In the schematics of the integration sites, horizontal lines represent chromosomal DNA. The DNA transposon TxD is represented as a white rectangle, with short inverted repeats indicated by oppositely oriented triangles. In the drawing of the rDNA locus, filled boxes represent rRNA genes, open boxes indicate external and internal transcribed spacer regions. NTS, non-transcribed spacer. Below each integration site, the exact nucleotide sequences of the elements' target sites are given. The bottom and top strand cleavage sites in each target DNA are represented by bent lines, encompassing the future TSDs. Stippled lines indicate different top strand cleavage sites.

1.5.1 L1 from *Homo sapiens*

1.5.1.1 Structure

To date, human L1 is the most thoroughly characterised APE-type non-LTR retrotransposon (Ostertag and Kazazian, 2001a; Moran and Gilbert, 2002). A complete, retrotransposition-competent full-length L1 element is 6 kb in length and carries two open reading frames (ORFs) (Fig. 6). The nucleotide sequence of a representative functional member of the L1-family, L1.3, is given in Appendix B.

The 5' untranslated region (UTR) of L1 is approximately 900 bp in length. A major polymorphism of L1 elements occurs within this region with a 131-bp sequence being either present or absent (Hattori *et al.*, 1985). The 5' UTR has been shown to house the promoter of

L1 (Swergold, 1990), and the first 155 bp were found to be most critical for L1 expression (Minakami *et al.*, 1992).

The L1 promoter is unusual in that it possesses features of both RNA polymerase II (Pol II) promoters, which control transcription of all protein-coding genes, and RNA polymerase III (Pol III) promoters that are responsible for synthesis of tRNA, 5S RNA and several small, stable RNAs. The L1 promoter creates a long, protein-encoding, polyadenylated transcript that contains several oligo(T) stretches. Since Pol III would terminate transcription at these signals, L1 is likely transcribed by Pol II. However, the promoter is internal, initiates transcription at position +1 of the L1 sequence (Swergold, 1990) and lacks features characteristic of Pol II promoters such as upstream TATA and CAAT boxes. Inhibition experiments yielded contradictory results, supporting sensitivity either to α-amanitin, a Pol II inhibitor, or to tagetitoxin, a Pol III inhibitor (Shafit-Zagardo *et al.*, 1983; Kurose *et al.*, 1995).

Due to the disease causing potential of L1, the host has an evolutionary advantage if transposition is downregulated in somatic cells (see 1.1). However, since L1 can only propagate by vertical transmission, L1 expression and transposition must occur in cells destined for the next generation, i.e. germ cells or early embryonal cells. Indeed, co-expression of the two L1-encoded proteins has recently been detected by immunohistological analyses in prespermatogonia of human fetal testis and in germ cells of human adult testis (Ergün *et al.*, 2004).

Several proteins have been shown to be involved in the transcriptional regulation of L1. Sox11, a member of the SRY family of transcription factors, is a positive regulator of L1 transcription (Tchenio *et al.*, 2000). The same is true for the 'runt-domain transcription factor' RUNX3 that binds to nucleotides +83 to +101 of the L1 5' UTR (Yang *et al.*, 2003). The ubiquitous transcription factor YY1 binds to nucleotides +13 to +26 of the L1 sequence (Becker *et al.*, 1993; Kurose *et al.*, 1995). Since YY1 is capable of both activating and repressing transcription, this protein may play a role in downregulating L1 transcription in some cell types, while activating it in others (Becker *et al.*, 1993).

The 5' UTR of L1 contains a heavily methylated CpG island (Woodcock *et al.*, 1997). In a study of eight cell lines, an inverse correlation was seen between ORF1 protein (ORF1p) expression and the methylation state of the 5' end of L1, indicating that methylation of this region plays a role in L1 regulation (Thayer *et al.*, 1993).

The first open reading frame of L1 (L1 ORF1) is 1017 bp in length and encodes a 338 *aa* protein called p40. Although ORF1p is clearly indispensable for the activity of APE-type retrotransposons (Moran *et al.*, 1996), the function of this protein is still not entirely understood. ORF1p has been shown to form cytoplasmic ribonucleoprotein-complexes with L1 RNA (Hohjoh and Singer, 1996; Hohjoh and Singer, 1997). ORF1p of mouse L1 was demonstrated to have nucleic acid chaperone activity *in vitro* (Martin and Bushman, 2001), indicating involvement in annealing processes during L1 replication (see discussion section, 4.3).

The initiator methionine of ORF2 in the human L1 element is separated from ORF1 by a 66-bp in-frame spacer region containing three stop codons. It is not clear how the separate translation of both ORFs from the bicistronic RNA is accomplished, a problem made even more intriguing by the fact that the spacer region is not conserved between L1 elements of different species. Suppression of the stop codons or ribosomal frameshifting to create a fusion protein could be experimentally excluded (Leibold *et al.*, 1990; McMillan and Singer, 1993). Therefore it was concluded that translation of ORF2 must be accomplished either by reinitiation of translation (Kozak, 1987) or internal initiation via an internal ribosomal entry site (IRES) (McMillan and Singer, 1993).

The second open reading frame (ORF2) of L1 codes for a protein of ~150 kDa containing 1275 *aa* (Scott *et al.*, 1987). This polyprotein harbours an N-terminal AP-like EN (see 1.4) as well as an RT domain (Mathias *et al.*, 1991). At the C-terminal end, there is a cysteine-rich region whose function is still unclear. However, it has been shown that mutations in this region abolish retrotransposition in cultured cells (Moran *et al.*, 1996).

The 3' UTR covers 205 bp, includes a polyadenylation signal, and terminates in a poly(A) tail (Grimaldi *et al.*, 1984). This portion of the L1 element is little conserved within and between species (Scott *et al.*, 1987), and no functional role of the 3' UTR has yet been documented. Interruption of this region by additional nucleotides does not seem to have severe effects on retrotransposition. This could be demonstrated in a reporter assay, where L1 tolerates marker genes of up to 3500 bp in length in its 3' untranslated region (Moran *et al.*, 1996; Ostertag *et al.*, 2000; Gilbert *et al.*, 2002; Symer *et al.*, 2002).

All specifications given above apply to full length copies of L1. However, only 5 % of the ~one million endogenous human L1 elements are 6 kb in length. The remaining 95 % are 5' truncated and/or internally rearranged (Szak *et al.*, 2002). Some of these damages may be the result of coincidental genomic rearrangements after integration of the retrotransposon, but the two major aberrations, 5' truncation and inversion, most probably occur during the retrotransposition process (Ostertag and Kazazian, 2001a). 5' truncations are generally thought to originate from low processivity of the reverse transcriptase. If the RT and the RNA template dissociate before completion of reverse transcription, the resulting insertion will be truncated at the 5' end (Ostertag and Kazazian, 2001a). In inverted L1 elements, the L1 sequence is not only 5' truncated, but the 5' part of the transposed sequence is oriented in the direction opposite to its 3' end. This structure is thought to the consequenc of a mechanism called 'twin priming' (Ostertag and Kazazian, 2001b), which will be described in detail in 4.3. Inversions are by no means rare events, they can be detected in about 25 % of insertions of members of the youngest L1 subset, the 'transcribed, active' Ta family (Ostertag and Kazazian, 2001a; Skowronski *et al.*, 1988).

3' transduction is another structural peculiarity of L1 elements. As the L1 polyadenylation signal is rather weak, it is often ignored by the RNA polymerase if a stronger signal is localised downstream of L1. This results in retrotransposition of a possibly truncated copy of the L1 sequence along with its 3' flanking genomic sequence (Pickeral *et al.*, 2000; Goodier *et al.*, 2000; Szak *et al.*, 2003).

L1 integrants are usually flanked by variable TSDs with lengths up to 60 bp (Szak *et al.*, 2002) which are the consequence of the replication mechanism of L1. It should be noted though, that some TSDs are difficult to identify, e.g. due to statistical uncertainties about the occurrence of short duplications or due to multiple mutations in TSDs of ancient integrants. Still, many L1 elements are not flanked by TSDs, which may be the result of integration into blunt end nicking sites (Van Arsdell and Weiner, 1984) or into a staggered double strand break with a 5' instead of a 3' overhang. The latter process causes a deletion of the target site instead of a duplication (Gilbert *et al.*, 2002).

1.5.1.2 Target site specificity

L1 elements accumulate in A+T-rich regions of the genome (Lander *et al.*, 2001) and generally transpose into the consensus sequence 5'-T/AAAA-3' (Jurka, 1997; Szak *et al.*, 2002). (It should be noted that integration sites are usually given in the same orientation and on the same strand as the coding strand of the inserted element. For the description of EN nicking sites however, it is more useful to refer to the actual recognition sequence on the non-coding strand, i. e. to the reverse complementary sequence.)

L1 was the first element with a direct correlation being observed between the insertion specificity of an APE-type retrotransposon and the nicking specificity of the EN it is coding for (Feng *et al.*, 1996). Feng and co-workers reported that the protein encoded by the amino-terminus of L1 ORF2 has nuclease activity but shows no preference for AP sites (Feng *et al.*, 1996) By mutating crucial residues in the human L1 EN, it could be demonstrated that its activity is required for active transposition in cultured cells (Feng *et al.*, 1996). *In vitro* assays showed that the specificity of purified L1 EN for the 5'-TTTT$_{\uparrow}$A-3' consensus sequence (Feng *et al.*, 1996; Cost and Boeke, 1998; Cost *et al.*, 2001) mirrors the sequence at the sites of *de novo* L1 insertion *in vivo* (Symer *et al.*, 2002; Gilbert *et al.*, 2002). This experimental evidence has been corroborated by computer analysis of the sites of pre-existing L1 and *Alu* element insertions in the human genome (Jurka, 1997; Szak *et al.*, 2002).

L1 EN was demonstrated to be specific for DNA within a range of structural and sequence parameters, with minor groove width being of particular importance. On free DNA, L1 EN nicks at kinkable regions of DNA present between regions of very stiff DNA. The DNA sequence that best correlates with these requirements is T_nA_n, with nicking occurring preferentially at the TpA and flanking phosphodiester bonds (Cost *et al.*, 2001). L1 EN recognition of the 5'(T_n) portion of this sequence is far more extensive and important for nicking than the rather minimally contacted 3'-half of the target DNA. Nucleotide substitutions which conserve the homopyrimidine or homopurine run are generally well tolerated.

In vivo, much of the genome exists in the form of chromatin or is undergoing biochemical transactions such as transcription, replication or repair, which may alter the accessibility of the DNA for the L1 transposition machinery. Thus, the effect of substrate chromatinisation on the nicking activity of L1 EN was examined (Cost *et al.*, 2001). It was found that nucleosomal wrapping of DNA renders it a less-efficiently-nicked substrate, but when so wrapped some phosphodiesters at specific positions in the nucleosome are nicked at an increased rate (Cost *et al.*, 2001). While the global choice of integration sites may be determined by the accessibility of DNA within chromatin, on a local scale the endonuclease domain is the primary determinant of the specificity of L1 integration (Cost and Boeke, 1998).

1.5.2 Tx1L from *Xenopus laevis*

1.5.2.1 Structure

Tx1L was first mentioned in the description of two complex families of transposable elements, Tx1 and Tx2, from the genome of the South African frog *X. laevis*. Both related families were described to consist of apparent cut-and-paste transposons (Tx1D or Tx2D) interrupted by non-LTR retrotransposons (Tx1L or Tx2L) (Garrett *et al.*, 1989). Further analysis showed that only 10% of the approximately 1500 copies of Tx1D and Tx2D carry a TxL element, indicating that TxL elements are autonomous non-LTR retrotransposons that specifically target their corresponding TxD element.

TxL elements were selected for this study since they are structurally and phylogenetically closely related to L1 (25% sequence identity of the EN domains), but exhibit a much higher sequence specificity. Besides, their EN domains have been studied in great detail. Since Tx1L is better characterised than Tx2L, I decided to use Tx1L for the planned experiments.

Full length Tx1L is 6.9 kb in length and has a 555-bp 5' UTR (Fig. 6). ORF1 encodes a protein of 775 *aa*, while the ORF2 protein comprises 1308 *aa*. The ORFs are not separated by a spacer region as in L1, but overlap by seven bp. The 133-bp 3’ UTR does not carry a classical polyadenylation signal. Thus, Tx1L is the only documented member of the L1 clade not ending in a genuine poly(A) tract. Still, it ends in an A-rich tail with the sequence AATAATATA, bearing some similarity to the $(TAA)_n$ 3’ repeats of I clade retrotransposons (Finnegan, 1997). Tx1L is flanked by a perfect TSD of the 23 bp sequence 5’-TCAGCTAATGAAAAATCAACACA-3’, which is part of the transposon Tx1D.

1.5.2.2 Target site specificity

A striking feature of the two closely related elements Tx1L and Tx2L is that, despite ~70% sequence identity of their target sequences, a cross-integration of Tx1L into Tx2D or Tx2L into Tx1D has not been found, even after screening dozens of elements. It was suggested that the endonucleases encoded by the TxL elements have sufficient specificity to enforce this segregation. In order to test this hypothesis, the EN domains of the two TxL elements were overexpressed in bacteria and analysed for their DNA nicking specificity (Christensen *et al.*, 2000; Christensen *et al.*, 2001). The activities of both ENs were tested on oligonucleotides representing the Tx1L- and Tx2L-specific insertion sites. Tx1L makes a specific nick in the bottom strand of its own target sequence precisely at the 5’ end of the presumed Tx1L TSD (Christensen *et al.*, 2000). In addition to the major nicking site, Tx1L cleaves a few other sites with a low frequency. However, when offered the Tx2 target DNA, Tx1L EN exhibited less sequence specificity. An attempt to define a consensus recognition sequence from the most prominent observed nicking sites yielded the rather compliant consensus 5'-YTGN/AR(T/A)T-3' (Christensen *et al.*, 2001). Tx2L EN also makes a strong nick at the expected site for TPRT and prefers its own target DNA, but on the whole it is less specific than Tx1L EN (Christensen *et al.*, 2001).

Neither EN shows sufficient specificity *in vitro* to account for the observation that neither element is found in the *X. laevis* genome outside its corresponding target sequence (Garrett *et al.*, 1989). This indicates that additional determinants might be necessary to achieve the sequence specific integration observed *in vivo*. However, it should be considered that the *in vitro* activity might not fully reflect the behaviour of the EN domains in their natural context as part of a polyprotein. Besides, for my experiments stringent specificity of Tx1L EN is not desirable as the exact 23-bp Tx1L target sequence does not exist in the human genome. However, the observed target site preference of the purified TxL EN domains is a strong indication that the endonuclease indeed is an important, if not the main, determinant of integration specificity.

1.5.3 R1 from *Bombyx mori* (R1Bm)

1.5.3.1 Structure

R1 elements are a family of non-LTR retrotransposons that interrupt the 28S rRNA genes in the rDNA loci of every arthropod lineage examined to date (Jakubczak *et al.*, 1991; Burke *et al.*, 1998; Burke *et al.*, 1993). In the genome of the silkmoth *Bombyx mori*, there are about 25 copies of R1Bm (Xiong and Eickbush, 1988). A full-length copy of this element is 5.1 kb long and carries two open reading frames (Fig. 6). ORF1 codes for a gag-like protein with 461 *aa*. ORF2 overlaps with ORF1 by 20 nucleotides in the +1 reading frame (mistakenly counted as 19 in Xiong and Eickbush, 1988) and is 1051 *aa* long. R1Bm is similar to Tx1L in that its 110-bp 3' UTR does not contain any polyadenylation signal or a poly(A) tail. However, in contrast to Tx1L, it does not even end in an adenine-rich sequence. R1Bm is flanked by the defined 14 bp target site duplication 5'-TGTCCCTATCTACT-3'.

1.5.3.2 Target site specificity

A number of retrotransposons, e.g. R1, R2, R6, R7, G, Mutsu (reviewed in Zingler *et al.*, in press), target ribosomal RNA genes. Several factors make the ribosomal RNA locus an excellent choice for a target site:

- Since rRNAs have universal and essential functions and work as RNA molecules, their functional regions are highly conserved at the nucleotide level.
- Interruption of a subset of RNA genes in a family of several hundred copies per genome will have less severe effects than insertion into an essential single-copy gene.
- Mobile elements that insert at random in the genome run the risk of inserting into regions where new copies will be expressed at too high or too low levels. In contrast, rDNA is constantly and uniformly transcribed, furnishing new insertions with a stable environment.
- Since recombination within the rDNA locus continually removes insertions, this process of concerted evolution means that only active elements will survive in the long run. The accumulation of defective elements in a genome has been suggested as a major reason why mobile element lineages are lost from a species (Charlesworth and Langley, 1989; Hartl *et al.*, 1997).

Thus by evolving specificity for a highly conserved repeated gene, rDNA-specific elements have eliminated the randomness associated with the insertion of new copies. Although insertion into rDNA does not *per se* exclude deleterious effects on the host, no such effects have been reported to date.

R1Bm occupies approx. 10% of the estimated 240 ribosomal DNA units in *B. mori* (Xiong and Eickbush, 1988). These insertions are always flanked by a defined, perfect TSD of 14 bp within the 28S rRNA gene (see above). However, R1Bm is not exclusively located in the rDNA locus. A *B. mori* genomic screen identified at least two R1Bm copies that had integrated into different sites. Interestingly, these non-rDNA targets exhibit similarity to the 28S target sequence (Xiong *et al.*, 1988). These findings imply that R1Bm should also be able to integrate into human rDNA, which differs in 1 bp from *B. mori* rDNA within the R1Bm recognition sequence. Furthermore, these results suggested that targeting of R1Bm depends on DNA sequence rather than interaction with rDNA-specific chromatin factors, which was later confirmed by the biochemical characterisation of the EN domain of R1Bm (Feng *et al.*, 1998). The bacterially expressed and purified EN was shown to cleave with sequence specificity precisely at positions in rDNA corresponding to the boundaries of the R1Bm target site duplication. However, additional cleavage products were observed on the top strand, indicating that cleavage by R1 EN is not absolutely sequence specific *in vitro*. Further nicking experiments suggested that R1 EN specificity can be altered by the effects of flanking sequences (Feng *et al.*, 1998).

1.6 Potential of the Human L1 Retrotransposon as Vector for Gene Delivery

Functional retrotransposons transport genetic information to other genomic loci. Like viruses, they can therefore serve as tools to modify the genome. They could be harnessed as insertional mutagens, cell lineage markers and, most importantly, as gene delivery vectors.

Gene therapy is defined as a medical intervention that changes the genetic material of living cells. To this purpose, DNA carrying a therapeutic gene has to be introduced into the target cells. To achieve a long term effect, stable integration of the transgene into the host cell genome is desirable. Also, the application of gene therapy should not have severe side effects. Initial experiments and clinical studies showed that introduction of DNA into host cells is best achieved with the help of natural "gene shuttles" like viruses (overview in http://www.wiley.co.uk/genetherapy/clinical). Depending on the virus type, the DNA is only transiently transduced and rapidly eliminated from proliferating cells (e.g. adenoviral vectors, [Volpers and Kochanek, 2004]), remains episomal and replicates in synchrony with the host cell (e.g. EBV-based vectors, [Delecluse and Hammerschmidt, 2000]) or is integrated into the host genome (e.g. retroviral vectors, [Coffin, 1996]). However, viral vectors lost much of their attractiveness after the death of a test person due to an anaphylactic shock in a clinical trial with adenoviral gene therapy (Raper *et al.*, 2003) and the occurrence of two instances of leukemia probably caused by insertional mutagenesis of retroviral gene therapy vectors (for details see 4.2, p.98 and Hacein-Bey-Abina *et al.*, 2003).

Retrotransposons are a potential alternative to the currently used gene delivery systems. Several features make them interesting candidates as gene delivery vehicles.

- If a gene delivery vector is based on the human retroelement L1, the danger of eliciting an immune response of the patient against the modified cells is minimal.
- LINEs are able to stably integrate into the genome, thus ensuring a long term therapeutic effect.
- Viruses change their host organism frequently and only have to ensure that the host survives the virus' replicative cycle. In contrast, several non-LTR retrotransposons have evolved intricate strategies to target specific innocuous genomic sites in order to ensure their long-term survival in the host genome. These strategies could also be applied in gene therapy to avoid adverse effects by insertional mutagenesis.

- The risk of uncontrolled replication, which is a safety issue with viral vectors, is very low with retrotransposons, as the inherent 5’ truncations (see 1.5.1.1, p.25) efficiently inactivate most retrotransposed copies.

However, retrotransposons cannot infect cells and therefore have to be introduced into their target cells by other means. Direct physical methods like lipofection are not recommended since they are very inefficient, allow no control of the copy number and may cause severe chromosomal aberrations upon transgene integration (Schmidt-Wolf and Schmidt-Wolf, 2003). Alternatively, gene delivery could be achieved by a combination of retrotransposons and virus shuttles. As proof-of-principle, the Kasahara laboratory recently demonstrated stable integration and expression of transgenes delivered by an L1-adenovirus hybrid vector (Soifer *et al.*, 2001). Since this system involves “gutless” vectors devoid of all coding viral genes that could be toxic or immunogenic, it unites high transduction efficiency and low immunogenicity of these helper-dependent adenoviral vectors with the advantages of retrotransposons mentioned above.

1.7 Aims of this Study

Non-LTR retrotransposons have had a major impact on the genomes of almost every known eukaryotic organism. In spite of their wide distribution, the significance of ‘long interspersed nuclear elements’ has long been underestimated due to their classification as “junk DNA”. Only in the last 15 years, systematic studies have been launched to examine their biology. Although great progress has been made, many fundamental questions about their origin, evolution, putative function and mechanism of replication are still only partially answered.

Although not thoroughly characterised yet, LINEs have been demonstrated to exhibit several features that make them potential tools for gene therapy. This thesis is focussing on three characteristics of non-LTR retrotransposons that are relevant for their usefulness as gene delivery vectors:

- For gene therapy purposes, long term expression of the delivered transgene is crucial. Since L1 elements are usually silenced in somatic cells, it was important to identify the mechanism that is responsible for the repression of L1. Although methylation of the promoter has been suggested as cause of repression, direct experimental evidence for this hypothesis was not available. Coexpression studies of L1 and methyl-CpG-binding proteins should elucidate whether methylation of L1 indeed correlates with its transcriptional regulation.
- When compared to conventional gene therapy vectors like retroviral vectors, one of the major advantages of non-LTR retrotransposons is the ability of a subset of these elements to insert into specific DNA sequences or defined loci of the host genome without substantially harming the host. One aim of this study was to define protein regions or structural motifs that influence target site recognition of non-LTR retrotransposons by exchanging the AP-like EN domain of the semi-site-specific human L1 element with ENs borrowed from closely related site-specific retrotransposons.
- The predisposition toward frequent and variable truncations at the 5’ end of integrated L1 copies is an ambivalent property in terms of gene therapy: While deletion of the promoter region in 95% of all retrotransposition events provides a built-in safety mechanism that prevents subsequent transposition of the transgene-containing L1 insertion, more extensive truncations damage or delete the therapeutic gene. Elucidation of the mechanism leading to 5’ truncations promises to yield new strategies to direct the truncation to a point where it would be possible to exploit the advantage without suffering the disadvantage.

2. MATERIALS AND METHODS

The methods used in the experimental part of this thesis were carried out as described in standard laboratory manuals (Ausubel *et al.*, 1999; Sambrook *et al.*, 1989). In this section, only modified methods are described in detail.

2.1 Chemicals

All chemicals were purchased at analytical grade from the following companies unless stated otherwise: Amersham Biosciences (Freiburg), Biochrom (Berlin), Biomol (Hamburg), Difco-Laboratories (Hamburg), Invitrogen (Karlsruhe), Merck (Darmstadt), neoLab (Heidelberg), Roche (Mannheim), Serva (Heidelberg) and Sigma (München). α-[^{32}P]-labelled radionucleotides were supplied by Hartmann Analytic (Braunschweig).

2.2 Enzymes

DNA modifying enzymes, e.g. restriction endonucleases, T4 ligase, calf intestinal phosphatase and *Taq* polymerase were obtained from Roche (Mannheim), Invitrogen (Karlsruhe), MBI Fermentas (St. Leon-Rot), New England Biolabs (Schwalbach), Stratagene (Heidelberg) and Qiagen (Hilden). Reactions were carried out according to manufacturer's directions.

2.3 Buffers and Solutions

TE buffer

10 mM Tris-HCl
1 mM Na$_2$EDTA

pH 8.0

50x TAE buffer

2.0 M Tris-Acetate
0.1 M Na$_2$EDTA

pH 8.2

Sample buffer for agarose gels

50 % (v/v) glycerol
0.25 % (w/v) Xylenecyanol FF
0.25 % (w/v) Bromophenolblue
in TE buffer

PBS (Phosphate Buffered Saline)

137.0 mM NaCl
2.7 mM KCl
6.5 mM Na$_2$HPO$_4$
1.5 mM K$_2$HPO$_4$

pH 7.4

STET-buffer

8 % (w/v) sucrose
5 % (v/v) Triton X-100
50 mM Tris
50 mM EDTA

pH 8.0

<u>LB Broth (Luria-Bertani-Medium)</u>

10 g tryptone
5 g yeast extract
5 g NaCl

ad 1 l dH$_2$O
autoclave

<u>LB Agar</u>

10 g NaCl
5 g tryptone
5 g yeast extract
20 g agar

ad 1 l H$_2$0
autoclave

<u>SOC-Medium</u>

20 g tryptone 2 % (w/v)
5 g Yeast extract 0,5 % (w/v)
10 mM NaCl
2.5 mM KCl
10 mM MgCl$_2$
10 mM MgSO$_4$
20 mM Glucose

ad 1 l H$_2$O
sterilise by filtration

<u>20x SSC</u>

3.0 M NaCl
0.3 M trisodium citrate

pH 7.0

2.4 Methods of Molecular Biology

2.4.1 Bacterial strains

strain	characteristics	reference
DH5α	F' φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F)U169 *deo*R *rec*A1 *end*A1 *hsd*R17(rk$^-$,mk$^+$) *pho*A *sup*E44 λ$^-$ *thi*-1 *gyr*A96 *rel*A1	Invitrogen (Karlsruhe)
XL1-Blue	*rec*A1, *end*A1, *gyr*A96, *thi*-1, *hsd*R17, *sup*E44, *rel*A1, *lac* {F', *pro*AB, *lac*I^{q}ZΔM15, Tn10(Tet^R)}	Stratagene (Heidelberg)
DH10B	F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80d*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ$^-$ *rps*L *nup*G	Invitrogen (Karlsruhe)
GM2163	F$^-$ *ara*-14 *leu*B6 *fhu*A31 *lac*Y1 *tsx*78 *gln*V44 *gal*K2 *gal*T22 *mcr*A *dcm*-6 *his*G4 *rfb*D1 *rpsL136 dam*13::Tn9*(*Cam^R*) xyl*A5 *mtl*-1 *thi*-1 *mcr*B1 *hsd*R2	New England Biolabs (Schwalbach)
Top10F'	F$^-$ {*lac*I^q Tn*10* (Tet^R)} *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G	Invitrogen (Karlsruhe)

Table 1: Bacterial strains. Names, genotypes and suppliers of the bacterial strains used in this work.

For standard cloning steps, the *E. coli* strains DH5α and XL1-Blue were used. Ligation reactions yielding large plasmids (>18 kb) were transformed into MAX EFFICIENCY DH10B Competent Cells (Invitrogen, Karlsruhe). Plasmids requiring digestion with methylation sensitive restriction enzymes were transformed into and reisolated from the dcm- and dam-negative strain GM2163 (New England Biolabs, Schwalbach). PCR products were cloned using the TA cloning kit (Invitrogen) with the supplied Top10F' bacteria. For plasmid rescue experiments, ElectroMax DH10B cells (Invitrogen) were electroporated.

2.4.2 Cultivation and storage of *E. coli*

Bacteria were cultivated in Luria-Bertani (LB)-broth or on LB-agar plates (see 2.3) at 37°C. For selection of transformed bacteria, broth and plates were supplied with 50 mg/l ampicillin or 50 mg/l kanamycin.

Freshly plated bacteria were viable for approximately one month when stored at 4°C. For long-term storage, glycerol stocks were prepared by mixing 500 µl of an overnight liquid culture with 500 µl of 15 % sterile glycerol. The stocks were stored at –80°C.

2.4.3 Transformation of DNA into *E. coli*

Chemically competent or electro-competent *E. coli* cells were obtained from Invitrogen (MAX EFFICIENCY DH10B, OneShot TOP10F' and ElectroMax DH10B) or prepared following standard $CaCl_2$-protocols (Ausubel *et al.*, 1999). Plasmids and ligation reactions were introduced into bacteria by heat-shock transformation according to the protocol supplied with MAX EFFICIENCY DH10B Cells or by electroporation at 1.9 kV, 200 Ω and 25 µF in a GenePulser Xcell electroporator (Bio-Rad, München).

2.4.4 Preparation of plasmid DNA from *E. coli*

2.4.4.1 Boiling method

For analytical purposes not requiring high quality DNA (e.g. screening of colonies), a modified boiling method (Holmes and Quigley, 1981) was used:

Cells from 1.5 ml liquid culture were pelleted and resuspended in 300 µl STET buffer. The cells were lysed by incubation with 10 µl of 10 mg/ml lysozyme for 5 minutes at room temperature and boiled for 3 minutes. Cell debris and denatured proteins were pelleted by centrifugation (13,000 rpm, 15 min) and removed with a toothpick. The DNA was isolated from the supernatant by isopropanol precipitation followed by a washing step with 70% ethanol. If further purification was necessary, the DNA was extracted with phenol/chloroform/isoamyl alcohol (25:24:1) and subsequently precipitated with ethanol.

2.4.4.2 Ion exchange purification

DNA needed for sequencing or transfection of HeLa cells was isolated with the commercially available ”Plasmid Purification Kits” (Qiagen, Hilden). Extraction and purification of the DNA was achieved by ion exchange columns (QIAGEN-tip 20, 100 or 500) following the user manual. The resulting DNA pellet was dissolved in TE buffer (pH 8.0). After photometrical determination of the yield, the isolated DNA was characterised by restriction and sequencing analysis.

2.4.5 PCR methods

The polymerase chain reaction (PCR) is a powerful method to amplify specific DNA fragments. Developed by K. B. Mullis and co-workers in 1988 (Saiki *et al.*, 1988), it is now a well established and versatile standard procedure in molecular biology. As a major part of my work is based on PCR methods, I will shortly specify the modifications introduced to adapt the PCR for various purposes.

The reactions were performed either in a T3 thermocycler (Whatman-Biometra, Göttingen) or in a GeneAmp 9700 (Applied Biosystems, Darmstadt).

2.4.5.1 Preparative PCR

If PCR products were needed for cloning purposes, the reactions were carried out using the proof-reading polymerase *Pfu* (Stratagene, Heidelberg).

Additional restriction sites needed for cloning were introduced either at the ends of the PCR product by primers containing the desired sites or within the PCR product using site overlap extension PCR (SOE-PCR) (Aiyar *et al.*, 1996) which is performed in two steps (Fig. 7):

In the first step, two primer pairs are used to amplify two fragments that are at least 30 bp complementary to each other in the region of the mutation to be introduced. In the second step, these two PCR products act as megaprimers which, assisted by the two outer primers, amplify the complete fragment with the new mutation within.

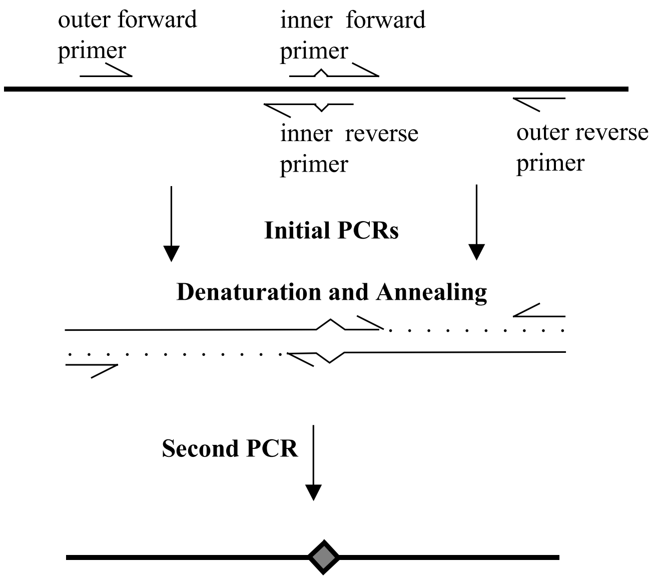


Fig. 7: Principle of site-overlap extension PCR (SOE-PCR) to introduce a mutation. Two primer pairs are used in initial PCRs to amplify two fragments overlapping by 30 bp in the region of the mutation to be introduced (symbolised by triangular structures). These fragments are used as megaprimers that are elongated (indicated by dotted lines) in the second PCR. The outer primers are added to increase the yield after the first overlap extension reactions have created the full-length product. thin lines, single stranded DNA; bold lines, double stranded DNA.

Initial denaturation of DNA:	96°C	1 min	
Exponential amplification:			25 cycles
Denaturation of DNA:	96°C	20 s	
Annealing of primers:	45-65°C	15 s	
Primer elongation by *Pfu*:	72°C	30-120 s	
Final DNA elongation:	72°C	7 min	

Reaction mix for the initial PCR:

5 µl 10 x Pfu Polymerase buffer (Stratagene)
10 pmol of each primer
20 fmol DNA template
0.2 mM of each dNTP
1 U Pfu Polymerase
ad 50 µl dest H_2O

Reaction mix for the second PCR:

equimolar amounts of each PCR product (about 100 fmol)
10 pmol of each outer primer
0.2 mM of each dNTP
1 U Pfu Polymerase
ad 50 µl dest H_2O.

The conditions chosen for the second PCR were the same as those described for the initial one.

2.4.5.2 Analytical PCR

In order to screen bacterial colonies resulting from the transformation of ligation reactions for clones containing the desired insert, single colonies are picked with a toothpick and swirled in 30 µl TE-buffer. This suspension is used instead of template DNA, since bacteria are lysed and liberate their plasmid DNA in the first denaturation step.

Bacterial lysis and initial denaturation of DNA:	96°C	4 min	
Exponential amplification:			
Denaturation of DNA:	96°C	20 s	25 cycles
Annealing of primers:	45-65°C	15 s	
Primer elongation by *Taq*:	72°C	30-120 s	
Final DNA elongation:	72°C	5 min	

Reaction mix:

2 µl of bacterial suspension (a single colony was picked with a toothpick and swirled in 30 µl TE-buffer)
4 pmol of each appropriate diagnostic primer
0.2 mM of each dNTP
0.5 u *Taq* Polymerase
2 µl "Yellow Sub"
ad 20 µl dest H_2O

"Yellow Sub" (Geneo BioProducts GmbH, Hamburg) acts as enhancer of annealing specificity as well as loading buffer substitute at the same time.

2.4.5.3 PCR on genomic DNA

PCR protocols optimised for plasmid templates can easily be adapted for genomic DNA as template by extending the initial denaturation temperature of the DNA (in order to ensure complete melting of the chromosomal DNA) and choosing highly specific primers with high annealing temperatures.

Initial denaturation of DNA:	96°C	4 min	
Exponential amplification:			
Denaturation of DNA:	96°C	20 s	25 cycles
Annealing of primers:	60-65°C	15 s	
Primer elongation by Taq:	72°C	30-120 s	
Final DNA elongation:	72°C	5 min	

2.4.5.4 Extension primer tag selection preceding solid-phase ligation-mediated PCR (EPTS/LM PCR)

A modified version of the previously published EPTS/LM-PCR protocol (Schmidt *et al.*, 2001) was applied to isolate and characterise L1 integration events from a large background of genomic DNA. 1-2 µg genomic DNA were digested with 6 U of restriction enzyme *Msc*I (New England Biolabs) overnight at 37°C and repurified by standard ethanol precipitation. As negative control, genomic DNA from mock-transfected HeLa cells was also digested. DNA from mock-transfected HeLa cells mixed with 1 ng of pSV2neo (Clontech, Heidelberg) served as a positive control for amplification of a sequence flanking a *neo*-gene. For primer extension (95°C for 5 min, 65°C for 30 min, 72°C for 30 min), restriction-digested DNA was added to the reaction mixture:

2.5 U *PfuTurbo* DNA polymerase (Stratagene),
5 µl 10x Cloned *Pfu* DNA polymerase reaction buffer,
250 µM dNTPs (Qiagen),
0.25 pmol of 5‘-biotinylated neo-specific primer GS177 (MWG Biotech, Ebersberg)
deionised water ad 50 µl

The reaction mixture was purified by the QIAquick PCR Purification Kit (Qiagen) to remove excess biotinylated primer. The extension product was eluted in 40 µl H_2O, mixed with 200 µg streptavidin-coated magnetic beads (Dynabeads M-280 Streptavidin, Dynal Biotech, Hamburg) diluted in 40 µl 2x BW buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl) and incubated overnight at room temperature. The captured DNA was exposed to a magnetic particle concentrator. The supernatant was discarded, and the captured DNA was washed twice with 100 µl H_2O and then resuspended in 10 µl ligation mixture (6 U T4 DNA ligase (New England Biolabs), 1 µl 10 x T4 DNA ligase buffer, 100 pmol annealed linker cassette (OC) and deionised water ad 10 µl). Ligation took place overnight at 16°C. The magnetic beads were washed twice with 100 µl H_2O and resuspended in 10 µl deionised water.

The first exponential PCR contained

1 µl suspended magnetic beads,
3 U “Expand Long Range Polymerase Mix” (Roche)
5 µl 10 x Expand Long Template PCR buffer 1,
200 µM dNTPs (Qiagen),
25 pmol primer GS94,
25 pmol linker-specific primer OCI
deionised water ad 50 µl

Initial denaturation of DNA:	96°C	5 min	
Exponential amplification:			30 cycles
Denaturation of DNA:	94°C	30 s	
Annealing of primers:	62°C	30 s	
Primer elongation by “Expand Long Range Polymerase Mix” (Roche)	68°C	150 s	
Final DNA elongation:	68°C	10 min	

If necessary, nested PCR was performed under identical cycling conditions using 25 pmol primer GS90 and 25 pmol linker-specific primer OCII on 1 µl of the first PCR reaction. PCR products were separated in a 0.8% agarose gel, isolated from the gel using the QIAquick Gel Extraction Kit (Qiagen), concentrated by standard ethanol precipitation and sequenced using ABI PRISM BigDye Terminators (Applied Biosystems, Darmstadt).

2.4.6 Construction of plasmids

2.4.6.1 Cloning procedure

Restriction digestion of plasmids was done according to the enzyme supplier’s instructions. When PCR products were cloned, care was taken that there were at least 6 bp between the restriction sites used for cloning and the PCR product’s end to ensure efficient enzymatic cleavage. Vectors were usually dephosphorylated and controlled for their religation potential by a mock ligation without insert.

2.4.6.2 Cloning strategies

The original retrotransposition reporter plasmid pJM101/L1.3 (Fig. 8A) (Moran *et al.*, 1996), the negative control reporter construct pJM105/L1.3 (Moran *et al.*, 1999), the CMV promoter deletion mutant pJM101L1.2ΔCMV (Moran *et al.*, 1996) and the plasmid rescue vector pCEP4/L1.3/ColE1/$mneoI_{400}$ (Gilbert *et al.*, 2002) were gifts from John Moran (University of Michigan, Ann Arbor). Plasmid pE1EN (Christensen *et al.*, 2000), containing the Tx1L-endonuclease coding region, was kindly supplied by Dana Carroll (University of Utah, Salt Lake City, USA). The R1Bm EN gene fragments were amplified from pGS405 (Feng *et al.*, 1998). Expression plasmids pGal4-TRD$_{\text{MeCP2}}$ and pMeCP2 (Yu *et al.*, 2001) were obtained from Prof. W. Strätling (Universitätsklinikum Hamburg-Eppendorf). Nucleotide sequences of the oligonucleotides used in this study are given in Appendix A.

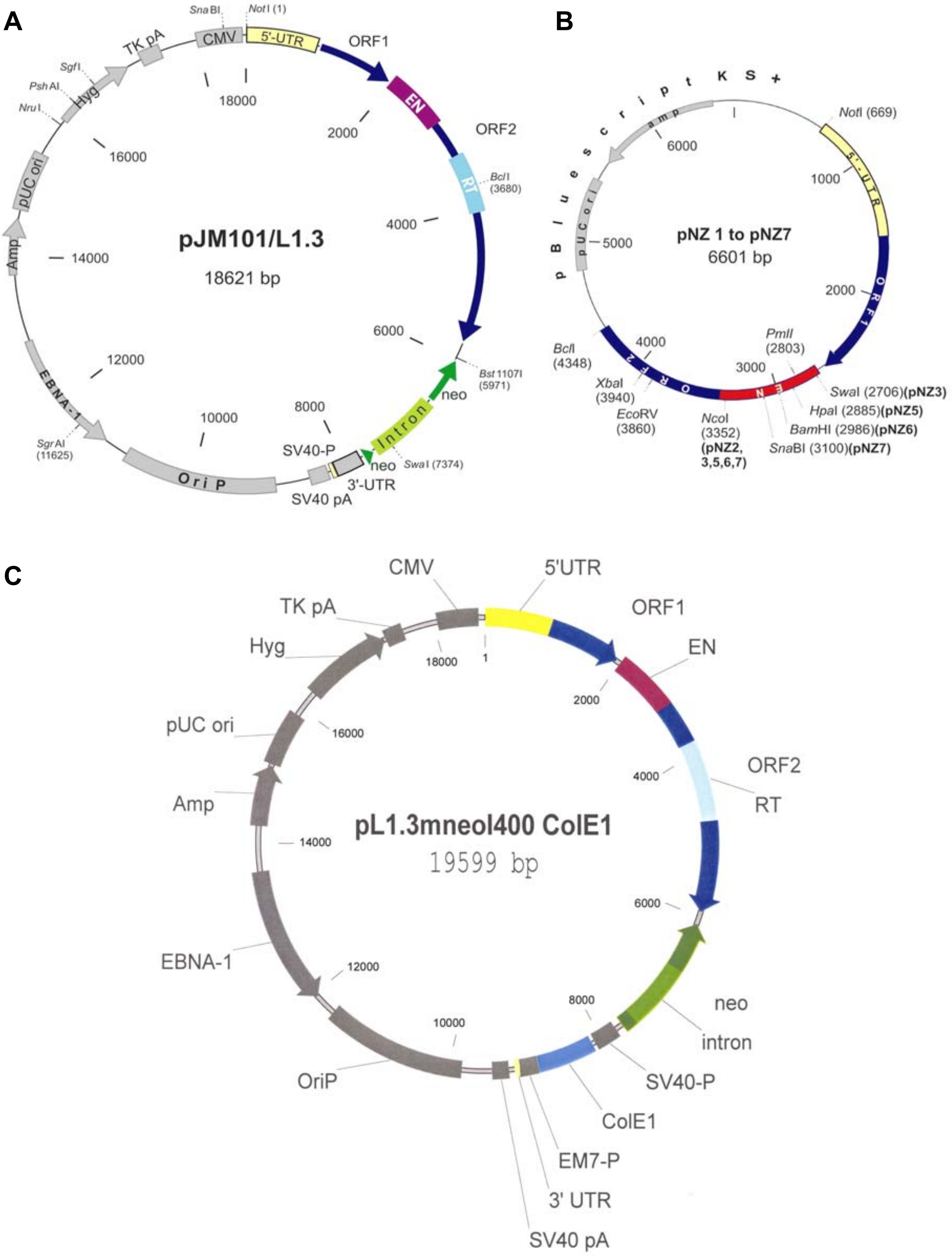
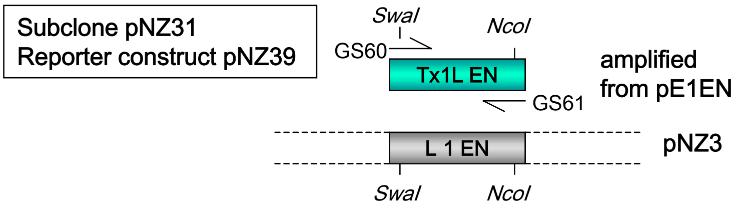


Fig. 8: Retrotransposition reporter construct pJM101/L1.3 (A), subclones pNZ1/2/3/5/6/7 (B) and plasmid rescue construct pL1.3*mneoI*$_{400}$ColE1 (C). Vector-derived sequences are represented in grey, elements of the retrotransposition cassette are colour-coded: yellow, untranslated regions (UTRs); dark blue, open reading frames (ORFs); red, endonuclease domain (EN); light blue, reverse transcriptase domain (RT); dark green, neomycin phosphotransferase gene (*neo*); light green, γ globin intron 2 (Intron). Unique cutters used for subcloning and EN manipulation (*Not*I, *Bcl*I, *Xba*I, *Eco*RV, *Pml*I) are indicated. If a restriction site is only present in a certain subclone, the respective name of this subclone is indicated in brackets.

EN block swaps (pNZ1-7 and pNZ33-39)

In order to enable easy modification of the L1 EN domain, the 3.7 kb *Not*I/*Bcl*I-fragment of pJM101/L1.3 (Fig. 8A) was subcloned into pBluescript KS+ to create pNZ1. A unique *Nco*I site was introduced at the 3‘ end of the EN coding region at position 3352 (position 2683 in L1.3) by site directed mutagenesis, resulting in the concomitant conversion of serine 231 into a glycine. This was considered an acceptable change as the amino acid corresponding to position 231 in L1 EN is not conserved in the family of AP endonucleases. The resulting plasmid pNZ2 was further modified by the introduction of a second unique restriction site at the positions where block swaps should take place. Thus, plasmids pNZ3 (*Swa*I site at position 2706), pNZ5 (*Hpa*I site at position 2885), pNZ6 (*Bam*HI site at position 2986) and pNZ7 (*SnaB*I site at position 3100) were created (Fig. 8B).

To prepare the chimeric L1/Tx1L element, the *Nco*I/*Swa*I-fragment of pNZ3 was replaced with the *Nco*I/*Swa*I digested Tx1L EN fragment amplified from plasmid pE1EN (Christensen *et al.*, 2000) with the primer pair GS60 and GS61. The resulting pBluescript based subclone was named pNZ31. Reintroduction of the *Not*I/*Bcl*I-fragment into pJM101/L1.3 yielded plasmid pNZ39.

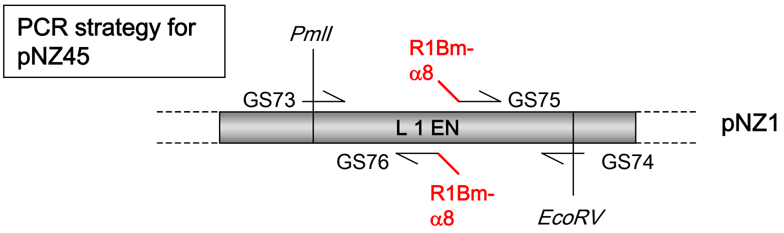


A similar strategy was used to create the reporter construct carrying the chimeric L1/R1Bm element: a *Nco*I/*Sma*I digested R1Bm EN fragment amplified from pGS405 (Feng *et al.*, 1998) with primers GS22 and GS23 replaced the *Nco*I/*Swa*I-fragment of pNZ3. After introduction of the modified *Not*I/*Bcl*I-fragment into pJM101/L1.3, the resulting pJM101/L1.3 derivative was named pNZ33.

Swapping truncated R1Bm EN-fragments (Fig. 16, p.66) was performed by the same principle. The fragments flanked by *Nco*I and the newly introduced restriction sites of pNZ5, 6 and 7 were replaced with the corresponding regions from pGS405 amplified with the primer pairs GS22/GS30, GS22/GS33 and GS22/GS36, respectively. After reintroduction into pJM101/L1.3, these swaps were named pNZ35, pNZ36 and pNZ37.

R1Bm α5 and α8 swaps (pNZ44-47)

The 24 nucleotides coding for the α5-region and the 27 nucleotides coding for the α8-region of R1Bm EN were directly introduced into pNZ1 using SOE-PCR (see 2.4.5.1). The α5-swap was generated using primers GS73, GS75, GS76 and GS74, the α8- swap was amplified with primers GS73, GS77, GS78 and GS74. Reintroduction into pJM101/L1.3 via the unique cutters *Pml*I and *Eco*RV yielded the plasmids pNZ44 (α5 swap), pNZ45 (α8 swap) and pNZ47 (α5/α8 double swap) (Fig. 19, p.71).



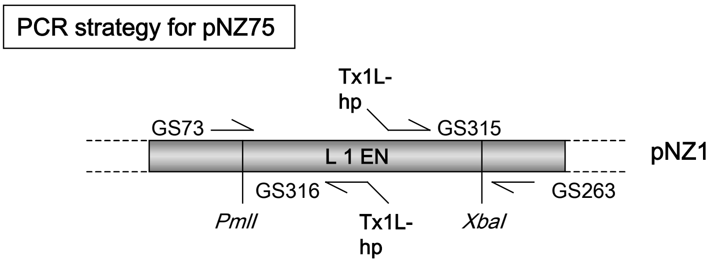
pNZ49-66

In order to exclude the possibility of *trans* complementation of the L1/Tx1L-chimera pNZ39 by endogenous elements (see 3.3.2.2, p.67), the control plasmid pNZ49 was prepared by introducing the *Not*I/*Bcl*I-fragment of pNZ39 into pJM105/L1.3. Thus, a LINE element containing Tx1L EN and a point mutation in the RT coding region was created.

As a negative control for Tx1L EN enzymatic activity, pNZ51 was generated by site directed mutagenesis of Tx1L EN in pNZ39. This was achieved by SOE-PCR (see 2.4.5.1) using primers GS263 and GS285 as outer primers. The primer pair GS265/GS266 was used to introduce the point mutation H230A. Since pNZ51 showed the same activity as pNZ39, plasmids pNZ63-pNZ66 were created to obtain different single, double and triple mutants of Tx1L EN (Fig. 17, p.68). The double mutant D143A/N145A (pNZ63) was generated using primers GS286 and GS287 with pNZ31 as template. The primer pair GS288/GS289 was used to introduce the point mutation D143A (pNZ64, pNZ65), and the mutation D205A was generated with the primer pair GS290/GS291 (pNZ66). Double and triple mutants were obtained by applying the SOE strategy on plasmids that already harboured a mutation. The PCR products were subcloned into pNZ31 via *Xba*I and *Mun*I, and the resulting mutated ENs were reintroduced into pJM101/L1.3.

Hairpin loop swaps (pNZ73-85)

Finally, the hairpin mutants pNZ73, pNZ75 and pNZ76 (Fig. 20, p.74) were generated using SOE-PCR with pNZ1 as template. The inner primers introducing the mutation were named GS311 and GS312 for the hairpin deletion mutant pNZ73. In this mutant the L1 hairpin loop Fig. 14, p.64) is replaced by two glycines. pNZ75 was generated using GS315 and GS316 (see figure below), carrying the Tx1L hairpin loop (Tx1L-hp) in place of the L1 hairpin structure.



GS317 and GS318 were used for the generation of pNZ76, the corresponding R1Bm hairpin loop swap. In all three cases, GS73 and GS263 were used as outer primers. They include the restriction sites *Pml*I and *Xba*I*,* which were used to clone the mutated SOE-PCR products back into pNZ1, resulting in the subclones pNZ69, 71 and 72. From these subclones, the hairpin mutants were introduced into the plasmid rescue vector pCEP4/L1.3/ColE1/*mneoI*$_{400}$ (Fig. 8, p.44)via the restriction sites *Not*I and *Bcl*I.

In order to generate a negative control construct for the retrotransposition reporter assay, the RT coding region of pCEP4/L1.3/ColE1/*mneoI*$_{400}$ was replaced by the *Bcl*I/*Bst*Z17I-fragment of pJM105 carrying the inactive D702A point mutant of RT. In analogy to pNZ63, an enzymatically inactive EN control mutant was generated for each hairpin swap. SOE-PCR with the primer pair GS323 and 324 introduced the two mutations D145A/N147A. GS73 and GS263 were used as flanking primers, and the resulting PCR product was transferred into the subclones pNZ69, 71 and 72 via *Pml*I and *Xba*I. Transfer of the *Not*I/*Bcl*I fragments into pCEP4/L1.3/ColE1/*mneoI*$_{400}$ yielded pNZ82, 84 and 85 (Fig. 20, p.74).

pG5JM101/L1.2ΔCMV

For L1 promoter studies, five Gal4 recognition motifs (G5) were isolated from pFR-Luc (Stratagene, Heidelberg) by restriction with *Bam*HI. The fragment was blunted with Klenow polymerase and inserted into the blunted unique *Not*I site of pJM101/L1.2ΔCMV (Moran *et al.*, 1996) to yield pG5JM101/L1.2ΔCMV (Yu *et al.*, 2001).

2.4.7 Preparation of genomic DNA from eukaryotic cells

Genomic DNA was isolated from approx. 10^7 cells by using DNazol® Genomic DNA Isolation Reagent (Molecular Research Center Inc., Cincinnati, USA) according to producer's instructions. The preparations were incubated with RNase A at 37°C for at least 3 hours in order to degrade co-isolated RNA. Subsequently, the DNA was tested for integrity by agarose gel electrophoresis.

Alternatively, DNeasy columns (Qiagen) were used to isolate genomic DNA from approx. 2 x 10^6 HeLa cells according to the manufacturer's instructions.

2.4.8 Southern blot analysis

2.4.8.1 Restriction digestion and electrophoresis of genomic DNA

15-20 µg genomic DNA were digested with an excess of a suitable restriction enzyme and separated on a 0.7% agarose gel overnight (field strength 1.3 V/cm).

2.4.8.2 Transfer of DNA onto nylon membranes

In order to facilitate blotting of large DNA fragments, the DNA was depurinised by treating the gel with 0.25 M HCl for 20 min. The DNA was then denatured in 0.2 N NaOH, 0.6 M NaCl for 30 min, and finally neutralised for 30 min in 0.24 M Tris-Cl, 0.6 M NaCl, pH 7.5.

Transfer of the DNA onto a nylon membrane (Biodyne B, 45µm, Pall, Portsmouth, U.K.) was accomplished by capillary blotting in 10x SSC. Cross-linking by UV-light (2x 120 mJ) in the UV-Stratalinker™ 1800 (Stratagene, Heidelberg)) fixed the DNA on the membrane irreversibly.

2.4.8.3 Radioactive labelling of DNA probes

Radioactive probes were labelled by random oligonucleotide-primed synthesis using the DecaLabel DNA Labelling Kit (MBI Fermentas, St. Leon-Rot). The reaction was carried out according to the manufacurer's protocol using 50 µCi (≅1,85 MBq) α-[^{32}P]-dCTP radionucleotide (specific radioactivity: 3000 Ci/mmol). Free nucleotides and primers were removed with MobiSpin S-300 columns (MoBiTec, Göttingen) following manufacturer's instructions.

2.4.8.4 Hybridisation of blotted DNA

After blotting, the membrane was pre-hybridised with salmon sperm DNA to saturate unspecific binding sites. For this purpose, the membrane was incubated for at least 1 h at 55°C with 10 ml preheated hybridisation buffer (50 % deionised formamide, 6x SSC, 0.5% SDS, 2.5x Denhardt-solution, 8% dextran sulfate) containing 100 µg/ml denatured salmon sperm DNA.

The radioactively labelled DNA probe was then denatured by incubation at 95 °C for 5 min and added to the pre-hybridisation buffer at a concentration of at least 1.5x10^6 cpm/ml. The membrane was incubated with the probe for 10-24 h at 55 °C with constant agitation. Subsequently, the membrane was washed twice with 2x SSC for 5 min at room temperature and twice with 0.1% SDS, 0.1x SSC for 1 h at 65°C. In order to allow a possible re-hybridisation, the membrane was prevented from drying completely. It was sealed in a plastic bag and exposed to a ^{Super}RX Medical X-Ray film (Fuji Photo Film, Düsseldorf) or a phosphoimager plate (Fuji Photo Film). Exposure times were typically around 4 to 24 h. The films were developed using a CP 1000 film processor (Agfa-Gevaert, Köln), the phosphoimager plates were scanned with a “BIO-Imaging analyser Fujix BAS 2000” (Fuji Photo Film) using the software TINA2.0 (Raytest, Straubenhardt).

2.4.8.5 Re-hybridisation of DNA

400 ml 0.1% SDS were heated to boiling temperature in a microwave. The membrane was washed briefly with 200 ml of the boiling solution, then covered with the remaining liquid and incubated at room temperature to cool down to 30-40°C. After repeating this procedure, the membrane was sealed in a plastic bag and placed on a phosphoimager plate overnight to test for complete removal of the previous probe. Subsequently, the membrane was pre- and re-hybridised as described above.

2.4.9 L1 plasmid rescue from eukaryotic cells

De novo L1 integrants derived from pNZ75 were isolated by a rescue procedure adapted from Gilbert *et al.*, (2002) and Symer *et al.*, (2002) (Fig. 8, Fig. 23). Genomic DNA from G418^R HeLa cells was prepared using DNAeasy columns (Qiagen, Hilden) and restricted to completion using *Hin*dIII. Fragments were ligated under extremely dilute conditions (0.5-1 ng/µl) to favour intramolecular circularisation; typically, 300 ng restricted genomic DNA was incubated with 4 U T4 DNA ligase (New England Biolabs) in a volume of 500 µl ligation buffer at 16°C overnight. The ligation mixture containing added glycogen (Invitrogen, Karlsruhe) was then ethanol precipitated and resuspended in water; the entire concentrated ligation was used to transform electro-competent DH10B cells (Electromax DH10B, Invitrogen) by electroporation in 1 mm gap cuvettes (see 2.4.3). Transformed cells were selected on kanamycin plates. After plasmid isolation, the clones were characterised by restriction digestion, analytical PCR and sequence analysis.

2.4.10 DNA sequencing analysis

All sequencing reactions were performed by the dideoxy method (Sanger *et al.*, 1977) using the Big Dye kit (Applied Biosystems, Darmstadt). Subsequent separation and analysis were done on a model 377 DNA sequencer (Applied Biosystems).

2.5 Tissue Culture

2.5.1 Cultivation of HeLa cells

HeLa cells (ATCC number CCL-2) were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 2 mM glutamine.

2.5.2 Long-term storage of HeLa cells

Approximately 5 x 10^6 HeLa cells were trypsinised, centrifuged and resuspended in a mixture of 90% FCS and 10% DMSO. The suspension was slowly cooled to -80°C in an isopropanol bath (cooling rate 1°C/min) and then stored in liquid nitrogen.

2.5.3 Determination of cell number

Cells were trypsinised, and a 20 µl aliquot was mixed with 20 µl of a 0.36 % trypan blue solution. After 3 minutes of staining, only dead cells turn blue, while living cells remain unstained. The latter were counted in a Neubauer chamber.

2.5.4 L1 retrotransposition reporter assay

Initially, retrotransposition assays were performed as described previously by (Moran *et al.*, 1996) except that instead of lipofectamine 3 µl Fugene-6 transfection reagent (Roche, Mannheim) were used. Later we switched to a more efficient modified version of the rapid, quantitative transient L1 retrotransposition assay described by (Wei *et al.*, 2000). The results obtained from the initial assays were reproduced using the new, faster assay to allow direct comparison of the activity of all chimeric constructs.

2×10^5 HeLa cells were plated in each well of a six-well dish and grown to 50-80% confluence in DMEM. The following day, triplicate dishes were transfected using 6 µl Fugene-6 transfection reagent (Roche) and 2 µg of a Qiagen DNA preparation per well. At 24 h post-transfection, the transfection mixture was removed and replaced by DMEM. At 72 h post-transfection, the medium was replaced with DMEM containing 400 µg/ml G418 (DMEM-G418). After 10-14 days, G418^R colonies were stained with 0.4% Giemsa for visualisation. Alternatively, single clones were isolated by limiting dilution or trypsinisation in cloning rings. They were expanded in DMEM-G418 and genomic DNA was prepared for subsequent analysis.

2.6 Computational Methods

2.6.1 Homology searches

DNA homology searches were performed with the BLAST and/or the BLAT program (http://www.ncbi.nlm.nih.gov and http://genome.ucsc.edu) (Altschul *et al.*, 1990; Kent, 2002). Two sequence alignments were done with the BLAST 2 sequences, MegAlign 4.00 (DNASTAR Inc., Madison, USA), and DNA Strider (CEA, Gif-sur-Yvette, France) programmes and, when needed, further adjustments were made by hand. Multiple alignments of both DNA and protein sequences were constructed by using the MegAlign 4.00 programme (DNASTAR Inc.) and further refined by hand. In protein motif alignments, conserved amino acid residues were classified according to the following physico-chemical properties: weakly hydrophilic (N, Q, S, T), acidic (D, E), basic (H, K, R), small (A, G), hydrophobic (I, L, M, V), and aromatic (F, W, Y).

2.6.2 Sequence logos

Sequence logos of consensus sequences were generated following the instructions on the website http://ep.ebi.ac.uk/EP/SEQLOGO (Schneider and Stephens, 1990).

2.6.3 Identification of endogenous L1 sequences flanked by TSDs

In collaboration with U. Willhöft from the Zentrum für Bioinformatik (Universität Hamburg), endogenous human L1 elements were identified by applying the programme TSDfinder (Szak *et al.*, 2002) on non-redundant human sequence contigs (NT_* records) assembled at NCBI (http://www.ncbi.nlm.nih.gov). A DNA reference sequence of the human genome constituting approximately 99% of the euchromatic genome (build 33 as of 14 April 2003) served as data set, with the exception that the file 'unplaced_contigs' was excluded from the study. For identification of L1 elements and their respective 5' and 3' flanking sequences we used the method described in (Szak *et al.*, 2002) with minor modifications: L1 elements were annotated using the Repeat Masker programme (http://repeatmasker.genome.washington.edu/cgi-bin/RM2_peq.pl, Smit and Green, unpublished) with the limitation to report only sequences >90 % identity. The custom library contained only the L1.3 sequence (GenBank accession number L19088 with modifications as cited in Szak *et al.*, [2002]). After introduction of minor modifications, TSDfinder was used to generate an output file with identification numbers, coordinates and further information for all matching sequences of the RepeatMasker result.

A Perl programme was written to parse information from the TSDfinder output for each hit and join it with sequence information from the DNA reference sequence of the human genome. This Perl program produced an output file in *fasta* format reporting the TSDfinder identifications number, GenBank accession numbers and the position of the elements within the human sequence contigs. Additionally it provided the sequence of the L1 element as well as additional 150 nucleotides 5' and 3' flanking sequence information. Another Perl programme parsed the sequence information of the TSD as assigned by the programme TSDfinder and the respective identification number for each element in a tabular fashion.

2.6.4 Identification of microhomologies localised at the junctions between 5’ ends of L1 insertions and 3’ ends of their TSDs

In collaboration with H.-P. Brose and V. Schoder from the Institut für Medizinische Biometrie und Epidemiologie, (Universitätsklinikum Hamburg-Eppendorf) we tested for reliability of the TSDfinder results by searching for TSD sequences in the output files. Only post-integration sites with the correct TSD at the 3’ end of each 150-bp 5’ flanking sequence window of the output sequences were used for further analysis. Besides, TSDs consisting exclusively of adenines were discarded due to statistical uncertainties. The adenines could be either real TSDs or fortuitous matches between the target site and the poly(A) tail. Next, it was examined whether any untemplated nucleotides were inserted between the 5’ end of the L1 sequence and the TSD. Therefore, 30 bp 3’ of the TSD were aligned with the L1 consensus sequence. We discarded all integration events that displayed less than 83% sequence identity with the L1 sequence in the best match. Moreover, a perfect match of the first 3 nucleotides directly adjacent to the TSD was required. To exclude grossly rearranged

L1 copies, we introduced a length criterion by assuming that the sum of the insertion length and the truncation position should lie between 6000 and 7000 bp. After this preliminary selection process, microhomologies were searched for by comparing the 3' end of the TSD with the sequence that lies 5' of the L1 truncation position in the L1 consensus.

In order to obtain an adequate number of full-length integrants, the complete analysis was repeated with two different L1 consensus sequences, each containing a major polymorphism at the transcriptional start site. These two alternative sequences commence with GAGGG and GGAGG instead of GGGGG.

2.6.5 Statistical analyses

For statistical analysis of the junctions between target site duplications (TSDs) and 5' ends of the L1 sequences, the following assumptions were made. In the case of endogenous full-length L1 insertions, the 5' end is defined by a G_5-stretch. Thus, if ties occured only by chance, the probability to observe exactly j consecutive ties is p^j x (1-p), where j = 0,1,2,… and p denotes the proportion of G in the target sequence. This means that the random variable X being defined as the number of ties until the first non-tie follows a geometric distribution with probability 1-p. This assumption can be made since the base at position 6 starting from the 5' end of full-length L1 is the first non-G and the probability $P(X>5)$ is almost 0.

Nevertheless, the assumption of the geometric distribution does hold true only if the five consecutive nucleotides at the 5' end remain Gs. Therefore, any integrants including insertions or deletions affecting the first three nucleotides were not included in computations.

In the case of truncated L1 insertions, statistical assumptions were made according to Roth *et al.*, 1985. Basically, the probability to observe a sequence of j homologies is computed as $P(X = j)$ = (j+1) x p^j x $(1\text{-}p)^2$ where p denotes the probability of one random homology and can be set to 0.25 assuming unbiased base composition of the target sequences. p was also estimated after calculation of the actual base composition of the DNA sequences flanking each TSD in a 20 nucleotide window.

In both cases we tested whether the data observed could originate from the distribution specified under the assumption of random events. For that purpose, we performed a Kolmogoroff-Smirnow-Test computing p-values and 99% confidence limits in a Monte Carlo Simulation. This simulation consisted of 100,000 independent draws from the hypothesised distribution (using estimated probabilities in the case of the truncated insertions) and for each draw the maximum absolute distance of the observed and the theoretical cumulative distribution function was calculated. The proportion of draws which exceeded the analogous distance observed in our data is reported as an unbiased estimator of the true p-value. Finally, 99% confidence limits for the p-value were computed using standard statistical techniques.

Simulations were performed with the software program S-Plus 4.5 (MathSoft Inc., Cambridge, USA).

3. RESULTS

3.1 Determining L1 Retrotransposition Frequencies

The present study investigates several aspects of the L1 retrotransposition mechanism in order to evaluate its potential as a gene delivery tool. For this purpose, it was crucial to employ an assay system that allows controlled manipulation of an L1 element as well as tracking the fate of its *de novo*-integrated copies. Therefore, all experiments described here are founded upon a cell-culture based genetic assay that permits both determination of the retrotransposition rate and elucidation of structure and sequence of integrated L1 copies.

In this assay, drug resistance is conferred to an L1-transfected cell only after retrotransposition took place within that cell (Moran *et al.*, 1996). L1 is tagged with an indicator gene by introducing the reporter cassette *mneoI* (Freeman *et al.*, 1994) into its 3' UTR (Fig. 8, p.44, Fig. 9). *mneoI* consists of a selectable marker gene (*neo*) in the antisense orientation which is flanked by an SV40 promoter (P') and an SV40 polyadenylation signal (A'). The *neo* gene is disrupted by an intron (IVS 2 of the γ-globin gene) in the opposite transcriptional orientation, i.e. the sense orientation relative to L1. mRNA transcripts originating from the CMV promoter driving L1 expression (P_{CMV}) are spliced, but contain an antisense copy of the *neo*-gene. Transcripts initiated from P' cannot be spliced and thus do not yield a functional neomycin phosphotransferase. $G418^R$ colonies only arise when a transcript originating from the CMV promoter is spliced, reverse transcribed and reintegrated into chromosomal DNA. Then the intact *neo* gene can be expressed from P' and renders the host cell resistant to G418 (Fig. 9).

The tagged L1 element is subcloned into the pCEP4 expression vector, a plasmid particularly suited for the purpose of the assay because it replicates as an extrachromosomal nuclear episome at moderate copy numbers in primate cells (Yates *et al.*, 1985). Moreover, it contains a hygromycin gene (*hyg*) for the selection of transfected cells and places the L1 element under the control of the well characterised cytomegalovirus (CMV) immediate early promoter (Boshart *et al.*, 1985) (Fig. 8).

In my initial experiments, the original assay (Moran *et al.*, 1996) was applied (see 2.5.4, p.50). However, in this protocol the cells are expanded several times in medium containing hygromycin between transfection and G418 selection in order to select for transfected cells.

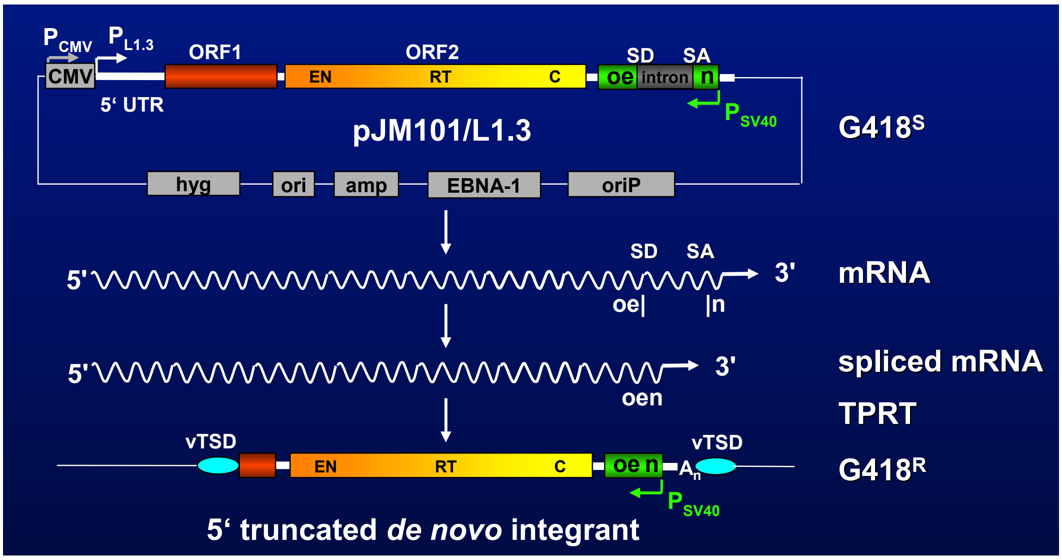


Fig. 9: Schematic representation of the retrotransposition reporter assay. L1.3 was cloned into pCEP4 to create pJM101/L1.3. pCEP4 contains an origin of replication (ori) and a selectable marker (amp) for prokaryotic cells as well as an origin of replication (oriP/EBNA-1) and a selectable marker (hyg) for eukaryotic cells. Thin white lines represent plasmid or genomic DNA, while broad white lines designate L1 UTRs. ORF1 is shown in red, ORF2 in orange and the *mneoI* cassette in green. P, promoter; vTSD, variable target site duplication. L1.3 was tagged with an indicator gene containing an antisense copy of the *neo*-gene disrupted by an intron in the sense orientation. The splice donor (SD) and splice acceptor (SA) of the intron are indicated. G418-resistant (G418^R) colonies arise only when the L1 transcript is spliced and integrated into chromosomal DNA by target primed reverse transcription (TPRT).

Since during this selection process retrotransposition can already take place, cells containing an early retrotransposition event proliferate and generate numerous G418^R colonies, while a late integration event yields only one single colony. Both the generation of multiple identical clones and the bias for early retrotransposition events hamper the subsequent analysis of integration specificity.

Therefore, all constructs were (re-)tested in a modified version of the rapid, transient retrotransposition assay described in Wei *et al.* (2000) (see 2.5.4). Highly reproducible and efficient transfection is a prerequisite for this faster assay, as there is no selection for cells bearing the episomal reporter plasmid. Cells are selected on G418 immediately after transfection. This makes the assay faster (2 weeks versus 5 weeks) and virtually ensures that each of the G418^R foci observed arise from an independent L1 retrotransposition event. The new assay is an excellent, reliable tool for rapidly comparing retrotransposition efficiencies of different reporter constructs.

In order to determine whether the G418^R foci obtained in the retrotransposition reporter assays resulted from independent retrotransposition events, single cells were isolated by the technique of limiting dilution, generating independent clonal cell lines. *De novo* L1 integration events were characterised by Southern blot analysis. As pre-existing L1 sequences are abundant in the human genome, hybridisation was carried out with a *neo*-specific probe, taking advantage of the unique sequence of the marker cassette. In clones obtained from the hygromycin-based original assay, one to five differently sized bands were detected in each lane, with fragment sizes ranging from 2.5kb to >10kb (Fig. 10A). Rehybridisation of the membrane with a probe detecting a single copy gene (*gmcsf*) yielded only one band per lane (data not shown), proving that the genomic DNA had been digested to completion. Therefore it could be concluded that the multiple bands observed with the *neo*-probe result from different *de novo* insertions of *neo*-tagged L1 elements. The prominent band of about 9 kb seen in 15 out of 22 clones migrates at the same height as the *Eco*RI fragment of pJM101/L1.3 containing the *neo*-cassette (10126 bp). This strongly suggests that in many clonal cell lines, integration of the complete reporter plasmid occurred. However, for the selected cells to become G418 resistant, each clone has to contain at least one retrotransposition event, i.e. one spliced *neo* copy.

Similar analysis of clones derived from the rapid transient retrotransposition assay revealed less integration events per *neo*^{R}-clone. In 76% of the clones, a single band was detected in Southern blot analysis, while the remainder displayed two or three bands (Fig. 10B).

In order to test for the loss of the γ–globin intron in the newly integrated L1 hybrid elements, PCR analysis of the genomic DNAs was performed (Fig. 10C). *Neo*-specific primers were used to amplify the *neo*-gene with or without intron, depending on its splice status. In 13 out of 22 characterised clonal cell lines derived from the original, hygromycin-based assay, only the 793-bp fragment was amplified. This indicated that the one to three *neo*-tagged integrations detected in each of these clones by Southern analysis are the result of different *de novo* retrotransposition events. Additional amplification of the 1694-bp product from nine clonal DNAs indicated that, in addition to retrotransposition, one or more genomic L1/Tx1L copies derived from integration (recombination) of the reporter plasmid into the genome. In clonal cell lines originating from the rapid transient retrotransposition assay, PCR amplification of the *neo*-cassette reproducibly yielded only the 793-bp fragment deriving from the spliced *neo*-gene (Fig. 10C).

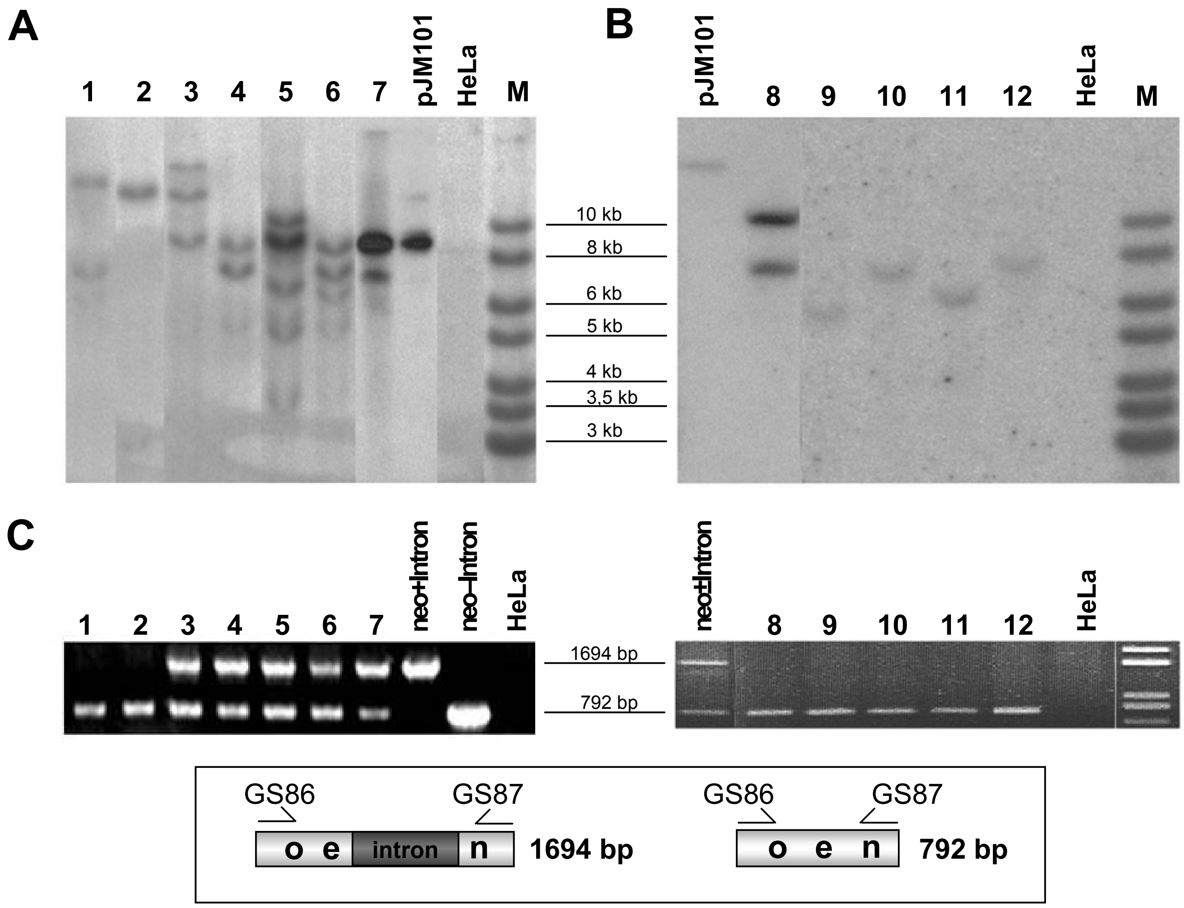


Fig. 10: Characterisation of *de novo* integration events derived from the original retrotransposition reporter assay and from the rapid transient assay. (A) Southern blot analysis of genomic DNA from G418^R clonal HeLa cells derived from the retrotransposition assay including selection for the episomal reporter plasmid pNZ39 (Fig. 17) with hygromycin. Genomic DNA from seven representative samples (lanes 1-7) was digested with *Eco*RI. A radiolabelled 700-bp *neo*-gene served as probe. Genomic DNA from mock-transfected HeLa cells was loaded as negative control ('HeLa') and plasmid DNA containing the *neo*-gene (pJM101/L1.3) was used as positive control ('pJM101'). The prominent band migrating at ~9 kb corresponds to the 10.1 kb *Eco*RI fragment of pJM101/L1.3 bearing the *neo*-gene. M, size marker. **(B)** Southern blot analysis of genomic DNA from G418^R clonal HeLa cells transfected with pNZ45 (Fig. 19). The hygromycin selection step was eliminated. Genomic DNA from five representative samples (lanes 8-12) was digested with *Bgl*II. A radio-labelled 700-bp *neo*-gene served as probe. Genomic DNA from mock-transfected HeLa cells was loaded as negative control ('HeLa') and plasmid DNA containing the *neo*-gene (pJM101/L1.3) was used as positive control ('pJM101'). M, size marker. **(C)** PCR with *neo*-specific primers GS86 and GS87 revealed a 792-bp DNA fragment diagnostic for loss of intron in each DNA preparation (lanes 1-12). In 68% of the clonal cell lines obtained from the hygromycin-based assay, an additional 1694-bp PCR product was detected. This unspliced *neo*-gene was never detected in clones derived from the transient assay. pJM101/L1.3 and bML3 served as positive controls for the presence of an unspliced and a spliced *neo*-gene ('neo+intron' and 'neo-intron').

These results show that multiple retrotransposition events in one cell are possible, and that they become more frequent if the episomal plasmid is given more time to "launch" new retrotransposition events by the hygromycin selection step. However, with prolonged duration of the retrotransposition assay, the risk of recombinational integration of the episomal reporter plasmid into the genome increases.

3.2 Identification of Methyl-CpG Binding Protein 2 as Major Regulator of Human L1 Retrotransposition

In order to use L1 as vector for gene therapy, it is indispensable to understand the regulation of its expression. In collaboration with Prof. W. Strätling's laboratory at the Universitätsklinikum Hamburg-Eppendorf, we explored the possible mechanism by which L1 elements are transcriptionally silenced in the genome (Yu, F., Zingler, N., Schumann, G. and Strätling, W.H., 2001). While three positive regulators of L1 transcription have already been identified (SOX11, RUNX3 and YY1, see 1.5.1.1, p.25), negative regulation was assumed to be conveyed by the strong methylation observed in the promoter region of L1 in somatic cells (Thayer *et al.*, 1993; Woodcock *et al.*, 1997; Yoder *et al.*, 1997). Repression by DNA methylation is thought to be established through binding of members of the methyl-CpG-binding domain (MBD) protein family, recruitment of histone deacetylases and local condensation of chromatin. This leads to the generation of a transcriptionally inactive chromatin structure that blocks binding of the Pol II transcription complex to the promoter region (Hendrich and Bird, 1998; Bird and Wolffe, 1999).

In transient transfection assays, our collaborators demonstrated that the transcriptional-repression domains (TRDs) of two MBD proteins, methyl-CpG-binding protein 2 (MeCP2) and MBD1, efficiently repress transcription from L1 promoter-driven luciferase constructs when targeted to the transcriptional start site via a linked Gal4 DNA-binding domain. In contrast, the TRD of another member of the MBD protein family, MBD2, had no significant effect on transcription (Fig. 11A) (Yu *et al.*, 2001).

In the experiments described above, the repressor domains were targeted to the promoter via the artificial Gal4 system. To test whether repression of L1 can also be achieved by binding of the complete MBD to methylated CpGs, subsequent experiments assessed the effects of co-expressed full length MBD proteins on transcription of an L1 promoter-driven luciferase construct in response to its methylation by *Hpa*II methylase. Interestingly, the transcription rate of the methylated promoter was strongly reduced (77%) only when MeCP2 was over-expressed. Full-length MBD1 and MBD2 did not influence the luciferase expression level significantly (Fig. 11B) (Yu *et al.*, 2001).

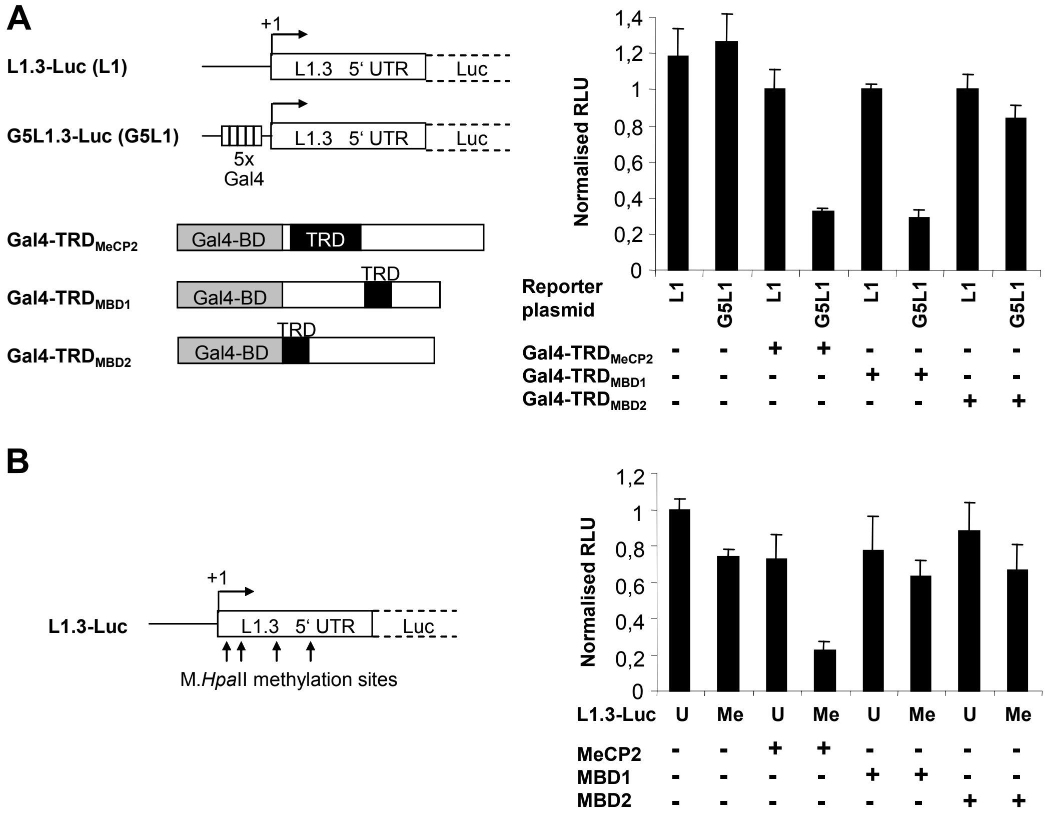


Fig. 11: The effect of methyl-CpG-binding proteins on L1-promoter-driven luciferase expression. (A) The transcriptional-repression domains (TRDs) of MeCP2 and MBD1 repress transcription controlled by an L1 promoter. Schematic maps of the reporters L1.3-Luc and G5L1.3-Luc and the expression constructs Gal4-TRD$_{MeCP2}$, Gal4-TRD$_{MBD1}$ and Gal4-TRD$_{MBD2}$ (Yu *et al.*, 2001) are shown. Grey bars represent the Gal4-DNA-binding domain, white bars MBD protein portions with the respective TRDs indicated by black bars. HEK293 cells were co-transfected with reporter constructs L1.3-Luc or G5L1.3-Luc and expression constructs Gal4-TRD$_{MeCP2}$, Gal4-TRD$_{MBD1}$ or Gal4-TRD$_{MBD2}$. Luciferase activities of L1.3-Luc co-transfected with the Gal4-TRD constructs (0.1µg each) were set as 1. **(B)** MeCP2 represses transcription from a methylated L1 promoter. Schematic representation of the sites (nucleotides 36, 101, 304 and 481) in the 5' UTR of L1.3 methylated by M.*Hpa*II methylase. HEK293 cells were co-transfected with unmethylated (U) and *Hpa*II-methylated (Me) reporter L1.3-Luc and expression constructs encoding full-length MeCP2, MBD1v3 or MBD2b (0.1µg each). Luciferase activity of the unmethylated reporter in the absence of co-expressed genes was set as 1. Black columns in (A) and (B) represent mean relative luciferase activities ± standard deviations of three to five independent experiments.

Based on these results, I tested whether the observed transcriptional regulation of the L1 promoter by MeCP2 is also able to affect L1 retrotransposition. For this purpose, the L1 reporter assay (see 3.1) was employed. In the standard assay, the L1 reporter is under control of a CMV promoter in addition to the internal L1 promoter. For the promoter studies described here, the external promoter (CMV) was deleted (pJM101/L1.2ΔCMV [Moran *et al.*, 1996]). While retrotransposition of pJM101/L1.2ΔCMV was not affected by a co-expressed Gal4-TRD$_{MeCP2}$ fusion protein, targeting the TRD of MeCP2 to the reporter through insertion of Gal4 DNA-binding sites (pG5JM101/L1.2ΔCMV) led to a drastic reduction of retrotransposition by 82% (Fig. 12).

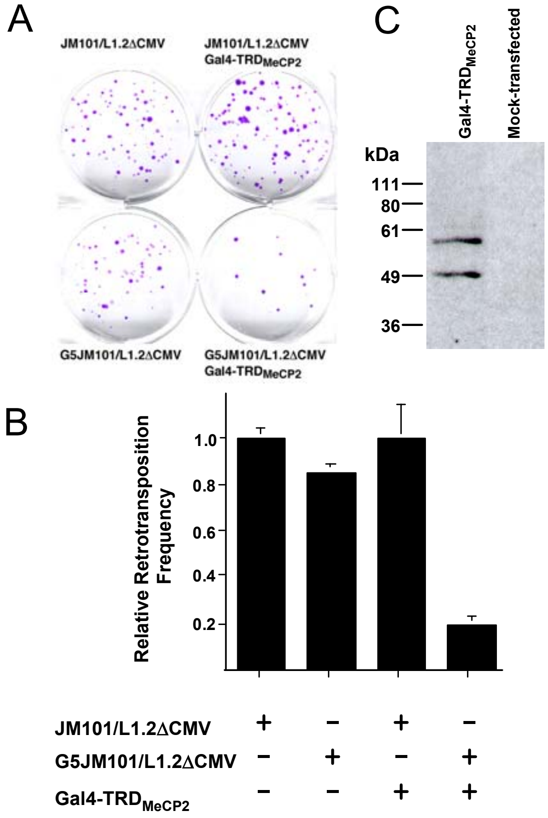


Fig. 12: The TRD of MeCP2 represses L1 retrotransposition. (A) Reporter construct pJM101/L1.2ΔCMV or pG5JM101/L1.2ΔCMV was co-transfected with the empty vector pcDNA1.1 or with expression construct Gal4-TRD$_{MeCP2}$ (Yu *et al.*, 2001) into HeLa cells. Results of a representative transposition assay are shown. **(B)** Effect of Gal4-TRD$_{MeCP2}$ on retrotransposition frequencies (n=6). **(C)** Expression of Gal4-TRD$_{MeCP2}$ 72 h post-transfection was controlled for by immunoblot analysis with anti-Gal4BD antibody. The upper 57 kDa band represents the full-length Gal4-TRD$_{MeCP2}$. The lower band likely results from cellular protease activity, since the C-terminal half of MeCP2 is sensitive to proteolysis (Lewis *et al.*, 1992).

Subsequently, the effects of over-expression of full-length MeCP2 on retrotransposition of methylated L1 reporter constructs were tested. *Hpa*II methylation of the reporter pJM101/L1.2ΔCMV reduced its ability to retrotranspose by 58% (Fig. 13). This is likely due to binding of endogenous methyl-CpG-binding proteins including MeCP2. Overexpression of FLAG-tagged MeCP2 (Fig. 13C) resulted in a further, although weak reduction of the retrotransposition frequency (30% relative to unmethylated pJM101/L1.2ΔCMV). Methylation-induced repression in the presence or absence of over-expressed MeCP2 did not differ significantly, probably due to binding of endogenous MeCP2. Furthermore, overexpression of MeCP2 slightly reduced transposition from the unmethylated reporter construct, either due to a weak affinity of MeCP2 to the unmethylated template (Weitzel *et al.*, 1997) or because over-expression of MeCP2 downregulates other factors needed for retrotransposition.

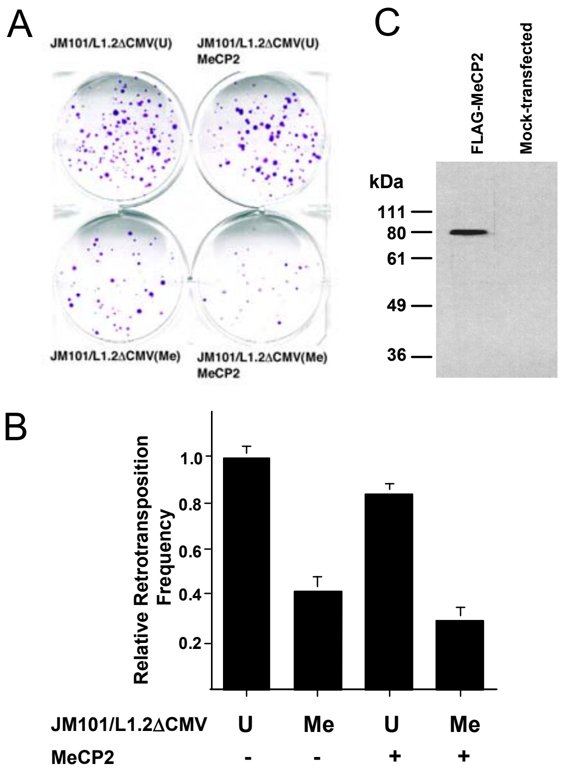


Fig. 13: Overexpressed full-length MeCP2 diminishes L1 retrotransposition frequency from a methylated reporter construct. (A) Unmethylated (U) or *Hpa*II-methylated (Me) reporter pJM101/L1.2ΔCMV was co-transfected with empty vector pcDNA1.1 or with the expression construct for FLAG-tagged MeCP2 (pFLAG-MeCP2, Yu *et al*, 2001) and subjected to G418 selection. **(B)** Effect of MeCP2 on relative retrotransposition frequencies of the methylated versus unmethylated L1.2 reporter (n=3). **(C)** Expression of FLAG-tagged MeCP2 (81 kDa) 72 h post-transfection was controlled for by immunoblot analysis with anti-FLAG antibody.

Summarising the results from the luciferase activity assays and the retrotransposition assays, the data support the conclusion that MeCP2 is recruited to the L1 promoter via methylated CpGs and can repress L1 retrotransposition. Since two other members of the MBD protein family, MBD1 and MBD2, failed to show a comparable repressive effect, MeCP2 seems to have a specific role in L1 regulation.

3.3 Altering the Target Site Specificity of L1

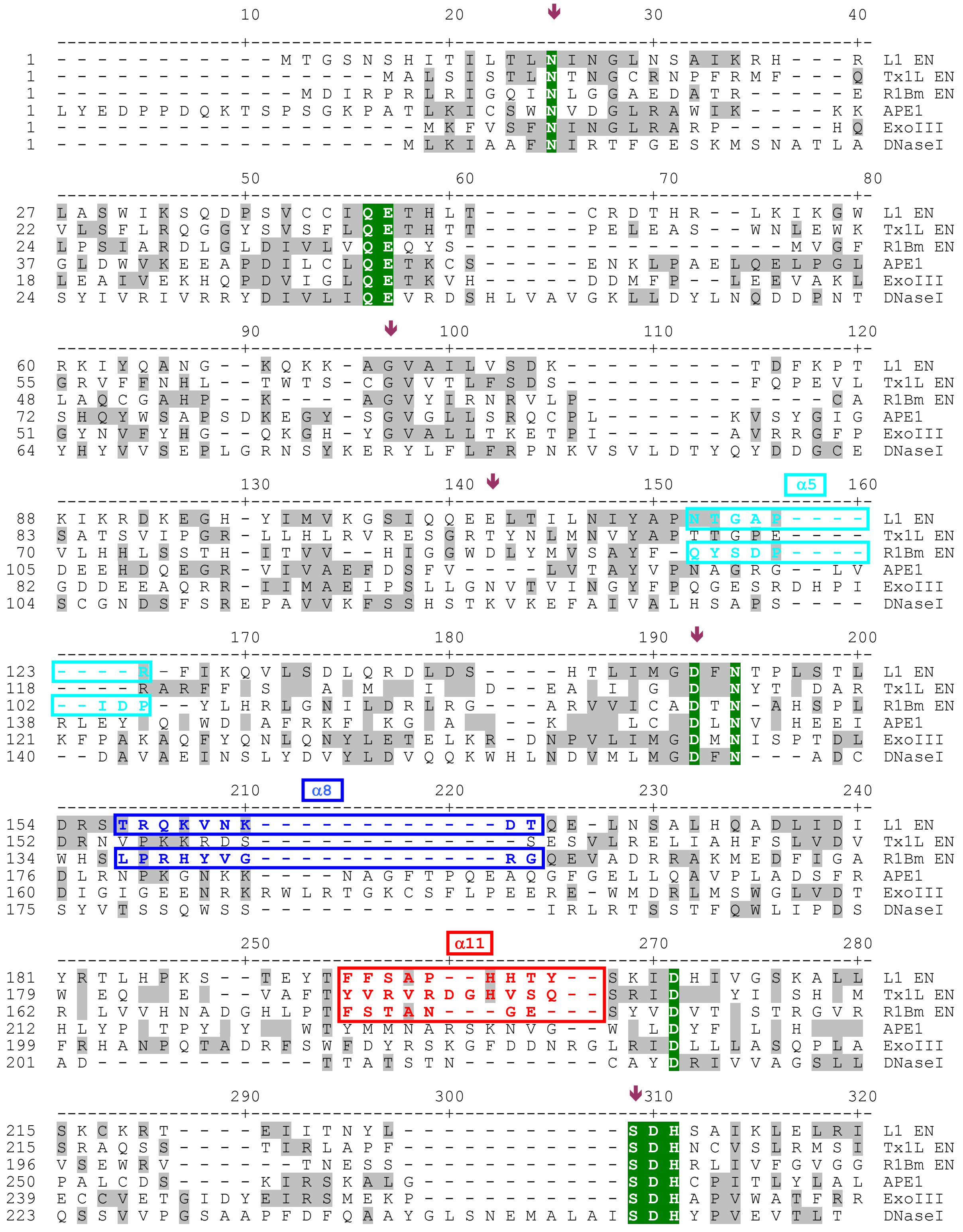
It has been proposed that the EN domain of LINE elements is the major determinant of their target site specificity (Luan *et al.*, 1993; Feng *et al.*, 1996; Takahashi and Fujiwara, 2002). In order to test this hypothesis and to evaluate the contribution of different moieties and structural features of EN to the elements' target site recognition, L1 EN was manipulated by exchanging regions of its coding sequence with corresponding regions of the sequence specific ENs of the related retrotransposons R1Bm and Tx1L. The rapid retrotransposition assay (Wei *et al.*, 2000) was used to assess chimeric L1 elements for their retrotransposition potential and capability for targeted integration.

3.3.1 Identification of EN regions probably involved in target site recognition

Modifying the EN domain of L1 reporter plasmids by swapping regions of the specifically integrating elements Tx1L and R1Bm into the corresponding L1 EN coding regions promised to yield insights into target site recognition of retrotransposons. As a first step it was necessary to align the peptide sequences of the three endonucleases encoded by L1, Tx1L and R1Bm in order to resolve two important questions:

- Which amino acid residues are critical for catalysis? - To ensure a high probability that the newly generated chimeric ENs remain functional, it was important not to disrupt the catalytic centre.
- Which regions might be involved in recognition of the respective target sequences of the three ENs? - Highly diversified polypeptides localised on the DNA binding surface of the enzymes were the best candidates for swapping experiments.

As the ENs encoded by L1, Tx1L and R1Bm belong to the family of AP-like ENs, they were not only aligned with each other, but also with three crystallised (Fig. 5, p.23) and enzymatically well characterised closely related enzymes, APE1, ExoIII and DNase I (Fig. 14). The alignment was performed with the programme MegAlign and manually edited where the programme could not perform well due to the considerably different lengths of the proteins. All highly conserved residues identified in APE1, ExoIII and DNase I ([Gorman *et al.*, 1997], labelled green in Fig. 14) could also be found at the same relative positions in the retrotransposon ENs (Feng *et al.*, 1996). Using these conserved residues as checkpoints, it was possible to assign amino acid sequences (boxed cyan and blue in Fig. 14) that should correspond to the α–loop structures α5 and α8 described to be involved in AP-site recognition of APE1 (Gorman *et al.*, 1997; Mol *et al.*, 2000) (Fig. 15). Nevertheless, the exact beginning and end of the loops could not be unambiguously defined as the sequences flanking the loops are not conserved. This difficulty was alleviated by the elucidation of the L1 EN crystal structure (Fig. 5, p.23 and Fig. 30, p.103) (Weichenrieder *et al.*, in press). It led to the identification of a hairpin loop structure (*aa* 192-202 in L1 EN, boxed red in Fig. 14) that is believed to directly interact with the target DNA and possibly contributes to target site recognition (Weichenrieder *et al.* in press). This hairpin loop corresponds to the α11 loop, the third loop protruding from the DNA binding surface of APE1 (Gorman *et al.*, 1997) (Fig. 15). The structural data allowed the identification of a highly conserved threonine anchoring the loop base (T192 in L1 ORF2p) which so far had not been recognised as conserved since the alignment requires manual editing in this region to allow for different sizes of the hairpin loops.



```
                       10                  20      ↓           30                  40
     ------------------+-------------------+-------------------+-------------------+-
  1  - - - - - - - - - - - M T G S N S H I T I L T L N I N G L N S A I K R H - - - R  L1 EN
  1  - - - - - - - - - - - - - - - - M A L S I S T L N T N G C R N P F R M F - - - Q  Tx1L EN
  1  - - - - - - - - - - - - M D I R P R L R I G Q I N L G G A E D A T R - - - - - E  R1Bm EN
  1  L Y E D P P D Q K T S P S G K P A T L K I C S W N V D G L R A W I K - - - - K K  APE1
  1  - - - - - - - - - - - - - - - - - - M K F V S F N I N G L R A R P - - - - - H Q  ExoIII
  1  - - - - - - - - - - - - - - - - - M L K I A A F N I R T F G E S K M S N A T L A  DNaseI

                       50                  60                  70                  80
     ------------------+-------------------+-------------------+-------------------+-
 27  L A S W I K S Q D P S V C C I Q E T H L T - - - - - C R D T H R - - L K I K G W  L1 EN
 22  V L S F L R Q G G Y S V S F L Q E T H T T - - - - - P E L E A S - - W N L E W K  Tx1L EN
 24  L P S I A R D L G L D I V L V Q E Q Y S - - - - - - - - - - - - - - - - M V G F  R1Bm EN
 37  G L D W V K E E A P D I L C L Q E T K C S - - - - - E N K L P A E L Q E L P G L  APE1
 18  L E A I V E K H Q P D V I G L Q E T K V H - - - - - D D M F P - - L E E V A K L  ExoIII
 24  S Y I V R I V R R Y D I V L I Q E V R D S H L V A V G K L L D Y L N Q D D P N T  DNaseI

                       90            ↓     100                 110                 120
     ------------------+-------------------+-------------------+-------------------+-
 60  R K I Y Q A N G - - K Q K K - A G V A I L V S D K - - - - - - - - - T D F K P T  L1 EN
 55  G R V F F N H L - - T W T S - C G V V T L F S D S - - - - - - - - - F Q P E V L  Tx1L EN
 48  L A Q C G A H P - - K - - - - A G V Y I R N R V L P - - - - - - - - - - - - C A  R1Bm EN
 72  S H Q Y W S A P S D K E G Y - S G V G L L S R Q C P L - - - - - - K V S Y G I G  APE1
 51  G Y N V F Y H G - - Q K G H - Y G V A L L T K E T P I - - - - - - A V R R G F P  ExoIII
 64  Y H Y V V S E P L G R N S Y K E R Y L F L F R P N K V S V L D T Y Q Y D D G C E  DNaseI

                       130                 140 ↓               150           α5    160
     ------------------+-------------------+-------------------+-------------------+-
 88  K I K R D K E G H - Y I M V K G S I Q Q E E L T I L N I Y A P N T G A P - - - -  L1 EN
 83  S A T S V I P G R - L L H L R V R E S G R T Y N L M N V Y A P T T G P E - - - -  Tx1L EN
 70  V L H H L S S T H - I T V V - - H I G G W D L Y M V S A Y F - Q Y S D P - - - -  R1Bm EN
105  D E E H D Q E G R - V I V A E F D S F V - - - - L V T A Y V P N A G R G - - L V  APE1
 82  G D D E E A Q R R - I I M A E I P S L L G N V T V I N G Y F P Q G E S R D H P I  ExoIII
104  S C G N D S F S R E P A V V K F S S H S T K V K E F A I V A L H S A P S - - - -  DNaseI

                       170                 180                 190 ↓               200
     ------------------+-------------------+-------------------+-------------------+-
123  - - - - R - F I K Q V L S D L Q R D L D S - - - - H T L I M G D F N T P L S T L  L1 EN
118  - - - - R A R F F   S     A   M     I     D - - E A   I   G   D   N Y T   D A R  Tx1L EN
102  - - I D P - - Y L H R L G N I L D R L R G - - - A R V V I C A D T N - A H S P L  R1Bm EN
138  R L E Y   Q   W D   A F R K F   K G   A     - - K       L C   D L N V   H E E I  APE1
121  K F P A K A Q F Y Q N L Q N Y L E T E L K R - D N P V L I M G D M N I S P T D L  ExoIII
140  - - D A V A E I N S L Y D V Y L D V Q Q K W H L N D V M L M G D F N - - - A D C  DNaseI

                       210      α8         220                 230                 240
     ------------------+-------------------+-------------------+-------------------+-
154  D R S T R Q K V N K - - - - - - - - - - - - D T Q E - L N S A L H Q A D L I D I  L1 EN
152  D R N V P K K R D S - - - - - - - - - - - - - S E S V L R E L I A H F S L V D V  Tx1L EN
134  W H S L P R H Y V G - - - - - - - - - - - - R G Q E V A D R R A K M E D F I G A  R1Bm EN
176  D L R N P K G N K K - - - - N A G F T P Q E A Q G F G E L L Q A V P L A D S F R  APE1
160  D I G I G E E N R K R W L R T G K C S F L P E E R E - W M D R L M S W G L V D T  ExoIII
175  S Y V T S S Q W S S - - - - - - - - - - - - - - I R L R T S S T F Q W L I P D S  DNaseI

                       250                  α11                270                 280
     ------------------+-------------------+-------------------+-------------------+-
181  Y R T L H P K S - - T E Y T F F S A P - - H H T Y - - S K I D H I V G S K A L L  L1 EN
179  W   E Q     E   - - V A F T Y V R V R D G H V S Q - - S R I D     Y I   S H   M  Tx1L EN
162  R   L V V H N A D G H L P T F S T A N - - - G E - - - S Y V D V T   S T R G V R  R1Bm EN
212  H L Y P   T P Y   Y     W T Y M M N A R S K N V G - - W   L D Y F   L   H        APE1
199  F R H A N P Q T A D R F S W F D Y R S K G F D D N R G L R I D L L L A S Q P L A  ExoIII
201  A D - - - - - - - - - - - T T A T S T N - - - - - - - C A Y D R I V V A G S L L  DNaseI

                       290                 300                ↓310                 320
     ------------------+-------------------+-------------------+-------------------+-
215  S K C K R T - - - - E I I T N Y L - - - - - - - - - - - S D H S A I K L E L R I  L1 EN
215  S R A Q S S - - - - T I R L A P F - - - - - - - - - - - S D H N C V S L R M S I  Tx1L EN
196  V S E W R V - - - - - - T N E S S - - - - - - - - - - - S D H R L I V F G V G G  R1Bm EN
250  P A L C D S - - - - K I R S K A L G - - - - - - - - - - S D H C P I T L Y L A L  APE1
239  E C C V E T G I D Y E I R S M E K P - - - - - - - - - - S D H A P V W A T F R R  ExoIII
223  Q S S V V P G S A A P F D F Q A A Y G L S N E M A L A I S D H Y P V E V T L T    DNaseI
```

Fig. 14: Alignment of the peptide sequences of the three APE-type endonucleases encoded by L1, Tx1L and R1Bm. The aligned amino acid sequences are from EN domains encoded by human L1.3 (L19088), Tx1L (M26915), and R1Bm (M19755) (see Appendix B), as well as from APE1 (M99703), ExoIII (D90818), and DNaseI (M60606). All residues that are identical at the corresponding position in at least two sequences are shaded grey, residues that are conserved in all sequences are shaded green. Purple arrows denote the junctions used for L1/R1Bm swapping experiments (see 3.3.2.1 and 3.3.2.2), while the helix loop regions α5, α8 and α11 (see 3.3.2.3 and 3.3.2.4) are boxed in cyan, blue and red.

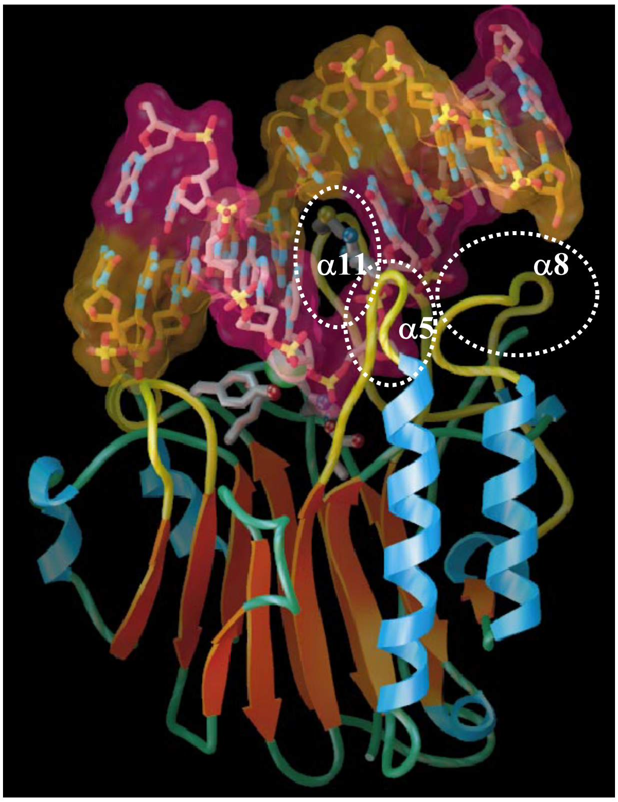


Fig. 15: Ribbon representation of the co-crystal structure of human APE1 bound to AP-DNA (from Mol *et al.*, 2000). The complex is viewed from the side that is roughly perpendicular to the kinked DNA helix axis oriented with the AP strand (pink) running 5' (left) to 3' (right). orange arrows, β-sheets; blue coils, α-helices; green tubes, coils; pink carbon polytubes and transparent surface, AP-DNA strand; orange polytubes and surface, opposing DNA strand. Peptide regions contacting the DNA are depicted in yellow, with the three loop regions α5, α8 and α11 indicated with white characters and white stippled lines.

3.3.2 Generation and activity assessment of chimeric L1 retrotransposons

3.3.2.1 L1 elements bearing chimeric L1/R1Bm block swap ENs are not retrotransposition competent

The first set of experiments was designed to determine which regions of EN are responsible for target site recognition. For this purpose, chimeric elements were generated by modifying the L1 EN coding region in a functional L1 reporter construct. L1 fragments of decreasing size were replaced by the homologous regions of R1Bm EN (Fig. 16). The resulting junctions between the swapped polypeptides and the preserved L1 EN sequences are indicated by purple arrows in Fig. 14.

Construct pNZ33 contains almost the entire R1Bm EN coding region (*aa* 14 to 209) in place of L1 EN (*aa* 15 to 230) in the L1.3 element (Sassaman *et al.*, 1997, Appendix B) encoded on retrotransposition reporter plasmid pJM101/L1.3 (Moran *et al.*, 1999) (Fig. 16A). Since amino acid substitutions in the N-terminal domain of L1 EN had indicated that the region from *aa* 1 to 14 of L1-ORF2 was essential for L1 retrotransposition (J. Moran, personal communication), this sequence was not modified in the chimeric EN.

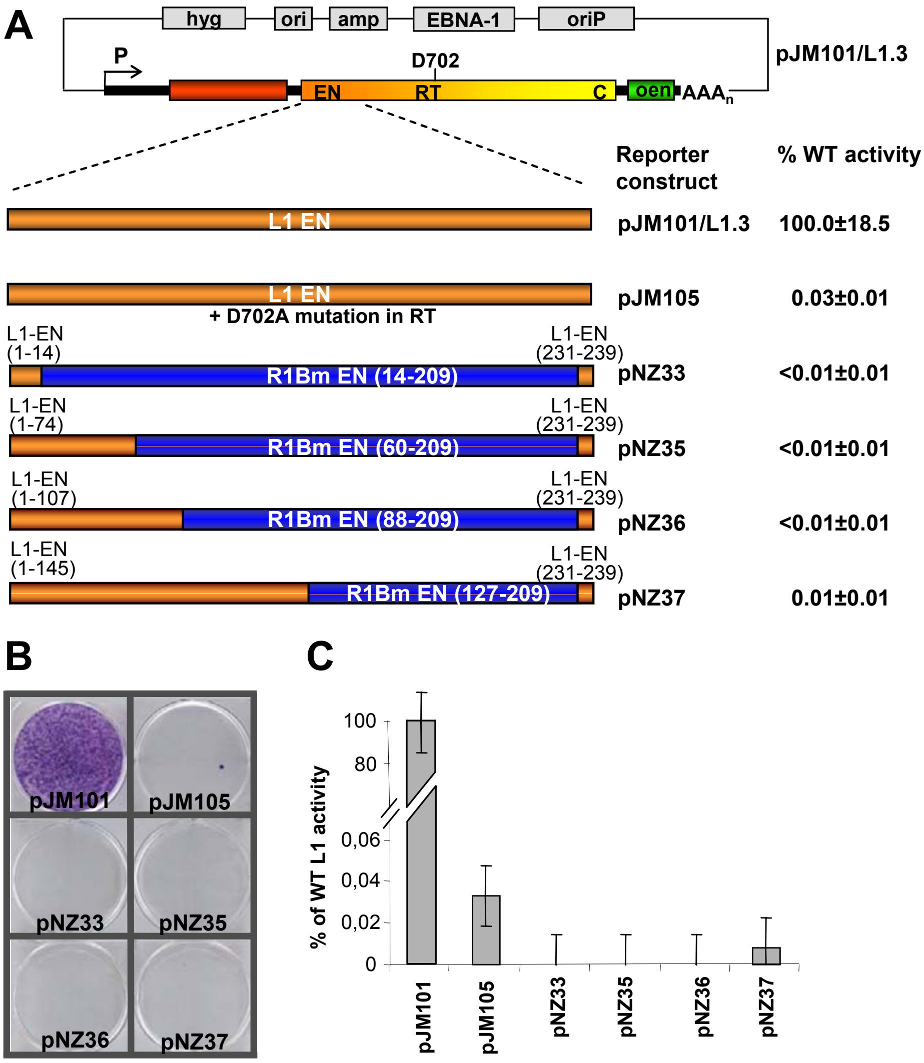


Fig. 16: Reporter constructs bearing chimeric L1/R1Bm ENs are not retrotransposition competent. (A) Schematic representation of the structures of L1/R1Bm EN chimeras incorporated in the L1 reporter construct. pJM101/L1.3 carrying a wild-type L1 EN (orange) was used as positive control. The negative control pJM105 is identical to pJM101/L1.3 in the EN domain, but harbours a D702 mutation in its RT domain. R1Bm sequences are depicted in blue. Numbers indicate amino acid positions of the exchanged polypeptides. The names of the reporter constructs and their activity relative to wild-type L1 are indicated on the right. **(B)** Representative results of retrotransposition assays performed with the indicated reporter constructs are shown. **(C)** Graphic representation of the results of the retrotransposition assays (n=3). Due to the differences in activity of wild-type and mutant L1s, the y-axis is split into two parts with different scales.

AP-like endonucleases share an SDH-motif (*aa* 228 to 230 of L1-ORF2) at their C-terminal ends which was shown to be critical for catalysis (Mol *et al.*, 1995; Feng *et al.*, 1996). Since the precise C-terminal end of the EN domain in the multifunctional ORF2p is not known, this motif was used as junction between R1 EN and L1 sequences in the chimeric constructs.

Another reason for maintaining the L1-specific C-terminal peptide sequence beyond amino acid position 230 of L1 EN was the possibility that proteolytic cleavage of L1-ORF2 in this region may be necessary to generate functional RT- and EN proteins.

Three additional chimeric L1/R1 elements (pNZ35, pNZ36, pNZ37) were generated by swapping R1Bm EN-blocks of decreasing lengths into L1 (Fig. 16A). To minimise the risk of destroying the EN's active site, conserved residues were used as hinges when possible (G-V in pNZ35, E/D-L in pNZ36, G/A-D in pNZ37, indicated by purple arrows in Fig. 14).

The chimeric L1/R1 EN elements were tested for their ability to retrotranspose using the transient retrotransposition assay adapted from Wei *et al.*, (2000) (see 3.1, p.55). Retrotransposition frequencies were ≤ 0.01% of wild-type L1 activity (Fig. 16) and thus even lower than the background activity (0.03%) displayed by the negative control construct pJM105 (Moran *et al.*, 1996). pJM105 is an originally functional L1 with the point mutation D702A inactivating its RT domain. It is highly probable that the proteins encoded by this construct are folded correctly. The mutated RT domain might retain some residual activity, giving rise to the occasional G418^R colonies seen in HeLa cells transfected with pJM105 (Moran *et al.*, 1996; Moran *et al.*, 1999 and own unpublished observations). In the L1/R1 chimeras, however, protein sequences from two different retrotransposons have been transplanted into a complex polyprotein. Steric clashes at the interface between the two mismatched polypeptides could scramble the three-dimensional structure of EN or the complete ORF2 protein. Furthermore, the R1Bm sequences could disable putative interactions with ORF1 or host factors. Therefore the inactivity of the R1 EN block swaps probably stems from steric incompatibility of L1 and R1 EN protein moieties.

3.3.2.2 An L1 element harbouring the Tx1L EN retrotransposes in an apparently EN-independent manner

Each of the two parallel β-sheets of AP-like ENs are formed by one half of the EN peptide chain. Partial EN swaps were therefore reasoned to lead to steric incompatibility at the extensive interface of the β-sheet structures. Consequently, only one block swap experiment was carried out with Tx1L EN: pNZ39 is a construct with almost the complete EN of L1.3 being substituted by the corresponding polypeptide sequence of Tx1L EN (Fig. 17A). In this setting, at least the EN domain should be able to fold correctly. Analogous to pNZ33, the EN substitution was limited to the region between residues 15 and 230 of L1 ORF2p (see 3.3.2.1, Fig. 16).

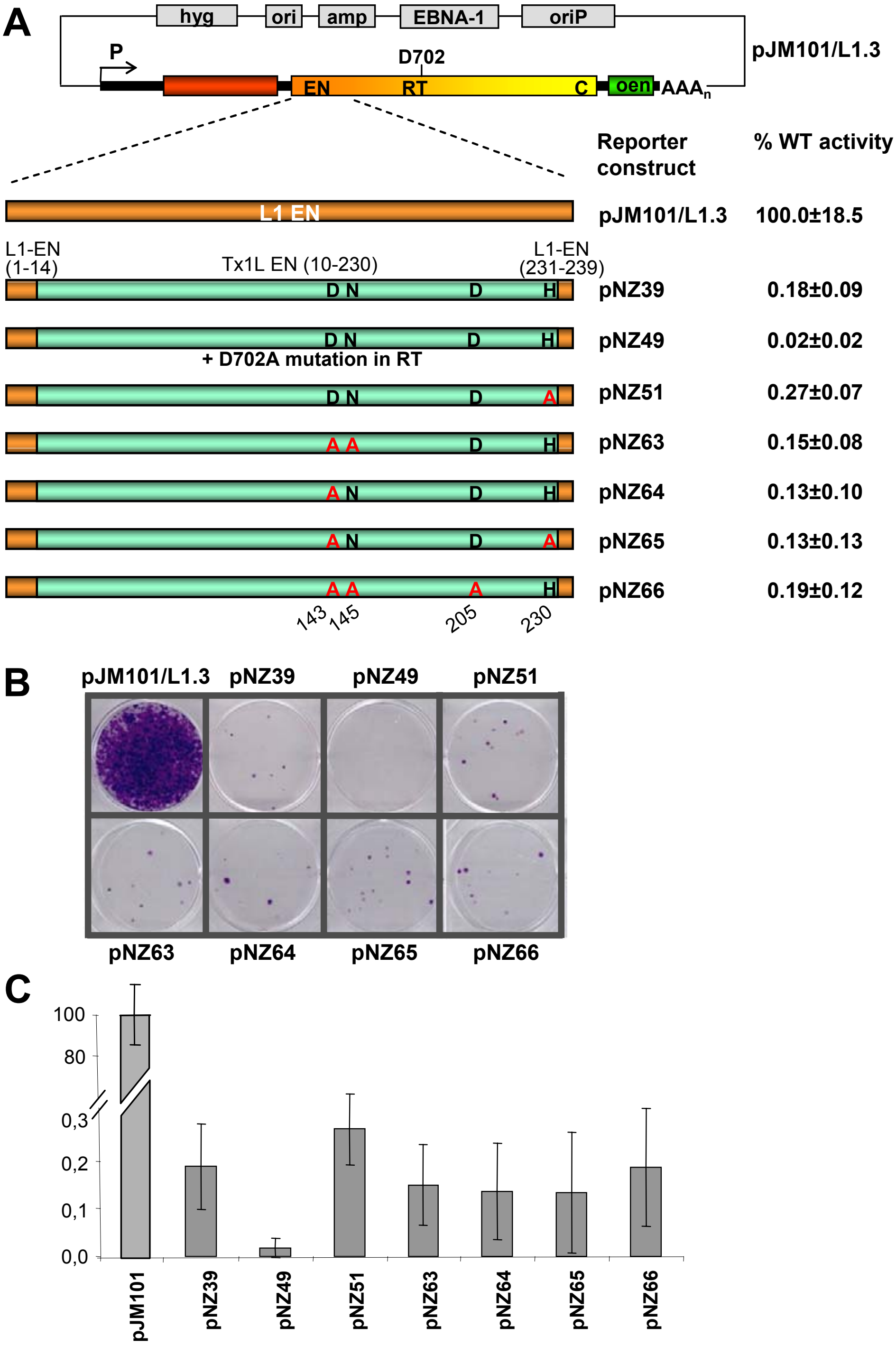


Fig. 17: Reporter construct pNZ39 harbouring an L1/Tx1L EN chimera is retrotransposition-competent but transposes in an apparently EN-independent manner. (A) Schematic representation of the structures of L1/Tx1L EN chimeras. pJM101/L1.3 carrying a wild-type L1 EN (orange) was used as positive control. Tx1L-derived sequences are coloured green. The four catalytical residues D143, N145, D205 and H230 are shown and their replacement by alanines is indicated in red. The names of the reporter constructs and their activity relative to wild-type L1 are indicated on the right. **(B)** Representative results of retrotransposition assays performed with the indicated reporter constructs are shown. **(C)** Graphic representation of the results of the retrotransposition assays (n=6). Due to the differences in activity of wild-type and mutant L1s, the y-axis is split into two parts with different scales.

In seven independent experiments, this construct reproducibly yielded a very low number (2-16) of G418 resistant colonies (Fig. 17B,C). This number was always above the levels of the negative control pJM105 and of the R1Bm block swaps (0.18±0.09% wild-type activity of pNZ39 versus 0.03% for pJM105 and ≤0.01% for pNZ33-37). This suggested that the L1 element with a transplanted Tx1L EN is able to retrotranspose.

As mentioned in the introduction, L1 elements usually act in *cis*, i.e. L1 proteins process predominantly the RNA molecule they are encoded on. This effect is probably due to spatial proximity of the RNA and the nascent protein, rather than recognition of specific sequences on the RNA. In rare cases, however, RNA from a retrotransposition-incompetent L1 element is retrotransposed by proteins derived from an active L1, a process known as *trans* complementation. Accordingly, fortuitous retrotransposition of foreign RNAs has been observed in an artificial cell culture system at a very low frequency (0.2-0.9% of wild-type L1 activity), but only if *trans*-acting L1 elements are highly over-expressed (Wei *et al.*, 2001). Since the low level of colony formation caused by the L1/Tx1L EN chimera pNZ39 ranged in the same order of magnitude, it was necessary to perform controls to address the question whether the observed transposition frequency is merely the consequence of *trans* complementation of an inactive L1 chimera.

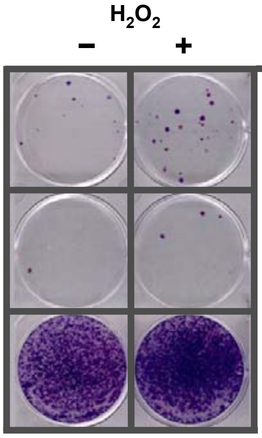
In order to control for a possible *trans* complementation of RT activity by endogenous L1 elements, the control plasmid pNZ49, a pNZ39 analogue carrying the RT missense mutation D702A, was generated. This construct exhibited no retrotransposition potential (0.02% relative to wild-type L1), showing that *trans* complementation of the RT does not occur. Nevertheless, a further control experiment was performed to test whether the EN is indeed responsible for the observed retrotransposition events. For this purpose, construct pNZ51 was generated, carrying the same chimeric L1/Tx1L element as pNZ39 except that His 230, a residue involved in catalysis, is replaced by Ala. This point mutation had been used before to destroy L1 EN activity (Wei *et al.*, 2001). The resulting mutated chimera was expected to be unable to transpose due to its inactivated EN. However, pNZ51 retained the same retrotransposition frequency as the parental construct pNZ39 (0.27 ± 0.07%, Fig. 17).

This particular histidine residue (H309 in APE1) is not directly involved in the catalytic cleavage of the scissile phosphate bond, but is one of three amino acids that fix the phosphate in the correct orientation (Mol *et al.*, 2000). As it is highly probable that the conformation of the chimeric L1/Tx1L EN differs at least slightly from L1 EN or Tx1L EN, it was reasoned that H230 might not be a critical residue in the chimera and therefore indifferent to substitution. Consequently, four additional mutants were generated, bearing one or more point mutations in the conserved residues D143 (the actual catalytic residue activating the attacking nucleophile), N145 and D205. The combination of mutations in the resulting constructs pNZ63-66 are indicated in Fig. 17.

Each point mutation had the same effect and did not result in a reduced retrotransposition frequency. Fig. 17 summarises the results obtained and illustrates that the changes of activity of the different mutants relative to pNZ39 are not statistically significant. Hence it must be concluded that the chimeric L1/Tx1L EN of pNZ39 itself is already inactive.

This result was rather unexpected. In an attempt to explain how the L1/Tx1L chimeras can transpose without a functional EN, the following theory was formulated: pNZ39 may be able to initiate retrotransposition without possessing a nucleolytically active EN because the chimeric EN can still bind to pre-existing nicks in the target DNA. It could thus recruit the element's ribonucleoprotein particle to the target DNA and initiate transposition with a higher frequency than for example the corresponding L1/R1 chimera pNZ33 (Fig. 16, p.66). If this were the case, an increased number of chromosomal nicks in the target cell should result in an augmented retrotransposition rate.

To test this hypothesis, the retrotransposition assay was performed under conditions that increased the amount of single-strand breaks in the host cell. Hydrogen peroxide (H_2O_2) is an oxidative reagent known to induce mainly single-strand DNA breaks when added to the medium of cultured cells (Dahm-Daphi *et al.*, 2000). First, HeLa cells were titrated with H_2O_2 in order to establish the maximum concentration not harming the cells under the assay conditions (10^{-5} M, causing an estimated 10^4 single-strand breaks/cell, data not shown). Subsequently, the L1/Tx1L EN chimera pNZ39 and its RT-mutant pNZ49 were tested in the retrotransposition assay in the presence or absence of 10^{-5} M H_2O_2. The results of this experiment are shown in Fig. 18. While the retrotransposition frequency of pNZ39 was indeed elevated by a factor of 2.5 under the influence of H_2O_2, the same was true for pNZ49. This latter construct, however, should not react to an increase of genomic nicks as its RT is inactivated. Increased retrotranspositional activity in response to H_2O_2 treatment was also observed for pJM101/L1.3. This effect can be ascribed to transcriptional upregulation of L1 elements as a result of irradiation (Servomaa and Rytömaa, 1990) or oxidative stress (G. Tolstonog, Heinrich-Pette-Institut, personal communication), leading to subsequent increased *trans* complementation. Due to this general effect, it was not possible to identify a pNZ39-specific effect of single-strand breaks.



Transfected construct	Retrotransposition frequency relative to pJM101 in the absence of H_2O_2 [%]		
	− H_2O_2	+ H_2O_2	x-fold increase
pNZ39	0.36±0.07	0.84±0.09	2.3
pNZ49	0.06±0.06	0.15±0.03	2.5
pJM101	100±14.8	200±50	2.0

Fig. 18: Treatment of HeLa cells with the oxidative reagent H_2O_2 leads to an increased retrotransposition frequency in both mutant and wild-type L1 elements. Representative results of retrotransposition assays performed with the indicated reporter constructs in the absence or presence of H_2O_2. Schematic drawings of pNZ39, pNZ49 and pJM101 are shown in Fig. 17. Relative retrotransposition frequencies were normalised for pJM101/L1.3 activity in the absence of H_2O_2 (n=3). "x-fold increase" quantifies the stimulating effect of H_2O_2 treatment on retrotransposition.

3.3.2.3 An L1 element bearing the R1α8-helix is retrotransposition competent, while the α5-swap is inactive

In order to reduce the probability of steric clashes that could affect the nucleolytic activity of L1 EN, I decided to change its structure as little as possible. Therefore, additional chimeras were created with only short polypeptides of L1 and R1 EN being swapped. Previous studies of the structure of APE1 and DNase I had implicated defined α5- and α8-helices (Fig. 15) to be involved in major groove interactions with the target DNA and showed that a helix transplant of α8 can generate an EN with altered specificity (Gorman *et al.*, 1997; Cal *et al.*, 1998). Since L1 EN and R1 EN display significant homologies to APE1, it was reasoned that the corresponding regions in all ENs are essential for target site recognition. The segments of the R1 enzyme that correspond to the α5 - and α8-helices of APE1 were grafted into the L1 EN sequence of the L1-retrotransposition reporter construct pJM101/L1.3 either separately (pNZ44, pNZ45) or in combination (pNZ47) (Fig. 19A).

The altered L1 elements' ability to retrotranspose was evaluated using the transient retrotransposition assay. While the α5 swap (pNZ44) and the double-swap (pNZ47) generated inactive hybrid retrotransposons, the α8 swap (pNZ45) resulted in a functional chimeric EN displaying a retrotransposition frequency of 5.4% relative to L1 wild-type activity (Fig. 19).

Although the level of activity of pNZ45 lies far above the reported *trans* complementation frequencies (see 3.3.2.2), construct pNZ50 was designed to test the possibility that the observed transposition frequency was a result of *trans* complementation. pNZ50 differs from pNZ45 exclusively in a point mutation in the RT domain of the chimeric L1, resulting in a D702A exchange. This construct displayed a drastic decrease of activity to 0.04 ± 0.05 % of L1 wild-type activity, indicating that transposition of the pNZ45 encoded chimera is dependent on its own ORF2p and not on *trans* complementation. Therefore, exchange of the L1α8 helix loop with its R1Bm counterpart yielded an actively retrotransposing element.

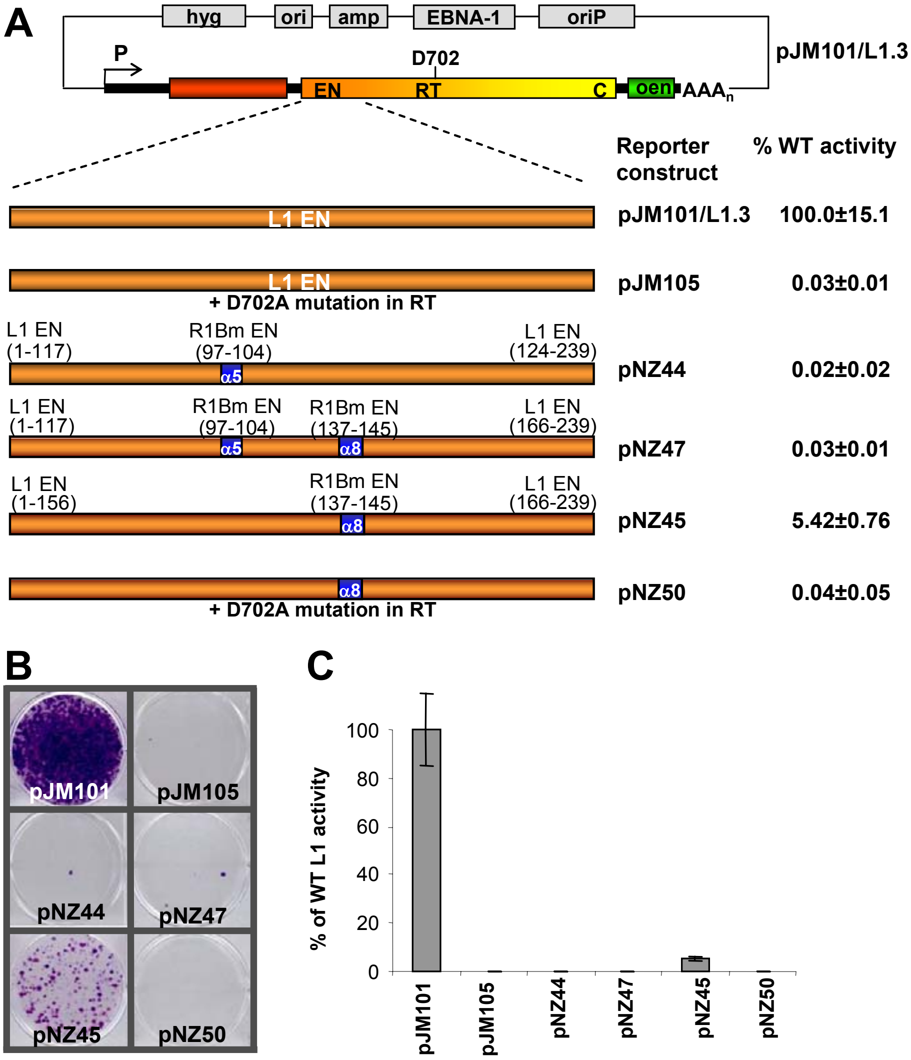


Fig. 19: Reporter construct pNZ45 harbouring an L1/R1α8 chimera is retrotransposition-competent. (A) Schematic representation of the structures of L1/R1Bm α5 and α8 chimeras. pJM101/L1.3 carrying a wild-type L1 EN (orange) was used as positive control. R1Bm-derived sequences are coloured blue. The names of the reporter constructs and their activity relative to wild-type L1 are indicated on the right. **(B)** Representative results of retrotransposition assays performed with the indicated reporter constructs are shown. **(C)** Graphic representation of the results of the retrotransposition assays (n=3).

3.3.2.4 Replacement of the L1α11 helix loop with its Tx1L counterpart leads to a highly active chimeric element

The elucidation of the crystal structure of L1 EN (Weichenrieder *et al.*, in press) allowed us to devise a fourth set of experiments. The three-dimensional structure of the L1 EN domain revealed a prominent hairpin-shaped loop (Fig. 5, p.23) corresponding to the α11-loop of APE1 (Fig. 15, p.65). This hairpin loop is anchored in the active site cleft by two highly conserved residues, T192 and S202, and protrudes from the putative DNA binding surface of L1 EN. Most probably it contacts the minor groove of the target DNA and bends it into the correct conformation for nicking (Weichenrieder *et al.*, in press). This notion fits nicely with data showing that L1 EN preferrably nicks targets with a stretch of pyrimidines followed by a polypurine tract, a sequence naturally taking a kinked form (Cost and Boeke, 1998).

The EN alignment (Fig. 14) suggests that the EN domains of the sequence-specific retrotransposons Tx1L and R1Bm also possess an α11 hairpin loop. The anchoring amino acids Thr and Ser are conserved, but the loop sequences themselves differ from each other and from the L1 sequence, possibly allowing for recognition of different target sites.

In order to verify the importance of the hairpin loop in enzymatic activity and specificity of the EN, three reporter constructs were generated (Fig. 20A): pNZ73 lacks the loop completely; the deletion is bridged by a highly flexible linker consisting of two glycine residues. The length of this linker was deduced from the structural data as being required to cover the distance between T192 and S202. In pNZ75 and pNZ76, the L1 hairpin loop was exchanged for the corresponding loops from Tx1L and R1Bm, respectively. The danger of destroying the correct conformation of the active site and the overall structure of L1 EN is minimal in these three constructs, as the loop sticks out from the bulk of the protein and does not interact with any other parts of the enzyme (Fig. 30, p. 103).

Constructs pNZ73, 75 and 76 are based on the plasmid rescue vector pCEP4/L1.3*mneoI*$_{400}$/ColE1 (Gilbert *et al.*, 2002), a vector displaying a retrotransposition efficiency of 18% of the standard reporter pJM101/L1.3 (see 3.3.3.2, p.79). To allow for direct comparison between constructs derived from pJM101/L1.3 and pCEP4/L1.3*mneoI*$_{400}$/ColE1, all retrotransposition rates were normalised against the activity of the respective wild-type construct. The negative control plasmid pNZ77, which differs from pCEP4/L1.3*mneoI*$_{400}$/ColE1 only in the RT point mutation D702A, was not retrotransposition-competent. pNZ73 and pNZ76 displayed comparable low retrotransposition frequencies slightly above background level (2.6% and 6.1% wild-type activity), but pNZ75 retrotransposed with a substantial retrotransposition frequency of ~25 % of pCEP4/L1.3*mneoI*$_{400}$/ColE1.

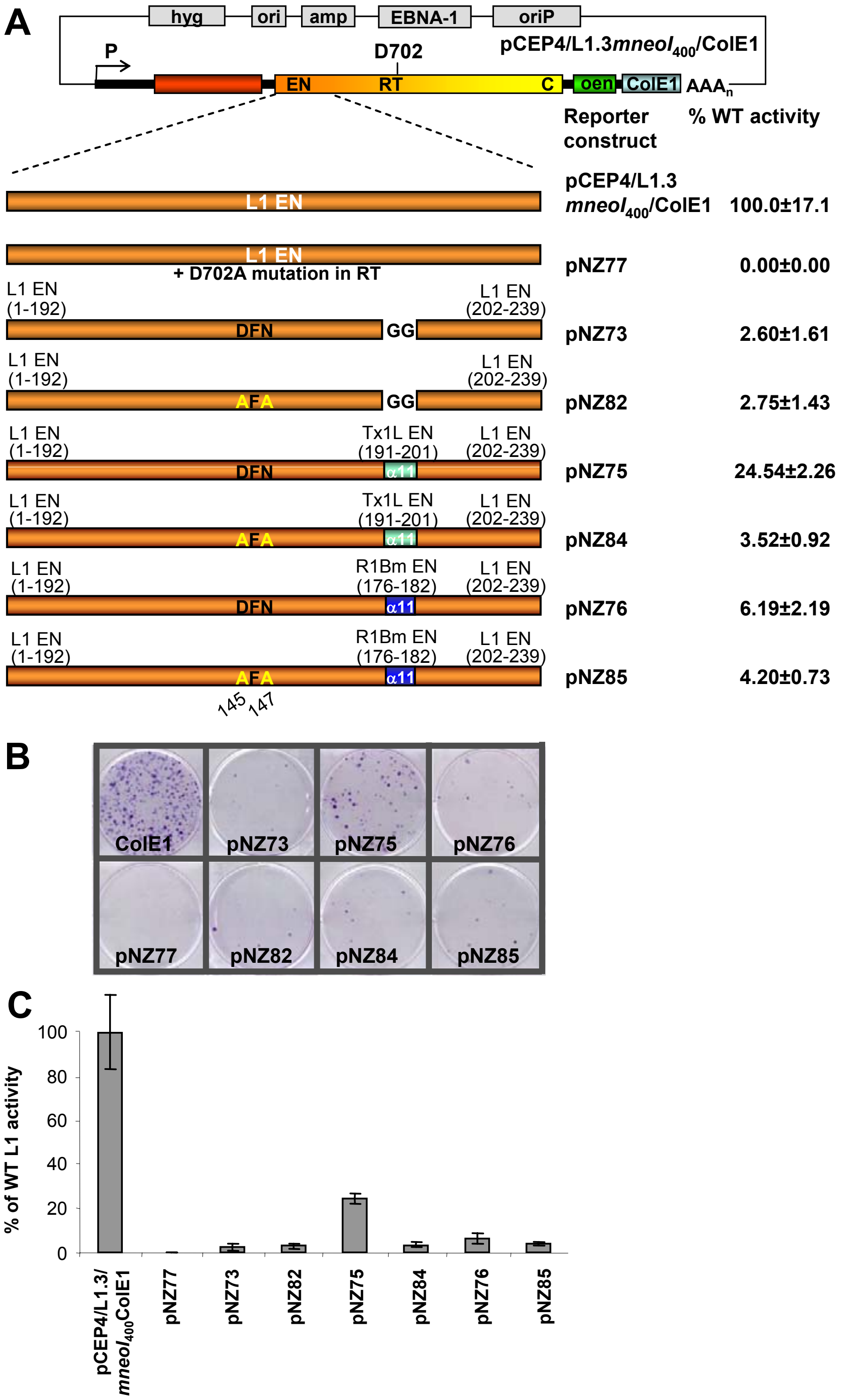


Fig. 20: The L1 reporter construct pNZ75 harbouring a Tx1L hairpin loop chimera is retrotransposition-competent. (A) Schematic representation of the structures of the hairpin loop modifications. pCEP4/L1.3*mneoI*$_{400}$/ColE1 carrying a wild-type L1 EN (orange) was used as positive control for retrotransposition. pNZ77, carrying an inactive RT domain, was used as negative control for retrotransposition. The two Gs in pNZ73 and pNZ82 indicate the substitution of the L1 hairpin structure by two glycine residues. Tx1L-derived sequences are coloured in green, while R1Bm-derived sequences are depicted in blue. Inactive variants harbouring exchanges of the catalytic residues D145 and N147 for alanines were created for each construct as indicated. The names of the reporter constructs and their activity relative to wild-type L1 are shown on the right. **(B)** Representative results of retrotransposition assays performed with the indicated reporter constructs. **(C)** Graphic representation of the results of the retrotransposition assay (n=3).

Analogous to pNZ39 (3.3.2.2, p.67), the contribution of *trans* complementation to the activity of the hairpin swaps was controlled for with the constructs pNZ82, 84 and 85. They contain the two point mutations D145A and N147A in addition to the hairpin modifications (Fig. 20A). EN inactivation did not significantly influence retrotransposition of the hairpin deletion mutant in pNZ82 and in the R1Bm hairpin chimera pNZ85, indicating that the observed low activities of pNZ73 and pNZ76 are EN-independent. In contrast, retrotransposition of the Tx1L hairpin chimera pNZ75 dropped to background levels when the EN mutations were introduced (pNZ84, 3.5% wild-type activity), arguing for EN-mediated autonomous transposition of pNZ75. Thus, only one of the three α11 hairpin modifications yielded a functional retrotransposon. However, this construct, the L1/Tx1Lα11 chimera, displayed the highest retrotransposition frequency of all examined hybrid elements. These findings confirmed two of our assumptions: 1) Inactivation of L1 EN by deletion of the hairpin implies a crucial role of this loop in DNA binding and/or cleavage. 2) The high activity of the L1/Tx1Lα11 hybrid indicates that suitable exchange of the protruding hairpin loop does not destroy the active conformation of L1 EN.

3.3.3 Methods used for sequence analysis of the integration sites

After the hybrid constructs had been tested for their potential to transpose, the target sequences of the active L1 chimeras had to be analysed for changes in integration specificity. Determining the insertion sites of L1 elements poses four inherent problems:

- The sequences flanking the new integrants are unknown. Thus, direct PCR methods using primers bracketing the region of interest cannot be used for amplification of the 3’ and 5’ junctions of *de novo* integrants.
- *De novo* L1 insertions in the human genome are particularly difficult to track due to the high copy number of homologous endogenous L1 sequences. Any method using an L1-specific primer is bound to fail due to the generation of a large background of PCR products. The only unique sequences present in the chimeric elements characterised in this study are the foreign EN sequences and the *mneoI* cassette. Tracking the fate of the chimeras in the genome thus has to start from these sequences.
- Once the 3’ junction of an integration event is identified, another difficulty encountered with non-LTR retrotransposons is the frequent 5’ truncation (see 1.5.1.1). Not only are the flanking sequences of new integration events unknown, but it is also impossible to predict the extent of the inserted L1 sequence.

- Last but not least, inversions of the 5'-half sequences of the newly inserted L1 copies are possible (see 1.5.1.1). This complicates the amplification of 5' junctions, even if the flanking sequence is known.

Initial attempts to identify flanking sequences of integrated chimeric L1 elements were carried out using inverse PCR (iPCR), a method which utilises circularised genomic DNA fragments as templates for a PCR reaction with primers binding to the known sequence and pointing in opposite directions (Ochman *et al.*, 1988). However, these attempts were unsuccessful. As this method works best with DNA templates whose sequence complexity is less than 10^9 bp (Ausubel *et al.*, 1999), iPCR is difficult to perform with total mammalian genomic DNA as template. HeLa cells contain a genome of varying polyploidy with even higher sequence complexity. The HeLa cells used in my experiments were karyotyped and proved to have a genome of 1.5 chromosome sets on average (data not shown), making iPCR virtually impossible.

However, two unrelated approaches to sequence the junctions of *de novo* integrants and their flanking DNA were successful. I first adapted and improved a PCR method originally designed to isolate rare retroviral integration events (Schmidt *et al.*, 2001) to fulfil the requirements of isolating L1 integrations, and then used a plasmid rescue vector recently developed (Gilbert *et al.*, 2002) explicitly for the isolation of *de novo* L1 insertions.

3.3.3.1 EPTS/LM-PCR was used to isolate sequences flanking *de novo* integrants derived from pNZ45 and pNZ39

New integrants resulting from retrotransposition of the L1/Tx1L hybrid (pNZ39) and the L1/R1α8-hybrid (pNZ45) were isolated by a method of non-target DNA removal via magnetic extension primer tag selection (EPTS) preceding solid-phase ligation-mediated (LM) PCR (Fig. 21). In order to perform EPTS/LM-PCR, genomic DNA harbouring an integration event is digested with a restriction enzyme not cleaving in the region of interest. The DNA fragment containing the 3' junction is selectively labelled using a biotinylated primer in a primer extension reaction. Then, the resulting labelled DNA-fragment is bound to magnetic streptavidin-beads and freed of the bulk of genomic DNA by washing. In this way, the problems with highly complex genomes encountered in iPCR are circumvented. Subsequently, a unidirectional oligonucleotide cassette (linker OC) is ligated to the DNA bound on the beads, and the unknown sequence between the linker cassette and the known sequence of the integrant can be amplified by PCR.

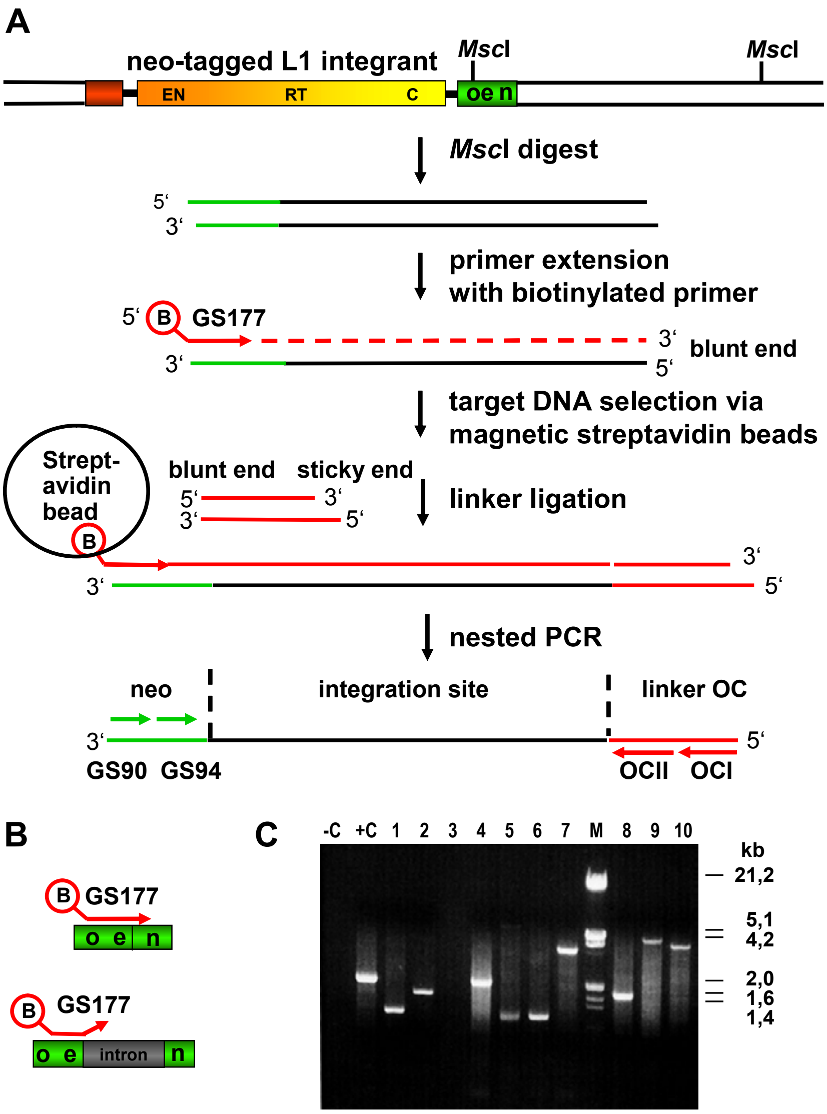


Fig. 21: Extension primer tag selection/ligation mediated PCR (EPTS/LM-PCR) (A) Schematic representation of the method of EPTS/LM-PCR. Genomic DNA containing a *neo*-tagged L1 integrant is digested with *Msc*I. Primer extension with a *neo*-specific biotinylated primer (GS177) is followed by target-DNA selection via magnetic streptavidin beads. After ligation of a linker molecule (OC) to the primer-extension product, PCR is performed with *neo*-and linker-specific primers to amplify the junction between the L1/*neo* sequence and its integration site. **(B)** The biotinylated primer GS177 binds to the *neo*-cassette only if the cassette is spliced. **(C)** Separation of EPTS/LM-PCR products obtained from 10 representative G418^{R} HeLa cell lines in an 0.8% agarose gel (lanes 1-10). –C, HeLa DNA; +C, HeLa DNA mixed with 1 ng of pSV2neo (Clontech, Heidelberg).

For the isolation of *de novo* chimeric L1 integration events, I took advantage of the unique *mneoI* cassette. *Msc*I was chosen to digest the genomic DNA and the primer extension reaction was performed with the biotinylated primer GS177 binding within the *neo*-sequence directly downstream of the *Msc*I site (position 6633 in pJM101/L1.3). In order to exclusively isolate genuine retrotransposition events, GS177 was designed in a way that it binds to the *neo*-cassette only when the cassette is spliced (Fig. 21B). Thus, chimeric reporter constructs that inserted into the genome by recombination cannot give rise to contaminating PCR products.

This method proved to be an efficient and very reliable means to isolate the 3’ flanking sequences of chimeric *de novo* L1 integrants. A representative example of an agarose gel loaded with the PCR products of ten EPTS/LM-PCR reactions is shown in Fig. 21C. Most clonal HeLa cell lines harbouring a chimeric L1 integration event yielded a PCR product ranging from one to seven kb. Empty lanes represent cell lines where the distance between the integrant and the nearest *Msc*I site was too long to be amplified in the PCR reaction (>7kb). In these cases, a fragment was obtained when the genomic DNA was digested with one or two additional enzymes (non-cutters in the L1 sequence between the *Msc*I site at position 6633 and the 3’ end of L1).

Sequencing of the resulting PCR bands was attempted from both directions. However, L1 copies usually end in a poly(A) tail that cannot be read through in a sequencing reaction. Only in rare cases of a short or non-existent poly(A) tail (see 3.3.4.1, p. 82), the 3’ junction could be sequenced with a primer binding in the 3’ UTR. The majority of flanking sequences were identified by sequencing with a primer (OCI or OCII, Fig. 21A) specific for the linker cassette.

Most sequencing reactions with primers OCI or OCII did not extend towards the 3’ junction between genomic DNA and the hybrid L1 copy. Therefore, the obtained sequences were used as probes in BLAT searches in the human genome working draft (HGWD) sequence available through the UCSC-web browser (http://genome.cse.ucsc.edu). In all but two instances, the flanking sequences matched unique sequences present in the HGWD with >99% identity, allowing to localise the genomic position of the new integrant. In order to exactly identify the 3' junctions, primers were designed that bind ~200 bp downstream of the presumed integration site as judged from the length of the EPTS/LM-PCR fragment (Fig. 21C). With these insertion-specific primers and a *neo*-specific primer (GS 90), the 3’ junctions could be amplified directly from genomic DNA. This technique yielded 23 unique, unambiguous 3’ flanking sequences for pNZ39 (Fig. 17, p.68) and nine 3’ flanks for pNZ45 (Fig. 19, p.72).

While 3’ junctions of L1 integrants can be easily isolated by EPTS/LM-PCR, this method is ill suited for the amplification of 5’ junctions. The *mneoI* cassette, the only sequence that can be used for selective primer extension, is located within the 3’ UTR. Even if a restriction enzyme could be found that cuts within the *neo* gene, but not within the entire upstream L1 sequence, the processivity of *Pfu* polymerase would limit the yield of the primer extension reaction drastically.

In order to identify the 5’ junctions of *de novo* integrants, it was acted on the assumption that the chimeric retrotransposons create integrants with a structure similar to wild-type L1, notably that they are flanked by small TSDs. Primers were designed that bind approximately 200 bp upstream of the genomic integration site identified at the 3’ junction. A comprehensive set of oligonucleotides (GS88, GS17, GS14, GS16, GS52, GS10 and GS189) spanning the entire L1-element was used as reverse primers in order to account for the different lengths of truncated insertions (Fig. 22). The 5' junctions were amplified directly from genomic DNA. This approach was successful in five out of nine cases of integrants derived from the L1/R1α8 hybrid pNZ45 and in 17 out of 23 cases of integrants derived from the L1/Tx1L EN chimera pNZ39.

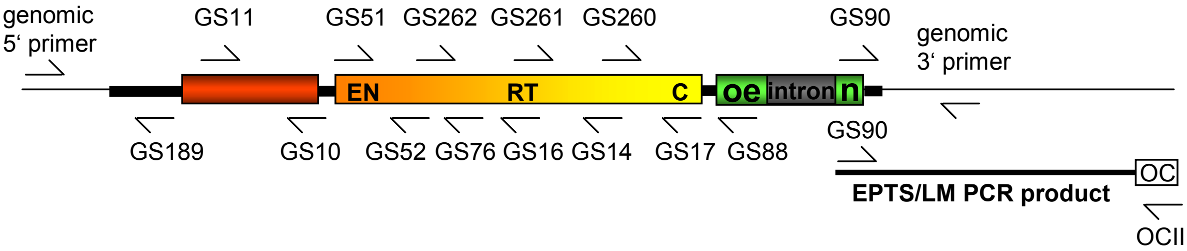


Fig. 22: Schematic representation of the binding sites of primers involved in the isolation of 5’ and 3’ junctions of *de novo* L1 integrants. The exact binding coordinates on L1.3 are given in Appendix A.

The remaining clones were screened for inversion events that would elude the former strategy. The primer specific for the expected genomic 5’-flanking sequence was used in conjunction with a set of L1-specific oligonucleotides pointing in the sense direction of L1. This led to the detection of two inverted integrants derived from pNZ45. In all other clones, the 5’ junctions could not be elucidated. Target site duplications or target site deletions larger than 200 bp may be responsible for this, as they lead to dislocation or deletion of the binding site for the designed 5’ primer. Although not impossible, resolution of such structures requires a disproportionate amount of time and was therefore not attempted.

3.3.3.2 A plasmid rescue procedure was used to isolate 3’ and 5’ junctions of *de novo* integrants derived from pNZ75

In 2002, the laboratories of J. Boeke and J. Moran published two independent L1 reporter plasmids that allow to directly clone individual marked *de novo* L1 integrants together with their flanking genomic DNA in bacteria (Symer *et al.*, 2002; Gilbert *et al.*, 2002). In these so-called plasmid rescue vectors, a bacterial origin of replication and a prokaryotic selectable marker is introduced into the 3’ UTR of L1 in the retrotransposition reporter construct pJM101/L1.3. The vectors differ only slightly in the arrangement of their components and the choice of marker. J. Moran kindly supplied us with the rescue vector pCEP4/L1.3*mneoI*$_{400}$/ColE1. In this construct, the *neo*-gene of the *mneoI* cassette is used as both eukaryotic and prokaryotic marker by inserting the bacterial EM7 promoter and a Shine-Dalgarno-sequence upstream of the *neo* initiator codon. Besides, a ColE1 origin of replication is added (Fig. 8, Fig. 23). It should be noted that these modifications reduce the retrotransposition efficiency ~6-fold when compared to pJM101/L1.3 (Gilbert *et al.*, 2002), probably due to the increased length of the retrotransposed sequence needed to confer G418-resistance to HeLa cells. This disadvantage, however, is more than compensated by the ease and speed with which new retrotransposition events can be recovered. Genomic DNA isolated from G418^R cell lines derived from the retrotransposition assay is digested with a restriction enzyme that does not cleave within the *mneoI*/ColE1 cassette, ligated under dilute conditions to form intramolecular circles and subsequently transformed into *E. coli* (Fig. 23).

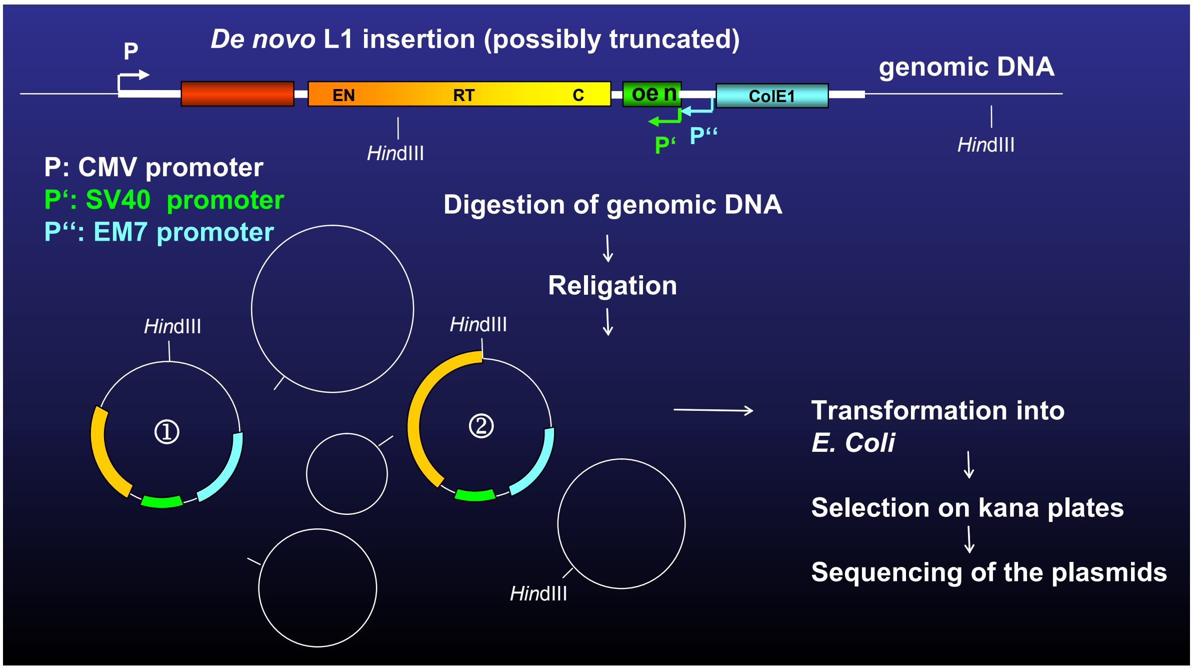


Fig. 23: Schematic drawing of the rescue procedure for integrants derived from pCEP4/L1.3*mneoI*$_{400}$/ColE1-based reporter constructs. Genomic DNA harbouring a tagged L1 integrant is digested with *Hin*dIII and religated under dilute conditions. The ColE1 origin of replication (ColE1 ori, blue bar) introduced downstream of the *neo*-cassette (green bar) converts the circular DNA bearing it into a replication-competent bacterial plasmid. The prokaryotic EM7 promoter (P’’) drives transcription and expression of the *neo*-gene in bacterial cells. Therefore, when the ligation products are transformed into *E. coli*, cells containing DNA derived from the 5’ end of a L1 integrant can be selected for on kanamycin plates. L1 copies that are truncated downstream of the L1-HindIII site can be recovered with both 3’ and 5’ flanking sequence (①), while integrants extending beyond that position merely allow identification of the 3’ junction (②).

As pCEP4/L1.3*mneoI*$_{400}$/ColE1 has been available only relatively recently, I only used it for generating the mutants with modified hairpin loops (pNZ73-76 and their control plasmids pNZ77 and pNZ82-85, Fig. 20). In the case of the highly active Tx1L hairpin chimera pNZ75, 23 independent insertion events were recovered, while the sporadic G418^R colonies obtained from plasmids pNZ73, 74 and 76 were not further analysed.

For the isolation of *de novo* integrants, the restriction enzyme *Hin*dIII was used. This enzyme cleaves the L1 element at position 3667. Therefore, only L1 copies that are 5' truncated downstream of the L1-*Hin*dIII cleavage site can be recovered with both 3'- and 5'-flanking sequences in one step (Fig. 23, ①). Integrants extending beyond that position result in rescued plasmids containing only the 3' half of the retrotransposon and its 3' flanking sequence (Fig. 23, ②).

Sequencing the 3' junctions with a primer binding to the *mneoI*/ColE1 cassette was not possible due to polymerase slippage on the poly(A) tail. The sequence located 5' of the L1 fragment had to be isolated first. As L1 integrants are variably truncated at their 5' end, the 5' junction in a rescued plasmid is localised between an L1 sequence of unknown length and a flanking sequence of unknown length. Therefore, a multiplex PCR with GS88 and a set of primers (GS260, GS261 and GS262) covering the L1 sequence 3' of the *Hin*dIII site was performed: based on the number of progressively longer PCR fragments, the length of the integrated L1 copy could be deduced. Using the corresponding primer (GS88, 17, 14, 16 or 76), pointing outward of the L1 sequence, the 5' junction of the new integrant was sequenced (see Fig. 22).

To obtain the remaining L1 sequence and the flanking genomic DNA sequence, strategies similar to the ones described in the previous chapter (3.3.3.1) were adopted. Pre-integration sites identified through homology searches in the HGWD were used to design oligonucleotide primers presumed to flank the retrotransposed L1 chimera. For characterisation of L1 integrants extending beyond the L1-specific *Hin*dIII site, the 5' primer was used in conjunction with GS76 to amplify the 5' flanking sequence by PCR from genomic DNA of cell lines harbouring the relevant insertions.

In this way, ten complete insertion events were characterised. In four cases, only the 3' flanking sequences were successfully recovered, while in four clones only the 5' flanks could be characterised. In five rescued plasmids, the L1 sequence ended with its *Hin*dIII site.

3.3.4 Characterisation of *de novo* retrotransposition events derived from chimeric L1 elements

In order to study the effects of the chimeric EN domains on target site specificity and on the mechanism of integration, the structures produced by retrotransposition events derived from hybrid elements were examined. I analysed pre- and post-integration sites of 55 retrotransposition events that were derived from the L1/R1α8 chimera pNZ45 (9 events), the L1/Tx1L EN hybrid pNZ39 (23 events) and the L1/Tx1Lα11 construct pNZ75 (23 events). Instead of describing each integrant individually, I will focus on several distinct features and characterise them summarily. The relevant data for each single clone are given in appendix C.

3.3.4.1 Chimeric *de novo* L1 integrants structurally resemble wild-type L1 elements

While element-encoded ENs have been implicated in target site selection of non-LTR retrotransposons and initiation of 'target primed reverse transcription' (TPRT), there is no indication for an additional role of the EN during retrotransposition. Therefore, modification of the EN in an otherwise unchanged L1 element was not expected to alter any structural features of the integrants. To test this assumption, I analysed retrotransposition events that were launched from the chimeric elements pNZ39, pNZ45 and pNZ75 for structural hallmarks like 5' truncations, inversions and poly(A) tails and compared them with wild-type elements present in the human genome (Szak *et al.*, 2002) and/or wild-type *de novo* L1 integrants derived from cell culture-based assays (Symer *et al.*, 2002; Gilbert *et al.*, 2002; Morrish *et al.*, 2002; Moran *et al.*, 1996; Moran *et al.*, 1999).

Of 55 isolated *de novo* integrants resulting from chimeric L1 retrotransposons generated in this study, 32 were sequenced completely, including their 3' and 5' junctions. Fig. 24 shows structures and extensions of these integrants.

Length distribution

Due to the need of retrotransposition of the *neo*-cassette, the minimal length of a retrotransposition event detectable with the retrotransposition assay is 1281 bp for pNZ45 and pNZ39 and 2329 bp in the case of pNZ75 (Fig. 24). As expected, all recovered chimeric integrants exceeded this length, but most of them were truncated within one kb upstream of the *neo*-cassette. Two clones (G1W6 and w3.1) harboured nearly full-length, but internally rearranged L1s. Examination of the DNA sequences at the inversion junctions revealed that "twin priming" might be responsible for their formation (Ostertag and Kazazian, 2001b). As a result of the isolation method, five pNZ75-derived integrants could only be sequenced up to the *Hin*dIII site (#5, 50, 54, 58, 60). These insertions are at least 4654 bp long; some of them might even represent full-length copies.

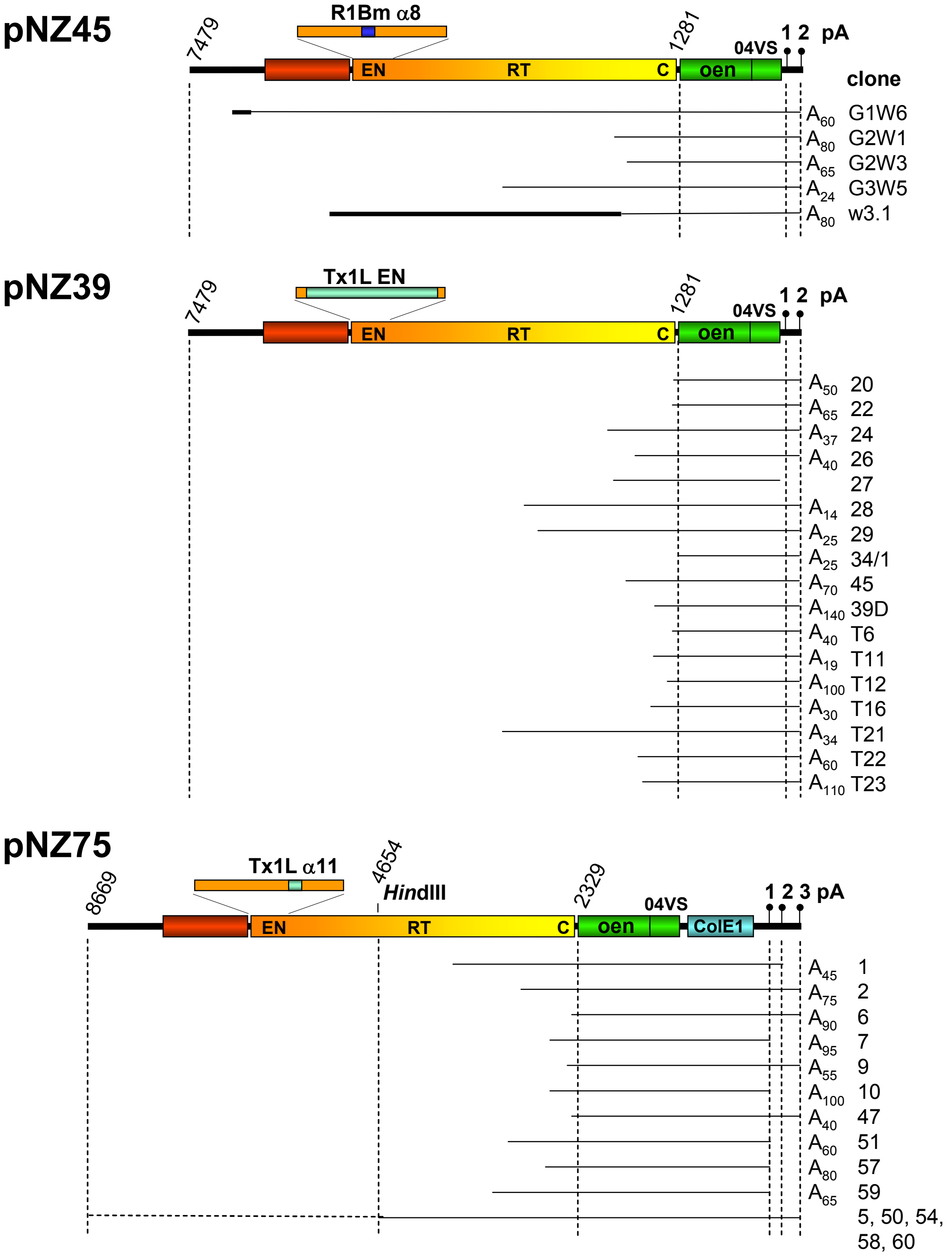


Fig. 24: Length distributions of retrotransposition events launched from the chimeric L1 reporter constructs pNZ45, pNZ39 and pNZ75. Schematics of the full-length reporters after retrotransposition are shown in colour. The relative positions of 5' UTR, ORF1, ORF2, the chimeric EN, the retrotransposed *mneoI* cassette - and the ColE1 ori in the case of pNZ75-derived integrants - are indicated. Lollipops represent the alternative polyadenylation signals present at the 3' ends of the reporter constructs (L1pA (1) and SV40pA (2) in pNZ45 and pNZ39 and SV40pA$_1$ (1), L1pA (2) and SV40pA$_2$ (3) in pNZ75). Lines of varying lengths below the schematics indicate the extensions of the retrotransposed elements. Bold black lines represent the relative sizes and and positions of inverted L1 fragments. The estimated size of the poly(A) tail and the names of the isolated clones are indicated at the right of the figure. The unknown extensions of the five integrants truncated at the *Hin*dIII site during the rescue procedure are indicated by a stippled line.

Considering the small sample size, this pattern of lengths is in good accordance with both the length distribution of pre-existing L1 elements in the human genome (Szak *et al.*, 2002) and *de novo* integrants from wild-type L1 elements (Symer *et al.*, 2002; Gilbert *et al.*, 2002). However, the fact that the average size of pNZ39-derived integrants is smaller than those of pNZ45 and pNZ75 could also reflect a phenomenon recently described by the Kazazian laboratory (Farley *et al.*, 2004): they found that the activity of a retrotransposon is correlated with the length of the resulting *de novo* insertions. Although the EN has not been specifically examined in this publication, the results do not exclude the contribution of EN activity to the frequency and extent of 5' truncations. A model explaining this putative connection will be proposed in the discussion (4.3).

Untemplated nucleotides

Seven examples of so-called "untemplated nucleotides" (several nucleotides whose origin is unclear) at the 5' junction between the chimeric L1 elements and genomic DNA were identified (Table 2).

		pNZ45	**pNZ39**	**pNZ75**
No. of 5' junctions sequenced		5	17	19
No. of integrants featuring extra nucleotides at the 5' junction		0	5	2
extra nucleotides	minimum length (bp)	-	2	1
	maximum length (bp)	-	16	1
	average length ± st. dev. (bp)	-	8.0 ± 5.8	1.0 ± 0.0

Table 2: Frequency and extension of untemplated nucleotide stretches at the 5' junction of *de novo* integrants

Such additional nucleotides were also found in earlier studies after *de novo* retrotransposition of wild-type L1 elements and it was suggested to use the term "extra nucleotides" since in some cases they might have been templated by other sequences (Symer *et al.*, 2002). With a maximum length of 16 nucleotides, however, the stretches of extra nucleotides described here were too short to identify any putative parental sequences. From the data summarised in Table 2, it might be inferred that the occurrence of extra nucleotides is inversely correlated with the activity of the respective EN. However, the frequent incidence of unknown bases at the 5' ends of integration events derived from the highly active wild-type L1.3 (11% of the integrants characterised in Symer *et al.*, [2002]) suggests that the observed bias is rather due to statistical fluctuations caused by the small sample size.

Poly(A) tails

Of 55 isolated *de novo* integrants derived from chimeric L1 retrotransposons, 3' junctions could be sequenced in 46 cases. The majority of the copies ended in a poly(A) tail, indicating that the integrants were indeed derived from an RNA intermediate. Only two copies of the Tx1L EN chimera pNZ39 were truncated at their 3' end and were therefore not polyadenylated (see below).

		pNZ45	pNZ39	pNZ75
No. of 3' junctions sequenced		10	23	16
No. of integrants ending in a poly(A) tail		10	21	16
poly(A) tails	minimum length (bp)	8	8	45
	maximum length (bp)	110	140	130
	average length ± st. dev. (bp)	61.7 ± 29.4	53.2 ± 36.5	80.9 ± 24.4

Table 3: Length distribution of poly(A) tail sizes. For A-stretches longer than approximately 20 bp, the size can only be estimated due to slippage of the polymerase in the sequencing reaction.

The poly(A) tails of integrants launched from the three chimeric constructs pNZ39, 45 and 75 (Fig. 24) displayed similar characteristics. On average, they were 50-80 bp long and varied in size from 8 to 140 bp. This is longer than the average length of poly(A) tails of endogenous L1 elements (18 ± 10 bp, Szak *et al.*, 2002), but consistent with data obtained for *de novo* L1 integrants from retrotransposition assays performed in other laboratories (88 ± 27 bp, Symer *et al.*, 2002, and ~60 bp, Gilbert *et al.*, 2002). These results could confirm the previous finding that the poly(A) tails of *Alu* elements and L1s become shorter with time after insertion into the genome (Arcot *et al.*, 1995; Ovchinnikov *et al.*, 2001). Alternatively, the long poly(A) tails could be attributed to the strong SV40 polyadenylation signal(s) present in the donor constructs (Fig. 24). With the exception of four cases, all integrants were polyadenylated at an SV40 polyadenylation signal (Fig. 24, Appendix C).

As mentioned above, two integrants derived from pNZ39 lacked a poly(A) tail. They were both characterised by a 3' truncation (52 bp [#27] and 6 bp [#34/2] upstream of the L1 polyadenylation signal) and a 3-bp overlap between their 3' end and the flanking genomic sequence. This 3' structure of retrotransposition events strongly resembles one previously described for endonuclease-independent integrations in XR-1 cells (Morrish *et al.*, 2002) and one obtained from *de novo* wild-type L1 retrotransposition (Gilbert *et al.*, 2002). Implications of these interesting data are discussed in 4.3.

Target site duplications and deletions

Endogenous human L1 elements are typically flanked by TSDs within the range of 7-21 bp (Szak *et al.*, 2002). TSDs generated after *de novo* retrotransposition of the chimeric elements displayed a much broader length distribution, ranging from three to 122 bp. Similarly extended TSDs were isolated in two parallel studies characterising a multitude of *de novo* integrants derived from wild-type L1s (Gilbert *et al.*, 2002; Symer *et al.*, 2002). Since the unusual TSD extensions observed in our studies are also found in *de novo* integrants derived from wild-type L1 elements it is reasonable to conclude that they are not the result of a chimeric EN.

		pNZ45	**pNZ39**	**pNZ75**
No. of integrants with both junctions sequenced		5	17	10
No. of integrants flanked by TSDs		5	13	6
No. of integrants causing target site deletions		0	4	4
TSDs	minimum length (bp)	3	3	3
	maximum length (bp)	20	122	71
	average length ± st. dev. (bp)	13.4 ± 6.3	49.7 ± 41.3	36.4 ± 27.1
target site deletions	minimum length (bp)	-	6	11
	maximum length (bp)	-	39	30
	average length ± st. dev. (bp)	-	18.8 ± 16.0	21.7 ± 9.7

Table 4: Effects of *de novo* integration of the chimeric retrotransposons on their genomic target sites

24 out of 32 fully characterised insertions were flanked by target site duplications. In the case of the remaining eight integrants we observed target site deletions of 6-39 nucleotides. Deletions at the integration site are not unusual since previous studies showed that wild-type L1 retrotransposition can result in the deletion of a number of nucleotides at the integration site (Kondo-Iida *et al.*, 1999; Narita *et al.*, 1993; Gilbert *et al.*, 2002; Symer *et al.*, 2002).

In a number of *de novo* integrants (23 out of 55), only one junction between the L1 sequence and its genomic target could be sequenced. Problems with the identification of the second L1 boundary might be attributed to large target site duplications or deletions. As explained in 3.3.3, the methods used for retrieval of integrants preferrably yield TSDs or target site deletions no longer than 100 bp. Thus the average size of both TSDs and target site deletions may be larger than indicated in Table 4.

3.3.4.2 Identification of a structural motif crucial for target sequence recognition of L1 EN

Due to the target primed reverse transcription (TPRT) mechanism initiating first-strand cDNA synthesis of L1 (see 1.3.2), knowledge of the sequence at the 3' junction is sufficient to identify the site where reverse transcription of the element's RNA started. This site is identical to the nicking site of the EN. The only uncertainty involved is caused by adenines at the target site: in the post-integration sequence, it is impossible to say whether a particular adenine is derived from the element's poly(A) tail or from a thymidine on the bottom strand of the target sequence. For the human L1 element, nicking is usually assumed to occur 3' of the ambiguous thymidine(s) on the bottom strand for two reasons. 1) Biochemical analysis of L1 EN cleavage specificity showed a preference for TpA sequences. Nicking between Ts or at an NpT site is rarely observed (Feng *et al.*, 1996; Cost and Boeke, 1998). 2) It has been postulated that thymidines on the partially melted bottom strand have to anneal to the RNA's poly(A) tail for successful TPRT (Jurka, 1997; Ostertag and Kazazian, 2001a).

However, both arguments apply only to L1 elements. Neither do Tx1L and R1Bm ENs preferentially nick TpA bonds, nor do these elements end in a poly(A) tail (see 1.5, p.24). But since the chimeric L1/Tx1L and L1/R1Bm reporter elements also end in poly(A) tails, I stuck to the convention when defining TSDs of L1/Tx1L and L1/R1Bm-derived integrants, and assigned any ambiguous As to the 3' part of the putative insertion site. If any bias is thus introduced into the sequence analysis, it is in favour of typical L1 nicking sites, setting stringent conditions for the detection of altered target site specificity of the chimeras.

To analyse the target site preferences of the chimeric constructs, primary genomic sequences of the unoccupied *de novo* target sites (presented as the plus-strand) were aligned so that four nucleotides to the left and 20 nucleotides to the right of the putative EN minus strand nicking site are shown (Fig. 25). The target sequences of integrants derived from pNZ45, pNZ39 and pNZ75 (Fig. 24, p. 83) were summarised in three sequence logos (Schneider and Stephens, 1990, http://ep.ebi.ac.uk/EP/SEQLOGO) and compared with the target sequences of *de novo* wild-type L1.3 integrants (Gilbert *et al.*, 2002) (Fig. 26).

pNZ45 (L1/R1Bm-α8-chimera):

L1 EN consensus target sequence:
`tt aaaa`

R1 EN nicking site:
`ccac tgtc`

clone	target sequence (5’→3’)	TSD
G1W6	gtat GAAATGTAAAATAAGicaca	yes (15)
G2W1	ttat aaGAGAATACTATGAATAat	yes (20 nt)
G3W5	atgt aGAAAACACAgaaaagagcg	yes (15 nt)
G2W3	cttt aaaaaggaagggaattctga	yes (3 nt)
w3.1	gaat GAAAACTAATGTTTiattgaa	yes (14 nt)
G1W4	ttgt aaaaacaaacggcgttgtct	n. d.
G3W3	aaat aaaaacttcacaaagtggtt	n. d.
G3W4/1	ccat aaaactatctaccaagaata	n. d.
G3W6	tctc aaaaaaaaaaaaaaaaaaaa	n. d.

pNZ39 (L1/Tx1L EN-chimera):

L1 EN consensus target sequence:
`tt aaaa`

Tx1L EN minor nicking site:
`taat gaaa`

Tx1L major nicking site:
`aact tcag`

clone	target sequence (5’→3’)	TSD
22	taat GAATGTTN$_{106}$ATTTTTTGic	yes (122 nt)
24	acag aaCCAAGN$_{20}$AGGAgggactg	yes (34 nt)
26	tctc aaaaaaaN$_{55}$GCAGGGAiagc	yes (91 nt)
28	caag aaaCCCTcaaaagtatttg	yes (10 nt)
29	ggac aaaaaaTN$_{89}$GAGTTTaatta	yes (106 nt)
34/1	cctc ataatcaaatgaccctaaat	yes (3 nt)
39D	taag aaGAAAAN$_{22}$ACCCTaataga	yes (36 nt)
T6	cccc aGATGAGN$_{48}$ACCCTTGiacc	yes (62 nt)
T11	agat aaaaTGAN$_{16}$TAATTAAigcc	yes (30 nt)
T16	ccat aaaaaatgatgagttcatgt	yes (18 nt)
T21	gggc aaaaGAAN$_{68}$TAGaagaaatt	yes (83 nt)
T22	aaac aaagagtttgtttgtaaaa	yes (5 nt)
T23	aact TAAAAATTACATAccacttc	yes (16 nt)
20	tctc attatattgcccaggctggt	no (Δ39)
27	aagt tgagtaattttaacactaaa	no (Δ17)
45	gttt aaaaaaaaaattagatgaaac	no (Δ6)
T12	tttt acttttttctgctattattc	no (Δ3)
21	tatg aaaaacaaatggaaaacatc	n. d.
34/2	gcat gttctcactcataggtggaa	n. d.
39E	tttt aaatatggaagtgtacatgt	n. d.
40	tctc aaaaaaaaaaaaaaaaaaaa	n. d.
T4	gtgt aaaaaataaaagagaaaatc	n. d.
T10	agcc aaaaaaaattaatatcaacc	n. d.

pNZ75 (L1/Tx1L-α11-chimera):

L1 EN consensus target sequence:
`tt aaaa`

Tx1L EN minor nicking site:
`taat gaaa`

Tx1L EN major nicking site:
`aact tcag`

clone	target sequence (5’→3’)	TSD
2	tttt aaaCAAATTTTAtagaggtg	yes (13 nt)
6	attt GAAATTCN$_{48}$CTCCcatcaac	yes (62 nt)
7	aaag AAGACATTTACGtggtcaac	yes (12 nt)
9	cctt AAAaattaaaaattactttg	yes (4 nt)
47	actt AGAAAAAN$_{40}$GTacatttgga	yes (54 nt)
57	acct AAGAAAATN$_{27}$TATTTCAigt	yes (41 nt)
1	ttag aagaattataagagacctaa	no(Δ9)
10	aaat aaaaaatatatatatataca	no(Δ18)
51	taag aagagaaagagagacctgag	no(Δ30)
59	gatt gaaagaatcaatataatgca	no(Δ10)
13	tgtg acaaatgcaatctcttatct	n. d.
73	acac agaaaatctcaaacttccca	n. d.
74	ctct aaaagttattttttactttt	n. d.
82	tttg aaaattttctacaataggca	n. d.

Fig. 25: Plus-strand sequences of the pre-integration sites of chimeric retrotransposition events derived from pNZ45, pNZ39 and pNZ75 (see Fig. 24). Gaps in the sequences indicate the bottom-

strand cleavage site. Blue shaded nucleotides are in agreement with the degenerate consensus sequence of the L1 EN cleavage site. Stretches of pyrimidines or purines that are 5’ or 3’ of the cleavage site are highlighted by gray shading. Minimal TSDs are indicated with underlined upper case letters, while microhomologies between insertions and flanking genomic sequences are shown as underlined lower case letters. Thus the largest possible TSD is underlined. Yellow i, insertion of untemplated nucleotides or inverted L1 segments.

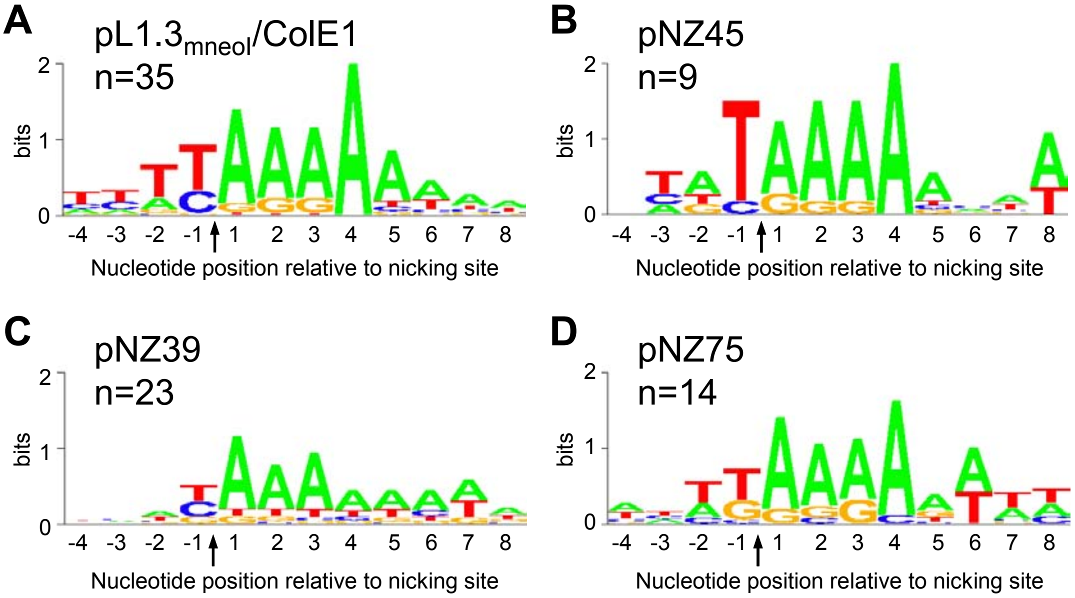


Fig. 26: Sequence logos of *de novo* integration sites of retrotransposition events launched from reporter plasmids pL1.3*mneoI*/ColE1 (wild-type L1, Gilbert *et al.*, 2002) and the reporter plasmids pNZ45, pNZ39 and pNZ75 (see Fig. 24). A sequence logo displays the frequencies of bases at each position as the relative height of letters, and the degree of sequence conservation as the total height of a stack of letters, measured in bits of information. Arrows indicate the nicking sites.

The sequences targeted by the L1/R1-α8-swap construct (pNZ45) did not differ from the consensus L1 EN cleavage site (Fig. 26A, B), indicating that the α8-swap did not affect target site specificity of the altered EN domain. The α8-swap is clearly not sufficient to modify EN specificity but instead results in a reduction of endonucleolytic activity to 5% which is reflected by a reduced transposition frequency (see 3.3.2.1).

In the case of pNZ39-derived integrants, the sequence logo (Fig. 26C) shows an overall decrease of specificity at every position in the consensus. When each integrant is analysed separately however (Fig. 25), it becomes clear that the deviations from the consensus are not introduced independently. Approximately 50% of the characterised elements (11 out of 23) have retained strict specificity for typical L1 insertion sites, while the target sequences of the remaining integrants typically differ from the L1 consensus 5’-T/AAA-3’ in two or more nucleotides.

This bimodal distribution may indicate the action of different, independent integration mechanisms. As my studies of the activity of pNZ39 and its various EN mutants had indicated that the chimeric L1/Tx1L EN is not nucleolytically active, the shift seen in target site selection does not stem from altered EN cleavage specificity. *Trans* complementation could lead to the insertions into L1-like sequences, while it is probable that the other *de novo* integrants resulted from endonuclease-independent retrotransposition, with integration occurring at pre-existing lesions in the DNA. It should be noted though that in three cases (#22, #T23, #27) the integration sites of the chimeric L1/Tx1L elements partially matched one of the two specific bottom-strand nicking sites of Tx1L EN (Christensen *et al.*, 2000) (Fig. 25). One of these nicking sites is identical to the integration site of the Tx1L retrotransposon in the *X. laevis* genome. It is therefore conceivable that the chimeric L1/Tx1L EN, although nucleolytically inactive, can still influence target site selection, e.g. by binding to pre-formed nicks resembling its own recognition sequence.

In this context, it is remarkable that the nucleotides downstream of the integration sites are usually adenines, even in target sequences deviating from the L1 consensus. Only 4 out of 23 integrants do not display an A in position +1 (#22, #T23, #27 and #34/2), and two of these are insertions lacking a poly(A) tail (#27 and #34/2, see Appendix C). This might indicate that complementary base pairing of the 3' end of the L1 RNA with a matching sequence of the nicked bottom strand DNA may be required for successful TPRT. That would explain the preference of polyadenylated L1/Tx1L EN hybrids for target sequences with poly(A) stretches. It implies that not only the specificity of EN but also complementarity of the target sequence with the 3' end of the retrotransposon RNA is crucial for integration specificity (see discussion, 4.3).

In contrast to the EN of pNZ39, the pNZ75-encoded L1/Tx1Lα11 EN has been shown to be active (3.3.2.4) and therefore responsible for target site selection. The sequences targeted by pNZ75 (Fig. 25, Fig. 26D) show similarity to the consensus L1 recognition sequence. The polypurine tract downstream of the nicking site is almost completely conserved. This may reflect either recognition of this sequence by the chimeric EN or, as discussed above, it may be a bias introduced by complementary base pairing during TPRT. However, at position -1 relative to the nicking site, a deviation from the L1 consensus is observed: in 36% of the integration sites, the usually strongly conserved pyrimidine is replaced with a G. As a 5'-GAAAA-3' sequence cannot form the typical kinked DNA structure recognised by wild-type L1 EN, this represents a considerable alteration in sequence specificity of the hybrid EN.

3.3.4.3 Chimeric L1 elements display frequent microhomologies between their 5' ends and adjacent TSD sequences

When analysing TSDs of *de novo* integrants, it was often difficult to define the extent of these duplications. Both at the 5' and at the 3' end of TSDs, ambiguous nucleotides (microhomologies) were frequently observed that could belong either to the genomic target or to the retrotransposon sequence (Fig. 25, underlined lower case letters; Fig. 27, coloured purple). According to the convention, TSDs were defined in a way that they comprised the maximum number of nucleotides (Fig. 27).

clone no.	no. of homologous nucleotides	TSD
pNZ45:		
G2W1	2	AAGAGAATACTATGAATAAT
G2W3	4	TAAA
G3W5	5	AGAAAACACAGAAAA
pNZ39:		
24	3	AACCAAGN_{20}AGGAGGG
28	3	AAACCCTCAA
29	2	AAAAAATAGAN_{86}GAGTTTAA
34/1	2	ATA
39D	2	AAGAAAAN_{22}ACCCTAA
T21	4	AAAAGGAN_{69}TAGAAGA
T22	3	AAA
T23	3	TAAAAATTACATACCA
pNZ75:		
2	1	AAACAAATTTTAT
6	3	GAAATTCN_{48}CTCCCAT
7	0	AAGACATTTACG
9	1	AAAA
47	5	AGAAAAAN_{40}GTACATT

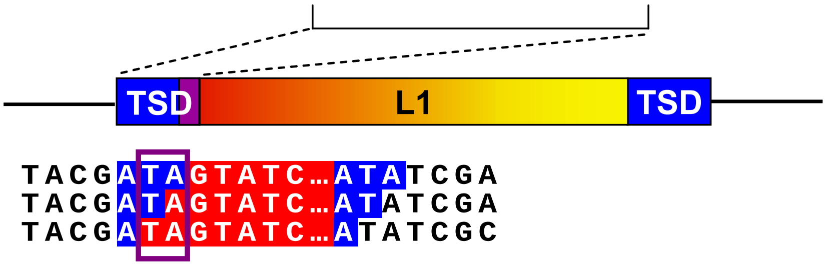


Fig. 27: 3' ends of most identified TSDs derived from chimeric L1 elements share one to five consecutive nucleotides with the 5' end of the adjacent *de novo* integrant. The schematic drawing and the nucleotide sequences below (derived from clone 34/1 as representative example) illustrate different interpretations of the observed microhomologies (purple box) as belonging either to the TSD (blue) or the L1 integrant (red).

At the TSDs' 5' ends, the ambiguous nucleotides were usually adenines (see 3.3.4.2, p.87). At the 3' ends, one to five overlapping nucleotides were observed in 94% of non-inverted integrants directly flanked by canonical TSDs. Inverted elements were excluded from analysis because they integrate by the mechanism of twin priming which has already been shown to cause microhomologies (Ostertag and Kazazian, 2001b). Insertion events harbouring untemplated nucleotides at the junction between L1 and its genomic target could not be analysed for microhomologies since sequence and extent of the untemplated insertions were unknown. For these reasons, only 16 canonical TSDs, derived from constructs pNZ45, pNZ39 and pNZ75, could be analysed (Table 5). Due to the small sample size, these data do not supply the statistical power to confirm the observed trend that microhomologies are longer and more frequent in the less active elements pNZ39 and pNZ45 than in the highly active construct pNZ75. However, taken together, the results provide convincing evidence for the importance of microhomologies in the integration process of the chimeric retrotransposons.

	pNZ45	**pNZ39**	**pNZ75**
No. of non-inverted integrants directly flanked by canonical TSDs	3	8	5
No. of integrants featuring microhomologies at their 5’ junction	3	8	4
homologies observed in	100 %	100 %	80 %

Table 5: Frequency of occurrence of microhomologies shared between the genomic target DNA and the 5’ end of *de novo* integrants.

3.4 Statistical Evaluation of the Microhomologies Observed at the 5’ Junctions of Endogenous, Pre-existing L1 Elements

Analysis of canonical TSDs flanking 5’ truncated *de novo* integrants revealed that in most cases, one to five contiguous nucleotides were shared between the 3’ end of a TSD and the 5’ end of its *de novo* integrant (Table 5). These microhomologies raised the possibility that a mechanism involving complementary base pairing might be responsible for their formation. The analysed *de novo* integration events resulted from L1 elements with chimeric EN domains, while similar microhomologies had not been reported before in pre-existing, endogenous L1 integrants. In order to evaluate whether the observed microhomologies were the result of chimeric EN activity or a general feature of L1 retrotransposition, a genome-wide analysis of TSDs from extant genomic L1 insertions was performed.

For this purpose I used a recently developed computer programme, TSDfinder, that is designed to precisely define retrotransposon boundaries (Szak *et al.*, 2002). Based on the RepeatMasker program (Smit and Green, unpubl.), TSDfinder identifies the location of repetitive sequences in the human genome and scans them for TSDs. In cooperation with U. Willhöft from the Zentrum für Bioinformatik (Universität Hamburg), TSDfinder was adapted to our purpose (see 2.6.3, p. 51). It was run against non-redundant human DNA-sequence contigs (NT_* records) assembled at NCBI by April 10, 2003 (build 33) constituting approximately 99% of the euchromatic genome. The programme identified 10,034 L1 insertions with an intact 3' end, a poly(A) tail and TSDs flanking the element. This data set was further refined with the help of algorithms developed by H.-P. Brose from the Institut für Medizinische Biometrie und Epidemiologie (Universitätsklinikum Hamburg-Eppendorf). All insertions that carried noncanonical features like inversions, untemplated nucleotides added to the 5' end of the L1 sequence, large internal insertions or deletions were discarded. Thus, a data set comprising 2724 truncated and 276 full length canonical L1 integrants was generated. Analysis of the TSDs of the truncated L1 insertions uncovered microhomologies between the 3' end of the TSD and the 5' end of their adjacent L1-insertion in 1583 out of 2724 cases (58%, Fig. 28). These regions of microcomplementarity comprised up to twelve consecutive nucleotides.

Subsequently, it was evaluated statistically whether the frequency of observed microhomologies was significantly higher than that of randomly occuring microhomologies. In cooperation with V. Schoder from the Institut für Medizinische Biometrie und Epidemiologie (Universitätsklinikum Hamburg-Eppendorf), we therefore applied a method similar to one described previously to evaluate viral/host junction sequences (Roth *et al.*, 1985). First, an unbiased base composition of the target sequences was assumed. It was used to calculate the distribution of microhomologies that would be expected to occur by chance alone (Fig. 28, white bars). In order to account for the insertion preference of L1 elements in A+T rich regions, the actual base composition of the flanking genomic sequences was determined in a 20-bp window surrounding each TSD. The resulting probabilities for the occurrence of each base were used to calculate a biased distribution pattern under the assumption of random integration (Fig. 28, grey bars). Statistical comparison of the two expected distributions with the observed distribution (Fig. 28, black bars) showed that for truncated L1 elements, the frequencies of contiguous overlapping nucleotides were significantly higher than expected by chance alone ($p<0.0001$ for both biased and unbiased base composition of the target sequences, Fig. 28A).

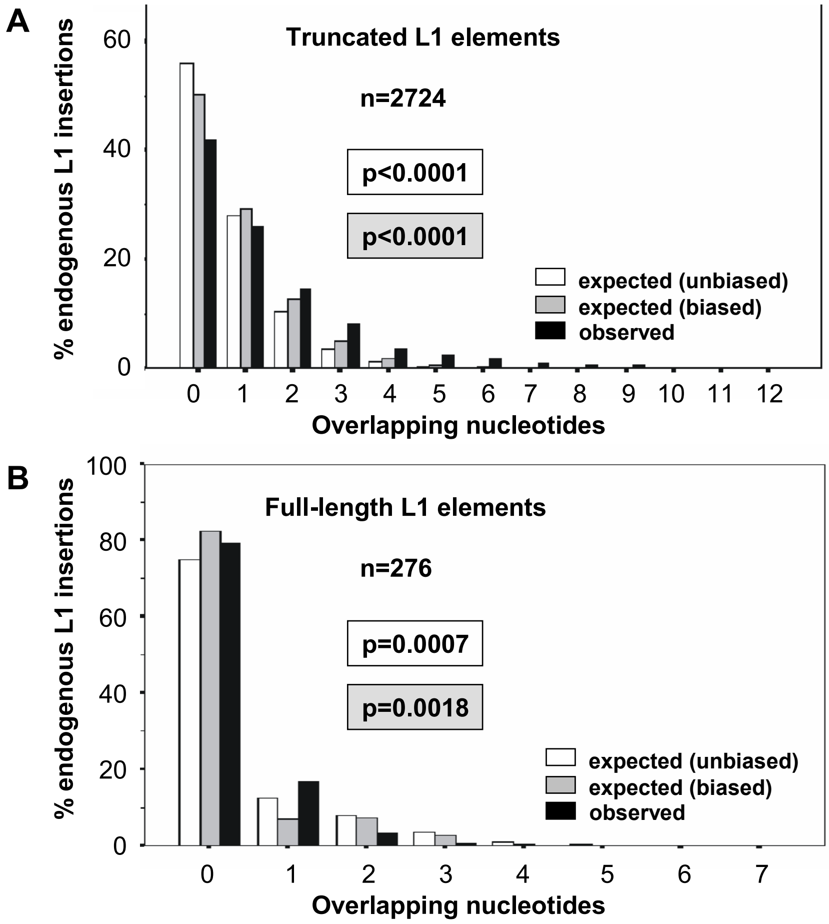


Fig. 28: Statistical analysis of homologies at the junctions between the 5’ ends of endogenous retrotransposons and the 3’ ends of their TSDs. Frequency of occurrence of microhomologies in **(A)** truncated L1 elements and **(B)** full-length L1 elements. Open bars represent the expected distribution, assuming an unbiased base composition at the integration site. Grey bars indicate the expected distribution after adjustment for the actual base composition of the L1 target sequences (biased). Base composition was calculated by determining the ATGC content of the 20 nucleotides flanking each TSD analysed. The occurence of the indicated numbers of overlapping nucleotides is represented as black bars. n, number of insertions analysed; open box, p-value for the comparison of the observed and expected (unbiased) distribution; grey box, p-value for the comparison of the observed and expected (biased) distribution.

Full-length L1 insertions were tested in a separate analysis, since the mathematical basis for calculation of the expected distribution of overlapping nucleotides differs between truncated and full-length insertions. As in full-length insertions, the sequence at one end of the junction is fixed, the distribution function shifts towards less microhomologies. Statistical analysis yielded the surprising result that the distribution of microhomologies in endogenous full-length integrants does not follow a clear trend. While insertions without overlapping nucleotides occur at the expected rate, one-nucleotide matches are more frequent than expected. In contrast, longer overlaps are less frequently found than expected by chance alone (p=0.0007 for unbiased and p=0.0018 for biased base composition of the target sequence, Fig. 28B). Although statistically significant due to the large sample size, this result may not be biologically relevant since no molecular mechanism is known that could select **against** random microhomologies. Thus it was concluded that integration of full-length L1 elements does not involve a microhomology-based mechanism.

A preference for microhomologies at the 5' end of truncated L1 elements is observed in endogenous L1 sequences (58%) as well as in *de novo* wild-type L1 integrants (68%, Symer *et al.*, 2002). However, this preference is most prominent in truncated L1 elements bearing chimeric ENs (94%). Therefore, microhomologies seem to be a general feature of truncated L1 elements, but modifications of the EN domain may support their formation. In contrast, endogenous full-length L1 elements do not show a significant number of overlapping nucleotides at their 5' end. The influence of EN mutations on the insertion pattern of full-length elements could not be evaluated because full-length integrants of the chimeric retrotransposons were not isolated.

The data imply a microhomology-related mechanism involved in the generation of truncated L1 elements. However, it cannot be concluded whether the mechanism is the cause or the consequence of truncation. In the discussion section (4.3), a comprehensive model is suggested to account for the phenomena described here.

4. DISCUSSION

In this study, the potential of non-LTR retrotransposons as gene therapy vectors was assessed by examining several aspects of L1 biology that have a crucial influence on their suitability as gene delivery systems. Investigation of the transcriptional regulation of L1 provided information on the long term activity of functional integrated retrotransposons. Systematic manipulation of the endonuclease domain of L1 served the purpose of identifying regions or structural motifs involved in target site recognition. Finally, analysis of pre- and postintegration sites yielded interesting insights into the replication mechanism of non-LTR retrotransposons.

4.1 Methyl-CpG-Binding Protein MeCP2 is a Major Repressor of L1 Retrotransposition

In order to explore the defence mechanism which represses L1 retrotransposition in somatic cells, we collaborated with Prof. Wolf Strätling's laboratory at the Universitätsklinikum Hamburg-Eppendorf. Several lines of evidence had previously indicated that the high methylation density of L1 elements plays a key role in their silencing (Yoder *et al.*, 1997; Thayer *et al.*, 1993). The three human methyl-CpG-binding proteins MeCP2, MBD1 and MBD2 were potential mediators of this transcriptional repression. Experiments with isolated transcriptional repression domains (TRDs) showed that only the TRDs of MeCP2 and MBD2, but not of MBD1 can efficiently repress L1-promoter-driven transcription (Fig. 11A, p.60). When examining the full-length MBD proteins, only MeCP2 was able to reduce transcription of L1 in a methylation dependent manner (Fig. 11B, Fig. 29). To my knowledge, this is the first description of selective repression caused by members of the MBD family. With solely *Hpa*II recognition sites being methylated, the experimental set-up does not reflect the *in vivo* situation which is characterised by many additional methylated CpG sequences (Woodcock *et al.*, 1997). Nevertheless, our approach for the first time provided direct experimental evidence that MBD proteins have different binding specificities and at least partially non-overlapping functions. As only MeCP2 could repress a moderately methylated L1 promoter, it might even have evolved as a specific regulator of retrotransposons.

L1 transcription (measured by L1 promoter-driven luciferase activity) and retrotransposition (measured by the retrotransposition frequency in the reporter assay) was also reduced by mere methylation of the reporter constructs, probably due to the action of abundant endogenous MeCP2 (Fig. 29). Moreover, overexpression of MeCP2 repressed transcription and transposition even of unmethylated reporter constructs. This might reflect either a weak affinity of MeCP2 to unmethylated DNA (Weitzel *et al.*, 1997) or secondary effects caused by MeCP2-mediated repression of endogenous proteins that influence L1 transcription. Interestingly, both L1 transcription and retrotransposition were repressed to the same degree in the different experimental set-ups (Fig. 29). This may suggest that L1 transcription is the rate-limiting step of retrotransposition.

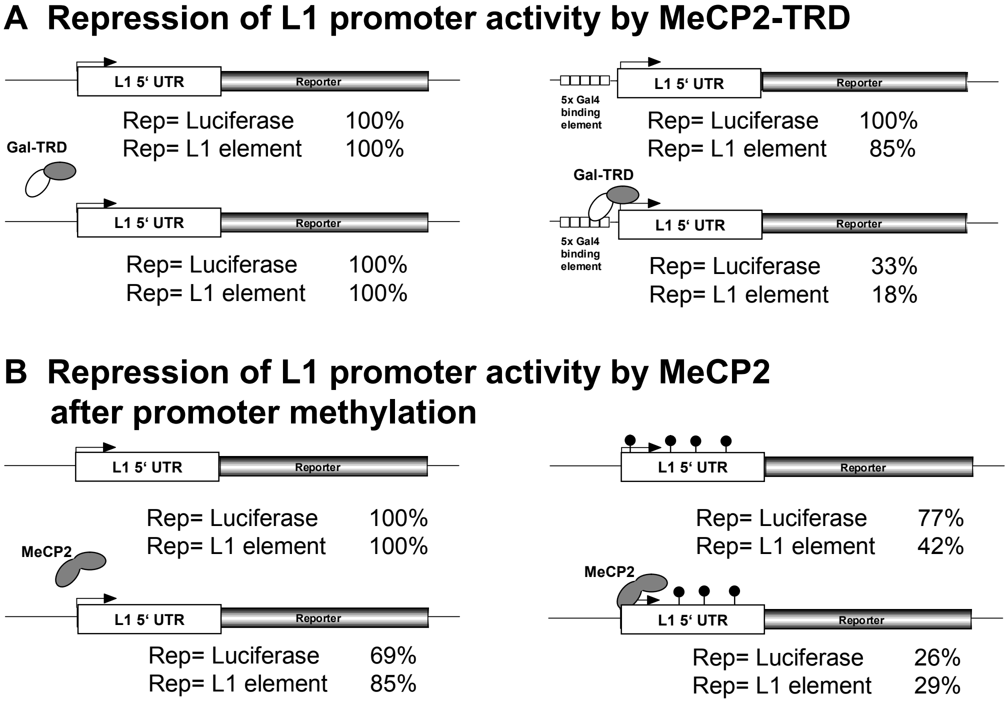


Fig. 29: The effects of MeCP2 on L1 promoter activity. (A) Repression of L1 promoter activity by MeCP2-TRD. **(B)** Repression of L1 promoter activity by MeCP2 after promoter methylation. White ovals represent the Gal4-DNA binding domain, grey ovals MeCP2-derived proteins. Methylated CpG sites are indicated as lollipops. Rep: reporter gene.

L1 transcription is strongly repressed in somatic cells, but the evolutionary genetics of L1 requires its expression and transposition in the germ line or at a very early stage of embryonic development (Ostertag and Kazazian, 2001a). The notion that endogenous L1 elements are repressed by MeCP2 implies an expression pattern of MeCP2 inverse to that of L1. Indeed, several studies at the mRNA as well as the protein level showed that MeCP2 is widely expressed in mammalian tissues, but almost undetectable in murine germ cells (Müller *et al.*, 2000; Reichwald *et al.*, 2000).

For the expression of a transgene in gene therapy, the L1 promoter could be a handicap. Similar to retroviral vectors (Bestor, 2000), retrotransposon-based vectors may be silenced rapidly and specifically by *de novo* methylation. Although expression of the transgene would be controlled by a separate promoter, repression of the L1 promoter might influence transgene expression due to long-range effects of methyl-CpG-mediated changes of chromatin structure (Richards and Elgin, 2002). However, regulation of L1 elements by the ubiquitous repressor MeCP2 via binding to 5-methylcytosines can also be advantageous with respect to the safety concern in gene therapy: When therapeutic DNA is transiently introduced into its target cell, it is usually unmethylated. Therefore, the L1 promoter in a possible retrotransposon-based vector is active in the beginning. New copies carrying the therapeutic gene can thus be launched from the vector with high frequency. As retrotransposons frequently undergo 5' truncation during the transposition process, the promoter will be deleted in most integrants, preventing further propagation of the vector. As a positive side-effect, the transgene will probably not be silenced to the same degree in 5' truncated integrants lacking the methylation-prone L1 promoter sequences as in the context of a complete L1. In rare cases where a full-length copy is transposed or the vector integrates via recombination, the process of *de novo* methylation will prevent re-mobilisation of this master copy in nearly all cell types. The only exception, the lack of methylation and MeCP2 in the germ line, is not an issue as genetic modification of germ line cells is not an aim of gene therapy due to ethical considerations.

In summary, L1 promoter repression by MeCP2 has both favourable and unfavourable effects on the suitability of L1 as gene delivery vector (safety vs. transgene silencing), but due to the process of 5' truncation, the advantages prevail.

4.2 Potential of the human L1 Retrotransposon as a Site-Specific Vector for Gene Delivery

If gene therapy should become a widely used and accepted method to cure diseases, it is indispensable that several prerequisites (sufficiently high expression level and cell-type specific long-term expression of the transgene) are fulfilled and side effects (immunogenicity, insertional mutagenesis) are reduced as much as possible (see 1.6, p.33). To date, retroviral and adenoviral vectors are the most commonly used vectors in gene therapy trials, accounting for 28% and 26% of the current trials (http://www.wiley.co.uk/genetherapy/clinical). Adenovirus-derived systems are well suited for transient expression of a therapeutic gene since they do not integrate into the host genome (Volpers and Kochanek, 2004). In contrast, long-term expression is most frequently attempted by the use of retroviral vectors which have many advantages. Most importantly, they show stable expression and/or transmission due to the fact that they integrate into the host genome, without causing significant chromosomal aberrations (Coffin, 1996). Integration occurs at predictable efficiencies with a predefined copy number (Kay *et al.*, 2001). Increasing knowledge regarding retroviral control elements has led to the development of novel vectors allowing quite precise adjustment of transgene expression levels (Baum *et al.*, 1996; Baum *et al.*, 1999). In addition, protocols for high-efficiency gene transfer have been developed for various target cells. For these reasons, retroviruses were the vectors used in the first successful gene therapy study: Nine children suffering from the X-linked "Severe combined immunodeficiency-X1" (SCID-X1) were subjected to an *ex vivo* infection of CD34+ cells with a replication-defective MoMuLV (Moloney Murine Leukemia Virus)-derived vector carrying a functional γ_c cytokine receptor subunit (Cavazzana-Calvo *et al.*, 2000; Hacein-Bey-Abina *et al.*, 2002).

However, the major drawback of retroviruses as gene shuttles is their propensity to integrate unspecifically throughout the genome. They even display a preference for active genes, i.e. open chromatin (Mooslehner *et al.*, 1990; Schröder *et al.*, 2002; Wu *et al.*, 2003). Although this had been discussed as a theoretical problem, numerous studies in mice had initially suggested that retrovirally transduced cells behaved normally *in vivo* (Anderson, 2000; Williams *et al.*, 2000). In 2002, Li and co-workers documented for the first time a case of leukemia caused by the integration of a retrovirally transduced transgene in mouse (Li *et al.*, 2002). One year later, follow-up examinations of the children cured from SCID-X1 by gene therapy showed that two out of nine children suffered from a proliferative disorder of their hematopoietic system caused by retroviral integration in proximity to the *LMO2* proto-oncogene promoter (Hacein-Bey-Abina *et al.*, 2003). Although both studies implied that the transgene itself contributed to the formation of the proliferative disorders (Hacein-Bey-Abina *et al.*, 2003; Baum *et al.*, 2003), these data underlined that insertional mutagenesis may represent a significant risk factor for any gene therapy approach based on retroviral vector insertion, and alternative vector systems are therefore urgently required (Baum and Fehse, 2003).

Several features make non-LTR retrotransposons interesting candidates as gene delivery vehicle. They are able to stably integrate into the genome at a controlled copy number and, as they cannot leave the cell, they are not infectious. Due to its endogenous nature, the human L1 element lacks proteins that are potentially immunogenic. Recent studies showed that L1 can retrotranspose from a chimeric adenoviral vector delivering marker genes to transformed cells in culture (Soifer *et al.*, 2001). A drawback that retrotransposons have in common with retroviruses is the infidelity of the reverse transcriptase, leading to a substantial mutation rate in the integrated transcripts (approx. one mutation in 10^4 bp) (Coffin, 1996; Baum *et al.*, 2003). However, in contrast to retroviral gene delivery systems, L1 elements show no integration preference for transcribed regions of the genome (Moran *et al.*, 1999; Gilbert *et al.*, 2002; Symer *et al.*, 2002). Several non-LTR retrotransposons even integrate site-specifically, thus minimising the risk of insertional mutagenesis.

The goal of the second part of this study was to utilise ENs from site-specific non-LTR retrotransposons closely related to the human L1 element to examine which protein regions are responsible for target site selection. I analysed the performance of newly generated chimeric L1 retrotransposons with part of L1 EN being replaced by corresponding polypeptides from closely related site-specific non-LTR retrotransposons.

4.2.1 L1 elements bearing chimeric ENs display all structural hallmarks of L1 retrotransposition

According to the experimentally verified 'target primed reverse transcription' (TPRT) mechanism, the EN domain of APE-type non-LTR retrotransposons is responsible for cleavage of the first strand of the target DNA and might even cleave the second strand (Cost *et al.*, 2002). However, AP-like ENs constitute a family of multifunctional enzymes, and any of their activities (see 1.4, p.22) could play an additional role in retrotransposition. Besides, some unknown function of retrotransposon-encoded AP-like ENs might influence the integration process.

Analysis of structural features of the integrants derived from actively transposing chimeric L1 elements with manipulated EN domains did not reveal any unforeseen effects attributable to novel EN functions. Neither did the lengths of the integrants significantly deviate from those obtained in similar studies with wild-type L1 elements (Symer *et al.*, 2002; Gilbert *et al.*, 2002) nor did the ratio of full-length, truncated and 'truncated and inverted' elements substantially change. Any differences observed here can be readily explained by statistical uncertainties due to the small sample size and a bias in the isolation method that favours characterisation of short, non-inverted insertions. Poly(A) tails of chimeric *de novo* integrants are 3.5 times longer on average than poly(A) tails of extant L1 elements. Long adenine tracts are discussed as common feature of new integrants that are later shortened by 'slippage' during replication (Ovchinnikov *et al.*, 2001). Alternatively, they might be a result of the strong ectopic SV40 polyadenylation signal present in the assay plasmid. In any case, they are not an effect of the altered EN domain since similar poly(A) sizes have been described for *de novo* wild-type L1 integrants before (Symer *et al.*, 2002; Gilbert *et al.*, 2002).

In summary, L1 elements with the chimeric ENs described in this study display all structural hallmarks of L1 retrotransposition, indicating that the described manipulation of the EN domain did not affect any mechanistic step other than target site recognition and cleavage.

4.2.2 R1Bm EN is not suitable to convey target-site specificity to the human L1 element

4.2.2.1 Swapping blocks of R1Bm EN into L1 leads to non-functional retrotransposons

Swapping the entire R1 EN coding region into a functional L1 element resulted in a chimeric L1/R1 hybrid element with no detectable retrotransposition frequency. Likewise, none of the three L1/R1 EN chimeras with increasing L1 EN moieties (Fig. 19) led to any transposition-competent element.

Perturbance of the functional conformation of the chimeric polyproteins might account for the significantly reduced retrotransposition frequency of the hybrid L1/R1 elements. Based on the assumption that R1Bm EN acts as part of a chimeric L1/R1-ORF2 polyprotein one could imagine that folding of the EN into an active conformation could be disturbed by its L1-ORF2p fusion partner. It is also possible that the successively increasing N-terminal L1 moieties of the hybrid enzymes and/or the C-terminal region comprising positions 230 to 239 of L1 hamper the correct folding of the R1Bm EN polypeptide.

In the swapping experiments, the R1Bm EN domain was separated from the larger R1Bm-ORF2 protein, while EN would naturally act as part of the polyprotein under wild-type conditions. Other regions of the intact ORF2p, the ORF1 protein and/or *B. mori*-specific host factors may also contribute to cleavage specificity and activity of the R1 EN either by contacting additional residues in the DNA or by affecting the structure of the EN domain. For example, ORF1p from murine L1Md was previously demonstrated to promote annealing of complementary DNA strands and to facilitate strand exchange to form the most stable RNA/DNA hybrids. Therefore it was suggested that ORF1p might play a role in TPRT (Martin and Bushman, 2001). It is questionable whether in the context of chimeric L1/R1 elements the corresponding L1-specific ORF1 and ORF2 proteins can completely compensate for the lack of other R1Bm-encoded proteins.

On the other hand, an increased target site specificity of the L1/R1 elements would reduce the number of possible integration sites and therefore also lead to a reduced retrotransposition frequency. Additionally, there is evidence that the specific enzymatic activity of R1 EN is only 1% of the L1 EN activity (Feng *et al.*, 1998). Even if R1 EN retained 100% of its endonucleolytic activity in the L1-context and there were as many possible R1 EN target sequences as there are L1 EN target sequences in the human genome, the retrotransposition rate of L1/R1 hybrid elements would be expected to drop to 1% of the L1 wild-type activity.

Targeting of the R1Bm specific integration site, the 28S rRNA gene, could lead to an additional problem: rDNA clusters are the constituents of nucleoli, distinct compartments in the nucleus of eukaryotic cells. They comprise the rRNA genes which are transcribed by RNA polymerase I. In contrast, the marker gene (*neo*) or a potential therapeutic gene is under control of a Pol II promoter (SV40 promoter in the case of the *mneoI* cassette). Since Pol II has been described to assume a random nucleoplasmid distribution with nucleolar exclusion (Singer and Green, 1997; Zeng *et al.*, 1997), it is theoretically conceivable that the L1/R1Bm chimera hits the R1Bm specific target, but the assay readout is hampered because the marker gene is not expressed.

4.2.2.2 Swapping R1Bm helix loops into L1 EN does not change target site specificity of the resulting hybrid elements

The helix-loop structures ‘α5’ and ‘α8’ of human APE1 are assumed to contact the major groove of the DNA and are therefore assumed to be involved in target site recognition. The ‘α8’-region was experimentally shown to be essential for AP target site recognition (Cal *et al.*, 1998).

Swapping R1Bm peptides corresponding to the ‘α5’ and ‘α8’ regions in APE1 resulted in three constructs (Fig. 19, p.72). While exchange of the ‘α5’-region inactivated the generated hybrid element pNZ44 and the double swap mutant pNZ47, replacing the ‘α8’-region yielded the functional chimeric L1/R1 element pNZ45. This element displayed a transposition rate that dropped to 5% of L1 wild-type activity. Analysis of the integration sites of the L1/R1α8 chimera (Fig. 26, p.89) demonstrated that it integrated almost exclusively at L1EN cleavage sites (Feng *et al.*, 1996; Cost and Boeke, 1998), indicating that the ‘α8-swap’ from R1 EN into L1 EN did not have any effect on target site specificity of the chimeric EN.

The most probable explanation for the transposition features of the three chimeric elements was that the replacement of the ‘α5-region’ hinders correct folding of the EN whereas the exchange of the ‘α8-region’ distorts the overall structure of the EN only little and does not lead to altered target site recognition. This hypothesis was later confirmed when the crystal structure of L1 EN was elucidated by Weichenrieder *et al.* (in press). Tracking the mutated peptides in the three-dimensional structure showed the principal but limited use of primary sequence alignments for the prediction of secondary structures. While the overall fold and the arrangement of the active site residues is conserved between L1 EN and APE1, slight differences are observed in the surface-loop regions. Therefore, the L1 sequences that, according to the sequence-based alignment, correspond to the α5- and α8-loops of APE1, are indeed located on the DNA binding surface of the enzyme, but they are both somewhat 'shifted' into the surrounding structural α-helices (Fig. 30A, exchanged loops are highlighted in cyan ['α5'] and blue ['α8']). The 'α5-region' partly enters the core of the protein (Fig. 30B), making misfolding caused by steric clashes of mismatched residues all the more likely. Alternatively, replacement of the 'α5-region' could abolish DNA binding. The 'α8-region' stays on the protein surface (Fig. 30B) and thus does not interfere with folding, but as it is rather remote from the active centre and probably contacts the DNA substrate only peripherally, it does not influence target site recognition.

In summary, modification of the loop region 'α5' inactivated the resulting hybrid ENs, while exchange of the L1α8 loop for its corresponding R1Bm sequence only impaired the activity of the chimeric EN, but had no effect on target site selection.

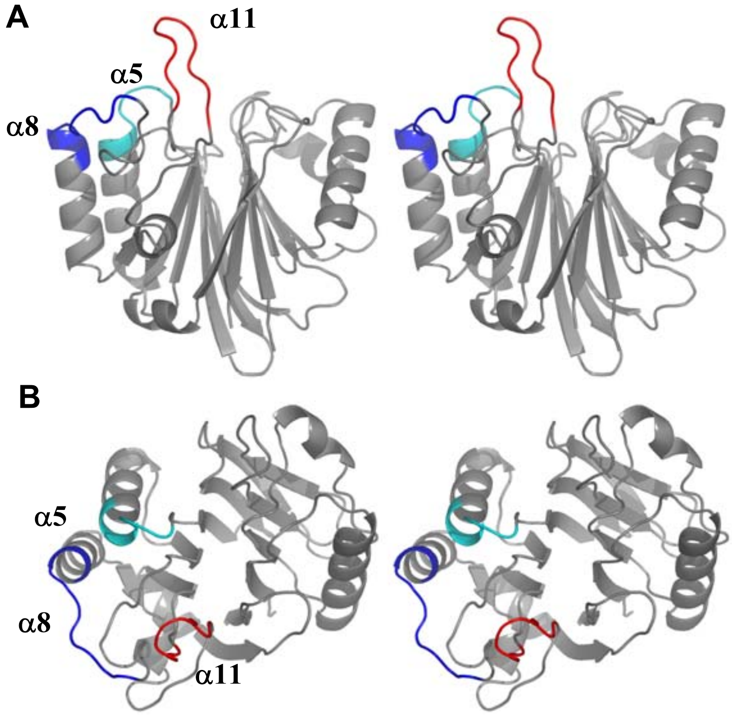


Fig. 30: Ribbon representation of the crystal structure of L1 EN in stereo (kindly supplied by O. Weichenrieder, The Netherlands Cancer Institute) The three loop regions α5, α8 and α11 are coloured cyan, blue and red. **(A)** side view; the DNA binding surface is on top; **(B)** top view directly onto the catalytic centre.

4.2.3 Tx1L EN is more compatible with L1 proteins than R1Bm EN

4.2.3.1 The L1/Tx1L EN hybrid retrotransposes in an apparently EN-independent manner

Replacing the EN domain of a functional human L1 element with the corresponding domain of the closely related site-specific non-LTR retrotransposon Tx1L from *X. laevis* resulted in an element (pNZ39, Fig. 17, p.68) with a greatly reduced retrotransposition frequency (0.14% of wild-type L1 levels). Although very low, this frequency reproducibly exceeded the residual background activities measured for the negative control pJM105 (0.04% of wild-type L1 activity) and the inactive R1Bm EN chimeras (see Fig. 17, p. 68). However, five mutated derivatives of pNZ39 containing either one or several missense mutations of residues described as essential for catalytic activity (Gorman *et al.*, 1997; Moran *et al.*, 1996) did not show a reduced retrotransposition frequency. These results might be explained by the possibility that all mutant ENs are less active than the pNZ39-EN, but still retain some residual activity. If nicking of the target site is not the rate-limiting step of retrotransposition, slow nicking by the mutants would yield the same outcome as the more efficient cleavage by the unmutated L1/Tx1L EN. However, this theory is largely based on improbable assumptions, and it is far more likely that the chimeric L1/Tx1L EN, as well as its mutated versions, are nucleolytically inactive. Basically the same reasons that were discussed above (4.2.2.1) as being responsible for the complete inactivation of the L1/R1Bm hybrids could also be the cause for inactivation of the L1/Tx1L EN. However, this theory raises the question how the G418^R colonies exceeding the background level were generated. In the following paragraphs, three alternative models on how the inactive L1/Tx1L EN may promote colony formation are discussed.

a) Trans complementation of EN-deficiency by the activity of functional endogenous L1 elements is the most obvious explanation for retrotransposition in spite of impaired target site cleavage. Wei *et al.* showed that retrotransposition of EN-deficient as well as RT-deficient L1 elements could be effectively rescued by co-expression of a replication-competent L1 (Wei *et al.*, 2001). This effect, which yielded retrotransposition frequencies of 0.2-0.9% relative to wild-type L1, could only be seen with simultaneous overexpression of a complementing functional element. Endogenous elements, which could be the only elements acting in *trans* in the experiments reported here, did not lead to significant colony formation (0.002-0.05% of wild-type activity) (Wei *et al.*, 2001). Besides, if retrotransposition by *trans* complementation played a role in our transient retrotransposition assays, we would expect to observe it also in the case of the reporter constructs pNZ33, pNZ35, pNZ36 and pNZ37 (Fig. 16, p. 66), since all constructs were tested in parallel under identical conditions. Given that four different chimeric L1/R1 elements did not cause any G418^R colony, it was concluded that *trans* complementation events by endogenous LINE elements of HeLa cells take place at a negligibly small rate. Still, pNZ39 could be a preferred target for *trans* complementation by persuing a strategy analogous to the one described for *Alu* elements (Boeke, 1997; Dewannieux *et al.*, 2003): By forming a specific RNA structure mimicking 7SL RNA, pNZ39 RNA could recruit endogenous nascent L1 proteins. However, this mechanism should also apply for pNZ49, the RT mutant of pNZ39. Interestingly pNZ49 proved to be inactive, arguing against this hypothesis.

b) Integration into pre-existing nicks of the genomic DNA offers an alternative explanation for the unusual behaviour of pNZ39. Endonuclease-independent L1 retrotransposition has been observed at near-wild-type levels in CHO cell lines deficient in nonhomologous end joining (NHEJ), but not in the NHEJ-competent parental CHO cells or in HeLa cells (Morrish *et al.*, 2002). To explain why pNZ39 shows an elevated retrotransposition rate compared to L1/R1Bm hybrids, it was postulated that the chimeric L1/Tx1L EN could be structurally contorted so that it is still capable of binding, but unable to cleave DNA. Thus, it might be possible that the EN domain recruits the pNZ39 ribonucleoprotein-complex to pre-existing nicks in the genome, which could then be used as primers for TPRT. The attempt to test this hypothesis by providing an increased number of potential targets through introduction of single strand breaks in the HeLa genome by hydrogen peroxide (Dahm-Daphi *et al.*, 2000) yielded inconclusive results (see 3.3.2.2), since the oxidative reagent had the same 2.5-fold stimulatory effect on retrotransposition rates of pNZ39, its RT mutant pNZ49 and the wild-type L1 element encoded on pJM101/L1.3. Enhanced transcriptional activity as a response to irradiation has been described for several transposable elements including L1 (Strand and McDonald, 1985; Morawetz, 1987; Servomaa and Rytömaa, 1990), and it is also observed after oxidative stress (G. Tolstonog, Heinrich-Pette-Institut, personal communication). It was probably this effect that led to the upregulation of expression of the reporter constructs, and/or to massive over-expression of endogenous elements, resulting in *trans* complemention. Any potential effect of the introduction of single strand breaks would have been masked by the general increase of retrotransposition. In order to unambiguously address this question, the experiments should be repeated with an agent inducing single strand breaks without inducing transcription of L1 (e.g. bleomycin or camptothecin [Liu, 1989]).

c) A third explanation is based on an unexpected observation by Cost *et al.* (2002). In a study analysing *in vitro* TPRT of the human L1 element, they observed a second endonuclease activity of ORF2p. This activity has not been characterised very thoroughly, yet there is some indication that it requires a form of EN-RT domain cooperation (Cost *et al.*, 2002). Such a cryptic activity in the L1/Tx1L chimera could explain both the low retrotransposition rate of construct pNZ39 and its lack of sensitivity to point mutations introduced at the main catalytic site.

While *trans* complementation (*a*) should lead to integration of pNZ39 derived copies into canonical L1 target sequences, utilisation of pre-existing nicks (*b*) or a cryptic EN activity (*c*) would lead to a different distribution of *de novo* integrants in the genome. Therefore, the integration sites of 23 integrants of the L1/Tx1L hybrid pNZ39 were analysed. In 11 cases, the target sequences matched the consensus 5'-T/AAAA-3' of L1 EN, three copies integrated into sequences resembling Tx1L *in vitro* nick sites, and nine integrants were located in genomic sequences dissimilar to both L1 and Tx1L EN recognition sites. This result allowed no decision for or against any of the hypotheses mentioned above. While the L1-like integration sites point to *trans* complementation, several integrants exhibited features typical of EN-independent retrotransposition events like 3' truncations or target site deletions. On the other hand, integration of some copies into Tx1L-specific sites indicates that the chimeric EN has maintained its capability of sequence-specific DNA binding. It is conceivable that a combination of two or even all of these mechanisms led to the formation of G418^R colonies derived from the L1/Tx1L EN hybrid pNZ39, thus setting it off from the completely inactive L1/R1Bm chimeras.

4.2.3.2 The Tx1Lα11 hairpin chimera shows altered target site recognition

When this project was started, the DNA binding 'α11' loop of human APE1 was not assumed to be crucial for sequence recognition as it contacts the DNA duplex in the minor groove (Gorman *et al.*, 1997), while DNA sequences are most easily distinguished in the major groove (Travers, 1993). The corresponding loop of L1 EN (highlighted in red in Fig. 30), however, is a much more prominent hairpin loop rigidified by multiple internal hydrogen bonds. It protrudes from the putative DNA binding surface of L1 EN and is anchored within the active site cleft (Weichenrieder *et al.*, in press). This loop probably recognises the kinked structure (Cost and Boeke, 1998) of the L1 target-sequence $Py_n{\uparrow}Pu_n$ and bends the DNA in the correct conformation for nicking. A structure-based alignment of the EN domains encoded by L1, Tx1L and R1Bm revealed that the other retrotransposon-encoded ENs share the threonine and serine residues that anchor the loop at its base, but the intervening sequences differ considerably. Modification of the ‘α11’-region thus promised to influence the nicking specificity of the resulting mutated ENs. A drastic change of conformation was not expected because interaction between the hairpin loop and the rest of the enzyme is minimal.

Deletion of the hairpin structure α11 (pNZ73, Fig. 20, p.74) and replacement of the L1 loop with the R1Bm loop (pNZ76) resulted in a drop of the retrotransposition rate by 97% and 94%, respectively. By introducing two point mutations (D145A and N147A in L1.3) into the modified EN domains, these activities were not further reduced, implying that the loop modifications inactivated the resulting ENs. In contrast, the chimeric L1 element containing the Tx1L hairpin loop instead of the L1-specific sequence (pNZ75) exhibited 25% wild-type L1 activity. This retrotransposition frequency is dependent on EN functionality, since mutation of the critical amino acids D145 and N147 reduced the activity of the resulting reporter construct to 3.5% of wild-type activity, comparable to the constructs pNZ73 and pNZ76 and their EN-deficient controls. The higher retrotransposition frequency of the L1/Tx1Lα11 chimera relative to the R1Bmα11 hybrid could be attributed to the phylogenetic relationship of L1, Tx1L and R1Bm. The overall primary sequence homology of 24.6% between the EN domains from L1 and Tx1L versus 17.3% in the case of L1 and R1 suggests that Tx1L EN is more compatible with L1 gene products than R1 EN.

Only the highly active Tx1L hairpin swap pNZ75 was further examined for alterations in target-site specificity by sequence analysis of the insertion sites. Although the nicking preference of the chimeric L1 element was not converted to recognition of sequences resembling the Tx1L-specific insertion site, a major influence of the hairpin modification on target site specificity was perceived: the requirement for a kinked DNA structure formed by a stretch of pyrimidines followed by several purines, which is almost compulsory for L1 target sequences, was alleviated in insertion sites of pNZ75. Although analysis of *de novo* integration sites of the chimeric retrotransposon revealed that the purine-stretch was conserved (probably due to the necessity for basepairing during TPRT, see 4.2.5 and 4.3), in 36% of the cases analysed, a guanosine was observed at position -1 that is usually occupied by pyrimidines (Fig. 25, p.88). Thus it was concluded that modification of the α11 hairpin loop of L1 EN influences sequence recognition at the most crucial position directly adjacent to the nicking site. It should be kept in mind that, as mentioned in 3.3.4.2 (p.87), adenines at the nicking site are by convention ascribed to the 3’ part of the putative nicking site. Therefore, it is conceivable that several integrants of pNZ75 are falsely identified as inserted into 5’-T↑AAAA-3’, but in fact inserted 3’ of an A (e.g. 5’-TA↑AAA-3). This would imply that the L1/Tx1Lα11 hybrid EN tolerates all nucleotides in position -1 of the nicking site.

These results show that the α11 hairpin loop of L1 EN is required for nucleolytic activity, and that it has an impact on target sequence selection. Careful modification of its amino acid sequence might allow to manipulate sequence recognition of L1 EN even further.

4.2.4 Is the EN domain the only determinant of target site specificity?

The results obtained from the various EN swapping-experiments performed in this study suggest that manipulation of the endonuclease domain alone is not sufficient to convert target site specificity entirely. However, Takahashi et al. (2002) presented data suggesting that the EN is the only determinant of target-site specificity.

This group concentrated their efforts on TRAS and SART, two families of telomeric repeat-specific non-LTR retrotransposons that coexist in various insect species (Okazaki *et al.*, 1995; Takahashi *et al.*, 1997; Takahashi and Fujiwara, 1999; Kubo *et al.*, 2001; Kojima and Fujiwara, 2003). The two elements insert at specific but different nucleotide positions in opposite orientation into the telomeric repeats of the same host organism (Okazaki *et al.*, 1993; Sasaki and Fujiwara, 2000).

By applying a novel retrotransposition assay, it was demonstrated that both SART1 and TRAS1 from *B. mor*i are capable of *in vivo* retrotransposition in *S. frugiperda* cells (Takahashi and Fujiwara, 2002). In order to answer the question whether the AP-like EN domain is responsible and sufficient for site-specific retrotransposition, the TRAS1 EN domain was swapped into a functional SART1 element and *in vivo* retrotransposition of the resulting chimeric element was characterised (Takahashi and Fujiwara, 2002). The target-site specificity of the modified SART1 element encoding the TRAS1 EN domain was completely converted to that of TRAS1.

This result represents convincing evidence that the EN domain is the **primary** determinant of target site selection. However, several facts argue against the notion that the EN is the **only** determinant of targeting specificity even in the case of TRAS and SART. Purified TRAS EN could generate specific nicks on both strands of the telomeric repeat sequence between T and A of the $(TT{\uparrow}AGG)_n$ bottom strand and between C and T of the $(CC^{\downarrow}TAA)_n$ top strand *in vitro*. These sites are consistent with insertion sites expected from the genomic structure of boundary regions of TRAS1. Still, with 10 bp representing the minimal structure to ensure endonucleolytic activity of TRAS1 EN (Anzai *et al.*, 2001), the EN domain does not exhibit sufficient target-site specificity to explain the exclusive localisation of TRAS1 in telomeres. Besides, insertions of both elements do not occur within 6 to 8 kb of the extreme end of the chromosome, despite the abundance of suitable DNA targets in these regions (Takahashi *et al.*, 1997).

The general location of the insertions is probably defined by chromatin structure or protein-protein interactions with the telosome, a complex of telomere-associated proteins. This may explain the success of the TRAS/SART swapping experiments. TRAS1 and SART1 are both telomere-specific elements, and besides, they are phylogenetically closely related. The presence of possible host factors participating in target-site selection is almost guaranteed since the experiments were performed in cells from *S. frugiperda* which belongs, like *B. mori*, to the order *Lepidoptera.* Targeting of TRAS1 and SART1 to the telomeres might therefore be achieved by a mechanism that is common to both elements (see below, '*Myb*-like domain'). Once localised to the chromosome ends, the EN domains have only a very limited choice of targets in the form of telomeric repeats. Thus, their imperfect cleavage specificity is sufficient to perform the fine-tuning by determining the exact insertion position (Takahashi and Fujiwara, 2002).

The aforementioned speculations are corroborated by *in vitro* analyses of the enzymatic activities of other site-specific AP-like ENs encoded by R1Bm, Tx1L and Tx2L. These biochemically characterised ENs all display a distinct selectivity for the target sequences expected from the TSD structures of their encoding elements. However, none of the ENs shows sufficient specificity to explain the restricted distribution of the respective elements in their host genomes (Feng *et al.*, 1998; Christensen *et al.*, 2000; Christensen *et al.*, 2001). This strongly suggests that additional determinants are necessary for targeted integration *in vivo*. One or more of the following candidate factors could play a role in site-specific integration:

Myb-like domain

By using a secondary structure prediction program, a three-helix-motif located between the EN and RT domain of TRAS elements was recently identified, which is typical of the transcriptional activator *c-myb* (Kubo *et al.*, 2001). Similar putative *myb*-like domains were found in the APE-type retrotransposons R1Bm, SART1, RT1Ag, TARTDm, and L1Hs (Kubo *et al.*, 2001). Binding of *c-myb* to the specific DNA sequence 5'-AACNG-3' is achieved by the cooperative action of at least two three-helix-bundles that can recognise the target sequence (Ogata *et al.*, 1994). Notably, many telomere binding proteins like RAP1, TAZ1, TRF1 and TRF2 share a limited amino acid similarity consisting of a *myb*-like three-helix-motif (König *et al.*, 1998). Thus the identified *myb*-like domain in retrotransposons might be involved in target-site recognition. In the case of the aforementioned elements TRAS1 and SART1 from *B. mori*, it was suggested that their *myb*-like domains might be responsible for the general targeting of each element to the telomeric regions (Takahashi and Fujiwara, 2002).

Cysteine-rich motifs

Cysteine-histidine motifs encoded by ORF1 and ORF2 of many APE-type retrotransposons are still poorly characterised. With a few exceptions, all elements code for an ORF1 protein which carries one to three CCHC-motifs of the consensus sequence $CX_2CX_4HX_4C$ that is also present in retroviral Gag proteins (Covey, 1986; Zingler *et al.*, in press). In retroviruses, this zinc knuckle region is implicated in binding retroviral RNA and in contributing to the interactions between Gag monomers (Gorelick *et al.*, 1999; Tanchou *et al.*, 1998).

ORF2-encoded proteins of many LINE-like elements carry at least one CCHC-motif at their carboxy-terminal end (mostly $CX_{1-3}CX_{7-8}HX_4C$ [Kajikawa *et al.*, 1997] or $CX_2CX_{12}HX_{3-5}H$ [Martín *et al.*, 1995]). Missense mutations within this motif in human L1 and TRAS1 rendered the resulting mutant retrotransposons inactive (Moran *et al.*, 1996; Takahashi and Fujiwara, 2002), indicating that the CCHC motif is essential for retrotransposition. However, some elements lack a Zn-finger domain in ORF2p and are still active (Kajikawa *et al.*, 1997 and references therein), or even integrate site-specifically (e. g. TART) (Danilevskaya *et al.*, 1994).

The function of the cysteine-rich region of ORF2 has not been elucidated yet, but in general, it is assumed that it interacts with the RNA-template and/or the genomic target-DNA. Nevertheless, the presence of a Zn-finger-like motif does not necessarily indicate interaction with nucleic acids: Zinc domains have also been implicated in protein-protein-interaction (Berg and Shi, 1996; Grishin, 2001). Thus the cysteine-rich region of ORF2p could also influence retrotransposition by co-factor binding.

ORF1 protein

Although ORF1p is clearly indispensable for the activity of APE-type elements (Moran *et al.*, 1996), this protein is much less understood than the functions of ORF2p. ORF1p of the human L1 element has been shown to form a ribonucleoprotein complex with L1 RNA (Hohjoh and Singer, 1996; Hohjoh and Singer, 1997), and ORF1p of the mouse LINE-1 was demonstrated to have nucleic acid chaperone activity *in vitro* (Martin and Bushman, 2001). However, in the two telomere-specific *Drosophila* elements TART and HeT-A, a very peculiar function of ORF1p (Gag) has been described recently. Both Gag proteins were shown to move into the nucleus efficiently, and HeT-A Gag even localises to characteristic Het dots that are preferentially associated with chromosome ends. In contrast, the ORF1 proteins of non-telomere-specific elements like *Doc, Jockey* and *I* stay in the cytosol (Rashkova *et al.*, 2002). Thus, TART and HeT-A are the only elements described so far with an ORF1p involved in intracellular targeting. Since ORF1p was demonstrated to be localised in Het dots, it might even contribute to target-site specificity (Rashkova *et al.*, 2003; Rashkova *et al.*, 2002).

Effects of chromatin

Taking into account that *in vivo,* genomic DNA is assembled as chromatin with many associated factors, other domains of retrotransposon proteins might be involved in target site selection through interaction with host chromatin proteins, as has been demonstrated for LTR retrotransposons Ty3 and Ty5 (Kirchner *et al.*, 1995; Xie *et al.*, 2001; Zhu *et al.*, 2003). Also, the macroscale distribution of retrotransposons in the genome is likely to depend on the accessibility of the chromosome to the transposition machinery. It was found that nucleosomal wrapping of DNA renders it a less efficiently nicked substrate, but when so wrapped some phosphodiesters at specific positions in the nucleosome are nicked at an increased rate (Cost *et al.*, 2001).

Spatial configuration of DNA

The effects of spatial configuration of the target DNA on target-site selection was studied by means of human L1 EN. It was shown that L1 EN target-site selection has its basis in the recognition of the unusual structural properties of the homopolymeric sequences T_nA_n and the junction formed between them (Cost and Boeke, 1998). Minor groove width was found to be an important factor for binding/cleavage by L1 EN. The TpA-junction of T_nA_n-tracts normally has a wide minor groove as a consequence of local sequence-dependent unwinding of the helix. When the substrate was further unwound, L1 EN activity increased (Cost and Boeke, 1998). This phenomenon may be relevant *in vivo*, as it was suggested that the genome is divided into torsionally constrained and differentially supercoiled segments (Kramer and Sinden, 1997 and references cited therein). Although poorly characterised, these regions may affect L1 element targeting by providing alternately favourable or poor substrates for L1 EN.

Host-encoded factors

Since tagged L1 elements localised on an episomal plasmid retrotranspose in some cell lines quite efficiently (HeLa, HCT116), but do not in others (Moran *et al.*, 1996; Symer *et al.*, 2002; Ostertag *et al.*, 2000; Ostertag and Kazazian, 2001a), it was concluded that HeLa cells express host factors that are essential for L1 retrotransposition. Host factors could contribute to cleavage specificity of the ENs either by contacting additional residues in the DNA or by affecting the structure of the EN domain or the entire ORF2p.

Retrotransposon-derived mRNA

The presence of element RNA and its interaction with ORF2p could lead to conformational changes of the polyprotein influencing cleavage-site specificity of the EN. Additionally, in order to initiate TPRT, the 3' end of the element's RNA has to form a primer-template complex with the 3' end of the nicked DNA strand, which is then extended by the RT-activity to form an RNA/DNA-hybrid. Complementary base pairing between the 3' end sequence of the RNA and the DNA target supports the formation of this primer-template complex (Feng *et al.*, 1998). Since the retrotransposon CR1 in chicken has been shown to preferentially integrate into sequences resembling its 3' repetitive sequence, it was suggested that the 3' end sequence of the element-encoded RNA could play a role in target site selection by hybridising to homologous sequences at nicked chromosomal sites (Burch *et al.*, 1993). This is supported by the observation that the genomic target sequence of Rex3 from *X. maculatus* also shows similarity to its $(GATG)_n$ 3' region (Volff *et al.*, 1999).

For the RE-type non-LTR retrotransposon R2Bm it has been demonstrated that sequence complementarity between co-transcript RNA and the target DNA increases the precision of TPRT even in the absence of cleavage precisely at the top strand TSD boundary (Luan and Eickbush, 1995). A similar mechanism has been suggested for R1Bm (Feng *et al.*, 1998) since a low level of co-transcription of R1Bm with its target 28S rDNA has been reported (Long and Dawid, 1979). The suggestion that complementary base pairing might be essential for first-strand synthesis by TPRT as well as for target-site selection is supported by experimental evidence reported in this study. It will be discussed in detail in paragraph 4.3.

4.2.5 Other Strategies of Site-Specific Integration

In order to achieve targeted integration into specific sites of the genome, additional very diversified approaches are pursued by different research groups. The following paragraph gives a short overview of the major directions as well as their advantages and disadvantages.

The ideal approach of gene therapy is the repair of the defective gene by specifically replacing the mutated exon sequence by the functional wild-type exon. This process, called gene-targeting, has been successfully achieved by homologous recombination in murine embryonic stem cells (Capecchi, 1989), but is very inefficient in human somatic cells (1 event in 10^6 transfected cells, Sedivy and Sharp, 1989). However, proviral DNA from adeno-associated virus (AAV) seems to trigger homologous recombination more efficiently than plasmid DNA (Russell and Hirata, 1998). Also, creation of a double-strand break (DSB) in the chromosomal target greatly enhances the frequency of localised recombination events. (Jasin, 1996; Donoho *et al.*, 1998). This shifts the problem from site-specific integration to site-specific cleavage of DNA. The most versatile ENs known to date are chimeric zinc finger nucleases containing a non-specific DNA cleavage domain linked to a modular DNA recognition domain. The latter domain is composed of three C_2H_2 zinc fingers each specific for defined DNA triplets. Thus, every conceivable 9-bp sequence can be specifically recognised and made a preferred target for homologous recombination (Porteus and Baltimore, 2003; Bibikova *et al.*, 2003). The major drawback of this strategy is that even under optimal conditions, efficiency of the whole process is still low (~1% of transfected cells, Porteus *et al.*, 2003).

Improving the already existing retroviral vectors in terms of sequence-specificity could theoretically be achieved by tethering a specific DNA binding domain to the unspecific integrase protein. Several studies reported proof-of-principle (Katz *et al.*, 1996; Bushman and Miller, 1997; Holmes-Son and Chow, 2000), but efficiency and specificity of this strategy are still unsatisfying (Bushman, 2002).

ZAM, an LTR-retrotransposon from *D. melanogaster*, is the only retrovirus-like element described so far that displays considerable sequence specificity. The mechanistic basis for recognition of its target sequence 5’-GCGCGC-3’ (Leblanc *et al.*, 1999) is not elucidated yet, but once characterised, the unique properties of ZAM integrase might facilitate the development of specifically integrating retroviral vectors.

A second naturally occurring site-specific integrase, phage ΦC31 integrase, may also be useful for gene therapy. Initial experiments in cell culture yielded promising results with unidirectional targeted integration occurring 10-fold more frequently than random integration. However, ΦC31 integrase works best with its natural target which is not present in the human genome (Thyagarajan *et al.*, 2001).

Insertion into virtually any desired site in bacterial genomes can be achieved with group II introns since they recognise their target by complementary base pairing and can be easily modified in the relevant sequence segment. Modified group II introns have been successfully used for targeting genes at near 100% efficiency in bacteria (Zhong *et al.*, 2003), but to date, an efficient method to utilise group II introns in eukaryotic cells has not been developed (Guo *et al.*, 2000).

None of the targeting strategies pursued so far have been applied in clinical studies. Apart from often insufficient target specificity, major obstacles are inefficient gene delivery and incompatibility of the targeting system and the human host. Therefore, alternative approaches like retrotransposon-based vectors, though far from being fully optimised yet, are a worthwhile object for further studies.

4.3 L1 Uses a Cellular Double-Strand Break Repair Pathway for Replication

In the third part of this study, a surprising feature of *de novo* L1 integrants isolated from the retrotransposition reporter assay was investigated: Almost every isolated *de novo* integrant (94%) flanked by canonical TSDs was characterised by microhomologies of one to five bp between the 3' end of the TSD and the 5' end of the inserted L1 sequence, making the precise assignment of the boundary ambiguous (Fig. 27, p.91). In a complete analysis of the human genome, a significant number (58%) of endogenous L1 insertions revealed similar regions of microcomplementarity between the sequences at the 3' end of the TSD and the 5' end of the predicted L1 transcript RNA beyond the point of truncation (Fig. 28, p.94). In contrast, full-length insertions did not show a bias for overlapping sequences at their 5' end, suggesting that two independent integration mechanisms might exist.

The phenomenon of retrotransposon-associated microhomologies has been described before in different contexts:

Schwarz-Sommer *et al.* noticed microhomologies of two to three nucleotides at the 5' ends of five out of six endogenous Cin4 insertions in the *Zea mays* genome. They proposed a staggered cut in the genomic DNA target producing a 5' overhang, which then hybridises with the Cin4 mRNA (Schwarz-Sommer *et al.*, 1987).

In an analysis of *de novo* wild-type L1 insertions, Symer *et al.* also observed microhomologies at the 5' end of L1 integrants, although at a lower frequency (68%, Symer *et al.*, 2002) than in our study. They offered an explanation in accordance to a model established by Martin and Bushman two years earlier, which was mainly founded on theoretical considerations (Martin and Bushman, 2001). In contrast to the suggestion by Schwarz-Sommer *et al*., they proposed a staggered cut creating a 3' overhang. This overhang is presumed to anneal to the cDNA copy of the L1 mRNA (Fig. 31A).

Ostertag *et al*. observed that at the inversion junctions of endogenous inverted L1 insertions, in most cases one to four nucleotides could have originated from either the non-inverted L1 sequence or the inverted sequence (Ostertag and Kazazian, 2001b). They suggested that after the process of twin priming deemed to be responsible for the formation of an inversion (Fig. 31B), the cDNA strands that are synthesised onto the two 3' ends of the target DNA pair at small regions of complementarity. This mechanism is identical to microhomology-driven single strand annealing (SSA), representing one pathway of non-homologous end joining (NHEJ) (Göttlich *et al.*, 1998). Microhomology-driven SSA can resolve double-strand breaks even when the extent of complementarity is limited to a single nucleotide match (Pfeiffer *et al.*, 1994).

Combining the most likely features of these theories, a model for second-strand synthesis following synthesis of the first cDNA strand by TPRT is suggested here (Fig. 31C). During TPRT the L1 reverse transcriptase uses the L1 RNA as template and the free 3' hydroxyl-end of the target DNA as primer to initiate reverse transcription. After completion of reverse transcription, the RNA template is removed and L1 endonuclease cleaves the top strand, generating an additional 3' hydroxyl group and a stretch of single-stranded DNA. The single-stranded top strand of the target DNA anneals to the L1-cDNA at regions of limited complementarity and primes L1 second-strand synthesis. The remaining DNA synthesis is controlled by the host's DNA repair mechanisms (Fig. 31C).

The advantage of this model is that it relies completely on mechanisms that have been observed in human cells. The drawback of the postulate is that it cannot explain the full-length insertions and the 42% of truncated L1 insertions where no microhomologies at the 5' end of the L1 sequence have been observed. So far no stringent theory as to how these integrations arise has been proposed. However, *in vitro* studies with R2Bm RT (Bibillo and Eickbush, 2002; Burke *et al.*, 1999) imply that a template jump of the RT from RNA to the single-stranded target DNA's 3' overhang might attach retrotransposon cDNA to its genomic integration site without the need for sequence homology. For L1, a similar 'double TPRT' mechanism has been suggested (Cost *et al.*, 2002). Alternatively, an unknown ligase activity might join the cDNA to the genomic target.

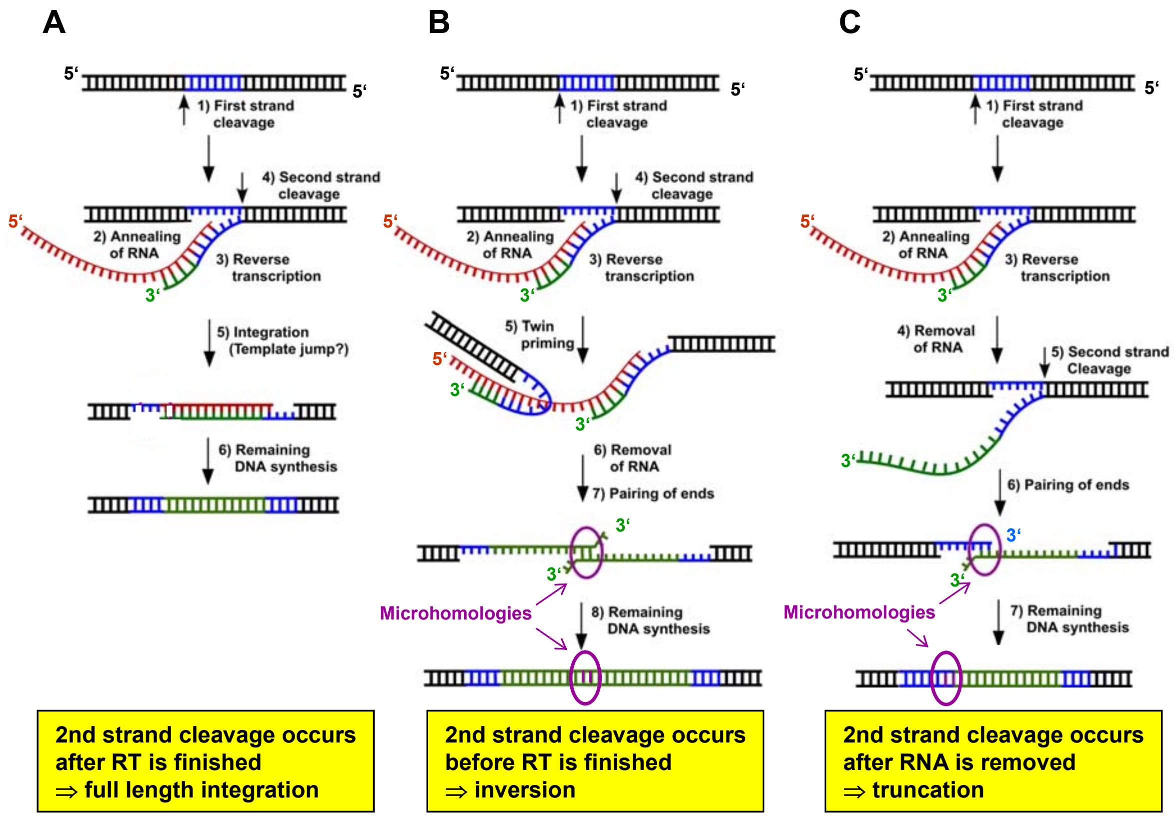

A
B
C
5‘
5‘
1) First strand cleavage
4) Second strand cleavage
2) Annealing of RNA
3) Reverse transcription
3‘
5) Integration (Template jump?)
6) Remaining DNA synthesis
5) Twin priming
6) Removal of RNA
7) Pairing of ends
8) Remaining DNA synthesis
4) Removal of RNA
5) Second strand Cleavage
6) Pairing of ends
7) Remaining DNA synthesis
Microhomologies
Microhomologies
2nd strand cleavage occurs after RT is finished ⇒ full length integration
2nd strand cleavage occurs before RT is finished ⇒ inversion
2nd strand cleavage occurs after RNA is removed ⇒ truncation

Fig. 31: Schematic representation of three alternative mechanisms of L1 integration. (A) TPRT is responsible for the initiation of L1 integration. After first strand cleavage (1), the bottom strand of the target site anneals to L1 mRNA (2) and primes reverse transcription (3). After second-strand cleavage (4), second-strand synthesis and integration of L1 are conveyed by an as yet unknown mechanism, possibly by L1 RT performing a template jump from the mRNA to the newly synthesised cDNA (Bibillo and Eickbush, 2002). **(B)** Twin priming creates inverted L1 integrants. When second-strand cleavage (4) takes place before reverse transcription has been completed, the upper strand of the target site serves as an internal primer that invades the L1 RNA and primes reverse transcription. After degradation of the RNA (6), the single-stranded cDNAs pair at a region of limited complementarity (7), and the remaining DNA synthesis is completed (8). **(C)** Microhomology-driven single strand annealing leads to the formation of truncated L1 elements. When second-strand cleavage (5) takes place after reverse transcription has been completed and the RNA has been degraded (4), the upper strand of the target site anneals to the cDNA at a region of limited complementarity (6) and primes second-strand synthesis (7).

Combining the proposed mechanism with the previously described “twin-priming” model (Ostertag and Kazazian, 2001b) and the “template-jump” model (Bibillo and Eickbush, 2002), I propose a compellingly simple mechanism to account for the generation of full-length, truncated, and inverted and truncated integrations that relies exclusively on the kinetics of second-strand cleavage: if the second strand is cleaved before reverse transcription is finished, twin priming is likely to occur (Fig. 31B) (Ostertag and Kazazian, 2001b). If second-strand cleavage takes place while the RT is still bound to the (full length or truncated) RNA, direct joining of the cDNA to the nicked target site is possible (Fig. 31A). If, however, the second strand is cleaved after reverse transcription is finished, the RT has dissociated from the

L1 cDNA and the RNA template has been degraded, microhomology-driven SSA might rescue and resolve these structures (Fig. 31C). This mechanism, leading to truncated elements, might be a “safety net” for unsuccessful retrotransposition events. Although there is no direct evidence for this theory, it is supported by the fact that microhomologies are more frequently found in hybrid-elements with impaired EN activity than in wild-type L1 integrants (94% vs. 58% of endogenous elements in the human genome (this study) vs. 68% of *de novo* wild-type L1 integrants (Symer *et al.*, 2002)). Additional evidence stems from data recently obtained in the Kazazian laboratory showing that point mutations that render L1 elements less active lead to shorter, truncated insertions (Farley *et al.*, 2004).

An annealing process might also be responsible for the crucial step of TPRT, the initiation of reverse transcription. Interestingly, this mechanism leads back to the target-site specificity of non-LTR retrotransposons (see 4.2.4): After nicking of the bottom strand of the target DNA, complementary base pairing between retrotransposon RNA and genomic DNA target may allow formation of a primer-template complex, which is then extended by the element’s RT activity (Jurka, 1997; Ovchinnikov *et al.*, 2001). This notion is supported by the observation that L1/Tx1L *de novo* integrants carrying a poly(A) tail inserted into sequences with a T-rich bottom strand while the two integrants lacking a poly(A) tail occupy target sites that are overlapping with their truncated 3’ ends.

4.4 Outlook

In the present study, the potential of non-LTR retrotransposons as site-specific gene delivery vectors was assessed. Althought experiments with heterologous EN domains did not yield a sequence-specifically integrating chimeric retrotransposon, they furnished several valuable clues for future projects in this line of research:

Swapping experiments showed that modification of the EN domain severely reduced the retrotransposition rate of the resulting hybrid constructs. This suggested that either the EN itself or its interaction with additional retrotransposon-encoded factors and/or host factors is very sensitive to changes in its three-dimensional structure. Besides, additional factors probably also contribute to sequence recognition since none of the active chimeras displayed a completely altered cleavage specificity. However, the so-called ‘α11’-hairpin loop was identified as a structural element that tolerates amino acid substitutions to a certain degree and at the same time influences target-site recognition. This region should therefore be an excellent starting point for systematic optimisation studies. Ideally, an *in vitro* evolution strategy should be adopted to select for sequence-specific ENs. However, inactive as well as hyperactive (i. e. unspecific) ENs would have to be selected against, which makes success of the selection procedure unlikely. Alternatively, the co-crystal structures of L1, Tx1L and R1Bm EN bound to DNA could be elucidated. A co-crystal allows the identification of amino acid residues directly contacting the DNA and could enable the design of a site-specific EN with minimal changes in the L1 EN amino acid sequence.

A second way to improve integration specificity of L1 elements can be inferred from the data obtained on the importance of microhomologies for integration. Sequence overlaps of the 3' ends of non-LTR retrotransposons with their target sequences have been implicated in the integration mechanism of several elements (Burch *et al.*, 1993; Luan and Eickbush, 1995; Volff *et al.*, 1999). Two *de novo* L1 integrants characterised in this study supply experimental evidence for this notion. Therefore, it could be worthwhile to borrow not only the EN domain, but also the corresponding 3' sequences from site-specifically integrating retrotransposons to render L1 integration more specific.

The microhomologies observed at the 5' end of *de novo* integrants could be exploited to improve two aspects of L1-mediated gene delivery, specificity and safety. To date, hotspots for the truncation of L1 elements have not been found (Szak *et al.*, 2002). For gene therapy, however, a truncation hotspot that is localised upstream of and directly adjacent to the transgene would be desirable. Thus, the therapeutic insert would not contain more vector sequence than necessary, and the L1 promoter which could re-mobilise the *de novo* integrant would not be part of the new L1 insertion. Provided that a sequence-specific EN will be developed, an array of several repeats of the sequence corresponding to the 3' end of the expected TSD could be introduced into the sequence 5' of the transgene. This could greatly increase the probability of 5' truncations in this region and add another level of sequence specificity.

In summary, the utilisation of L1 elements in target-site specific gene therapy seems feasible. Although some of the hurdles may be difficult to leap, the fact that many site-specific retrotransposons that are closely related to L1 manage to target integration encourages optimism for modifying human L1 retrotransposons.

5. BIBLIOGRAPHY

Aiyar, A., Xiang, Y. and Leis, J. (1996). Site-directed mutagenesis using overlap extension PCR. *Methods Mol Biol*, **57,** 177-91.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol*, **215,** 403-10.

Anderson, W.F. (2000). Gene therapy. The best of times, the worst of times. *Science*, **288,** 627-9.

Anzai, T., Takahashi, H. and Fujiwara, H. (2001). Sequence-specific recognition and cleavage of telomeric repeat $(TTAGG)_n$ by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol Cell Biol*, **21,** 100-8.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L. and Batzer, M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, **29,** 136-44.

Arkhipova, I.R. and Morrison, H.G. (2001). Three retrotransposon families in the genome of Giardia lamblia: two telomeric, one dead. *Proc Natl Acad Sci U S A*, **98,** 14497-502.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1999). Short Protocols in Molecular Biology. John Wiley & Sons, Inc., New York.

Barzilay, G. and Hickson, I.D. (1995). Structure and function of apurinic/apyrimidinic endonucleases. *Bioessays*, **17,** 713-9.

Baum, C., Dullmann, J., Li, Z., Fehse, B., Meyer, J., Williams, D.A. and von Kalle, C. (2003). Side effects of retroviral gene transfer into hematopoietic stem cells. *Blood*, **101,** 2099-114.

Baum, C., Eckert, H.G., Stockschlader, M., Just, U., Hegewisch-Becker, S., Hildinger, M., Uhde, A., John, J. and Ostertag, W. (1996). Improved retroviral vectors for hematopoietic stem cell protection and in vivo selection. *J Hematother*, **5,** 323-9.

Baum, C. and Fehse, B. (2003). Mutagenesis by retroviral transgene insertion: risk assessment and potential alternatives. *Curr Opin Mol Ther*, **5,** 458-62.

Baum, C., Richters, A. and Ostertag, W. (1999). Retroviral vector-mediated gene expression in hematopoietic cells. *Curr Opin Mol Ther*, **1,** 605-12.

Becker, K.G., Swergold, G.D., Ozato, K. and Thayer, R.E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum Mol Genet*, **2,** 1697-702.

Berg, J.M. and Shi, Y. (1996). The galvanization of biology: a growing appreciation for the roles of zinc. *Science*, **271,** 1081-5.

Bestor, T.H. (2000). Gene silencing as a threat to the success of gene therapy. *J Clin Invest*, **105,** 409-11.

Bibikova, M., Beumer, K., Trautman, J.K. and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. *Science*, **300,** 764.

Bibillo, A. and Eickbush, T.H. (2002). The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol*, **316,** 459-73.

Biemont, C., Tsitrone, A., Vieira, C. and Hoogland, C. (1997). Transposable element distribution in Drosophila. *Genetics*, **147,** 1997-9.

Bird, A.P. and Wolffe, A.P. (1999). Methylation-induced repression - belts, braces, and chromatin. *Cell*, **99,** 451-4.

Boeke, J.D. (1997). LINEs and Alus - the polyA connection. *Nat Genet*, **16,** 6-7.

Boeke, J.D. and Pickeral, O.K. (1999). Retroshuffling the genomic deck. *Nature*, **398,** 108-111.

Boshart, M., Weber, F., Jahn, G., Dorsch-Hasler, K., Fleckenstein, B. and Schaffner, W. (1985). A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. *Cell*, **41,** 521-30.

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

Burch, J.B., Davis, D.L. and Haas, N.B. (1993). Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. *Proc Natl Acad Sci U S A*, **90,** 8199-203.

Burke, W.D., Eickbush, D.G., Xiong, Y., Jakubczak, J. and Eickbush, T.H. (1993). Sequence relationship of retrotransposable elements R1 and R2 within and between divergent insect species. *Mol Biol Evol*, **10,** 163-85.

Burke, W.D., Malik, H.S., Jones, J.P. and Eickbush, T.H. (1999). The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.*, **16,** 502-511.

Burke, W.D., Malik, H.S., Lathe, W.C., 3rd and Eickbush, T.H. (1998). Are retrotransposons long-term hitchhikers? *Nature*, **392,** 141-2.

Bushman, F. (2002). Targeting retroviral integration? *Mol Ther*, **6,** 570-1.

Bushman, F.D. and Miller, M.D. (1997). Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J Virol*, **71,** 458-64.

Cal, S., Tan, K.L., McGregor, A. and Connolly, B.A. (1998). Conversion of bovine pancreatic DNase I to a repair endonuclease with a high selectivity for abasic sites. *EMBO J.*, **17,** 7128-38.

Capecchi, M.R. (1989). Altering the genome by homologous recombination. *Science*, **244,** 1288-92.

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., Selz, F., Hue, C., Certain, S., Casanova, J.L., Bousso, P., Deist, F.L. and Fischer, A. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288,** 669-72.

Charlesworth, B. and Langley, C.H. (1989). The population genetics of Drosophila transposable elements. *Annu Rev Genet*, **23,** 251-87.

Christensen, S., Pont-Kingdon, G. and Carroll, D. (2000). Target specificity of the endonuclease from the *Xenopus laevis* non-long terminal repeat retrotransposon, Tx1L. *Mol. Cell. Biol.*, **20,** 1219-1226.

Christensen, S., Pont-Kingdon, G. and Carroll, D. (2001). Comparative studies of the endonucleases from two related *Xenopus laevis* retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. *Genetica*, **110,** 245-256.

Coffin, J.M. (1996). Retroviridae: The viruses and their replication. In *Fundamental Virology*, Fields, B.N., Knipe, D.M. and Howley, P.M. (eds) pp. 763-844. Lippincott Raven, Philadelphia.

Cold Spring Harbor Laboratory. (2000). The complete sequence of a heterochromatic island from a higher eukaryote. The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems Arabidopsis Sequencing Consortium. *Cell*, **100,** 377-86.

Cost, G.J. and Boeke, J.D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37,** 18081-18093.

Cost, G.J., Feng, Q., Jacquier, A. and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription *in vitro*. *EMBO J.*, **21,** 5899-5910.

Cost, G.J., Golding, A., Schlissel, M.S. and Boeke, J.D. (2001). Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res.*, **29,** 573-577.

Covey, S.N. (1986). Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. *Nucleic Acids Res*, **14,** 623-33.

Curcio, M.J. and Derbyshire, K.M. (2003). The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol*, **4,** 865-77.

Dahm-Daphi, J., Sass, C. and Alberti, W. (2000). Comparison of biological effects of DNA damage induced by ionizing radiation and hydrogen peroxide in CHO cells. *Int J Radiat Biol*, **76,** 67-75.

Danilevskaya, O., Slot, F., Pavlova, M. and Pardue, M.L. (1994). Structure of the Drosophila HeT-A transposon: a retrotransposon-like element forming telomeres. *Chromosoma*, **103,** 215-24.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, Oxford.

Delecluse, H.J. and Hammerschmidt, W. (2000). The genetic approach to the Epstein-Barr virus: from basic virology to gene therapy. *Mol Pathol*, **53,** 270-9.

Demple, B. and Harrison, L. (1994). Repair of oxidative damage to DNA: enzymology and biology. *Annu Rev Biochem*, **63,** 915-48.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet*, **35,** 41-8.

Donoho, G., Jasin, M. and Berg, P. (1998). Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Mol Cell Biol*, **18,** 4070-8.

Eickbush, T.H. (2002). R2 and related site-specific non-long terminal repeat retrotransposons. In *Mobile DNA II*, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds) pp. 813-835. American Society for Microbiology, Washington, D.C.

Eickbush, T.H. and Malik, H.S. (2002). Origins and evolution of retrotransposons. In *Mobile DNA II*, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds) pp. 1111-1144. American Society for Microbiology, Washington, D.C.

Ergün, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Strätling, W.H. and Schumann, G.G. (2004). Cell type-specific expression of LINE-1 ORF1 and ORF2 in fetal and adult human tissues. *J Biol Chem.*, epub ahead of print.

Esnault, C., Maestre, J. and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.*, **24,** 363-367.

Evans, A.R., Limp-Foster, M. and Kelley, M.R. (2000). Going APE over ref-1. *Mutat Res*, **461,** 83-108.

Farley, A.H., Luning Prak, E.T. and Kazazian, H.H., Jr. (2004). More active human L1 retrotransposons produce longer insertions. *Nucleic Acids Res*, **32,** 502-10.

Feng, Q., Moran, J., Kazazian, H.H., Jr. and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87,** 905-916.

Feng, Q., Schumann, G. and Boeke, J.D. (1998). Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proc. Natl. Acad. Sci. USA*, **95,** 2083-2088.

Finnegan, D.J. (1997). Transposable elements: how non-LTR retrotransposons do it. *Curr Biol*, **7,** R245-8.

Freeman, J.D., Goodchild, N.L. and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. *Biotechniques*, **17,** 46-52.

Garrett, J.E., Knutzon, D.S. and Carroll, D. (1989). Composite transposable elements in the *Xenopus laevis* genome. *Mol. Cell. Biol.*, **9,** 3018-3027.

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

Goodier, J.L., Ostertag, E.M. and Kazazian, H.H., Jr. (2000). Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.*, **9,** 653-657.

Goodwin, T.J., Ormandy, J.E. and Poulter, R.T. (2001). L1-like non-LTR retrotransposons in the yeast Candida albicans. *Curr Genet*, **39,** 83-91.

Gorelick, R.J., Gagliardi, T.D., Bosche, W.J., Wiltrout, T.A., Coren, L.V., Chabot, D.J., Lifson, J.D., Henderson, L.E. and Arthur, L.O. (1999). Strict conservation of the retroviral nucleocapsid protein zinc finger is strongly influenced by its role in viral infection processes: characterization of HIV-1 particles containing mutant nucleocapsid zinc-coordinating sequences. *Virology*, **256,** 92-104.

Gorman, M.A., Morera, S., Rothwell, D.G., de La Fortelle, E., Mol, C.D., Tainer, J.A., Hickson, I.D. and Freemont, P.S. (1997). The crystal structure of the human DNA repair endonuclease HAP1 suggests the recognition of extra-helical deoxyribose at DNA abasic sites. *EMBO J.*, **16,** 6548-58.

Göttlich, B., Reichenberger, S., Feldmann, E. and Pfeiffer, P. (1998). Rejoining of DNA double-strand breaks in vitro by single-strand annealing. *Eur J Biochem*, **258,** 387-95.

Gregory, T.R. and Hebert, P.D. (1999). The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res*, **9,** 317-24.

Grimaldi, G., Skowronski, J. and Singer, M.F. (1984). Defining the beginning and end of KpnI family segments. *Embo J*, **3,** 1753-9.

Grishin, N.V. (2001). Treble clef finger - a functionally diverse zinc-binding structural motif. *Nucleic Acids Res*, **29,** 1703-14.

Guo, H., Karberg, M., Long, M., Jones, J.P., 3rd, Sullenger, B. and Lambowitz, A.M. (2000). Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science*, **289,** 452-7.

Hacein-Bey-Abina, S., Fischer, A. and Cavazzana-Calvo, M. (2002). Gene therapy of X-linked severe combined immunodeficiency. *Int J Hematol*, **76,** 295-8.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., Sorensen, R., Forster, A., Fraser, P., Cohen, J.I., de Saint Basile, G., Alexander, I., Wintergerst, U., Frebourg, T., Aurias, A., Stoppa-Lyonnet, D., Romana, S., Radford-Weiss, I., Gross, F., Valensi, F., Delabesse, E., Macintyre, E., Sigaux, F., Soulier, J., Leiva, L.E., Wissler, M., Prinz, C., Rabbitts, T.H., Le Deist, F., Fischer, A. and Cavazzana-Calvo, M. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302,** 415-9.

Hartl, D.L., Lohe, A.R. and Lozovskaya, E.R. (1997). Modern thoughts on an ancyent marinere: function, evolution, regulation. *Annu Rev Genet*, **31,** 337-58.

Hattori, M., Hidaka, S. and Sakaki, Y. (1985). Sequence analysis of a KpnI family member near the 3' end of human beta-globin gene. *Nucleic Acids Res*, **13,** 7813-27.

Hendrich, B. and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol*, **18,** 6538-47.

Hilliker, A.J., Appels, R. and Schalet, A. (1980). The genetic analysis of *D. melanogaster* heterochromatin. *Cell*, **21,** 607-19.

Hohjoh, H. and Singer, M.F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.*, **15,** 630-639.

Hohjoh, H. and Singer, M.F. (1997). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *Embo J*, **16,** 6034-43.

Holmes, D.S. and Quigley, M. (1981). A rapid boiling method for the preparation of bacterial plasmids. *Anal Biochem*, **114,** 193-7.

Holmes-Son, M.L. and Chow, S.A. (2000). Integrase-lexA fusion proteins incorporated into human immunodeficiency virus type 1 that contains a catalytically inactive integrase gene are functional to mediate integration. *J Virol*, **74,** 11548-56.

Jakubczak, J.L., Burke, W.D. and Eickbush, T.H. (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci U S A*, **88,** 3295-9.

Jasin, M. (1996). Genetic manipulation of genomes with rare-cutting endonucleases. *Trends Genet*, **12,** 224-8.

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

Kajikawa, M., Ohshima, K. and Okada, N. (1997). Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. *Mol Biol Evol*, **14,** 1206-17.

Katz, R.A., Merkel, G. and Skalka, A.M. (1996). Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. *Virology*, **217,** 178-90.

Kay, M.A., Glorioso, J.C. and Naldini, L. (2001). Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. *Nat Med*, **7,** 33-40.

Kazazian, H.H., Jr. (1998). Mobile elements and disease. *Curr Opin Genet Dev*, **8,** 343-50.

Kent, W.J. (2002). BLAT- the BLAST-like alignment tool. *Genome Res*, **12,** 656-64.

Kirchner, J., Connolly, C.M. and Sandmeyer, S.B. (1995). Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science*, **267,** 1488-91.

Kojima, K.K. and Fujiwara, H. (2003). Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol Biol Evol*, **20,** 351-61.

Kondo-Iida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y. and Toda, T. (1999). Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). *Hum Mol Genet*, **8,** 2303-9.

König, P., Fairall, L. and Rhodes, D. (1998). Sequence-specific DNA recognition by the myb-like domain of the human telomere binding protein TRF1: a model for the protein-DNA complex. *Nucleic Acids Res*, **26,** 1731-40.

Kozak, M. (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol Cell Biol*, **7,** 3438-45.

Kramer, P.R. and Sinden, R.R. (1997). Measurement of unrestrained negative supercoiling and topological domain size in living human cells. *Biochemistry*, **36,** 3151-8.

Kubo, Y., Okazaki, S., Anzai, T. and Fujiwara, H. (2001). Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. *Mol Biol Evol*, **18,** 848-57.

Kurose, K., Hata, K., Hattori, M. and Sakaki, Y. (1995). RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. *Nucleic Acids Res*, **23,** 3704-9.

Labrador, M. and Corces, V.G. (1997). Transposable element-host interactions: regulation of insertion and excision. *Annu Rev Genet*, **31,** 381-404.

Lahm, A. and Suck, D. (1991). DNase I-induced DNA conformation. 2Å structure of a DNase I-octamer complex. *J Mol Biol*, **222,** 645-67.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409,** 860-921.

Leblanc, P., Dastugue, B. and Vaury, C. (1999). The integration machinery of ZAM, a retroelement from Drosophila melanogaster, acts as a sequence-specific endonuclease. *J Virol*, **73,** 7061-4.

Leibold, D.M., Swergold, G.D., Singer, M.F., Thayer, R.E., Dombroski, B.A. and Fanning, T.G. (1990). Translation of LINE-1 DNA elements *in vitro* and in human cells. *Proc. Natl. Acad. Sci. USA*, **87,** 6990-6994.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, **69,** 905-14.

Li, T.H. and Schmid, C.W. (2001). Differential stress induction of individual Alu loci: implications for transcription and retrotransposition. *Gene*, **276,** 135-41.

Li, Z., Dullmann, J., Schiedlmeier, B., Schmidt, M., von Kalle, C., Meyer, J., Forster, M., Stocking, C., Wahlers, A., Frank, O., Ostertag, W., Kühlcke, K., Eckert, H.G., Fehse, B. and Baum, C. (2002). Murine leukemia induced by retroviral gene marking. *Science*, **296,** 497.

Liu, L.F. (1989). DNA topoisomerase poisons as antitumor drugs. *Annu Rev Biochem*, **58,** 351-75.

Long, E.O. and Dawid, I.B. (1979). Expression of ribosomal DNA insertions in Drosophila melanogaster. *Cell*, **18,** 1185-96.

Lovsin, N., Gubensek, F. and Kordi, D. (2001). Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Mol Biol Evol*, **18,** 2213-24.

Luan, D. and Eickbush, T.H. (1995). RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell. Biol.*, **15,** 3882-3891.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72,** 595-605.

Malik, H.S., Burke, W.D. and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*, **16,** 793-805.

Martín, F., Maranon, C., Olivares, M., Alonso, C. and Lopez, M.C. (1995). Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. *J Mol Biol*, **247,** 49-59.

Martin, S.L. and Bushman, F.D. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.*, **21,** 467-475.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D. and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science*, **254,** 1808-10.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci USA*, **36,** 344-355.

McMillan, J.P. and Singer, M.F. (1993). Translation of the human LINE-1 element, L1Hs. *Proc. Natl. Acad. Sci. USA*, **90,** 11533-11537.

Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M. and Sakaki, Y. (1992). Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res*, **20,** 3139-45.

Mol, C.D., Izumi, T., Mitra, S. and Tainer, J.A. (2000). DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination. *Nature*, **403,** 451-6.

Mol, C.D., Kuo, C.-F., Thayer, M.M., Cunningham, R.P. and Tainer, J.A. (1995). Structure and function of the multifunctional DNA repair enzyme exonuclease III. *Nature*, **374,** 381-386.

Mooslehner, K., Karls, U. and Harbers, K. (1990). Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. *J Virol*, **64,** 3056-8.

Moran, J.V., DeBerardinis, R.J. and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. *Science*, **283,** 1530-1534.

Moran, J.V. and Gilbert, N. (2002). Mammalian LINE-1 retrotransposons and related elements. In *Mobile DNA II*, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds) pp. 836-869. American Society for Microbiology, Washington, D.C.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D. and Kazazian, H.H., Jr. (1996). High-frequency retrotransposition in cultured mammalian cells. *Cell*, **87,** 917-927.

Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A. and Perlman, P.S. (1994). Splicing defective mutants of the *COX1* gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. *Nucl. Acids Res.*, **22,** 2057-2064.

Morawetz, C. (1987). Effect of irradiation and mutagenic chemicals on the generation of ADH2-constitutive mutants in yeast. Significance for the inducibility of Ty transposition. *Mutat Res*, **177,** 53-60.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A. and Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genet.*, **31,** 159-165.

Müller, C., Readhead, C., Diederichs, S., Idos, G., Yang, R., Tidow, N., Serve, H., Berdel, W.E. and Koeffler, H.P. (2000). Methylation of the cyclin A1 promoter correlates with gene silencing in somatic cell lines, while tissue-specific expression of cyclin A1 is methylation independent. *Mol Cell Biol*, **20,** 3316-29.

Narita, N., Nishio, H., Kitoh, Y., Ishikawa, Y., Minami, R., Nakamura, H. and Matsuo, M. (1993). Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest*, **91,** 1862-7.

Nekrutenko, A. and Li, W.H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends Genet*, **17,** 619-21.

Ochman, H., Gerber, A.S. and Hartl, D.L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120,** 621-3.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. and Nishimura, Y. (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, **79,** 639-48.

Ohno, S. (1972). So much "junk" DNA in our genome. In *Evolution of genetic systems*, Smith, H.H. (ed) pp. 366-370. Gordon and Breach, New York.

Okazaki, S., Ishikawa, H. and Fujiwara, H. (1995). Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. *Mol Cell Biol*, **15,** 4545-52.

Okazaki, S., Tsuchida, K., Maekawa, H., Ishikawa, H. and Fujiwara, H. (1993). Identification of a pentanucleotide telomeric sequence, $(TTAGG)_n$, in the silkworm Bombyx mori and in other insects. *Mol Cell Biol*, **13,** 1424-32.

Ostertag, E.M. and Kazazian, H.H., Jr. (2001a). Biology of Mammalian L1 retrotransposons. *Annu. Rev. Genet.*, **35,** 501-538.

Ostertag, E.M. and Kazazian, H.H., Jr. (2001b). Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.*, **11,** 2059-2065.

Ostertag, E.M., Luning Prak, E.T., DeBerardinis, R.J., Moran, J.V. and Kazazian, H.H., Jr. (2000). Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.*, **28,** 1418-1423.

Ovchinnikov, I., Troxel, A.B. and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: Evidence supporting random insertion. *Genome Res.*

Pagel, M. and Johnstone, R.A. (1992). Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proc R Soc Lond B Biol Sci*, **249,** 119-24.

Pardue, M.L., Danilevskaya, O.N., Lowenhaupt, K., Slot, F. and Traverse, K.L. (1996). Drosophila telomeres: new views on chromosome evolution. *Trends Genet*, **12,** 48-52.

Pfeiffer, P., Thode, S., Hancke, J. and Vielmetter, W. (1994). Mechanisms of overlap formation in nonhomologous DNA end joining. *Mol Cell Biol*, **14,** 888-95.

Pickeral, O.K., Makalowski, W., Boguski, M.S. and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.*, **10,** 411-415.

Porteus, M.H. and Baltimore, D. (2003). Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300,** 763.

Porteus, M.H., Cathomen, T., Weitzman, M.D. and Baltimore, D. (2003). Efficient gene targeting mediated by adeno-associated virus and DNA double-strand breaks. *Mol Cell Biol*, **23,** 3558-65.

Raper, S.E., Chirmule, N., Lee, F.S., Wivel, N.A., Bagg, A., Gao, G.P., Wilson, J.M. and Batshaw, M.L. (2003). Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. *Mol Genet Metab*, **80,** 148-58.

Rashkova, S., Athanasiadis, A. and Pardue, M.L. (2003). Intracellular targeting of Gag proteins of the Drosophila telomeric retrotransposons. *J Virol*, **77,** 6376-84.

Rashkova, S., Karam, S.E. and Pardue, M.L. (2002). Element-specific localization of Drosophila retrotransposon Gag proteins occurs in both nucleus and cytoplasm. *Proc Natl Acad Sci U S A*, **99,** 3621-6.

Reichwald, K., Thiesen, J., Wiehe, T., Weitzel, J., Poustka, W.A., Rosenthal, A., Platzer, M., Strätling, W.H. and Kioschis, P. (2000). Comparative sequence analysis of the MECP2-locus in human and mouse reveals new transcribed regions. *Mamm Genome*, **11,** 182-90.

Richards, E.J. and Elgin, S.C. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell*, **108,** 489-500.

Rizzon, C., Marais, G., Gouy, M. and Biemont, C. (2002). Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. *Genome Res*, **12,** 400-7.

Roth, D.B., Porter, T.N. and Wilson, J.H. (1985). Mechanisms of nonhomologous recombination in mammalian cells. *Mol Cell Biol*, **5,** 2599-607.

Russell, D.W. and Hirata, R.K. (1998). Human gene targeting by viral vectors. *Nat Genet*, **18,** 325-30.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239,** 487-91.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning*. Cold Spring Harbor Laboratory, New York.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci Usa*, **74,** 5463-5467.

Saporito, S.M., Smith-White, B.J. and Cunningham, R.P. (1988). Nucleotide sequence of the *xth* gene of Escherichia coli K-12. *J Bacteriol*, **170,** 4542-7.

Sasaki, T. and Fujiwara, H. (2000). Detection and distribution patterns of telomerase activity in insects. *Eur J Biochem*, **267,** 3025-31.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D. and Kazazian, H.H., Jr. (1997). Many human L1 elements are capable of retrotransposition. *Nat. Genet.*, **16,** 37-43.

Schmidt, M., Hoffmann, G., Wissler, M., Lemke, N., Mussig, A., Glimm, H., Williams, D.A., Ragg, S., Hesemann, C.U. and von Kalle, C. (2001). Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum Gene Ther*, **12,** 743-9.

Schmidt-Wolf, G.D. and Schmidt-Wolf, I.G. (2003). Non-viral and hybrid vectors in human gene therapy: an update. *Trends Mol Med*, **9,** 67-72.

Schneider, T.D. and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18,** 6097-100.

Schröder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R. and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110,** 521-9.

Schwarz-Sommer, Z., Leclercq, L., Göbel, E. and Saedler, H. (1987). Cin4, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. *EMBO J.*, **6,** 3873-3880.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D. and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics*, **1,** 113-25.

Sedivy, J.M. and Sharp, P.A. (1989). Positive genetic selection for gene disruption in mammalian cells by homologous recombination. *Proc Natl Acad Sci U S A*, **86,** 227-31.

Servomaa, K. and Rytömaa, T. (1990). UV light and ionizing radiations cause programmed death of rat chloroleukaemia cells by inducing retropositions of a mobile DNA element (L1Rn). *Int J Radiat Biol*, **57,** 331-43.

Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983). Transcription of the KpnI families of long interspersed DNAs in human cells. *Nature*, **304,** 277-80.

Siegel, V. and Walter, P. (1988). Each of the activities of signal recognition particle (SRP) is contained within a distinct domain: analysis of biochemical mutants of SRP. *Cell*, **52,** 39-49.

Singer, R.H. and Green, M.R. (1997). Compartmentalization of eukaryotic gene expression: causes and effects. *Cell*, **91,** 291-4.

Skowronski, J., Fanning, T.G. and Singer, M.F. (1988). Unit-length LINE-1 transcripts in human teratocarcinoma cells. *Mol Cell Biol*, **8,** 1385-97.

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev*, **6,** 743-8.

Smit, A.F. and Riggs, A.D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A*, **93,** 1443-8.

Soifer, H., Higo, C., Kazazian, H.H., Jr., Moran, J.V., Mitani, K. and Kasahara, N. (2001). Stable integration of transgenes delivered by a retrotransposon-adenovirus hybrid vector. *Hum Gene Ther*, **12,** 1417-28.

Strand, D.J. and McDonald, J.F. (1985). Copia is transcriptionally responsive to environmental stress. *Nucleic Acids Res*, **13,** 4401-10.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.*, **10,** 6718-6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G. and Boeke, J.D. (2002). Human L1 retrotransposition is associated with genetic instability in vivo. *Cell*, **110,** 327-38.

Szak, S.T., Pickeral, O.K., Landsman, D. and Boeke, J.D. (2003). Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome Biol*, **4**.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D. and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biol*, **3,** research0052.

Takahashi, H. and Fujiwara, H. (1999). Transcription analysis of the telomeric repeat-specific retrotransposons TRAS1 and SART1 of the silkworm Bombyx mori. *Nucleic Acids Res*, **27,** 2015-21.

Takahashi, H. and Fujiwara, H. (2002). Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *Embo J*, **21,** 408-17.

Takahashi, H., Okazaki, S. and Fujiwara, H. (1997). A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori. *Nucleic Acids Res*, **25,** 1578-84.

Tanchou, V., Decimo, D., Pechoux, C., Lener, D., Rogemond, V., Berthoux, L., Ottmann, M. and Darlix, J.L. (1998). Role of the N-terminal zinc finger of human immunodeficiency virus type 1 nucleocapsid protein in virus structure and replication. *J Virol*, **72,** 4442-7.

Tchenio, T., Casella, J.F. and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res*, **28,** 411-5.

Thayer, R.E., Singer, M.F. and Fanning, T.G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. *Gene*, **133,** 273-7.

Thomas, C.A. (1971). The Genetic Organization of Chromosomes. *Annu Rev Genet*, **5,** 237-256.

Thyagarajan, B., Olivares, E.C., Hollis, R.P., Ginsburg, D.S. and Calos, M.P. (2001). Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol Cell Biol*, **21,** 3926-34.

Tilford, C.A., Kuroda-Kawaguchi, T., Skaletsky, H., Rozen, S., Brown, L.G., Rosenberg, M., McPherson, J.D., Wylie, K., Sekhon, M., Kucaba, T.A., Waterston, R.H. and Page, D.C. (2001). A physical map of the human Y chromosome. *Nature*, **409,** 943-5.

Travers, A.A. (1993). *DNA-Protein Interactions*. Chapman and Hall, London, UK.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K. and Lenz, J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol*, **11,** 1531-5.

Ullu, E. and Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature*, **312,** 171-2.

Van Arsdell, S.W. and Weiner, A.M. (1984). Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3' truncation. *Nucleic Acids Res*, **12,** 1463-71.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. *Annu Rev Genet*, **19,** 253-72.

Volff, J.N., Korting, C., Sweeney, K. and Schartl, M. (1999). The non-LTR retrotransposon Rex3 from the fish Xiphophorus is widespread among teleosts. *Mol Biol Evol*, **16,** 1427-38.

Volpers, C. and Kochanek, S. (2004). Adenoviral vectors for gene transfer and therapy. *J Gene Med*, **6** Suppl 1**,** S164-71.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Jr., Boeke, J.D. and Moran, J.V. (2001). Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol Cell Biol*, **21,** 1429-1439.

Wei, W., Morrish, T.A., Alisch, R. and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accomodate multiple LINE-1 retrotransposition events. *Anal. Bioch.*, **284,** 435-438.

Weichenrieder, O., Repanas, K. and Perrakis, A. (in press). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure*.

Weichenrieder, O., Wild, K., Strub, K. and Cusack, S. (2000). Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*, **408,** 167-73.

Weitzel, J.M., Buhrmester, H. and Strätling, W.H. (1997). Chicken MAR-binding protein ARBP is homologous to rat methyl-CpG-binding protein MeCP2. *Mol Cell Biol*, **17,** 5656-66.

Williams, D.A., Nienhuis, A.W., Hawley, R.G. and Smith, F.O. (2000). Gene Therapy 2000. *Hematology* (Am Soc Hematol Educ Program)**,** 376-393.

Woodcock, D.M., Lawler, C.B., Linsenmeyer, M.E., Doherty, J.P. and Warren, W.D. (1997). Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem*, **272,** 7810-6.

Wu, X., Li, Y., Crise, B. and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300,** 1749-51.

Xie, W., Gai, X., Zhu, Y., Zappulla, D.C., Sternglanz, R. and Voytas, D.F. (2001). Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol Cell Biol*, **21,** 6606-14.

Xiong, Y., Burke, W.D., Jakubczak, J.L. and Eickbush, T.H. (1988). Ribosomal DNA insertion elements R1Bm and R2Bm can transpose in a sequence specific manner to locations outside the 28S genes. *Nucleic Acids Res.*, **16,** 10561-10573.

Xiong, Y. and Eickbush, T.H. (1988). The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol. Cell. Biol.*, **8,** 114-123.

Xiong, Y. and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo J*, **9,** 3353-62.

Yang, J., Malik, H.S. and Eickbush, T.H. (1999). Identification of the endonuclease domain encoded by R2 and other site- specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A*, **96,** 7847-52.

Yang, N., Zhang, L., Zhang, Y. and Kazazian, H.H., Jr. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res*, **31,** 4929-40.

Yates, J.L., Warren, N. and Sugden, B. (1985). Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. *Nature*, **313,** 812-5.

Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, **13,** 335-40.

Yu, F., Zingler, N., Schumann, G. and Strätling, W.H. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res.*, **29,** 4493-4501.

Zeng, C., Kim, E., Warren, S.L. and Berget, S.M. (1997). Dynamic relocation of transcription and splicing factors dependent upon transcriptional activity. *Embo J*, **16,** 1401-12.

Zhong, J., Karberg, M. and Lambowitz, A.M. (2003). Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic Acids Res*, **31,** 1656-64.

Zhu, Y., Dai, J., Fuerst, P.G. and Voytas, D.F. (2003). Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A*, **100,** 5891-5.

Zingler, N., Weichenrieder, O. and Schumann, G. (in press). APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenetic and Genome Research*.

Zou, S. and Voytas, D.F. (1997). Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. *Proc Natl Acad Sci U S A*, **94,** 7412-6.

A. LIST OF OLIGONUCLEOTIDES

Primer Name	Nucleotide Sequence (5'→3')	Coordinates (if applicable)
GS10	TTTCTTCCTAGTCTCCATGGTCTTTAC	L1.3: 1949-1923
GS11	GTATCAGCCATGGAAGATGAAATGAATG	L1.3: 1261-1288
GS14	GTGTTTTGGCCATGGAGTCCTTGCCC	L1.3: 4984-4959
GS16	TTTCTTTCTGCAGTGGTTTGTAGTTCTC	L1.3: 4363-4336
GS17	GTGTCCATGTGAATTCATTGTTCAATTCC	L1.3: 5838-5810
GS22	ATGAGCCCATGGTCGCTCGATGATTCATTAGTTACACG	R1 EN: 635- 598
GS23	ATCAACCCGGGTGGTGCAGAGGATGC	R1 EN: 34- 59
GS30	GGTGTATACATCCGCAATAGGGTGCTCCC	R1 EN: 171- 200
GS33	ACTATGAGATCTATATATGGTGTCTGCGTAC	R1 EN: 255- 285
GS36	ATATCATTTAAATGCCCACTCGCCATTGTGGC	R1 EN: 372- 403
GS51	CCCATATGCAGGATCAAATTCACACATAAC	L1.3: 1986-2015
GS52	CCGGATCCAATCCTGAGTTCTAGTTTGATTGC	L1.3: 2715-2683
GS60	CACTTAATTTAAATGGCTGTCGGAATCC	Tx1L EN: 20- 47
GS61	ACACAACCATGGTCTGAGAATGGTGC	Tx1L EN:698-673
GS73	GGAAACCCATCTCACGTG	L1.3: 2115-2132
GS74	GGGATCGGTGGTGATATC	L1.3: 3210-3184
GS75	GAATGGGTCGATCGGGTCACTATACTGGGGTGCATAAATATTTAGG	-
GS76	CCCCAGTATAGTGACCCGATCGACCCATTCATAAAGCAAGTCCTCAG	-
GS77	CTGCCCAGGCACTACGTAGGTCGGGGTCAGGAATTGAACTCAGCTC	-
GS78	ACCCCGACCTACGTAGTGCCTGGGCAGTGATCTGTCTAATGTTGAC	-
GS86	GAAGAACTCGTCAAGAAGGCGATAGAAGG	-
GS87	GCCATTGAACAAGATGGATTGCACGCAGG	-
GS88	CCTTCTATCGCCTTCTTGACGAGTTCTTC	-
GS90	TTCCACACCCTAACTGACACACATTCC	-
GS94	GGACAGGTCGGTCTTGACAAAAAGAACCG	-
GS117	GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG	-
GS118	CCTAACTGCTGTGCCACTGAATTCAGATCTCCCG	-
GS177	bio-CCAGCCACGATAGCCGCGCTGCCTCGTCCTGAAGCTC	-
GS189	GCTGGTGAGGAACTGCGTTCCTTTGG	L1.3: 996- 970
GS190	GTGTCTCTGCCCGGCTTTGGTATCAG	L1.3: 3586-3543
GS260	CAGGTGCTGGAGAGGATGCGGAG	L1.3: 5373-5395
GS261	CCTCAGAAATAATGCCGCATATC	L1.3: 4700-4722
GS262	CTAGAAAACCCCATCGTCTCAGC	L1.3: 4111-4133
GS263	GTGTCGAGGAATGTATCC	L1.3: 3293-3275
GS265	TCAGACGCCGGCGCAATCAAACTAGAACTC	-
GS266	GATTGCGCCGGCGTCTGAGAATGGTGCCAATC	-
GS285	GACAGGATCAAATTCACACATAAC	-
GS286	GGGTGTAAGCAAAGGCGCCCCCTATAATCAAGG	-
GS287	GGGGGCGCCTTTGCTTACACCCTTGATGCTCG	-
GS288	GGGTGTAATTAAAGGCGCCCCCTATAATCAAGG	-
GS289	GGGGGCGCCTTTAATTACACCCTTGATG	-
GS290	ATATATATCCTCGCGATCCGGGATTGAGAAAC	-
GS291	CAATCCCGGATCGCGAGGATATATATATCGAGC	-
GS311	GTCAATTTTGGATCCTCCGGTATATTCTGTTGATTTG	-
GS312	ACAGAATATACCGGAGGATCCAAAATTGACCACATAG	-
GS315	GAGAAACATGGCCATCTCTCACCCTGACATAGGTATATTCTGTTGATTTG	-
GS316	ATGTCAGGGTGAGAGATGGCCATGTTTCTCAATCCAAAATTGACCACATAG	-
GS317	TGGATTCTCCGTTCGCCGTACTGAAGGTATATTCTGTTGATTTG	-
GS318	ATACCTTCAGTACGGCGAACGGAGAATCCAAAATTGACCACATAG	-
GS323	AATAATGGGCGCCTTTGCCACCCCACTGTCAACATTAG	L1.3: 2412-2444
GS324	CAGTGGGGTGGCAAAGGCGCCCATTATTAATGTGTGG	L1.3: 2439-2403
OCI	GACCCGGGAGATCTGAATTC	-
OCII	AGTGGCACAGCAGTTAGG	-

B. Nucleotide Sequence of L1.3 (Genbank accession number L19088) and Amino Acid Sequence of its ORFs

```
   1  gggggaggag ccaagatggc cgaataggaa cagctccggt ctacagctcc cagcgtgagc gacgcagaag acggtgattt
  81  ctgcatttcc atctgaggta ccgggttcat ctcactaggg agtgccagac agtgggcgca ggccagtgtg tgtgcgcacc
 161  gtgcgcgagc cgaagcaggg cgaggcattg cctcacctgg gaagcgcaag ggtcaggga gttccctttc tgagtcaaag
 241  aaaggggtga cggtcgcacc tggaaaatcg ggtcactccc acccgaatat tgcgcttttc agaccggctt aagaaacggc
 321  gcaccacgag actatatccc acacctggct cggagggtcc tacgcccacg gaatctcgct gattgctagc acagcagtct
 401  gagatcaaac tgcaaggcgg caacgaggct gggggagggg cgcccgccat tgcccaggct tgcttaggta aacaaagcag
 481  ccgggaagct cgaactgggt ggagcccacc acagctcaag gaggcctgcc tgcctctgta ggctccacct ctgggggcag
 561  ggcacagaca aacaaaaaga cagcagtaac ctctgcagac ttaagtgtcc ctgtctgaca gctttgaaga gagcagtggt
 641  tctcccagca cgcagctgga gatctgagaa cgggcagaca gactgcctcc tcaagtgggt ccctgactcc tgacccccga
 721  gcagcctaac tgggaggcac cccccagcag gggcacactg acacctcaca cggcagggta ttccaacaga cctgcagctg
 801  agggtcctgt ctgttagaag gaaaactaac aaccagaaag gacatctaca ccgaaaaccc atctgtacat caccatcatc
 881  aaagaccaaa atagataaa accacaaaga tggggaaaaa acagaacaga aaaactggaa actctaaaac gcagagcgcc
   1                          ORF1:    M  G  K   K  Q  N  R   K  T  G   N  S  K   T  Q  S  A

 961  tctcctcctc caaaggaacg cagttcctca ccagcaacgg aacaaagctg gatggagaat gattttgacg agctgagaga
  18    S  P  P   P  K  E   R  S  S  S   P  A  T   E  Q  S   W  M  E  N   D  F  D   E  L  R

1041  agaaggcttc agacgatcaa attactctga gctacgggag gacattcaaa ccaaaggcaa agaagttgaa aactttgaaa
  44  E  E  G  F   R  R  S   N  Y  S   E  L  R  E   D  I  Q   T  K  G   K  E  V  E   N  F  E

1121  aaaatttaga agaatgtata actagaataa ccaatacaga gaagtgctta aaggagctga tggagctgaa aaccaaggct
  71   K  N  L   E  E  C  I   T  R  I   T  N  T   E  K  C  L   K  E  L   M  E  L   K  T  K  A

1201  cgagaactac gtgaagaatg cagaagcctc aggagccgat gcgatcaact ggaagaaagg gtatcagcaa tggaagatga
  98    R  E  L   R  E  E   C  R  S  L   R  S  R   C  D  Q   L  E  E  R   V  S  A   M  E  D

1281  aatgaatgaa atgaagcgag aagggaagtt tagagaaaaa agaataaaaa gaaatgagca aagcctccaa gaaatatggg
 124  E  M  N  E   M  K  R   E  G  K   F  R  E  K   R  I  K   R  N  E   Q  S  L  Q   E  I  W

1361  actatgtgaa aagaccaaat ctacgtctga ttggtgtacc tgaaagtgat gtggagaatg gaaccaagtt ggaaaacact
 151   D  Y  V   K  R  P  N   L  R  L   I  G  V   P  E  S  D   V  E  N   G  T  K   L  E  N  T

1441  ctgcaggata ttatccagga gaacttcccc aatctagcaa ggcaggccaa cgttcagatt caggaaatac agagaacgcc
 178    L  Q  D   I  I  Q   E  N  F  P   N  L  A   R  Q  A   N  V  Q  I   Q  E  I   Q  R  T

1521  acaaagatac tcctcgagaa gagcaactcc aagacacata attgtcagat tcaccaaagt tgaaatgaag gaaaaaatgt
 204  P  Q  R  Y   S  S  R   R  A  T   P  R  H  I   I  V  R   F  T  K   V  E  M  K   E  K  M

1601  taagggcagc cagagagaaa ggtcgggtta ccctcaaagg aaagcccatc agactaacag tggatctctc ggcagaaacc
 231   L  R  A   A  R  E  K   G  R  V   T  L  K   G  K  P  I   R  L  T   V  D  L   S  A  E  T

1681  ctacaagcca gaagagagtg gggccaata ttcaacattc ttaaagaaaa gaattttcaa cccagaattt catatccagc
 258    L  Q  A   R  R  E   W  G  P  I   F  N  I   L  K  E   K  N  F  Q   P  R  I   S  Y  P

1761  caaactaagc ttcataagtg aaggagaaat aaaatacttt atagacaagc aaatgttgag agattttgtc accaccaggc
 284  A  K  L  S   F  I  S   E  G  E   I  K  Y  F   I  D  K   Q  M  L   R  D  F  V   T  T  R

1841  ctgccctaaa agagctcctg aaggaagcgc taaacatgga aaggaacaac cggtaccagc cgctgcaaaa tcatgccaaa
 311   P  A  L   K  E  L  L   K  E  A   L  N  M   E  R  N  N   R  Y  Q   P  L  Q   N  H  A  K

1921  atgtaaagac catcgagact aggaagaaac tgcatcaact aatgagcaaa atcaccagct aacatcataa tgacaggatc
        M  -                                                               ORF2:    M  T  G

2001  aaattcacac ataacaatat taactttaaa tataaatgga ctaaattctg caattaaaag acacagactg gcaagttgga
   4  S  N  S  H   I  T  I   L  T  L   N  I  N  G   L  N  S   A  I  K   R  H  R  L   A  S  W

2081  taaagagtca agacccatca gtgtgctgta ttcaggaaac ccatctcacg tgcagagaca cacataggct caaaataaaa
  31   I  K  S   Q  D  P  S   V  C  C   I  Q  E   T  H  L  T   C  R  D   T  H  R   L  K  I  K

2161  ggatggagga agatctacca agccaatgga aaacaaaaaa aggcaggggt tgcaatccta gtctctgata aaacagactt
  58    G  W  R   K  I  Y   Q  A  N  G   K  Q  K   K  A  G   V  A  I  L   V  S  D   K  T  D

2241  taaaccaaca aagatcaaaa gagacaaaga aggccattac ataatggtaa agggatcaat tcaacaagag gagctaacta
  84  F  K  P  T   K  I  K   R  D  K   E  G  H  Y   I  M  V   K  G  S   I  Q  Q  E   E  L  T

2321  tcctaaatat ttatgcaccc aatacaggag cacccagatt cataaagcaa gtcctcagtg acctacaaag agacttagac
 111   I  L  N   I  Y  A  P   N  T  G   A  P  R   F  I  K  Q   V  L  S   D  L  Q   R  D  L  D

2401  tcccacacat taataatggg agactttaac accccactgt caacattaga cagatcaacg agacagaaag tcaacaagga
 138    S  H  T   L  I  M   G  D  F  N   T  P  L   S  T  L   D  R  S  T   R  Q  K   V  N  K

2481  tacccaggaa ttgaactcag ctctgcacca agcagaccta atagacatct acagaactct ccaccccaaa tcaacagaat
 164  D  T  Q  E   L  N  S   A  L  H   Q  A  D  L   I  D  I   Y  R  T   L  H  P  K   S  T  E

2561  ataccttttt ttcagcacca caccacacct attccaaaat tgaccacata gttggaagta aagctctcct cagcaaatgt
 191   Y  T  F   F  S  A  P   H  H  T   Y  S  K   I  D  H  I   V  G  S   K  A  L   L  S  K  C

2641  aaaagaacag aaattataac aaactatctc tcagaccaca gtgcaatcaa actagaactc aggattaaga atctcactca
 218    K  R  T   E  I  I   T  N  Y  L   S  D  H   S  A  I   K  L  E  L   R  I  K   N  L  T

2721  aagccgctca actacatgga aactgaacaa cctgctcctg aatgactact gggtacataa cgaaatgaag gcagaaataa
 244  Q  S  R  S   T  T  W   K  L  N   N  L  L  L   N  D  Y   W  V  H   N  E  M  K   A  E  I

2801  agatgttctt tgaaaccaac gagaacaaag acaccacata ccagaatctc tgggacgcat tcaaagcagt gtgtagaggg
 271   K  M  F   F  E  T  N   E  N  K   D  T  T   Y  Q  N  L   W  D  A   F  K  A   V  C  R  G

2881  aaatttatag cactaaatgc ctacaagaga aagcaggaaa gatccaaaat tgacacccta acatcacaat taaaagaact
 298    K  F  I   A  L  N   A  Y  K  R   K  Q  E   R  S  K   I  D  T  L   T  S  Q   L  K  E

2961  agaaaagcaa gagcaaacac attcaaaagc tagcagaagg caagaaataa ctaaaatcag agcagaactg aaggaaatag
```

```
 324  L  E  K  Q   E  Q  T   H  S  K   A  S  R  R   Q  E  I   T  K  I   R  A  E  L   K  E  I
3041  agacacaaaa aacccttcaa aaaatcaatg aatccaggag ctggtttttt gaaaggatca acaaaattga tagaccgcta
 351   E  T  Q   K  T  L  Q   K  I  N   E  S  R   S  W  F  F   E  R  I   N  K  I   D  R  P  L
3121  gcaagactaa taaagaaaaa aagagagaag aatcaaatag acacaataaa aaatgataaa ggggatatca ccaccgatcc
 378    A  R  L   I  K  K   K  R  E  K   N  Q  I   D  T  I   K  N  D  K   G  D  I   T  T  D
3201  cacagaaata caaactacca tcagagaata ctacaaacac ctctacgcaa ataaactaga aaatctagaa gaaatggata
 404  P  T  E  I   Q  T  T   I  R  E   Y  Y  K  H   L  Y  A   N  K  L   E  N  L  E   E  M  D
3281  cattcctcga cacatacact ctcccaagac taaaccagga agaagttgaa tctctgaata gaccaataac aggctctgaa
 431   T  F  L   D  T  Y  T   L  P  R   L  N  Q   E  E  V  E   S  L  N   R  P  I   T  G  S  E
3361  attgtggcaa taatcaatag tttaccaacc aaaaagagtc caggaccaga tggattcaca gccgaattct accagaggta
 458    I  V  A   I  I  N   S  L  P  T   K  K  S   P  G  P   D  G  F  T   A  E  F   Y  Q  R
3441  catggaggaa ctggtaccat tccttctgaa actattccaa tcaatagaaa aagagggaat cctccctaac tcattttatg
 484  Y  M  E  E   L  V  P   F  L  L   K  L  F  Q   S  I  E   K  E  G   I  L  P  N   S  F  Y
3521  aggccagcat cattctgata ccaaagccgg gcagagacac aaccaaaaaa gagaatttta gaccaatatc cttgatgaac
 511   E  A  S   I  I  L  I   P  K  P   G  R  D   T  T  K  K   E  N  F   R  P  I   S  L  M  N
3601  attgatgcaa aaatcctcaa taaaatactg gcaaaccgaa tccagcagca catcaaaaag cttatccacc atgatcaagt
 538    I  D  A   K  I  L   N  K  I  L   A  N  R   I  Q  Q   H  I  K  K   L  I  H   H  D  Q
3681  gggcttcatc cctgggatgc aaggctggtt caatatacgc aaatcaataa atgtaatcca gcatataaac agagccaaag
 564  V  G  F  I   P  G  M   Q  G  W   F  N  I  R   K  S  I   N  V  I   Q  H  I  N   R  A  K
3761  acaaaaacca catgattatc tcaatagatg cagaaaaagc ctttgacaaa attcaacaac ccttcatgct aaaaactctc
 591   D  K  N   H  M  I  I   S  I  D   A  E  K   A  F  D  K   I  Q  Q   P  F  M   L  K  T  L
3841  aataaattag gtattgatgg gacgtatttc aaaataataa gagctatcta tgacaaaccc acagccaata tcatactgaa
 618    N  K  L   G  I  D   G  T  Y  F   K  I  I   R  A  I   Y  D  K  P   T  A  N   I  I  L
3921  tgggcaaaaa ctggaagcat tccctttgaa aaccggcaca agacagggat gccctctctc accgctccta ttcaacatag
 644  N  G  Q  K   L  E  A   F  P  L   K  T  G  T   R  Q  G   C  P  L   S  P  L  L   F  N  I
4001  tgttggaagt tctggccagg gcaatcaggc aggagaagga aataaagggt attcaattag gaaaagagga agtcaaattg
 671   V  L  E   V  L  A  R   A  I  R   Q  E  K   E  I  K  G   I  Q  L   G  K  E   E  V  K  L
4081  tccctgtttg cagacgacat gattgtatat ctagaaaacc ccatcgtctc agcccaaaat ctccttaagc tgataagcaa
 698    S  L  F   A  D  D   M  I  V  Y   L  E  N   P  I  V   S  A  Q  N   L  L  K   L  I  S
4161  cttcagcaaa gtctcaggat acaaaatcaa tgtacaaaaa tcacaagcat tcttatacac caacaacaga caaacagaga
 724  N  F  S  K   V  S  G   Y  K  I   N  V  Q  K   S  Q  A   F  L  Y   T  N  N  R   Q  T  E
4241  gccaaatcat gggtgaactc ccattcgtaa ttgcttcaaa gagaataaaa tacctaggaa tccaacttac aagggatgtg
 751   S  Q  I   M  G  E  L   P  F  V   I  A  S   K  R  I  K   Y  L  G   I  Q  L   T  R  D  V
4321  aaggacctct tcaaggagaa ctacaaacca ctgctcaagg aaataaaaga ggacacaaac aaatggaaga acattccatg
 778    K  D  L   F  K  E   N  Y  K  P   L  L  K   E  I  K   E  D  T  N   K  W  K   N  I  P
4401  ctcatgggta ggaagaatca atatcgtgaa aatggccata ctgcccaagg taatttacag attcaatgcc atccccatca
 804  C  S  W  V   G  R  I   N  I  V   K  M  A  I   L  P  K   V  I  Y   R  F  N  A   I  P  I
4481  agctaccaat gactttcttc acagaattgg aaaaaactac tttaaagttc atatggaacc aaaaaagagc ccgcattgcc
 831   K  L  P   M  T  F  F   T  E  L   E  K  T   T  L  K  F   I  W  N   Q  K  R   A  R  I  A
4561  aagtcaatcc taagccaaaa gaacaaagct ggaggcatca cactacctga cttcaaacta tactacaagg ctacagtaac
 858    K  S  I   L  S  Q   K  N  K  A   G  G  I   T  L  P   D  F  K  L   Y  Y  K   A  T  V
4641  caaaacagca tggtactggt accaaaacag agatatagat caatggaaca gaacagagcc ctcagaaata atgccgcata
 884  T  K  T  A   W  Y  W   Y  Q  N   R  D  I  D   Q  W  N   R  T  E   P  S  E  I   M  P  H
4721  tctacaacta tctgatcttt gacaaacctg agaaaaacaa gcaatgggga aaggattccc tatttaataa atggtgctgg
 911   I  Y  N   Y  L  I  F   D  K  P   E  K  N   K  Q  W  G   K  D  S   L  F  N   K  W  C  W
4801  gaaaactggc tagccatatg tagaaagctg aaactggatc ccttccttac accttataca aaaatcaatt caagatggat
 938    E  N  W   L  A  I   C  R  K  L   K  L  D   P  F  L   T  P  Y  T   K  I  N   S  R  W
4881  taaagattta aacgttaaac ctaaaaccat aaaaacccta gaagaaaacc taggcattac cattcaggac ataggcgtgg
 964  I  K  D  L   N  V  K   P  K  T   I  K  T  L   E  E  N   L  G  I   T  I  Q  D   I  G  V
4961  gcaaggactt catgtccaaa acaccaaaag caatggcaac aaaagacaaa attgacaaat gggatctaat taaactaaag
 991   G  K  D   F  M  S  K   T  P  K   A  M  A   T  K  D  K   I  D  K   W  D  L   I  K  L  K
5041  agcttctgca cagcaaaaga aactaccatc agagtgaaca ggcaacctac aacatgggag aaaatttcg caacctactc
1018    S  F  C   T  A  K   E  T  T  I   R  V  N   R  Q  P   T  T  W  E   K  I  F   A  T  Y
5121  atctgacaaa gggctaatat ccagaatcta caatgaactt aaacaaattt acaagaaaaa aacaaacaac cccatcaaaa
1044  S  S  D  K   G  L  I   S  R  I   Y  N  E  L   K  Q  I   Y  K  K   K  T  N  N   P  I  K
5201  agtgggcgaa ggacatgaac agacacttct caaaagaaga catttatgca gccaaaaaac acatgaagaa atgctcatca
1071   K  W  A   K  D  M  N   R  H  F   S  K  E   D  I  Y  A   A  K  K   H  M  K   K  C  S  S
5281  tcactggcca tcagagaaat gcaaatcaaa accactatga gatatcatct cacaccagtt agaatggcaa tcattaaaaa
1098    S  L  A   I  R  E   M  Q  I  K   T  T  M   R  Y  H   L  T  P  V   R  M  A   I  I  K
5361  gtcaggaaac aacaggtgct ggagaggatg cggagaaata ggaacacttt tacactgttg gtgggactgt aaactagttc
1124  K  S  G  N   N  R  C   W  R  G   C  G  E  I   G  T  L   L  H  C   W  W  D  C   K  L  V
5441  aaccattgtg gaagtcagtg tggcgattcc tcagggatct agaactagaa ataccatttg acccagccat cccattactg
1151   Q  P  L   W  K  S  V   W  R  F   L  R  D   L  E  L  E   I  P  F   D  P  A   I  P  L  L
5521  ggtatatacc caaatgagta taaatcatgc tgctataaag acacatgcac acgtatgttt attgcggcac tattcacaat
1178    G  I  Y   P  N  E   Y  K  S  C   C  Y  K   D  T  C   T  R  M  F   I  A  A   L  F  T
5601  agcaaagact tggaaccaac ccaaatgtcc aacaatgata gactggatta agaaaatgtg gcacatatac accatggaat
1204  I  A  K  T   W  N  Q   P  K  C   P  T  M  I   D  W  I   K  K  M   W  H  I  Y   T  M  E
```

```
5681  actatgcagc cataaaaaat gatgagttca tatcctttgt agggacatgg atgaaattgg aaaccatcat tctcagtaaa
1231   Y  Y  A   A  I  K  N   D  E  F   I  S  F   V  G  T  W   M  K  L   E  T  I   I  L  S  K

5761  ctatcgcaag aacaaaaaac caaacaccgc atattctcac tcataggtgg gaattgaaca atgagatcac atggacacag
1258    L  S  Q   E  Q  K   T  K  H  R   I  F  S   L  I  G   G  N  -

5841  gaaggggaat atcacactct ggggactgtg gtggggtcgg gggagggggg agggatagca ttgggagata tacctaatgc
5921  tagatgacac attagtgggt gcagcgcacc agcatggcac atgtatacat atgtaactaa cctgcacaat gtgcacatgt
6001  accctaaaac ttaaagtata ataaa
```

Nucleotide and Amino Acid Sequence of Tx1L EN

(GenBank accession number M26915)

```
  1  atggccttga gtataagcac acttaatact aatggctgtc ggaatccttt ccgaatgttt caggtactct cctttcttcg
  1    M  A  L   S  I  S   T  L  N  T   N  G  C   R  N  P   F  R  M  F   Q  V  L   S  F  L

 81  tcaaggaggg tactctgtga gtttcctcca agagacccac accactccag agcttgaagc aagctggaat ctggagtgga
 27  R  Q  G  G   Y  S  V   S  F  L   Q  E  T  H   T  T  P   E  L  E   A  S  W  N   L  E  W

161  agggaagggt ctttttaat cacctcactt ggacatcatg cggggtggtg accctttct cagattcctt ccagccagag
 54   K  G  R   V  F  F  N   H  L  T   W  T  S   C  G  V  V   T  L  F   S  D  S   F  Q  P  E

241  gtcctgagtg ctacctctgt catccctggc cgtctattgc atcttcgggt ccgggagtca ggtagaacat ataatctaat
 81    V  L  S   A  T  S   V  I  P  G   R  L  L   H  L  R   V  R  E  S   G  R  T   Y  N  L

321  gaatgtgtat gctcctacta ccggaccaga gagggcacgg ttctttgaaa gtttgtcagc ctacatggag acaattgact
107  M  N  V  Y   A  P  T   T  G  P   E  R  A  R   F  F  E   S  L  S   A  Y  M  E   T  I  D

401  ctgatgaagc cttgattata gggggtgatt ttaattacac ccttgatgct cgagatcgca atgtacccaa gaaaagagac
134   S  D  E   A  L  I  I   G  G  D   F  N  Y   T  L  D  A   R  D  R   N  V  P   K  K  R  D

481  tcgtctgagt ccgttttgcg agaactaatt gctcatttct ccttggttga tgtctggaga gaacagaacc cagagacggt
161    S  S  E   S  V  L   R  E  L  I   A  H  F   S  L  V   D  V  W  R   E  Q  N   P  E  T

561  tgcctttacc tatgtcaggg tgagagatgg tcatgtttct caatcccgga ttgataggat atatatatcg agccatctca
187  V  A  F  T   Y  V  R   V  R  D   G  H  V  S   Q  S  R   I  D  R   I  Y  I  S   S  H  L

641  tgtcacgagc ccagtcgagc accattagat tggcaccatt ctcagaccac aattgtgtat ccctgagaat gtcaatcaga
214   M  S  R   A  Q  S  S   T  I  R   L  A  P   F  S  D  H   N  C  V   S  L  R   M  S  I  R

721  ggatct
241    G  S
```

Nucleotide and Amino Acid Sequence of R1Bm EN

(GenBank accession number M19755)

```
  1  atggatatta ggccccgact tcgtattggc caaatcaatc tgggtggtgc agaggatgcg acgagggagc taccctccat
  1    M  D  I   R  P  R   L  R  I  G   Q  I  N   L  G  G   A  E  D  A   T  R  E   L  P  S

 81  tgcacgggat ctcggcctgg atattgttct tgtacaggaa caatattcca tggtcgggtt cctagcccaa tgtggagcac
 27  I  A  R  D   L  G  L   D  I  V   L  V  Q  E   Q  Y  S   M  V  G   F  L  A  Q   C  G  A

161  accccaaggc gggtgtgtat atccgcaata gggtgctccc ctgcgcggtt ctgcaccacc ttagcagcac acatataacg
 54   H  P  K   A  G  V  Y   I  R  N   R  V  L   P  C  A  V   L  H  H   L  S  S   T  H  I  T

241  gtagtgcaca ttgggggtg ggacttatat atggtgtctg cgtacttcca gtatagtgac cctattgacc catacctgca
 81    V  V  H   I  G  G   W  D  L  Y   M  V  S   A  Y  F   Q  Y  S  D   P  I  D   P  Y  L

321  ccggctcggg aatattcttg accggctgcg gggggctcgg gtcgttatct gcgcagacac taatgcccac tcgccattgt
107  H  R  L  G   N  I  L   D  R  L   R  G  A  R   V  V  I   C  A  D   T  N  A  H   S  P  L

401  ggcactcgct gcccaggcac tacgtcggtc ggggtcagga agtggctgac cgccgcgcca agatggagga tttcattggg
134   W  H  S   L  P  R  H   Y  V  G   R  G  Q   E  V  A  D   R  R  A   K  M  E   D  F  I  G

481  gcgaggcggt tggtcgtcca taacgcggat ggccacctgc cgaccttcag tacggcgaac ggagaatctt atgtcgatgt
161    A  R  R   L  V  V   H  N  A  D   G  H  L   P  T  F   S  T  A  N   G  E  S   Y  V  D

561  cacgctgtct acgcggggag tacgcgtgtc tgaatggcgt gtaactaatg aatcatcgag cgatcaccgg ctcattgtgt
187  V  T  L  S   T  R  G   V  R  V   S  E  W  R   V  T  N   E  S  S   S  D  H  R   L  I  V

641  ttggggtggg gggcggt
214   F  G  V   G  G  G
```

C. TABLE OF CHARACTERISED INTEGRATION EVENTS

pNZ45 (α8-swap):

clone	Chr., arm	TSD	length of TSD [bp]	length of poly(A) [bp]	length of L1-ins. [bp]	5'-homology
G1W6	1q23.1	gtat GAAATGTAAAATAAGcaca	15	60	6513+ 226inv	n. a.
G2W1	1p35.3	ttat aaGAGAATACTATGAATAat	20	80	2206	AT
G3W5	5p15.31	atgt aGAAAACACAgaaaagagcg	15	24	3521	GAAAA
G2W3	16q22.2	cttt aaaaaggaagggaattctga	3	65	2039	TAAA
w3.1	5q35.1	gaat GAAAACTAATGTTTattgaa	14	80	2125+ 3452inv	n. a.
G1W4	5p15.31	ttgt aaaaacaaacggcgttgtct	n. d.	80	n. d.	n. d.
G3W3	3q24	aaat aaaaacttcacaaagtggtt	n. d.	50	n. d.	n. d.
G3W4/1	18q22.3	ccat aaaactatctaccaagaata	n. d.	110	n. d.	n. d.
G3W6	15q22.31	tctc aaaaaaaaaaaaaaaaaaaa	n. d.	60	n. d.	n. d.

pNZ39 (Tx1-EN):

clone	Chr., arm	TSD	length of TSD [bp]	length of poly(A) [bp]	length of L1-ins. [bp]	5'-homology
22	12q23.1	taat GAATGTTN$_{106}$ATTTTTTGic	122 + i 11	65	1500	n. a.
24	4p14	acag aaCCAAGN$_{20}$AGGAgggactg	34	37	2265	GGG
26	10q11.21	tctc aaaaaaaN$_{55}$GCAGGGAiagc	91 + i 2	40	1944	n. a.
28	21q22.3	caag aaaCCCTcaaaagtattttg	10	14	3248	CAA
29	Xq21.2	ggac aaaaaaTN$_{89}$GAGTTTaatta	106	25	3092	AA
34/1	11q14.3	cctc ataatcaaatgaccctaaat	3	25	1436	TA
39D	12q24.11	taag aaGAAAAN$_{22}$ACCCTaataga	36	140	1710	AA
T6	1p34.1	cccc aGATGAGN$_{48}$ACCCTTGiacc	62 + i 8	40	1502	n. a.
T11	4p15.2	agat aaaaTGAN$_{16}$TAATTAAigcc	30 + i 16	19	1723	n. a.
T16	3q12.1	ccat aaaaaatgatgagttcatgt	18	30	1872	n. a.
T21	9p21.2	gggc aaaaGAAN$_{69}$TAGaagaaatt	83	34	3509	AAGA
T22	19p13.12	aaac aaagagttttgtttgtaaaa	5	60	1907	AAAGA
T23	3q13.33	aact TAAAAATTACATAccacttc	16	110	1848	CCA
20	17p13.3	tctc attatattgcccaggctggt	Δ 39	50	1486	n. a.
27	14q21.3	aagt tgagtaattttaacactaaa	Δ 24	-	1966	n. a.
45	2q36.1	gttt aaaaaaaaattagatgaaac	Δ 6/i 3	70	2057	n. a.
T12	13q14.2	tttt actttttctgctattattc	Δ 6	100	1561	n. a.
21	7q11.23	tatg aaaaacaaatggaaaacatc	n. d.	8	n. d.	n. d.
34/2	12q22	gcat gttctcactcataggtggaa	n. d.	-	n. d.	n. d
39E	5q14.3	tttt aaatatggaagtgtacatgt	n. d.	90	n. d.	n. d.
40	6p22.3	tctc aaaaaaaaaaaaaaaaaaaa	n. d.	70	n. d.	n. d.
T4	2q22.3	gtgt aaaaaataaaagagaaaatc	n. d.	50	n. d.	n. d.
T10	2q32.3	agcc aaaaaaaattaatatcaacc	n. d.	120	n. d.	n. d.

pNZ75 (α11-swap):

Clone	Chr., arm	TSD	Length of TSD [bp]	Length of polyA [bp]*	Length of L1-int. [bp]	5' homology
2	11p14.1	tttt aaaCAAATTTTAtagaggtg	13	75 (3)	3301	T
6	Xq25	attt GAAATTCN$_{48}$CTCCcatcaac	61	90 (3)	2698	CAT
7	4q32.3	aaag AAGACATTTACGtggtcaac	12	95 (1)	2963	--
9	4q13.1	cctt AAAaattaaaaattactttg	3	55 (3)	2758	A
47	1p22.1	actt AGAAAAAN$_{40}$GTacatttgga	54	40 (3)	2729	ACATT
57	1p31.3	acct AAGAAAATN$_{27}$TATTTCAigt	41 + i 1	80 (1)	3023	n.a.
1	1p31.1	ttag aagaattataagagacctaa	Δ 11	45 (2)	4050	n. a.
10	1p31.2	aaat aaaaaatatatatatataca	Δ 24	100 (1)	2959	n. a.
51	20q13.33	taag aagagaaagagagacctgag	Δ 30	60 (1)	3462	n. a.
59	6p12.3	gatt gaaagaatcaatataatgca	Δ 10	65 (1)	3623	n. a.
13	4p15.33	tgtg acaaatgcaatctcttatct	n. d.	65	n. d.	n. d.
73	11p15.4	acac agaaaatctcaaacttccca	n. d.	100	n. d.	n. d.
74	6p25.3	ctct aaaagttatttttttacttt	n. d.	70	n. d.	n. d.
82	8q11.21	tttg aaaatttctacaataggca	n. d.	50	n. d.	n. d.
5	3q26.2	n. d.	n. d.	n. d.	>4654	n. d
8	22q11.22	n. d.	n. d.	130 (1)	3075	n. d
50	19q13.33	n. d.	n. d.	n. d.	>4654	n. d
53	15q13.1	n. d.	n. d.	n. d.	2718	n. d
54	6q22.1	n. d.	n. d.	>10 (3)	>4654	n. d.
55	8q13.2	n. d.	n. d.	n. d.	4082	n. d
56	3p25.3	n. d.	n. d.	n. d.	3863	n. d
58	12p11.22	n. d.	n. d.	>20 (3)	>4654	n. d
60	5p13.3	n. d.	n. d.	120 (2)	>4654	n. d

* numbers in brackets indicate which of the alternative poly(A) signals is used: SV40pA$_1$ (1), L1pA (2) or SV40pA$_2$ (3).

i: insertion of untemplated nucleotides
inv: inversion
Δ: target site deletion
n. a.: not applicable
n. d.: not determined

D. GEFAHRENMERKMALE UND SICHERHEITSRATSCHLÄGE FÜR VERWENDETE GEFAHRSTOFFE

3-Methylbutanol (Isoamylalkohol)
R10: Entzündlich.
R20: Gesundheitsschädlich beim Einatmen.
S24/25: Berührung mit den Augen und der Haut vermeiden.

Chloroform

R22: Gesundheitsschädlich beim Verschlucken.
R38: Reizt die Haut.
R40: Irreversibler Schaden möglich.
R48/20/22: Gesundheitsschädlich: Gefahr ernster Gesundheitsschäden bei längerer Exposition durch Einatmen oder durch Verschlucken.
S36/37: Bei der Arbeit geeignete Schutzhandschuhe und Schutzkleidung tragen.

Ethidiumbromid

R23: Giftig beim Einatmen.
R68: Irreversibler Schaden möglich.
S36/37: Bei der Arbeit geeignete Schutzhandschuhe und Schutzkleidung tragen.
S45: Bei Unfall oder Unwohlsein sofort Arzt zuziehen (wenn möglich, das Etikett vorzeigen).

Phenol

R24/25: Giftig bei Berührung mit der Haut und beim Verschlucken.
R34: Verursacht Verätzungen.
S28: Bei Berührung mit der Haut sofort abwaschen mit viel Wasser.
S45: Bei Unfall oder Unwohlsein sofort Arzt zuziehen (wenn möglich, das Etikett vorzeigen).

Wasserstoffperoxid

R34: Verursacht Verätzungen
S28: Bei Berührung mit der Haut sofort abwaschen mit viel Wasser.
S36/39: Bei der Arbeit geeignete Schutzkleidung und Schutzbrille/Gesichtsschutz tragen.
S45: Bei Unfall oder Unwohlsein sofort Arzt zuziehen (wenn möglich, das Etikett vorzeigen).

E. CURRICULUM VITAE

Nora Zingler

Geburt: 1. April 1974 in München

Schulausbildung:

9/1980 - 8/1984 Grundschule Ismaning

9/1984 - 5/1993 Werner-Heisenberg-Gymnasium Garching
Abitur

Wissenschaftlicher Werdegang:

10/1993 - 2/1999 Studium der Chemie an der Technischen Universität München

8/1996 – 2/1997 Auslandssemester an der University of St. Andrews, Scotland

9/1998 - 2/1999 Diplomarbeit am Lehrstuhl für Organische Chemie und Biochemie der Technischen Universität München (Vorstand Prof. Dr. Dr. A. Bacher)

Thema: "Die Domänenstruktur der Riboflavinsynthase"

3/1999 – 6/1999 Wissenschaftliche Mitarbeiterin am Lehrstuhl für Organische Chemie und Biochemie der Technischen Universität München (abschließende Arbeiten zur Veröffentlichung der Ergebnisse der Diplomarbeit)

8/1999 – Dissertation
in der Abteilung für Zell- und Virusgenetik (Vorstand Prof. Dr. W. Ostertag) am Heinrich-Pette-Institut für experimentelle Virologie und Immunologie an der Universität Hamburg,
am Universitätsklinikum Hamburg-Eppendorf, Hamburg und
am Paul-Ehrlich-Institut, Bundesamt für Sera und Impfstoffe, Langen
unter Leitung von PD Dr. Gerald G. Schumann

Thema: "Mechanisms Involved in Target Sequence Recognition and Integration of Human LINE-1 Retrotransposons"

F. PUBLIKATIONSLISTE

Publikationen:

Eberhardt, S., Zingler, N., Kemter, K., Richter, G., Cushman, M. and Bacher, A. (2001).
Domain structure of riboflavin synthase. *Eur J Biochem*. **268**(15), 4315-23.

Yu, F., Zingler, N., Schumann, G. and Strätling, W.H. (2001).
Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res.*, **29**, 4493-4501.

Zingler, N., Weichenrieder, O. and Schumann, G. (in press).
APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenetic and Genome Research*.

Zingler, N., Willhöft, U., Schoder, V., Brose, H.-P., Schumann, G.
Comparison of wild-type and mutant LINE-1 integrants suggests two alternative pathways for second-strand synthesis.
(Manuskript in Vorbereitung)

Tagungsbeiträge und Vorträge:

N. Zingler und G. G. Schumann
„Targeting LINE-1 Retrotransposition – Hybrid Retrotransposon L1/Tx1L is Transposition Competent in Cultured Cells“
Poster im Rahmen der Jahrestagung der Gesellschaft für Virologie 2002, Erlangen, 8.-10.4.02

N. Zingler und G. G. Schumann
“Altering the Target Site Specificity of the Human LINE-1 element, a Member of the Family of APE-type Non-LTR Retrotransposons”
Vortrag am Netherlands Cancer Institute, Amsterdam, Niederlande, 17.4.03

N. Zingler
„The human LINE1 retrotransposon utilizes a cellular DSB repair pathway for replication.”
Vortrag im Rahmen des DNA-Repair Network: Workshop “Recombination and Repair”, München, 5.-6.9.03

N. Zingler, U. Willhöft, V. Schoder, H.-P. Brose, und G. G. Schumann
“The Human LINE-1 Retrotransposon Utilizes a Cellular Double Strand Break Repair Pathway for Replication”
Poster im Rahmen der Jahrestagung der Gesellschaft für Virologie 2004, Tübingen,
17.-20.3.04

DANKSAGUNG

Ich möchte allen danken, die durch ihren Beitrag das Entstehen dieser Arbeit ermöglicht haben.

PD Dr. Gerald G. Schumann danke ich für die interessante Themenstellung und die sorgfältige Betreuung dieser Arbeit, die ohne sein erfolgreiches Einwerben von Drittmitteln nicht möglich gewesen wäre. Die anregenden, zuweilen hitzig geführten Diskussionen mit ihm ließen immer ausreichend Spielraum für eigene Wege.

Prof. Dr. Wolfram Ostertag danke ich für die Übernahme der formellen Betreuung, interessante Denkanstöße und sein stetes Interesse am Fortgang meiner Arbeit.

Prof. Dr. H. Marquardt möchte ich für die freundliche Unterstützung und die Vertretung meiner Arbeit gegenüber dem Fachbereich Chemie der Universität Hamburg danken.

Prof. Dr. W. Strätling bin ich für die vorübergehende Aufnahme in sein Labor sehr zu Dank verpflichtet. Außerdem danke ich ihm und Dr. Fang Yu für die produktive Zusammenarbeit am MeCP2-Projekt.

Bei unserem Kooperationspartner Dr. Oliver Weichenrieder möchte ich mich für viele anregende Diskussionen und die Bereitstellung einiger Abbildungen für diese Arbeit bedanken.

Dr. Ute Willhöft, Hans-Peter Brose und Volker Schoder bin ich für die Zusammenarbeit und Unterstützung bei Datenbankauswertungen und statistischen Analysen zu Dank verpflichtet.

Allen Mitgliedern der Arbeitskreise von W. Ostertag, J.-M. Buerstedde, W. Strätling und R. Löwer danke ich für die freundliche Aufnahme und das stets hervorragende Arbeitsklima. Anne Buchwald, Isabella Serafin und Heike Strobel haben mich bei Klonierungs- und anderen Laborarbeiten tatkräftig unterstützt – danke! My special thanks to Vrun – without your help, I'd probably still be cloning….

Frau Dr. Carol Stocking möchte ich für ihre stete Hilfs- und Diskussionsbereitschaft danken, unter anderem auch bei der Korrektur dieser Arbeit. Ihre Anregungen und ihre Unterstützung waren für mich immer sehr wertvoll.

Herrn Dr. Boris Fehse danke ich für seine Freundschaft- und natürlich für's Korrekturlesen.

Auch Frau Dr. Gabriele Rauch möchte ich herzlich für die Durchsicht meiner Arbeit danken.

Mein besonderer Dank gilt jedoch meinen Eltern, die mich immer uneingeschränkt unterstützt haben.

Danke!

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, daß ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfaßt, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den verwendeten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ferner versichere ich, daß ich diese Dissertation noch an keiner anderen Universität eingereicht habe, um ein Promotionsverfahren eröffnen zu lassen.

Hamburg, den

Nora Zingler