Structural and Functional Studies on the LINE-1 Retrotransposon Endonuclease

Structurele en functionele studies over LINE-1 retrotransposon endonuclease

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. S.W.J. Lamberts

and in accordance with the decision of the Doctorate Board

The public defence shall be held on

Thursday 20 September 2007 at 9.00 hrs

by

Konstantinos Repanas

born in Kozani, Greece

Doctoral Committee

Promotor :

Prof.dr. T.K. Sixma

Other members:

Prof.dr. C.P. Verrijzer

Prof.dr. R. Kanaar

Dr. A.M.J.J Bonvin

Copromotor :

Dr. A. Perrakis

Table of Contents

Chapter 1

Opening LINEs: General Introduction

Opening LINEs : *General Introduction*

Historical Background: The discovery of mobile genetic elements

It is generally accepted that we owe the discovery of mobile genetic elements to American scientist Barbara McClintock. However, others before her did indeed encounter transposons, but without being able to recognize them. The early work of maize geneticist Emerson (Emerson, 1917) put him ahead of his peers because he managed to understand what others failed to; how an unstable mutation responsible for stripes of dark red pigmentation on the kernels of corn behaves within the Mendelian paradigm. He realized that a certain inhibiting factor was affecting the pigmentation gene, but could not exactly describe what this factor could be. This was followed up by Marcus Rhoades (Rhoades, 1945) who found out that a stable null mutation on the A locus of maize could be converted to an unstable one if combined with a locus he named Dotted.

It was this type of unstable mutation that Barbara McClintock managed to explain by predicting the existence of jumping genes or mobile genetic elements. While studying chromosome breakage in corn, she noticed the repeated loss of a fragment of chromosome 9, and termed this breakage site the Dissociation (Ds) locus. She further observed that another locus was essential for chromosome breakage activation, the Activator (Ac) locus, and that chromosome breakage was linked to frequency and timing of mutations leading to variegation of maize kernels. Her genetic experiments led to the realization that Ds was indeed moving to a new location on chromosome 9 and she named such an element "Controlling element". Her "Controlling elements" were transposon insertions disrupting certain genes, causing the type of unstable mutation that Emerson and Rhoades were trying to explain.

Barbara McClintock, in contrast to general belief argued that transposition must have a more important function than to turn maize color genes on and off. She was the first to hypothesize that transposable elements provide a mechanism to rapidly reorganize the genome in response to environmental stress. In this sense mutations produced by mobile genetic elements would be seen as a source of variation to drive the evolutionary process. Current day large-scale sequencing projects, made possible by recent technological advances, come as an extra proof that Barbara McClintock was not really "…out of her mind" back in the 40's; we cannot claim her ideas were very well received at the time. Compensation was of course offered in the form of a nobel prize in 1983, although she was often referring to it as being rather a "complication" (Fedoroff, 2001; McClintock, 1950; McClintock, 1956; McClintock, 1984; http://nobelprize.org).

Types and evolutionary origin of mobile elements

If the origins of life are in an "RNA world" which at some point was reverse transcribed into DNA, then the various mobile genetic elements must have an instrumental role in genome evolution. Mammalian genomes have been colonized in the course of evolution by transposable elements that now account for more than a half of the mass of the human genome. The estimations vary, but most researchers believe that it must be even more, since ancient mobile elements that have been inactivated, have diverged by mutation to the point where they are unidentifiable. Researchers involved in genome sequencing projects must be by now pretty well acquainted with transposable elements, since the latter over-flood the formers' results (Brosius, 1991; Brosius, 2005; Kazazian, 2004; Wessler, 2006).

The classic transposons that operate via a 'cut and paste' mechanism on DNA level are a minority among these 'intragenomic parasites' (Orgel & Crick, 1980). The vast majority of mobile genetic elements are retrotransposons that propagate via an RNA intermediate and are still active in the human genome, constituting the most important transposable element class. Retroelements are present in the genomes of virtually all eukaryotes and can be subdivided into two general structural classes. Long terminal repeat (LTR) retrotransposons, which resemble simple retroviruses, and non-LTR retrotransposons, that lack LTRs and generally terminate in a polyadenylic acid (polyA) tail. Non-LTR retrotransposons can be further classified in autonomous or non-autonomous elements; depending on whether they encode their own retrotransposition machinery, which includes reverse transcriptase (RT) and endonuclease (EN) activities or whether they rely on an external source for RT-EN. Non-LTR retrotransposons are thought to have a very long history of about 500-600 million years (Weiner

et al, 1986; Kazazian & Moran, 1998; Kazazian, 2004) (Table 1).

DNA Transposons

DNA transposons are mobile sequences found mainly in bacteria, but also in humans where their almost 300.000 copies form 3% of our genome. Their propagation does not involve an RNA intermediate, since they jump from one genomic location to the other by a "cut and paste" mechanism. Typically 1-3 kilobases in length, they consist of a central transposase region that is flanked by Inverted Terminal Repeats (ITR). Integration specificity only involves a few nucleotides, which means that insertions happen at many different sites. Most new insertions though occur near the parental element, an observation called "local hopping" (Kazazian, 2004; Babushok & Kazazian, 2007).

Excision and integration is driven by the encoded transposase; it binds the ITRs and the target DNA and by a breakage/joining reaction moves the element to a new site. Upon integration target site duplications are formed from the host DNA. The target site duplication remains in the genome as a transposon signature, when the transposon moves from one place to another. Representative elements in humans include MER1-Charlie, MER2-Tigger and Mariner. The, by now famous, pair of synthetic DNA transposons Sleeping Beauty and the Frog Prince able to transpose in humans, also belong to this category of elements and may prove invaluable tools for gene therapy (Fedoroff, 1989; Plasterk, 1996; Ivics et al, 1997; Miskey et al, 2003; Ivics & Izsvák, 2006).

Long Terminal Repeat Retrotransposons

Mechanistically different from DNA transposons are the retrotransposons that are flanked by long terminal repeats (LTRs). This class of elements makes up 9% of the human genome with about 443.000 copies and includes retrotransposons and endogenous retroviruses like HERVK10, ERVL and MaLR. Full-length elements are typically 6-11 kb in length, but many of them are truncated, especially at the 5′ end (Kazazian, 2004; Bannert & Kurth, 2004; Babushok & Kazazian, 2007).

During their propagation they transcribe an RNA intermediate, which then gets reverse transcribed and reintegrated, producing a complete double-stranded cDNA copy of the original retrotransposon. The final integration step is reminiscent of the action of transposases that mobilise DNA transposons. Based on sequence analysis it has been illustrated that LTR retrotransposons have probably resulted from the combination of activities of an ancient non-LTR retrotransposon and an ancient DNA transposon (Malik & Eickbush, 2001).

LTR retrotransposons and retroviruses are very similar. They contain gag (group specific antigen), pol (polymerase) and prt (protease) genes. Gag encodes the viral particle coat and pol includes reverse transcriptase (RT) activity for first and second strand DNA synthesis, ribonuclease H (RH) activity for cleavage of RNA in the RNA/DNA hybrid after first strand synthesis and integrase (IN) activity that cleaves target DNA and ligates the retrovirus into the cleaved site. The basic difference is that retroviruses contain an env gene, making the envelope protein that enables them to move from one cell to the other. On the contrary, LTR retrotransposons do not code for an envelope protein and they are bound to insert only to the genome of origin. LTR elements share many features with the HIV (human immunodeficiency virus) and MLV (mouse leukaemia virus) (Bushman, 2003; Kazazian, 2004).

Autonomous non-LTR Retrotransposons

This type of mobile genetic elements lack long terminal repeats (LTRs) and their main and only active representative in the human genome is the Long Interspersed Nuclear Element-1 (LINE-1) or L1 retrotransposon. L1 elements have successfully populated and modified eukaryotic genomes for hundreds of millions of years and are currently actively propagating in the human genome, aiding in its expansion. It is estimated that almost a quarter of the mass of the human genome has resulted from L1 retrotransposition and about one-third of mammalian genomes have been created directly or indirectly by the same force (Han & Boeke, 2005). Non-LTR retrotransposons propagate via an RNA intermediate, and in humans the most common elements are the autonomous L1 elements and the non-autonomous primate-specific Alu elements (Kazazian & Moran, 1998; Kazazian & Goodier, 2002).

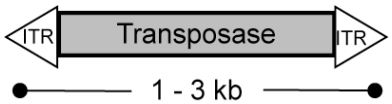
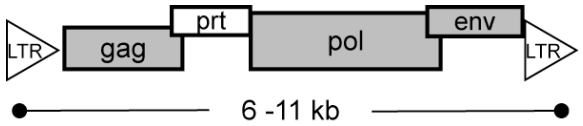
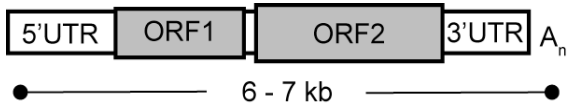
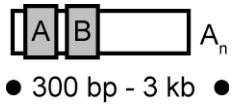
These genetic elements encode factors needed for autonomous movement via an RNA transcript, and the process by which they move to different genomic locations is termed target-primed reverse transcription (TPRT). In the human genome, L1s have accumulated over time to hundreds of thousands of copies of various ages and structures, but most L1 elements are retrotransposition defective. Nonetheless, there are approximately 80-100 full-length elements capable of retrotransposition in the diploid genome. There are about 868.000 copies of this family in the human genome, accounting for an impressive 21% (Han & Boeke, 2005; Kazazian, 2004).

Most identified retrotransposons encode an endonuclease (EN) that is similar to apurinic/apyrimidinic (AP) DNA repair endonucleases. Based on phylogenetic studies APE-type non-LTR retrotransposons have been divided into four groups and eleven clades. The four groups, named after the first elements to be discovered, are L1, RTE, I and Jockey. Elements relevant for this work that are discussed in more detail are: L1 and Tx1L retrotransposons that belong to the L1 clade (L1 group); R1Bm and TRAS1 retrotransposons that belong to the R1 clade (I group) (see also alignments in Chapter 2) (Lovsin et al, 2001; Zingler et al, 2005a; Eickbush & Malik, 2002).

Non-autonomous retrotransposons in the human genome

Short interspersed nuclear elements (SINEs) and processed pseudogenes belong to this category of mobile elements. The main mechanistic difference from other mobile elements is that these elements have no protein coding potential and hence need to proceed in hijacking the proteins from other elements, like for instance the

Table 1. Representative types of mobile genetic elements.

Interspersed repeat	Representative structure	Mechanism of integration	Contribution to human genome	Examples
DNA transposons		Excision and integration using encoded transposase gene and ITRs	294.000 copies 3%	MER1-Charlie, MER2-Tigger, Mariner
LTR retrotransposons		Retroviral-like mechanism, with reverse transcription using the encoded pol gene, primed by tRNA	443.000 copies 9%	ERVL, MaLR HERVK10
LINEs		Encoded ORF2 protein transcribes RNA template using host strand as primer	868.000 copies 21%	LINE-1, LINE-2
SINEs		Do not encode proteins; they hijack L1 proteins for integration	1.558.000 copies 14%	Alu, SVA, MIR

Adapted from *Babushok & Kazazian, 2007, Human Mutation*

ones from L1s, in order to retrotranspose to a new genomic location. The only nonautonomous retrotransposons in the human genome with evidence of current activity are the Alu elements and the SVA elements (Ostertag et al, 2003).

Elements of this class are normally 300bp to 3000bp in length and the 1.558.000 SINE copies, the majority of which are Alu elements, account for 14% of the human genome. Representative elements include Alu and SVA (discussed in more detail later), and MIR. A typical element is depicted in the shematic drawing (Fig. 1); the element ends with the characteristic poly(A) tail and the A and B boxes contain sequences vital for RNA polymerase III-mediated transcription (Babushok & Kazazian, 2007).

Alu elements

Alus are approximately 300 bp in length, are currently active in the human genome and their name comes from a single recognition site for the restriction enzyme AluI located almost in the middle of the Alu element. Human chromosomes contain about 1,000,000 Alu copies, which account for 10% of the total genome. Alu elements are derived from the 7SL RNA gene that encodes the RNA component of the signal recognition particle (SRP), which labels proteins for export from the cell. An initial deletion of the central sequence was most likely followed by a duplication and at a final stage the acquirement of a polyA stretch important for retrotransposition (Mighell et al, 1997; Cordaux et al, 2006).

The biochemical characterisation and the crystal structure of an Alu RNA in complex with the SRP9/14 heterodimeric protein, shed new light in the ways of Alu retrotransposition (Weichenrieder et al, 2000; Weichenrieder et al, 2001). The suggested dimeric Alu RNP retrotransposition intermediate is likely to bind to the translating ribosome. This could enable Alu RNA to compete efficiently with L1 RNA for nascent EN-RT and hijack the protein by providing an alternative polyA tail in the ribosomal neighbourhood. This way an Alu element can force the ORF2 protein to reverse transcribe and integrate its RNA and not the LINE-1 mRNA. This hypothesis could explain the retrotranspositional success of these 'parasites of parasites' (Boeke, 1997). Further studies revealed that the poly-A stretch at the 3' end of Alus is essential for mobility, L1 elements are required for transposition and the rate of retrotransposition is 100-1000 times higher for Alu transcripts than for control mRNAs. This high retrotransposition frequency could account for the high mutational activity of Alu elements in humans (Dewannieux et al, 2003).

SVA elements

SVA is a composite retrotransposon named after its main components SINE-R element, variable-number-tandem-repeats (VNTR) and Alu. It was Shen and coworkers that used the term "SVA" (SINE-R, VNTR, and Alu) to describe this type of retrotransposon (Shen et al, 1994). Roughly 2000 to 5000 copies of this element type exist in the human genome. Many characteristics of SVA insertions are reminiscent of L1 insertions. Some of them are 5' truncated, they end in poly(A) tails preceded by a poly(A) signal and they are flanked by TSDs similar in length to the ones of L1 elements. With an estimated age of 18 to 25 million years old, SVA represents the youngest family of retrotransposons in the primate order. They are currently active in the human genome, mobilized in trans by active L1 elements and known to cause disease in humans (Ostertag et al, 2003). At their 5′ ends, full-length SVA elements have hexameric (CCCTCT) repeats. This region is followed by an antisense Alu sequence, a VNTR region containing multiple copies of a 35–50 bp repeat, a SINE-R sequence, and a polyadenylation signal and a poly A tail (Ostertag et al, 2003).

Processed pseudogenes

This type of nonautonomous elements make up about 0.5% of the human chromosome 21 sequence (Hattori et al, 2000) and they resemble retrotransposed RNA polymerase II transcripts since they lack introns and promoters, end in poly(A) tail and are flanked by TSDs of variable length (Vanin, 1985). It is known that most processed pseudogenes are not expressed and represent dead genes; some of them however are expressed and are possibly functional in the human genome, as well as in the mouse and the rat genomes (Brosius, 1999). In general, processed pseudogenes are not considered products of gene duplication (Makałowski, 2001).

The Gene structure of L1 elements

An active L1 element

L1s are 6.0 kiloBases long and contain a 5' untranslated region (UTR) that harbors an internal promoter. They also contain two non-overlapping open reading frames, ORF1 and ORF2, and a 3' UTR ending in a characteristic poly(A) tail (Figure 1) (Dombroski et al, 1991; McMillan & Singer, 1993). Biochemical studies revealed that ORF1 encodes a novel 40 kiloDalton RNA binding protein (ORF1p) (Hohjoh & Singer, 1996) that seems to specifically bind L1-RNA. ORF2 on the other hand, encodes a multifunctional protein (ORF2p) of 150 kiloDalton that contains endonuclease (EN) and reverse transcriptase (RT) domains and has a carboxy-terminal cysteine-rich domain (C) of unknown function (Mathias et al, 1991; Feng et al, 1996). Another feature of L1s is that they are always flanked by variable-length target site duplications (TSD), which are considered as characteristic hallmarks of the integration process (Jurka, 1997).

ORF1p

Biochemical studies revealed that ORF1 encodes a 40 kDa RNA binding protein that specifically binds L1 RNA, since it appears to co-localise with L1 RNA in cytoplasmic RNPs (Hohjoh & Singer, 1996; Hohjoh & Singer, 1997). In addition, the protein is necessary for the retrotransposition of unrelated mRNAs in processed pseudogene formation (Esnault et al, 2000; Wei et al, 2001). Further studies indicate that ORF1p exhibits "nucleic acid chaperone" activity, reminiscent of retroviral gag proteins, facilitating strand exchange during the retrotransposition process (Martin & Bushman, 2001).

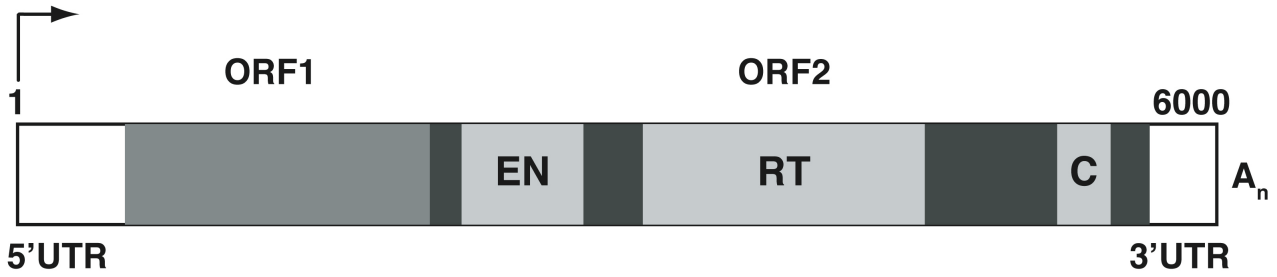


Figure 1. The primary structure of an active L1 element. Transcription starts from an internal promotor (arrow) and there are two non-overlapping open reading frames (grey and dark grey) with the known endonuclease (EN), reverse transcriptase (RT) and C-terminal cysteine-rich (C) domain highlighted in silver.

A point mutation in the mouse ORF1p that abolishes chaperone activity without affecting RNA or DNA binding, also eradicates the retrotransposition process (Martin et al, 2005). In a similar manner, the same mutation in human L1 is able to stop retrotransposition, but not enough to hamper RNP formation (Kulpa & Moran, 2005).

An interesting but also surprising finding was that the otherwise essential ORF1p could be spared when Alu elements hijack the L1 machinery, courtesy of ORF2p. A possible explanation could be that the SRP9/14 protein can in that case take over the role of ORF1p (Dewannieux et al, 2003). A general bottleneck in the study of ORF1p is the fact that it lacks similarity to proteins of known function, and the only useful outcome from sequence predictions was the presence of a long coiled-coil domain, which in human L1 contains a leucine zipper motif, typically associated with protein-protein interactions (Holmes et al, 1992). Finally, it has been proposed that the coiled-coil domain of both human and mouse proteins is responsible for the formation of ORF1p multimers, with trimer being the biologically relevant state in the mouse (Martin et al, 2003). In summary ORF1p binds RNA strongly, forms RNPs with L1 RNA, is a nucleic acid chaperone essential for retrotransposition and possibly engages in higher order complexes with ORF2p during TPRT.

ORF2p

ORF2 encodes a multifunctional 150 kiloDalton protein that contains endonuclease (EN) and reverse transcriptase (RT) activities and in addition has a carboxy-terminal cysteine-rich domain (C) of unknown function (Mathias et al, 1991; Fanning & Singer, 1987; Feng et al, 1996).

Endonuclease domain

The EN domain of the L1 retrotransposon was the main focus of this work and is thoroughly discussed in chapters 2-5. The L1 endonuclease is responsible for more than 1.5 million retrotransposon integration events in the history of the human genome, both direct L1 and L1-mediated SINE insertions. This enzyme belongs to the extended family of metal-dependent phosphohydrolases, together with AP-like retrotransposon ENs, AP DNA repair ENs, sphingomyelinases and polyphosphate phosphatases (Chapter 5). The isolated 28 kiloDalton EN domain has been expressed in E.coli and has been shown to cleave target-site DNA in vitro. The main function of L1-EN is to recognize and cleave its target substrate DNA sequence, during the beginning of the TPRT process, priming the synthesis of a new LINE DNA strand by the L1 reverse transcriptase. L1-EN is AP-like but lacks preference for AP sites and is essential for active retrotransposition (Feng et al, 1996).

Its crystal structure, the first of a retrotransposon-encoded protein, is presented in Chapter 2. Biochemical studies show that L1-EN has moderate specificity for T+A rich targets and that DNA sequence and structural parameters like minor groove width are particularly important (Cost & Boeke, 1998). The target selectivity of human L1-EN is apparently determined by factors both on the side of DNA substrate, but also by certain residues or loops on the protein surface. Identification of such elements and how their manipulation can alter EN specificity is the topic of Chapter 3. It is thought that the nicking specificity of the isolated EN domain, the EN behavior in the context of the full-length ORF2p and finally the integration specificity of an active L1 element are clearly related but further work is needed for building a conclusive model. What appears to be crucial for nicking specificity are surface loops that are also found in related ENs from other retrotransposons. That issue, together with how L1-EN binds to a double-stranded DNA target, was addressed by a combination of experimental and computational methods in Chapter 4. On a more global scale however, and depending on the chromatin environment in vivo, it is believed that certain genomic locations might be more susceptible to L1 retrotransposition (Cost et al, 2001; Chen et al, 2005; Conley et al, 2005).

Reverse Transcriptase domain

In 1986, phylogenetic analyses indicated that ORF2p has homology to reverse transcriptase (RT) (Hattori et al, 1986; Loeb et al, 1986). Further analysis showed these RTs to be most similar to those encoded by non-LTR retrotransposons, hence representing a distinct lineage to RTs encoded by LTR-retrotransposons and retroviruses (Xiong & Eickbush, 1990; Malik et al,

1999). Conclusive studies illustrated that the human L1 ORF2 protein is able to encode an active RT. This RT activity could in the presence of divalent ions extend homopolymer/oligonucleotide primer template complexes (Mathias et al, 1991; Dombroski et al, 1994).

The closest sequence homologue to non-LTR retrotransposon-encoded RT is telomerase RT (TERT). TERT is known to form a stable functional RNP complex with telomerase RNA and there are likely mechanistic parallels between TPRT and the maintenance of chromosomal DNA ends via telomere elongation (Lingner et al, 1997; Nakamura et al, 1997). Clements and Singer (Clements & Singer, 1998) reported the expression of full-length ORF2p in yeast, observing that both the wild-type protein and EN or C domain deletion variants retain RT activity.

The reverse transcriptase activity of ORF2p has a remarkable cis-preference, in which case the nascent protein associates into a ribonucleoprotein particle (RNP) preferentially with the polyA tail of its own mRNA (Kimberland et al, 1999; Esnault et al, 2000). However, as biochemical activities owing to ORF2p have been difficult to detect in cells, it was not until very recently that L1 RNA, ORF1p and ORF2p were shown to co-localize for the first time in a putative RNP retrotransposition intermediate. In the same work it was further demonstrated that ORF2p prefers using its encoding RNA as a template for reverse transcription, thus providing the first biochemical evidence for cis-preferential action of the L1-RT (Kulpa & Moran, 2006). Finally, there are data demonstrating that RT from the human L1 element is a highly processive polymerase among RT enzymes, and that missing RNase H activity for the L1 ORF2 protein in vitro, is distinguishing L1 RT from retroviral RTs (Piskareva & Schmatchenko, 2006).

C domain

Lastly, the ORF2 protein contains at its carboxy terminus a conserved cysteine-histidine-rich domain (Fanning & Singer, 1987). It is termed domain C and its function is thus far unknown. When conserved histidine and cysteine residues within this domain are mutated, L1 retrotransposition in cultured cells is reduced by two orders of magnitude, without disturbing ORF2p-RT activity (Dombroski et al, 1994; Moran et al, 1996). The above observation led to the belief that the function of C domain is necessary for retrotransposition in a distinct way from EN and RT domains. There is increasing speculation that conserved amino acids in the C domain function in ORF2p nucleic acid binding (Moran & Gilbert, 2002).

L1-like retrotransposons (from other organisms) that are target specific

L1-like elements exist in many other organisms, and many of them are distinct from L1s because they are highly specific for the target sequence they recognize and integrate. For instance, Tx1L, R1Bm and TRAS1 are all non-LTR elements that have very similar features to the L1 elements including sequence size, two open reading frames ORF1 and ORF2, and an EN-RT domain in the second ORF. It is interesting to note how these different endonucleases compare to each other and to L1-EN in terms of specificity, and also if this specificity can be manipulated to affect retrotransposon integration (discussion follows in the next chapters).

Tx1L retrotransposon

Tx1L is a non-LTR retrotransposon found dispersed in the genome of the South African frog Xenopus laevis. These elements are highly similar to L1 retrotransposons and are evolutionary classified as members of the L1 clade. All copies of Tx1L are always found inserted in specific sites within a family of DNA transposons, the Tx1D family, indicating that Tx1L is a specific retro-element. Like in L1s, the ORF2 of Tx1L also encodes an active EN, but the behavior of the latter in vitro and the 10 base-pair minimal sequence that it recognizes do not seem enough to explain the highly specific nature of the elements' integration. It is thought that in the context of the ORF2p in vivo, a longer sequence might be recognized or additional host factors might aid in ensuring specificity (Christensen et al, 2000a; Christensen et al, 2000b; Garrett & Carroll, 1986; Garrett et al, 1989).

R1Bm retrotransposon

With distinct evolutionary origins, R1Bm elements belong to the R1 clade of the I group of non-LTR retrotransposons. R1Bm is found inserted in the genome of the silkworm Bombyx mori, showing preference for a specific position in the 28S ribosomal DNA (rDNA) of the silkworm. An EN activity is contained in the ORF2p, able to make precise and paired nicks on target DNA in vitro. Bottom strand cleavage takes place before the top strand is nicked, in agreement with the TPRT model discussed in detail for L1, but there is speculation that R1Bm may be also using pre-formed double-stranded breaks in 28S rDNA for integration (Xiong & Eickbush, 1988; Feng et al, 1998; Anzai et al, 2005).

TRAS1 retrotransposon

The TRAS1 element is much closer, regarding evolutionary origins, to R1Bm since it also belongs to the R1 clade of group I. TRAS1 is a telomere-associated element and encodes a well characterized, highly specific EN that bears little sequence but considerable structural similarity to L1-EN. TRAS1-EN cleaves telomeric repeats at sites that are consistent with integration sites and like in R1Bm bottom strand cleavage happens first. Retrotransposition of TRAS1 in vivo has been illustrated experimentally. Interestingly, the available data fuel speculation that elements like TRAS1 compensate repressed telomerase activity, consequently playing an important role in telomere maintenance in the silkworm (Okazaki et al, 1995; Takahashi & Fujiwara, 1999; Anzai et al, 2001; Maita et al, 2004; Fujiwara et al, 2005).

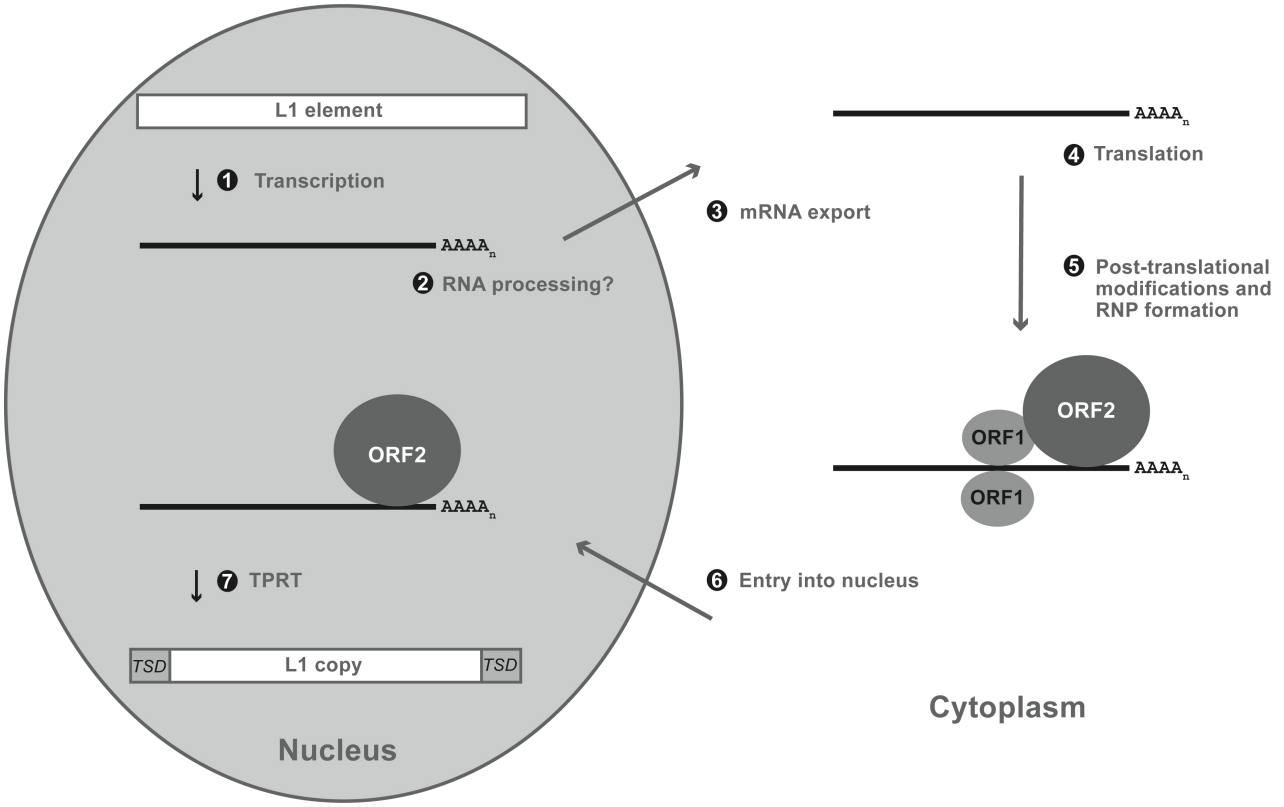


Figure 2. The seven steps of L1 retrotransposition: 1.Transcription of an active L1 element results in a bicistronic mRNA 2.The mRNA possibly undergoes RNA processing and 3.exits the nucleus via a currently unknown pathway 4.In the cytoplasm, ORF2 and ORF1 proteins are translated and associate with the (cis) RNA that transcribed them 5.A ribonucleoprotein (RNP) complex is formed, consisting of one OFR2 molecule, one RNA molecule and one or more ORF1 molecules 6.This retrotransposition intermediate will then enter the nucleus and 7.the L1 RNA will be reverse transcribed and inserted in a new location by a procedure termed target primed reverse transcription (TPRT), creating a copy of the original element flanked by target site duplications (TSD).

The L1 Retrotransposon Life Cycle

The current model regarding the steps in the L1 life cycle is a combination of studies on the L1 element structure, as well as understanding how similar retrotransposons operate in other organisms. However, it bears repeating that despite the functional and cell-culture assays already established, there are still many aspects of the retrotransposition mechanism waiting their turn to be understood (Figure 2).

Transcription

As indicated in the structure of the active L1 element the 5'UTR has internal promoter activity, independent of upstream sequences (Swergold, 1990). The life cycle of L1 begins with transcription of L1 DNA by either RNA polymerase II or RNA polymerase III and standard maturation into a bicistronic mRNA molecule. Conflicting data are available that make the question which machinery is transcribing L1 in vivo an unanswered one (Kurose et al, 1995; Moran et al, 1996; Woodcock et al, 1996; Paule & White, 2000). The "internal promoter" idea makes sense from a retrotransposons point of view, since L1s have to take their promoters with them to be able to generate an active copy when they insert in new genomic locations (Han & Boeke, 2005). Recently, various transcription factor-binding sites have been revealed: SRY-family-binding sites and a RUNX3-binding site, important for activation and a YY-1-binding site directing accurate transcription initiation (Tchénio et al, 2000; Yang et al, 2003; Athanikar et al, 2004) (Figure 2, Step 1).

RNA Processing and Nuclear Export

Typical processing of RNA Polymerase II (Pol II) transcripts includes: a) cleavage and addition of poly(A) tail, b) addition of 7-methylguanosine cap and c) intron splicing (Tollervey & Caceres, 2000). L1s contain a functional AATAAA poly(A) signal and possibly use cleavage and polyadenylation machinery of Pol II transcripts. Often L1s bypass their own poly(A) signal and use a downstream one, which results in retrotransposition of genomic sequence 3' of L1 in a process termed L1-mediated 3' transduction (Moran et al, 1999; Ostertag & Kazazian, 2001a). It is currently unclear if further modifications take place, but most likely the "intronless" L1 does not require splicing. However our understanding of L1 life cycle initiation could very well change since new

theories continuously arise, like for instance with reports that claim L1s might indeed contain functional splice sites (Belancio et al, 2006).

The L1 mRNA then exits the nucleus and is transported into the cytoplasm via a currently unknown pathway. Normally, unspliced RNAs are retained in the nucleus, denied exit by splicing factors termed commitment factors (Cullen, 2000). However, L1s have apparently found a way to cross to the cytoplasm. This could either mean that they have developed cis-acting elements to facilitate export, similar to certain viruses or that they indeed undergo splicing and coupled export, in the case that they contain functional splice sites (Ostertag & Kazazian, 2001a; Belancio et al, 2006) (Figure 2, Steps 2 & 3).

Translation

After transcription and transport in the cytoplasm, the full-length ORF1 and ORF2 proteins have to be translated. L1 mRNAs are atypical of mammalian mRNAs because they are bicistronic. It is necessary that both ORF1p and ORF2p are encoded in cis for full activity. This generally means that the L1 proteins preferentially associate with the RNA that transcribed them (Esnault et al, 2000; Wei et al, 2001) and recent biochemical evidence support this cis-preference (Kulpa & Moran, 2006). The mechanistic details of translation are another missing piece from the retrotransposition puzzle (Figure 2, Step 4).

Posttranslational Modifications and Ribonucleoprotein Formation

While posttranslational modifications of the two L1 proteins constitute a mystery, it is believed that OFR2p, L1 RNA and one or more ORF1 molecules come together to form a ribonucleoprotein (RNP) particle that serves as a retrotransposition intermediate (Martin, 1991). Recent data confirm that RNP formation in L1 is important but not sufficient for retrotransposition and indicate that ORF1p may function at downstream steps in the L1 retrotransposition pathway (Kulpa & Moran, 2005). Work by the same people showed for the first time the co-localization of L1 RNA, ORF1p and ORF2p to a putative ribonucleoprotein retrotransposition intermediate (Kulpa & Moran, 2006).

It is further believed that ORF1p is able to form higher order multimers, with the protein-protein interactions mediated by its leucine zipper in humans (Holmes et al, 1992; Hohjoh & Singer, 1996; Martin et al, 2003). The RNP particle is predominantly cytoplasmic but a part or subcomplex has to be transported in the nucleus (Han & Boeke, 2005) (Figure 2, Step 5).

Entry into the Nucleus

The retrotransposition intermediate RNP comlex must somehow gain access into the nucleus for the final step of L1 retrotransposition. Transport to the nucleus could be facilitated by a functional nuclear/nucleolar localization signal (NLS) that was recently mapped to the EN domain of ORF2p (Goodier et al, 2004). Data from the same study suggest that the EN-NLS is suppressed by interaction with the ORF2p C-terminus. When a C-terminal fragment is deleted, ORF2p is able to move to the nucleolus. Passive diffusion through the nuclear pore can be ruled out as a transport method, since ORF2p alone is way over the ~60 kiloDalton limit (Görlich & Kutay, 1999). What could happen in the case of active transport is that the NLS could be recognized by importins that would aid in entering the nucleus. Alternatively, entry could take place during nuclear membrane breakdown at mitosis or meiosis, especially since cell division appears to be a requirement for active L1 retrotransposition in human cultured cells (Shi et al, 2007) (Figure 2, Step 6).

L1 moves in mysterious ways: The process of Target Primed Reverse Transcription

In the final step of the L1 life cycle, elements are reverse transcribed and integrated into the host genome via a coupled reverse transcription/integration process termed target primed reverse transcription (TPRT). TPRT was originally described for the related non-LTR retrotransposon R2 in Bombyx mori (Luan et al, 1993; Yang et al, 1999). During this process the L1 RNA will be reverse transcribed into DNA and inserted in a new genomic location, creating a copy of the original element flanked by target site duplications. Initial stages of TPRT have been successfully reconstituted in vitro (Cost et al, 2002) (Figure 2, Step 7).

Target selection is most likely driven by L1-EN since there is a good agreement in EN target sequences and L1 integration sites (Berry et al, 2006). The endonuclease activity from the ORF2 protein is first in action, nicking the bottom DNA strand of the target sequence that is normally A+T rich (Figure 3). At the target site, the sequence is usually similar to the consensus 5'TTTTAA3' and cleavage occurs between the T and A nucleotides; cleavage however is not restricted to that spot only, owing to the promiscuous nature of the L1-EN (Feng et al, 1996; Cost & Boeke, 1998). The cleaved strand then dissociates and binds to the poly(A) tail of an L1 mRNA. A 3´ OH group of the DNA strand, freed as a result of the EN action, will then prime cDNA first strand synthesis by the coupled activity of L1-RT. Depicted in figure 5 is a schematic drawing of how ORF2p activities EN and RT could bind to each other, the target DNA and the L1 mRNA in order to start producing the first strand of the daughter element during TPRT. It is also thought that ORF1p might have a role for strand transfer during L1 reverse transcription (Martin et al, 2005).

Cleavage of the second DNA strand is thought to occur 7-20 nucleotides downstream of the first cut and the free 3´ OH group generated by this event is used to prime the second strand synthesis of L1 cDNA. Depending on the relative positions of first and second strand nicks, target site duplications or target site deletions might occur (Jensen et al, 1994). However, the mechanism of second strand cleavage and synthesis is not yet completely elucidated. The whole process ends by stable integration of a new double-stranded L1 DNA copy, in a novel position in the genome. It is interesting to note that the vast majority of L1 insertions in vivo are severely 5'

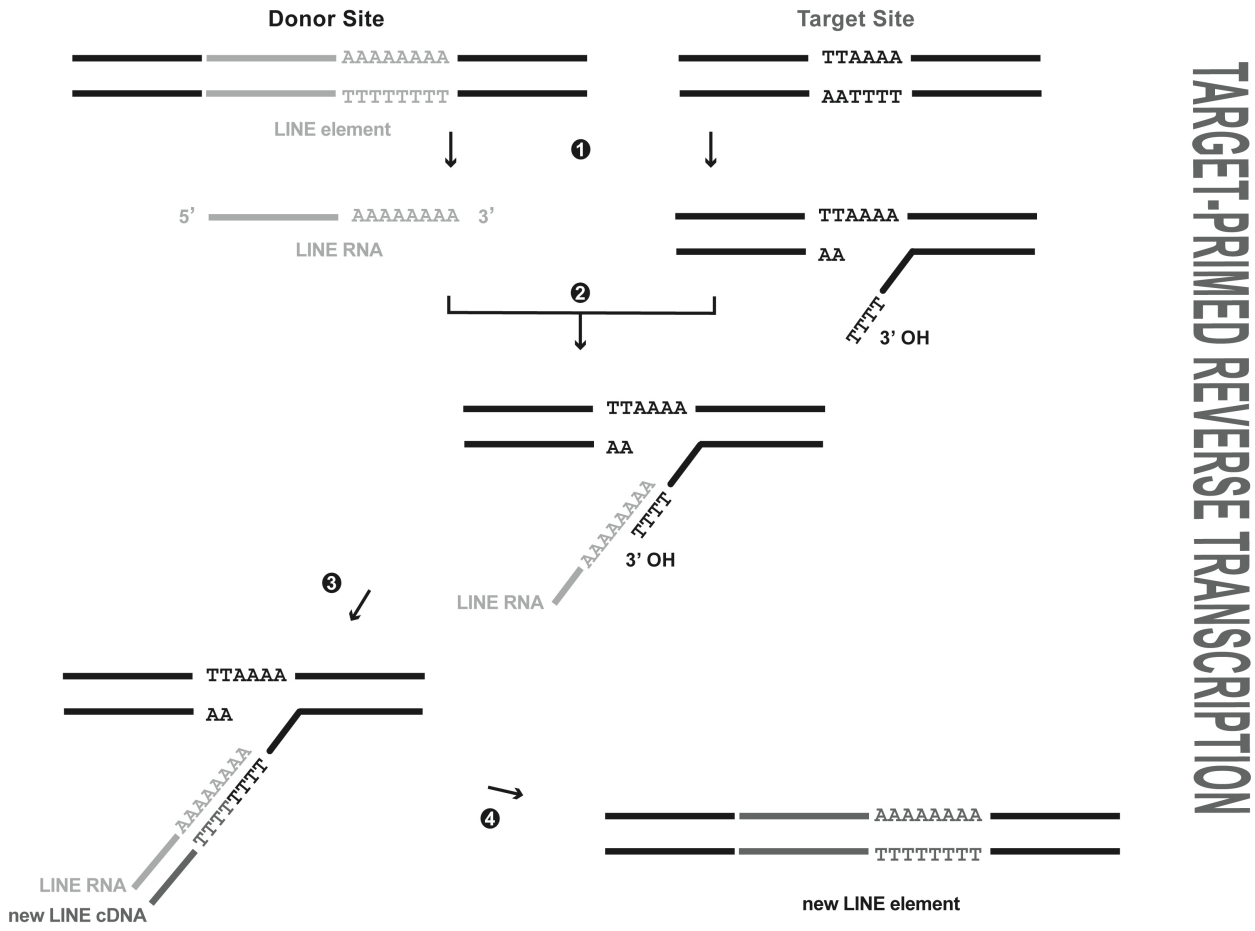


Figure 3. The process of Target-primed Reverse Transcription (TPRT) in four easy steps.

truncated or inverted and thus incapable of further retrotransposition, making the average insertion length about 1000 base pairs (Boissinot et al, 2000; Lander et al, 2001; Szak et al, 2002). This can be attributed to the rather non-processive nature of L1-RT, which is probably unable to finish copying the entire L1 RNA before dissociating from it.

At the moment three models have been proposed for explaining the less well-understood aspects of L1 TPRT and integration. The "Twin priming" model suggests creating an element containing a 3' direct region joined to a 5' inverted region, involving two ORF2p molecules (Ostertag & Kazazian, 2001b). Zingler and colleagues proposed a TPRT model for 5'-end attachment requiring microhomology-mediated end-joining (Zingler et al, 2005b). And lastly the "Template jumping" model that tries to combine and explain all of the characteristic features of L1 retrotransposition (Babushok et al, 2006). A common point that connects all three models is the belief that cellular enzymes need to be brought into play for final resolution and double-stranded nick repair. Very recently, the DNA repair enzyme ataxia telangiectasia mutated (ATM) was reported as being essential for L1-induced double-strand break repair and retrotransposition (Gasior et al, 2006) (Figure 4).

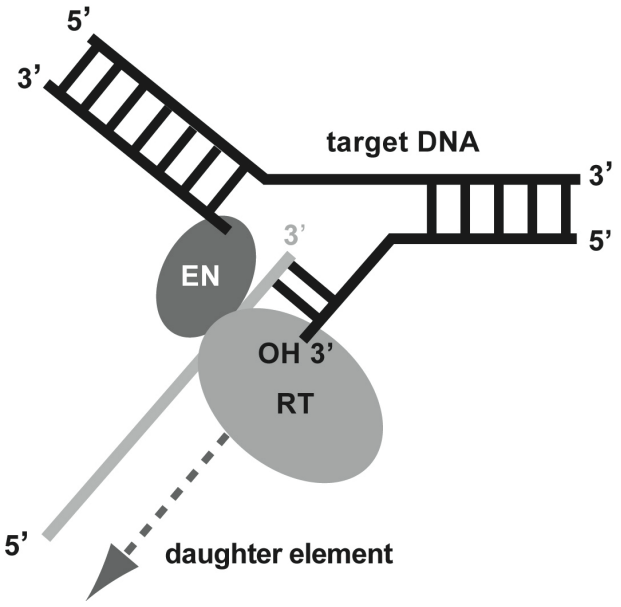


Figure 4. Endonuclease **(EN)** and reverse transcriptase **(RT)** actions and interactions with target DNA and each other during the TPRT process.

Impact of L1 and related elements on the host genome

The human genome is flooded by retrotransposons, most common of which are the autonomous L1 elements and the non-autonomous Alu elements. Together, they presently occupy about 28% of the total genome. (Lander et al, 2001; Venter et al, 2001; Babushok & Kazazian, 2007). The L1 retrotransposon is the major source of insertional mutagenesis and is able to maintain "genome fluidity" (Kazazian, 2000). It is responsible for the formation of processed pseudogenes (Esnault et al, 2000), exon shuffling in a process called 3' transduction (Moran et al, 1999) and the mobilization of Alu elements and other SINEs (Jurka, 1997; Dewannieux et al, 2003).

In summary, L1 retrotransposons can drive genome evolution in both destructive and constructive ways (Biémont & Vieira, 2006). The first category includes cis-insertional mutagenesis by L1s, trans-insertional mutagenesis by Alus/SVAs/processed pseudogenes, deletions/duplications due to unequal homologous recombination between Alus or L1s, deletions/inversions due to L1 rearrangements and finally duplications at new chromosomal sites. On the side of constructive mechanisms, we find repair of double-strand breaks by EN-independent L1 insertion, 3' or 5' transduction, formation of chimeric retrogenes like the U6/L1 chimera, presence of L1/Alu sequences in protein-coding exons, gene expression 5' of full-length L1s via an antisense promoter in L1 and lastly transcript termination at strong poly(A) signals in L1s (Moran & Gilbert , 2002; Kazazian, 2004; Han & Boeke, 2005).

Less than 5% of the human genome is protein-coding sequence; in itself this fact alone represents a decent defence against retrotransposons. Still, cells have developed unclear mechanisms to suppress the potentially hazardous effects of mobile elements. The cytosine methylation system might have evolved to repress L1 expression and retrotransposition (Yu et al, 2001; Yoder et al, 1997; Bestor, 2003; Bourc'his & Bestor, 2004), and additionally co-suppression mediated by natural siRNAs could also function as an RNAi pathway responsible for L1 control (Yang & Kazazian, 2006; Soifer & Rossi, 2006). A final possibility is that the family of apolipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) proteins, is also involved in retrotransposition control possibly via DNA/RNA deamination. Current ambiguities should soon be clarified given the "hyper-activity" of researchers in trying to address this open question (Chen et al, 2006a; Esnault et al, 2006; Stenglein & Harris, 2006; Bogerd et al, 2006; Muckenfuss et al, 2006; Hulme et al, 2007), which very recently yielded the crystal structure of APOBEC2 protein (Prochnow et al, 2007).

L1 Retrotransposons are associated with causing human genetic disease

The very first retrotransposition events reported in mammals, were associated with human disease. Two cases of de novo L1 insertions were reported in 1988. The insertion of a truncated L1 element into exon 14 of the factor VIII gene was found in two unrelated patients with haemophilia A and the presence of the L1 element was shown to be causative of the disease (Kazazian et al, 1988). Since those initial reports, a plethora of different examples of L1-mediated retrotranspositional events have been proven to be the cause of human disease. Around 40 of these events are termed simple, because they do not account for the loss of target gene material. A further 10 cases involve genomic deletions. Curiously, the frequency of L1 retrotransposition events per individual has been estimated to be one insertion in every 2-30 individuals, an estimation that makes us particularly vulnerable to this mobile element (Brouha et al, 2003).

A total of 53 disease-causing insertions have been identified thus far, 17 of which are caused directly by L1, while the rest are L1-mediated: 29 involving Alu and 4 involving SVA elements. The final 3 cases are also L1-mediated insertions of simple poly(A) repeats. Recombination between L1s has caused Alport syndrome, phosphorylase kinase deficiency of liver and muscle, and ataxia-telangiectasia. In addition L1-driven recombination between Alus is known to have caused diseases like familial hypercholesterolemia, a-thalassemia, type VI variant of Ehlers-Danlos syndrome and Tay Sachs disease (Druker & Whitelaw, 2004; Chen et al, 2005; Chen et al, 2006b; Babushok & Kazazian, 2007; Ostertag & Kazazian, 2006).

The most recent example of a large deletion triggered by an L1 element was associated with erasing exons 3 to 9 of the pyruvate dehydrogenase complex component X (PDHX) gene. The authors provide evidence of how an active L1 retrotransposon operates in the human genome by identifying a 46 kb genomic deletion in the PDHX gene, owing directly to the intronic insertion of a full-length L1 element in a patient with pyruvate dehydrogenase complex (PDHc) deficiency. Within the gene, both bottom and top strand cleavage sites agree with the consensus L1-EN target sequence 5'-TTTT/A-3' and the full-length element ends in a 67-bp poly(A) tail (Miné et al, 2007).

L1-mediated SINE mobilization is also known to be involved in disease causing cases. An ancient SINE insertion causes Fukuyama-type congenital muscular dystrophy (FCMD), one of the most common genetic disorders in Japan. This was the first report comfirming that even ancient retrotransposons might be a source of disease (Kobayashi et al, 1998). In another SINE case, point mutations or small inertions/deletions in the chromodomain helicase DNA-binding protein gene CHD7 are the main cause of CHARGE syndrome. For the first time, an Alu-mediated exonic deletion of CHD7 was found in a CHARGE syndrome patient (Udaka et al, 2007).

L1 Retrotransposons and their role in cancer

Although L1s can be beneficial in shaping mammalian genomes, L1 retrotransposition is also involved in cancer either directly or by mobilizing the Alu elements. In many cancer cells an elevated expression of L1 and Alu RNAs has been reported, possibly due to defects in the cytosine methylation or other suppressor systems (Sinnett et al, 1992; Singer et al, 1993; Thayer et al, 1993). Three are the main indications that L1s might have a role in cancer cells: first L1 sequences become

hypomethylated, second proteins or transcripts can be detected and third retrotransposition occurs at sites of breakage and recombination. Up to now the most solid data exist for L1 hypomethylation that was detected in bladder cancer, prostate cancer, colon cancer, gastric cancer, ovarian carcinoma and liver carcinoma (Schulz, 2006). L1 ORF1p is expressed in pediatric malignant germ cell tumor, where L1 retrotransposition is thought to be a common event but with unclear implications (Su et al, 2007).

Also in healthy tissue both germline and somatic insertions might cause cancer (Kazazian & Moran, 1998). Documented cases are breast and colon cancer linked to the de novo insertion of an Alu sequence into the BRCA2 gene and an L1 sequence into the APC gene, respectively (Miki et al, 1992; Miki et al, 1996). Finally, somatic recombination events have caused ALL1 rearrangement leukemias and BRCA1-associated familial breast cancer (Deininger & Batzer, 1999).

Possible applications of L1 retrotransposons

Phylogenetic markers

L1s' long history in colonizing genomes, combined with the fact that some elements are still active, renders them good candidates for phylogenetic markers. Old insertions can be useful in phylogenetic studies between species (Nikaido et al, 1999). Recent polymorphic insertions on the other hand, are suitable for the study of human population dynamics (Santos et al, 2000; Sheen et al, 2000; Konkel et al, 2007). Since Alu insertions have many similarities with L1, they can also be used to study human diversity (Watkins et al, 2001). Recent examples from the literature include: the study of L1NE sequences to elucidate the evolutionary origin of the MHC class I genomic region in primates (Fukami-Kobayashi et al, 2005), L1/Alu insertion polymorphism markers for human cell line fingerprinting (Ustyugova et al, 2005), investigating genetic affinities between Native American and East Asian populations using polymorphic Alu and L1 insertions (Mateus Pereira et al, 2005).

System for random mutagenesis

L1 elements are not very specific for their target site integration, meaning that they move rather randomly within the genome without any bias against inserting into gene sequences (Moran et al, 1999). Along those lines, an enhanced green fluorescent protein (EGFP)-based retrotransposition casette was developed for facilitating random mutagenesis systems (Ostertag et al, 2000). Finally, ORFeus is a synthetic L1 retrotransposon, which is much more active than normal L1s introduced in mice and could be developed into a tool for in vivo mutagenesis (Han & Boeke, 2004; An et al, 2006).

Vectors for gene delivery

There is great interest for using L1 elements as gene delivery vectors. The characteristics of L1 that make it suitable for such an application are the ability for stable integration in the genome, the ability to mobilize sequences via 3' transduction, and the absence of proteins that could be possibly immunogenic for the host. A big step towards this direction was taken with the creation of a hybrid L1/helper-dependent adenovirus vector. This vector mediates long-term gene expression by a two-stage mechanism: first the helper-dependent adenovirus serves as a vehicle for efficient delivery and expression of its encoded L1/transgene cassette, and second the L1 retro-element and its associated transgene permanently integrate into the genome of the adenovirus-transduced cells (Soifer et al, 2001; Soifer & Kasahara, 2004).

Other relevant recent advances

Very recently an L1 retrotransposon-based system was developed that allows delivery of small interfering RNA (siRNA) and stable silencing in human cells. This system demonstrated long-term siRNA expression and reduction in exogenous and endogenous gene expression. Controlled retrotransposition ensured that only one RNA interference (RNAi)-cassette got integrated into the host genome, sufficient for strong interference. Such a system could achieve stable gene silencing, with potential applications for ex vivo and in vivo molecular therapy (Yang et al, 2005). RNAi techniques have been used to stably suppress L1 retrotransposon expression in A-375 melanoma cell lines. Lower L1 expression resulted in lower proliferation rate, differentiated morphology and lower tumorigenicity when inoculated in nude mice, confirming that L1 silencing modulates gene expression (Oricchio et al, 2007). Yet another recent report combines microarray and biochemical techniques to indicate that modifying transcriptional activity of L1 retrotransposons in rat hearts may represent a novel anti-ischemic therapeutic strategy (Lucchinetti et al, 2006).

REFERENCES

An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, Ye P, Triplett C, Boeke JD (2006) Active retrotransposition by a synthetic L1 element in mice. Proc Natl Acad Sci U S A 103: 18662-18667.

Anzai T, Osanai M, Hamada M, Fujiwara H (2005) Functional roles of 3'-terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. Nucleic Acids Res 33: 1993-2002.

Anzai T, Takahashi H, 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: 100-108.

Athanikar JN, Badge RM, Moran JV (2004) A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res 32: 3846-3855.

Babushok DV, Kazazian HH (2007) Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat .

Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH (2006) L1 integration in a transgenic mouse model. Genome Res 16: 240-250.

Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101 Suppl 2: 14572-14579.

Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34: 1512-1521.

Berry C, Hannenhalli S, Leipzig J, Bushman FD (2006) Selection of target sites for mobile DNA integration in the human genome. PLoS Comput Biol 2: e157.

Bestor TH (2003) Cytosine methylation mediates sexual conflict. Trends Genet 19: 185-190.

Biémont C., Vieira C. (2006) Genetics: junk DNA as an evolutionary force. Nature 443: 521-524.

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

Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, Cullen BR (2006) Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc Natl Acad Sci U S A 103: 8780-8785.

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17: 915-928.

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

Brosius J (1991) Retroposons--seeds of evolution. Science 251: 753.

Brosius J (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107: 209-238.

Brosius J (2005) Echoes from the past--are we still in an RNP world? Cytogenet Genome Res 110: 8-24.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100: 5280-5285.

Bushman FD (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135-138.

Chen H, Lilley CE, Yu Q, Lee DV, Chou J, Narvaiza I, Landau NR, Weitzman MD (2006) APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. Curr Biol 16: 480-485.

Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN (2005) Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat 25: 207-221.

Chen JM, Férec C, Cooper DN (2006) LINE-1 Endonuclease-Dependent Retrotranspositional Events Causing Human Genetic Disease: Mutation Detection Bias and Multiple Mechanisms of Target Gene Disruption. J Biomed Biotechnol 2006: 56182.

Christensen S, Pont-Kingdon G, 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, Carroll D (2000) Comparative studies of the endonucleases from two related Xenopus laevis retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. Genetica 110: 245-256.

Clements AP, Singer MF (1998) The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. Nucleic Acids Res 26: 3528-3535.

Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HH (2005) Two independent retrotransposon insertions at the same site within the coding region of BTK. Hum Mutat 25: 324-325.

Cordaux R, Hedges DJ, Herke SW, Batzer MA (2006) Estimating the retrotransposition rate of human Alu elements. Gene 373: 134-137.

Cost GJ, Boeke JD (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 GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. EMBO J 21: 5899-5910.

Cost GJ, Golding A, Schlissel MS, Boeke JD (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29: 573-577.

Cullen BR (2000) Nuclear RNA export pathways. Mol Cell Biol 20: 4181-4187.

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67: 183-193.

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35: 41-48.

Dombroski BA, Feng Q, Mathias SL, Sassaman DM, Scott AF, Kazazian HH, Boeke JD (1994) An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Mol Cell Biol 14: 4485-4492.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH (1991) Isolation of an active human transposable element. Science 254: 1805-1808.

Druker R, Whitelaw E (2004) Retrotransposon-derived elements in the mammalian genome: a potential source of disease. J Inherit Metab Dis 27: 319-330.

Eickbush H. , Malik S. (2002) Origins and evolution of retrotransposons. In Craig N, Craggie R, Gellert M, Lambowitz A (eds) Mobile DNA II pp 1111-1144. Washington D.C.: ASM Press

Emerson RA (1917) Genetical Studies of Variegated Pericarp in Maize. Genetics 2: 1-35.

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

Esnault C, Millet J, Schwartz O, Heidmann T (2006) Dual inhibitory effects of APOBEC family proteins on retrotransposition of mammalian endogenous retroviruses. Nucleic Acids Res 34: 1522-1531.

Fanning T, Singer M (1987) The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res 15: 2251-2260.

Fedoroff N (2001) How jumping genes were discovered. Nat Struct Biol 8: 300-301.

Fedoroff NV (1989) About maize transposable elements and development. Cell 56: 181-191.

Feng Q, Moran JV, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905-916.

Feng Q, Schumann G, Boeke JD (1998) Retrotransposon R1Bm endonuclease cleaves the target sequence. Proc Natl Acad Sci U S A 95: 2083-2088.

Fujiwara H, Osanai M, Matsumoto T, Kojima KK (2005) Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. Chromosome Res 13: 455-467.

Fukami-Kobayashi K, Shiina T, Anzai T, Sano K, Yamazaki M, Inoko H, Tateno Y (2005) Genomic evolution of MHC class I region in primates. Proc Natl Acad Sci U S A 102: 9230-9234.

Garrett JE, Carroll D (1986) Tx1: a transposable element from Xenopus laevis with some unusual properties. Mol Cell Biol 6: 933-941.

Garrett JE, Knutzon DS, Carroll D (1989) Composite transposable elements in the Xenopus laevis genome. Mol Cell Biol 9: 3018-3027.

Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol 357: 1383-1393.

Goodier JL, Ostertag EM, Engleka KA, Seleme MC, Kazazian HH (2004) A potential role for the nucleolus in L1 retrotransposition. Hum Mol Genet 13: 1041-1048.

Görlich D, Kutay U (1999) Transport between the cell nucleus and the cytoplasm. Annu Rev Cell Dev Biol 15: 607-660.

Han JS, Boeke JD (2004) A highly active synthetic mammalian retrotransposon. Nature 429: 314-318.

Han JS, Boeke JD (2005) LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? Bioessays 27: 775-784.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, Groner Y, Soeda E, Ohki M, Takagi T, Sakaki Y, Taudien S, Blechschmidt K, Polley A, Menzel U, Delabar J, Kumpf K, Lehmann R, Patterson D, Reichwald K, Rump A, Schillhabel M, Schudy A, Zimmermann W, Rosenthal A, Kudoh J, Schibuya K, Kawasaki K, Asakawa S, Shintani A, Sasaki T, Nagamine K, Mitsuyama S, Antonarakis SE, Minoshima S, Shimizu N, Nordsiek G, Hornischer K, Brant P, Scharfe M, Schon O, Desario A, Reichelt J, Kauer G, Blocker H, Ramser J, Beck A, Klages S, Hennig S, Riesselmann L, Dagand E, Haaf T, Wehrmeyer S, Borzym K, Gardiner K, Nizetic D, Francis F, Lehrach H, Reinhardt R, Yaspo ML, Chromosome 21 mapping and sequencing consortium (2000) The DNA sequence of human chromosome 21. Nature 405: 311-319.

Hattori M, Kuhara S, Takenaka O, Sakaki Y (1986) L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature 321: 625-628.

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

Hohjoh H, Singer MF (1997) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. EMBO J 16: 6034-6043.

Holmes SE, Singer MF, Swergold GD (1992) Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. J Biol Chem 267: 19765-19768.

Hulme AE, Bogerd HP, Cullen BR, Moran JV (2007) Selective inhibition of Alu retrotransposition by APOBEC3G. Gene 390: 199-205.

Ivics Z, Hackett PB, Plasterk RH, Izsvák Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91: 501-510.

Ivics Z, Izsvák Z (2006) Transposons for gene therapy!. Curr Gene Ther 6: 593-607.

Jensen S, Gassama MP, Heidmann T (1994) Retrotransposition of the Drosophila LINE I element can induce deletion in the target DNA: a simple model also accounting for the variability of the normally observed target site duplications. Biochem Biophys Res Commun 202: 111-119.

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94: 1872-1877.

Kazazian HH (2000) Genetics. L1 retrotransposons shape the mammalian genome. Science 289: 1152-1153.

Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632.

Kazazian HH, Goodier JL (2002) LINE drive. retrotransposition and genome instability. Cell 110: 277-280.

Kazazian HH, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19: 19-24.

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332: 164-166.

Kimberland ML, Divoky V, Prchal J, Schwahn U, Berger W, Kazazian HH (1999) Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet 8: 1557-1560.

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, Hamano K, Sakakihara Y, Nonaka I, Nakagome Y, Kanazawa I, Nakamura Y, Tokunaga K, Toda T (1998) An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature 394: 388-392.

Konkel MK, Wang J, Liang P, Batzer MA (2007) Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. Gene 390: 28-38.

Kulpa DA, Moran JV (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet 14: 3237-3248.

Kulpa DA, Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol 13: 655-660.

Kurose K, Hata K, Hattori M, 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: 3704-3709.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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 JP, 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, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. Science 276: 561-567.

Loeb DD, Padgett RW, Hardies SC, Shehee WR, Comer MB, Edgell MH, Hutchison CA (1986) The sequence of a large L1Md element reveals a tandemly repeated 5' end and several features found in retrotransposons. Mol Cell Biol 6: 168-182.

Lovsin N, Gubensek F, Kordi D (2001) Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. Mol Biol Evol 18: 2213-2224.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (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.

Lucchinetti E, Feng J, Silva R, Tolstonog GV, Schaub MC, Schumann GG, Zaugg M (2006) Inhibition of LINE-1 expression in the heart decreases ischemic damage by activation of Akt/PKB signaling. Physiol Genomics 25: 314-324.

Maita N, Anzai T, Aoyagi H, Mizuno H, Fujiwara H (2004) Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. J Biol Chem 279: 41067-41076.

Makałowski W (2001) The human genome structure and organization. Acta Biochim Pol 48: 587-598.

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

Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11: 1187-1197.

Martin SL (1991) Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol 11: 4804-4807.

Martin SL, Branciforte D, Keller D, Bain DL (2003) Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci U S A 100: 13815-13820.

Martin SL, Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol 21: 467-475.

Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, Hodges RS, Williams MC (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol 348: 549-561.

Mateus Pereira LH, Socorro A, Fernandez I, Masleh M, Vidal D, Bianchi NO, Bonatto SL, Salzano FM, Herrera RJ (2005) Phylogenetic information in polymorphic L1 and Alu insertions from East Asians and Native American populations. Am J Phys Anthropol 128: 171-184.

Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. Science 254: 1808-1810.

McCLINTOCK B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36: 344-355.

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

McClintock B (1984) The significance of responses of the genome to challenge. Science 226: 792-801.

McMillan JP, Singer MF (1993) Translation of the human LINE-1 element, L1Hs. Proc Natl Acad Sci U S A 90: 11533-11537.

Mighell AJ, Markham AF, Robinson PA (1997) Alu sequences. FEBS Lett 417: 1-5.

Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y (1996) Mutation analysis in the BRCA2 gene in primary breast cancers. Nat Genet 13: 245-247.

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res 52: 643-645.

Miné M, Chen JM, Brivet M, Desguerre I, Marchant D, de Lonlay P, Bernard A, Férec C, Abitbol M, Ricquier D, Marsac C (2007) A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Hum Mutat 28: 137-142.

Miskey C, Izsvák Z, Plasterk RH, Ivics Z (2003) The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. Nucleic Acids Res 31: 6873-6881.

Moran JV, DeBerardinis RJ, Kazazian HH (1999) Exon shuffling by L1 retrotransposition. Science 283: 1530-1534.

Moran JV, Gilbert N (2002) Mammalian LINE-1 retrotransposons and related elements. In Craig N, Craggie R, Gellert M, Lambowitz A (eds) Mobile DNA II pp 836–869. Washington, DC: ASM Press

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87: 917-927.

Muckenfuss H, Hamdorf M, Held U, Perkovic M, Löwer J, Cichutek K, Flory E, Schumann GG, Münk C (2006) APOBEC3 proteins inhibit human LINE-1 retrotransposition. J Biol Chem 281: 22161-22172.

Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH, Lingner J, Harley CB, Cech TR (1997) Telomerase catalytic subunit homologs from fission yeast and human. Science 277: 955-959.

Nikaido M, Rooney AP, Okada N (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci U S A 96: 10261-10266.

Okazaki S, Ishikawa H, Fujiwara H (1995) Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. Mol Cell Biol 15: 4545-4552.

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284: 604-607.

Oricchio E, Sciamanna I, Beraldi R, Tolstonog GV, Schumann GG, Spadafora C (2007) Distinct roles for LINE-1 and HERV-K retroelements in cell proliferation, differentiation and tumor progression. Oncogene .

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73: 1444-1451.

Ostertag EM, Kazazian HH (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet 35: 501-538.

Ostertag EM, Kazazian HH (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11: 2059-2065.

Ostertag EM, Kazazian HH Jr (2006) Retrotransposition and Human Disorders. In Encyclopedia of Life Sciences. Chichester : John Wiley & Sons Ltd, http://www.els.net/[doi: 10.1038/npg.els.0005492]

Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH (2000) Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 28: 1418-1423.

Paule MR, White RJ (2000) Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res 28: 1283-1298.

Piskareva O, Schmatchenko V (2006) DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. FEBS Lett 580: 661-668.

Plasterk RH (1996) The Tc1/mariner transposon family. Curr Top Microbiol Immunol 204: 125-143.

Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS (2007) The APOBEC-2 crystal structure and functional implications for the deaminase AID. Nature 445: 447-451.

Rhoades MM (1945) On the Genetic Control of Mutability in Maize. Proc Natl Acad Sci U S A 31: 91-95.

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A, Singh L, Destro-Bisol G, Novelletto A, Qamar R, Mehdi SQ, Adhikari R, de Knijff P, Tyler-Smith C (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9: 421-430.

Schulz WA (2006) L1 retrotransposons in human cancers. J Biomed Biotechnol 2006: 83672.

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res 10: 1496-1508.

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 269: 8466-8476.

Shi X, Seluanov A, Gorbunova V (2007) Cell divisions are required for l1 retrotransposition. Mol Cell Biol 27: 1264-1270.

Singer MF, Krek V, McMillan JP, Swergold GD, Thayer RE (1993) LINE-1: a human transposable element. Gene 135: 183-188.

Sinnett D, Richer C, Deragon JM, Labuda D (1992) Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. J Mol Biol 226: 689-706.

Soifer H, Higo C, Kazazian HH, Moran JV, Mitani K, Kasahara N (2001) Stable integration of transgenes delivered by a retrotransposon-adenovirus hybrid vector. Hum Gene Ther 12: 1417-1428.

Soifer HS, Kasahara N (2004) Retrotransposon-adenovirus hybrid vectors: efficient delivery and stable

integration of transgenes via a two-stage mechanism. Curr Gene Ther 4: 373-384.

Soifer HS, Rossi JJ (2006) Small interfering RNAs to the rescue: blocking L1 retrotransposition. Nat Struct Mol Biol 13: 758-759.

Stenglein MD, Harris RS (2006) APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J Biol Chem 281: 16837-16841.

Su Y, Davies S, Davis M, Lu H, Giller R, Krailo M, Cai Q, Robison L, Shu XO (2007) Expression of LINE-1 p40 protein in pediatric malignant germ cell tumors and its association with clinicopathological parameters: A report from the Children's Oncology Group. Cancer Lett 247: 204-212.

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

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3: research0052.

Takahashi H, Fujiwara H (1999) Transcription analysis of the telomeric repeat-specific retrotransposons TRAS1 and SART1 of the silkworm Bombyx mori. Nucleic Acids Res 27: 2015-2021.

Tchénio T, Casella JF, Heidmann T (2000) Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res 28: 411-415.

Thayer RE, Singer MF, Fanning TG (1993) Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. Gene 133: 273-277.

Tollervey D, Caceres JF (2000) RNA processing marches on. Cell 103: 703-709.

Udaka T, Okamoto N, Aramaki M, Torii C, Kosaki R, Hosokai N, Hayakawa T, Takahata N, Takahashi T, Kosaki K (2007) An Alu retrotransposition-mediated deletion of CHD7 in a patient with CHARGE syndrome. Am J Med Genet A .

Ustyugova SV, Amosova AL, Lebedev YB, Sverdlov ED (2005) Cell line fingerprinting using retroelement insertion polymorphism. Biotechniques 38: 561-565.

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19: 253-272.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291: 1304-1351.

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB (2001) Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet 68: 738-752.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21: 1429-1439.

Weichenrieder O, Stehlin C, Kapp U, Birse DE, Timmins PA, Strub K, Cusack S (2001) Hierarchical assembly of the Alu domain of the mammalian signal recognition particle. RNA 7: 731-740.

Weichenrieder O, Wild K, Strub K, Cusack S (2000) Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature 408: 167-173.

Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55: 631-661.

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A 103: 17600-17601.

Woodcock DM, Williamson MR, Doherty JP (1996) A sensitive RNase protection assay to detect transcripts from potentially functional human endogenous L1 retrotransposons. Biochem Biophys Res Commun 222: 460-465.

Xiong Y, Eickbush TH (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, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9: 3353-3362.

Yang J, Malik HS, Eickbush TH (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: 7847-7852.

Yang N, Kazazian HH (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol 13: 763-771.

Yang N, Zhang L, Kazazian HH (2005) L1 retrotransposon-mediated stable gene silencing. Nucleic Acids Res 33: e57.

Yang N, Zhang L, Zhang Y, Kazazian HH (2003) An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic Acids Res 31: 4929-4940.

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

Yu F, Zingler N, Schumann G, Strätling WH (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, Schumann GG (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. Cytogenet Genome Res 110: 250-268.

Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Löwer J, Schumann GG (2005) Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. Genome Res 15: 780-789.

Chapter 2

Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon

Kostas Repanas*, Oliver Weichenrieder* and Anastassis Perrakis

Structure, Vol. 12, 1-20, June 2004

Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon

Kostas Repanas*, Oliver Weichenrieder* and Anastassis Perrakis

ABSTRACT

The human L1 endonuclease (L1-EN) is encoded by the non-LTR retrotransposon LINE-1 (L1). L1 is responsible for more than 1.5 million retrotransposition events in the history of the human genome, contributing more than a quarter to human genomic DNA (L1 and Alu elements). L1-EN is related to the well-understood human DNA repair endonuclease APE1 and its nicking specificity is a major determinant for retrotransposon integration site selection. The crystal structure of human L1 endonuclease is the first of a retrotransposon-encoded protein and a prototype for retrotransposon-encoded endonucleases involved in target-primed reverse transcription. Structure-based endonuclease alignments reveal a conserved threonine in addition to previously identified invariant residues and suggest DNA recognition to proceed via the accommodation of an extra-helical nucleotide within a pocket of the enzyme. The present analysis will help to refine phylogenetic and functional relationships among metal-dependent phosphohydrolases and provides a basis for manipulating non-LTR retrotransposon integration site selection.

*These authors contributed equally to the work.

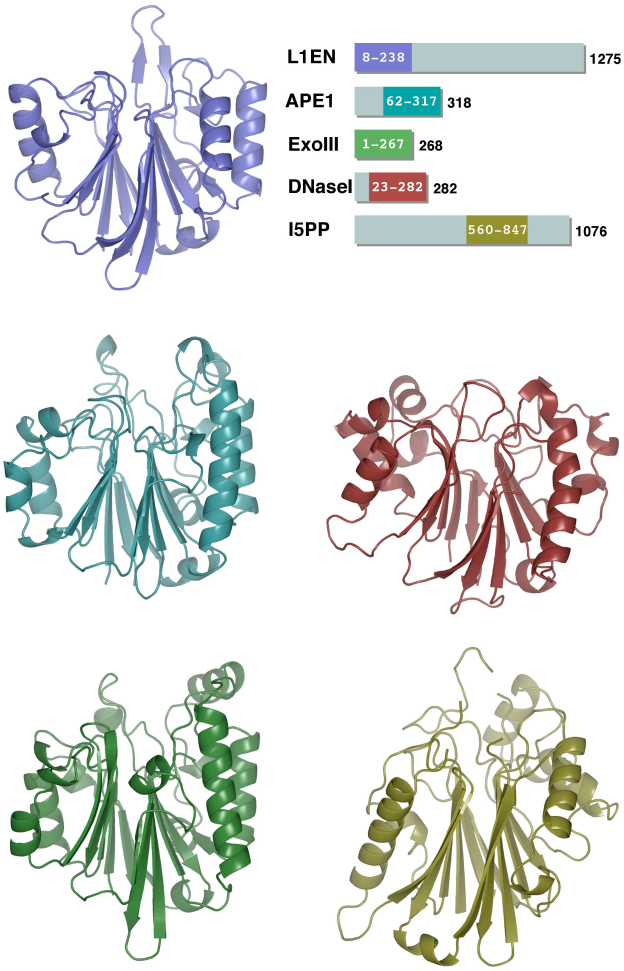
INTRODUCTION

The continuing insertion of 'selfish' retrotransposons has generated much of the so-called 'junk DNA' within eukaryotic genomes. This process is an important factor contributing to the 'fluidity' and evolution of genomes but it represents a major challenge for the respective species as it has both beneficial and disastrous consequences (Brosius, 2003; Eickbush and Malik, 2002). Retrotransposons can damage genes by insertional mutagenesis, shuffle exons to new genomic locations, co-mobilize other retroelements and assist in pseudogene-formation (Ostertag and Kazazian, 2001). They can also cause genomic rearrangements either as a direct consequence of the integration process or, indirectly, by offering sites for homologous recombination (Deininger et al., 2003).

Retrotransposons - in contrast to DNA transposons - propagate via a 'copy and paste' mechanism. They give rise to an RNA intermediate that is used to generate a DNA copy with the help of an element-encoded reverse transcriptase. Depending on the DNA integration mechanism two classes of retrotransposons are distinguished. The first class contains long terminal repeat (LTR) retrotransposons and retroviruses. These retroelements use an integrase that recognizes the LTRs of the double-stranded DNA copy. The second, much larger and more ancient class includes all non-LTR retrotransposons. Those are thought to integrate via target-primed reverse transcription (TPRT), a process in which reverse transcription and integration are coupled (Eickbush and Malik, 2002; Kazazian, 2004). An endonuclease that is part of the same polypeptide chain as the reverse transcriptase nicks the genomic DNA and hands over the resulting ribose 3'-hydroxyl end as a primer for reverse transcription of associated template RNA (Cost et al., 2002; Luan et al., 1993).

Most non-LTR retrotransposons encode an endonuclease located N-terminally of the reverse transcriptase. This endonuclease bears similarity to the human DNA repair endonuclease APE1 that recognizes apurinic and apyrimidinic (AP) sites (Mol et al., 2000). A minority of non-LTR retrotransposons encodes an endonuclease that is located C-terminally of the reverse transcriptase and that is rather similar to certain restriction enzymes. The APE-type non-LTR retrotransposons have been subdivided into 10-11 clades, based on a phylogenetic sequence analysis of their reverse transcriptases. Almost every eukaryotic genome contains at least one active element from at least one of these clades (Eickbush and Malik, 2002; Lovsin et al., 2001; Malik et al., 1999). The impact of any given element on its host genome depends largely on its insertion site specificity. High fidelity integration into repetitive telomeric sequences or into or next to multi-copy genes is tolerated more easily by the host, because this has little or no impact on the rest of the genome. It is believed that such highly specific integration is not only governed by the nicking specificity of the endonuclease but also by other element or host-cell specific targeting factors. In contrast, less specifically integrating elements can have more dramatic consequences for the host. They often lead to a significant increase in the amount of genomic DNA and are more likely to interfere with essential gene functions, because they get interspersed throughout the genome (Zingler et al.).

The human L1 element has been characterized as a long interspersed nuclear element (LINE). It is a member of the L1 clade of APE-type non-LTR retrotransposons and integrates into the frequent consensus target sequence 5'TTTT-AA3', where the dash represents the scissile bond on the nicked strand (Gilbert et al., 2002; Jurka, 1997; Symer et al., 2002)). The L1 element contains an



internal promoter and two open reading frames (ORFs). The first protein

Figure 1. Crystal structure of human L1 endonuclease (L1-EN) and comparison to members of the enzyme family of metal-dependent phosphohydrolases. Bars represent full-length proteins containing phosphohydrolase domains at the colored positions. The respective structures are drawn as ribbon diagrams juxtaposed in the same orientation with the substrate-binding surface on top. A common, central β-sandwich is surrounded by individual α-helices and surface loops. For abbreviations and PDB codes see methods

(ORF1p) binds to L1 RNA in multiple copies, forming large ribonucleoprotein particles (Hohjoh and Singer, 1996). The second protein (ORF2p) is modular, consisting of an N-terminal AP-like endonuclease, a central reverse transcriptase and a C-terminal domain of unknown function (Cost and Boeke, 1998; Feng et al., 1996; Ostertag and Kazazian, 2001). The human ORF2p is thought to recognize the L1 RNA primarily via the poly(A)-tail and it displays a remarkable cis-preference, i.e. it binds preferentially to the RNA molecule from which it is being translated. This assures that only full-length L1 RNA encoding functional proteins participates in retrotransposition and it prevents other poly(A) containing RNAs in the cell from competing (Wei et al., 2001). Alu RNAs, however, that are part of Alu ribonucleoprotein particles (Weichenrieder et al., 2000), seem to efficiently interfere with this cis-preference of ORF2p. Alu particles are believed to recruit ORF2p as well primarily via their poly(A)-tail, because it is the only obvious component common to both L1 and Alu RNAs (Boeke, 1997; Dewannieux et al., 2003). Consequently, the L1 endonuclease is responsible not only for approximately 520.000 L1 integrations, but also for more than 1.090.000 Alu integrations in the human genome, accounting for more than a quarter of its mass (The human genome sequencing consortium, 2001).

L1-EN belongs to an enzyme family of metal-dependent phosphohydrolases that share the same fold and active site residues and that cleave a large variety of phosphoester substrates (Dlakic, 2000; Hofmann et al., 2000). Next to the AP-like retrotransposon-encoded endonucleases (Eickbush and Malik, 2002; Lovsin et al., 2001) there are a number of other subfamilies. These include nucleases like the AP DNA repair endonucleases (Barzilay and Hickson, 1995), secreted DNases (Lara-Tejero and Galan, 2000) and single-stranded RNA deadenylases (Dupressoir et al., 2001), but also other enzyme families like inositol polyphosphate phosphatases (Whisstock et al., 2000) and sphingomyelinases (Goni and Alonso, 2002). Members of some subfamilies have been characterized on a molecular level, with crystal structures of enzyme-substrate complexes yielding detailed insight into substrate recognition and clues on the catalytic mechanism (Mol et al., 1995; Mol et al., 2000; Tsujishita et al., 2001; Weston et al., 1992). Structural information on retrotransposon-encoded endonucleases, however, has been lacking so far, despite the fact that many of them are well characterized functionally and biochemically (Zingler et al.).

Here we report the first crystal structure of a retrotransposon-encoded protein, the human L1 endonuclease. We generated a precise, structure-based alignment with the other structurally determined members of the phosphohydrolase enzyme family via multiple three-dimensional superpositions. Additionally, we used our knowledge of the L1-EN structure in the alignment of representative endonuclease sequences that we retrieved or reconstructed from raw database entries and that comprise all known clades of APE-type non-LTR retrotransposons. The analysis of the L1-EN crystal structure in the light of these alignments enables us to discuss the DNA nicking mechanism in the context of a newly identified conserved residue. Furthermore, the comparison to the structures of DNA complexes from other members of the phosphohydrolase enzyme family together with the biochemical information on L1-EN allows us to speculate on the functions of elements that are involved in DNA substrate recognition. This is particularly interesting with respect to the possibility of manipulating substrate specificity and with the goal of using non-LTR retrotransposons as a genetic tool (Soifer et al., 2001). Finally, we traced the conservation of structural elements and of particular side chains throughout the whole phosphohydrolase family, thereby providing information for the refinement of phylogenetic and functional relations.

RESULTS AND DISCUSSION

Expression, characterization and structure solution

Recombinant human L1-EN (residues 1-239 of L1-ORF2p) was expressed in *Escherichia coli* without tags. The apparent hydrodynamic radius of the purified protein indicates it to be monomeric at concentrations up to 10 μM. Therefore, a possible dimerization of ORF2p as observed for viral reverse transcriptases (Kohlstaedt et al., 1992) does not seem to apply to the isolated endonuclease domain. In a plasmid-based nicking assay our untagged version of the protein behaves similarly to a previously characterized, tagged version (Feng et al., 1996). The slow enzymatic turnover observed in both cases might be due to an inhibition of product release in the absence of RNA template and reverse transcriptase similar to the effect observed for the APE1 enzyme that is involved in DNA repair (Mol et al., 2000). This would indicate co-operation in the process of target-primed reverse transcription between the nicking activity of the endonuclease and subsequent enzymatic steps. Crystals of L1-EN diffracted up to 1.8 Å resolution and the structure was solved by molecular replacement using the structure of APE1 as a search model. The final model of L1-EN was built automatically and refined to an R factor of 22.0 % (Table 1). Various attempts to co-crystallize L1-EN with substrate DNA were unsuccessful.

Description of the structure

The crystal structure of L1-EN is the first representative of an endonuclease from a non-LTR retrotransposon and the fifth of a protein from the phosphohydrolase enzyme family. Like the other members of the enzyme family, L1-EN is a two-layered α-β sandwich with, approximately, two-fold internal symmetry (Figure 1). Figure 2 shows details of the L1-EN structure and includes an idealized, common topology for the enzyme family with the L1-EN sequence superimposed. In this idealized topology the two halves A and B of the enzyme face each other via the two six-stranded β-sheets, each of which is flanked by two α-helices on the outside. The N- and C- termini are always located between β-strands βA2 and βA3 of half A, but apart from that the individual members of the enzyme family deviate from this common double βββαβαββ topology to various degrees. The connecting loops between these idealized secondary structure elements are quite variable. They occasionally contain additional strands or helices and define two surfaces on opposite sides of the molecule, one of which binds the substrate and contains the active site cleft.

The present model of L1-EN contains amino acids 3-238 of L1-ORF2p. The β-strand βB6 from the idealized topology is interrupted by an α-helix and the α-helix αA2 is replaced by a loop and a 3-10 helix. For convenience, we nevertheless refer to these elements as βB6 and αA2. Furthermore, loop βB6-βB5 adopts a particular and rigid hairpin structure that protrudes from the putative DNA binding surface of L1-EN. No metal ions were identified in the structure, but there are several sulfate ions. One of them is coordinated by the side chains of H45 and N19 on the DNA binding surface and another one by Y115.

Table I. Data collection and refinement statistics

Data collection		
Resolution, Å	20-1.8	
Cell dimensions, Å	a=91.0, b=126.5, c=43.0	
Space group	$P2_12_12$	
R_{merge}, % [a]	6.2	(44.2)
Completeness, % [a]	97.5	(90.2)
I/σ(I) [a]	17.2	(2.4)
No. of reflections		
Unique observed	45765	
Total measured	685558	
Refinement		
R_{cryst}, %	18.5	
R_{free}, %	21.9	
Number of		
molecules per asymmetric	2	
unit	4266	
atoms		
ions	10	
glycerol molecules	2	
water molecules	454	
Ramachandran plot		
Most favored regions, %	90.1	
Allowed regions, %	9.9	
R.m.s.d. from ideal geometry		
Bond lengths, Å	0.016	
Bond angles, °	1.68	
Optical resolution, Å	1.5	
(Vaguine et al., 1999)		

[a] Values in parentheses correspond to those in the outer resolution shell (1.86-1.8 Å)

by a loop and a 3-10 helix. For convenience, we nevertheless refer to these elements as βB6 and αA2. Furthermore, loop βB6-βB5 adopts a particular and rigid hairpin structure that protrudes from the putative DNA binding surface of L1-EN. No metal ions were identified in the structure, but there are several sulfate ions. One of them is coordinated by the side chains of H45 and N19 on the DNA binding surface and another one by Y115. Quite likely, they occupy positions that can be taken by backbone phosphates of the DNA substrate. These positions were used in the docking of an NMR model of substrate DNA (Stefl et al., 2004) to L1-EN (Figure 5).

Structure-based similarity search and alignments

To assign functional significance to elements within the L1-EN crystal structure we compared it in detail to related structures and complexes as well as to related sequences. This comparison was done on three levels: the level of the enzyme family of metal-dependent phosphohydrolases, the level of the subfamily of retrotransposon-encoded endonucleases and the level of closely related mammalian-type L1 endonucleases (from mammals and fish) that presumably nick the same DNA target sequence.

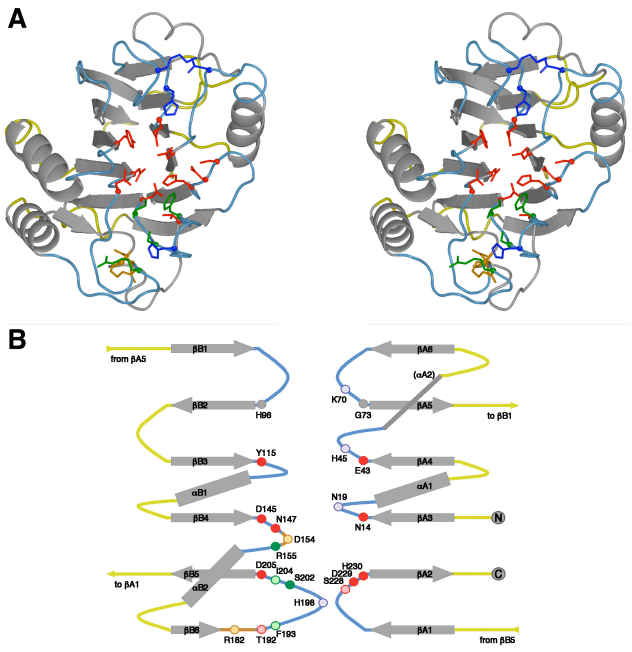


Figure 2. Structural details and topology of L1-EN. A. Ribbon diagram of L1-EN with the loops on the DNA binding side in cyan and the loops on the opposite side in yellow (stereo, top view). Selected side chains are drawn as balls-and-sticks and colored as in (B). B. Idealized topology diagram of the phosphohydrolase enzyme family, adapted to L1-EN. The diagram emphasizes the pseudo two-fold symmetry relating the two halves A and B of the molecule. Structural elements are labeled according to the respective half, and consecutively in space, not sequence. The prominent βB6-βB5 hairpin loop is enlarged and connections to the N and C termini are indicated. Selected residues are drawn as circles at their approximate positions and color-coded. Red: Residues conserved among all phosphohydrolases, that are catalytically (filled) or structurally (half-filled) important; Green: Residues proposed to recognize the extrahelical nucleotide via the ribose (half-filled) and the base (filled); Blue: Putative peripheral DNA binding residues; Orange: Salt-bridge restricted to AP DNA repair endonucleases and mammalian-type L1 endonucleases (half-filled).

First, a structure-based search (DALI, (Holm and Sander, 1993)) allowed us to retrieve four protein structures with significant (Z-score >2) similarity to L1-EN. Most similar are the human and bacterial (*Escherichia coli*) DNA repair endonucleases APE1 and ExoIII (Z-scores 26.1 and 25.9, respectively) followed by bovine DNaseI and yeast inositolpolyphosphate-5-phosphatase (IPP5) (Z-scores 19.9 and 15.5, respectively). The proteins belong to the phosphohydrolase enzyme family and function as isolated enzymes or in the co-ordinated context of multi-domain proteins (Figure 1). An accurate, structure-based alignment was derived from the superposition of these five structures. It reveals a minimal, common core of structural elements and of individual catalytic residues that are present in all five structures (Figure 3).

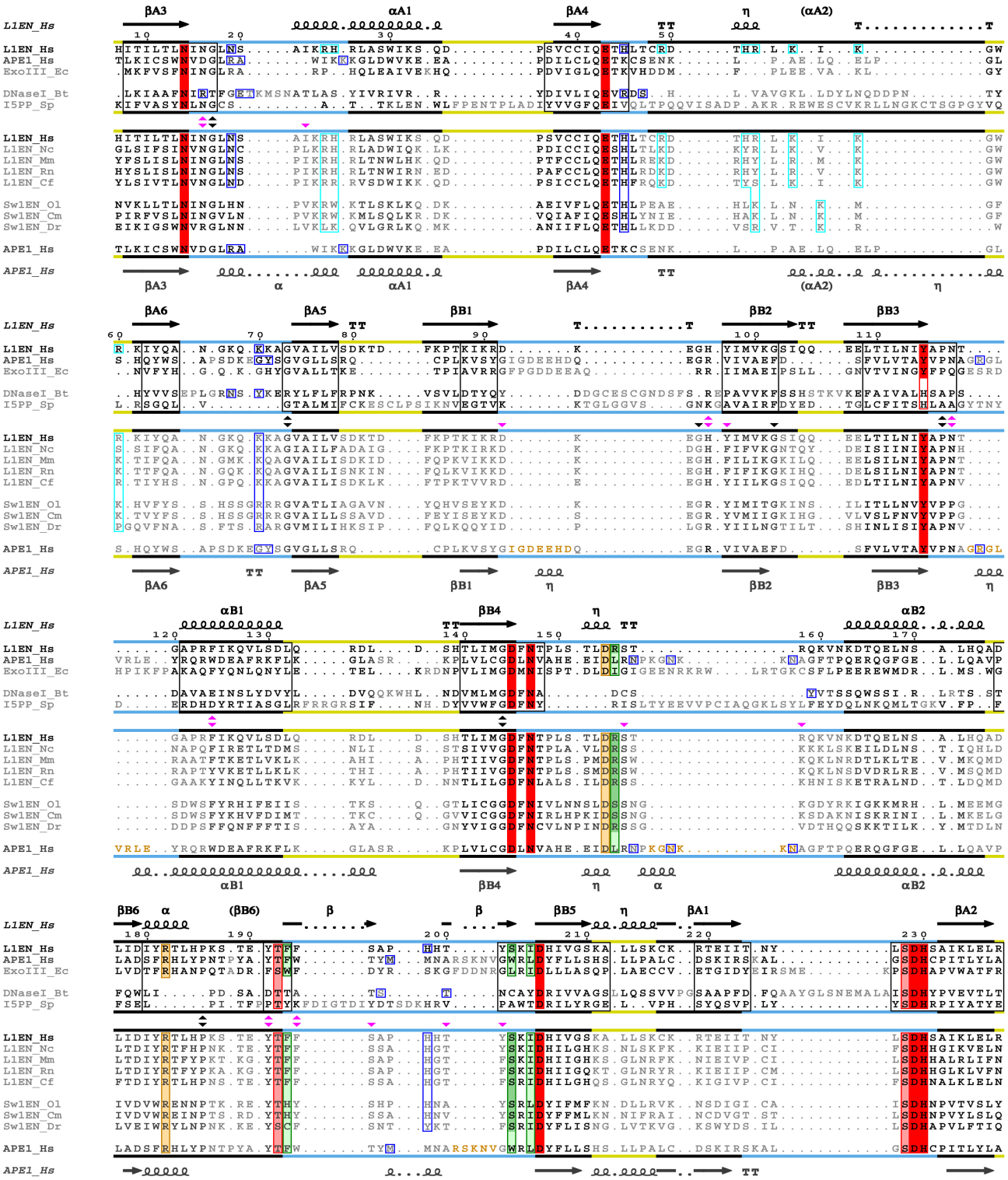
Second, a computer-assisted multiple sequence alignment of retrotransposon-encoded endonucleases was generated that covers all presently known clades of non-LTR retrotransposons (Eickbush and Malik, 2002; Lovsin et al., 2001). It takes into account not only sequence information (Clustal W, (Thompson et al., 1994)), but also structural information from the first, structure-based alignment alignment. In contrast to the first alignment all endonucleases within the present group are expected to be structurally adapted to participate in the mechanism of target-primed reverse transcription (Supplement).

Third, an alignment of a subset of closely related, mammalian-type L1 endonucleases was analyzed for conserved features that are not present in the other two alignments and that might be connected to functions restricted to this group of sequences only. Elements from this group presumably integrate into the same consensus DNA target site and, due to *cis*-preference in retrotransposition, their ORF2 protein is expected to bind the poly(A) tail of template RNA particularly tightly (Figure 3).

Conservation of the APE1-like active site and discussion of the DNA nicking mechanism

L1-EN is closely related to APE1. In contrast to L1-EN, the DNA substrate and product complexes of APE1 have been structurally determined, and a detailed reaction mechanism for this enzyme has been described (Mol et al., 2000). The nature and position of all the side chains in APE1 that are proposed to be involved in the orientation and cleavage of the scissile phosphoester bond are strictly conserved in L1-EN (Figure 4). It is therefore likely that also the catalytic mechanism *per se* is conserved between APE1 and L1-EN. In strict analogy to APE1, residue E43 in L1-EN (together with N14 and D229 (Beernink et al., 2001; Weston et al., 1992)) would be involved in the co-ordination of a magnesium ion that could stabilize the ribose 3'O leaving group while residues Y115, D145, N147, D205 and H230 would preorient the scissile bond and generate the attacking nucleophile. A role for N118 (N174 in APE1) in stabilizing the pentacovalent phosphate transition state remains possible in human L1-EN, but the residue is otherwise not conserved among AP-like retrotransposon-encoded endonucleases (Figure 3). D205 (D283 in APE1) forms a hydrogen bond to H230 (H309 in APE1), elevating its pK_a value. Therefore, this histidine was originally proposed to generate the attacking nucleophile (Gorman et al., 1997; Mol et al., 1995). However, the geometry in the subsequently solved complex of APE1 and abasic DNA suggested this role to rather be fulfilled by the equivalent of D145 (D210 in APE1, (Mol et al., 2000)). Finally, yet another mechanism was proposed for APE1, in which the attacking nucleophile is generated by a second metal ion that, in L1-EN, would be coordinated by residues D145, N147 and H230 (Beernink et al., 2001). It is still an open question which of these proposed mechanisms takes place and if they are necessarily mutually exclusive for all phosphohydrolases and for all their proposed functions. In this context, Y115 in L1-EN is particularly interesting. While all other catalytic residues are strictly conserved among the phosphohydrolase enzyme family, Y115 is conserved only among

Figure 3. Sequence alignments. Top five sequences: Structure-based alignment of phosphohydrolases with known crystal structures. Positions are shaded according to whether they can be aligned and superimpose (black), whether they can be aligned but do not superimpose (dark grey) or whether they cannot be aligned (light grey) with positions from the human L1 endonuclease. Positions that can be aligned and superimpose in all five sequences define the conserved core (boxed black). Bottom nine sequences: Automatic align-



ment of retrotransposon-encoded endonucleases, with structure-assisted manual adjustments. Only the subset of mammalian-type L1 endonucleases that nick the same DNA target sequence is shown, and, for comparison, APE1. Positions are shaded according to whether they, in the full alignment of retrotransposon-encoded endonucleases, can always be aligned automatically (black), whether they can always be aligned with manual adjustment (dark grey) or whether the alignment is not possible for all sequences (light grey). Residues in APE1 that lie on the DNA-binding surface and have no correspondence in L1-EN are highlighted (orange). Secondary structure elements are labeled according to the idealized phosphohydrolase topology (Figure 2). Horizontal bars mark the DNA-binding, top surface (cyan) and the opposite, bottom surface (yellow) of the molecule. Selected positions are colour-boxed. Red: Residues conserved among all phosphohydrolases, that are catalytically (filled or empty) or structurally (half-filled) important; Green: Residues proposed to recognize the extra-helical nucleotide via the ribose (half-filled) and the base (filled); Blue: Peripheral and putative peripheral DNA binding residues (empty); Cyan: Residues that are part of a positively charged patch of molecular surface that is restricted to (certain) L1 endonucleases. Orange: Salt-bridge restricted to AP DNA repair endonucleases and mammalian-type L1 endonucleases (half-filled). Structurally important glycines and prolines (black triangles) and conserved surface residues that might be functionally important (magenta triangles) are indicated. Sequences are labeled according to clade, individual name and organism (abbreviations and accession numbers see methods). α, alpha-helix; β, beta-strand; η, 3-10 helix; TT, beta-turn.

AP-like retrotransposon-encoded endonucleases and AP DNA repair endonucleases and is otherwise replaced by histidine. This illustrates the phylogenetic proximity of those two enzyme subfamilies and points to a common, yet unidentified function shared exclusively by them (Figure 3, Supplement). Unfortunately, due to the absence of substrate and metal ions in the present structure, the mechanistic details of DNA nicking by L1-EN cannot be resolved any further here.

Instead, we shall focus on the role of T192 in L1-EN, which has not been identified as a highly conserved residue previously. It plays an important structural role as a 'cornerstone' at the base of the aforementioned and prominent βB6-βB5 loop. This loop inserts into and partially collides with the wide minor groove of the docked DNA substrate, and we presume it to bend or unwind the DNA downstream (3') of the cleavage site (Figure 4, 5A). The backbone of T192 is fixed by multiple hydrogen bonds and the side chain oxygen receives weak hydrogen bonds from the main chain nitrogens of I204 and D205 at the other end of the βB6-βB5 loop. This anchors the bottom of the loop with respect to the active site. Interestingly however, the side chain oxygen of T192 also has the potential to donate a hydrogen bond to the same side chain oxygen of D205 that also receives the hydrogen bond from the catalytic H230 (Figure 4A, dotted red). As the angle, and hence the strength of this T192-D205 hydrogen bond varies considerably throughout the known crystal structures, one might speculate that the bond could be weakened by the transitory strain that the presence of uncleaved substrate DNA puts on the βB6-βB5 loop. A weakened bond between T192 and D205 would ultimately elevate the pK_a of H230. This could trigger the subtraction of a proton from a water molecule, generating the nucleophilic hydroxyl ion and/or help in the orientation of the scissile phosphate (Figure 4A, dotted green and cyan). Although quite speculative, this potential mechanochemical coupling in addition to the structural role would neatly explain the conservation of T192 as a threonine or serine in all metal-dependent phosphohydrolases (Figure 3, Supplement).

Comparison of L1-EN and APE1 regarding their potential to recognize an extra-helical nucleotide

The DNA recognition specificity of the endonuclease is the major determinant for the selection of a new integration site by a human L1 retrotransposon. Biochemical data and statistical sequence analysis indicate L1-EN to nick DNA at a 5' TTTT-AA 3' consensus sequence that is found at the junction of two opposing A-tracts (Feng et al., 1996; Jurka, 1997). In particular, L1-EN seems to recognize the special geometry of the A-tract upstream (5') of the scissile bond, and access to the DNA minor groove is thought to be important for phosphodiester hydrolysis. The protein seems to sense the flexibility of the DNA at the T-A step, where base-stacking is minimal (Cost and Boeke, 1998; Mack et al., 2001; Stefl et al., 2004). This mode of DNA recognition is due to both central and peripheral residues on the DNA-binding surface of L1-EN. Their role and individual contributions can be partially understood from a comparison of the structure of L1-EN to the structures of the DNA complexes of APE1 and DNaseI.

The active site cleft of APE1 not only contains the catalytic residues but also a hydrophobic pocket that accommodates the abasic ribose downstream (3') of the scissile bond in an extra-helical conformation (Figure 4B). In the APE1-DNA complex the flipped, abasic ribose rests on the small L282 and is flanked and positioned by F266 and W280. The bulky W280, together with residues from the βB4-αB2 loop, has the additional role of restricting the size of the hydrophobic pocket to fit only to an apurinic or an apyrimidinic residue. This prevents any flippable purine or pyrimidine nucleotide from entering and initiating unnecessary DNA repair. In the L1-EN structure L282, F266 and W280 are replaced by I204, F193 and S202, respectively (Figure 3). Because the I204 is small enough for a flipped ribose to be placed, and because the absence of the tryptophan at position 202 allows space for even a purine base, it is quite possible that L1-EN accommodates an extra-helical adenine downstream (3') of the scissile bond (Figure 4B, 4C, 5B). This hypothesis is supported by the exceptional mobility of the respective adenine, which results from very little stacking overlap at the junction of the two DNA A-tracts. Experimental disruption of the adenine mobility at the T-A step reduces DNA hydrolysis, whereas a widening of the minor groove by DNA bending increases adenine mobility and DNA hydrolysis (Cost and Boeke, 1998; Cost et al., 2001). It is likely that the extra-helical adenine is specifically recognized by L1-EN, compensating for the energetic cost of breaking its Watson-Crick hydrogen bonds and its remaining stacking interactions. S202, which is strictly conserved in the alignment of mammalian-type L1 elements, and R155 are likely residues to form hydrogen bonds with the extra-helical base (Figure 4C).

It seems that all retrotransposon-encoded endonucleases are able to accommodate an extra-helical nucleotide downstream (3') of the scissile bond. This can be concluded from the fact that the respective ribose can always rest on a small hydrophobic residue located at the position corresponding to I204 in L1-EN. The respective base seems to have space in all cases, although occasional exceptions might exist, where DNA target recognition proceeds without base-flipping. This would then be similar to the situation found in the DNaseI-DNA complex, where the downstream (3') ribose cannot be flipped due to the presence of a bulky tyrosine in the place of I204. Importantly however, the geometry of the proposed cleavage reaction is not affected by this, because the phosphate directly downstream (3') of the scissile bond is still bound in the same orientation in both the APE1 and the DNaseI complexes. An extra-helical ribose downstream (3') of the scissile bond is therefore no prerequisite for the cleavage mechanism to proceed. Nevertheless, the possibility to flip a base upon DNA bending and the subsequent recognition of the extra-helical ribose/base can contribute significantly to the specificity of the cleavage reaction.

Comparison of L1-EN and APE1 with respect to the general mode of DNA binding

Judging from the conserved core of both L1-EN and APE1 one can conclude that both enzymes follow a

very similar mechanism of substrate binding and cleavage. The main reason why L1-EN does not show any cleavage preference for abasic sites is therefore probably that it recognizes its target in a much more restricted context of DNA structure and sequence. Outside of the active site cleft many of the DNA binding residues in APE1 are not conserved in L1-EN or not even alignable (Figure 3). They are found in surface loops of individual structure and sequence, rendering a direct extrapolation towards DNA binding of L1-EN difficult. Nevertheless, some general conclusions are possible, because the L1-EN protein structure is not expected to change significantly upon DNA binding (Mol et al., 2000; Weston et al., 1992) and because the model of docked A-tract substrate DNA is of sufficient quality (Figure 5).

The regions of the L1-EN and APE1 proteins that anchor the intensively recognized DNA upstream (5') of the scissile bond (loops βA3-αA1, βA4-αA2, βA6-βA5) are structurally more similar to each other than the regions of the proteins that fix the DNA downstream (3') of the scissile bond (loops βB3-αB1, βB4-αB2, βB6-βB5). Furthermore, the alignment of mammalian-type L1 endonucleases that are believed to cleave the same 5' TTTT-AA 3' consensus sequence shows several surface side-chains within that 5' binding region to be conserved (Figure 6). K70/K71 correspond in position to the DNA binding residues G127/Y128 in APE1 and to Y76/K77 in DNaseI (Figure 3). H45 and N19 bind a sulfate ion (in all three available structures of L1-EN), which, extrapolating from the APE1-DNA structure, could very well take the place of the DNA backbone phosphate four base-pairs upstream (5') of the scissile bond, where the width of the narrow A-tract DNA minor groove seems to be sensed by the protein (Cost and Boeke, 1998). APE1 and, in particular, ExoIII also possess general 3'exonuclease, 3'phosphodiesterase and RNaseH activities. The absence of a downstream (3') DNA duplex in these cases re-illustrates the importance of upstream (5') over downstream (3') duplex binding. Retrotransposon-encoded endonucleases normally do not show any of these activities, although the present structure would not preclude them *a priori* (Figure 5).

With the upstream (5') DNA locked onto the L1-EN surface via the presumed K70/H45 contacts on one side and the active site contact on the other side the orientation, bendability and minor groove width of the downstream (3') DNA can be explored by the rest of the L1-EN DNA binding surface, in particular the loops βB3-αB1, βB4-αB2 and βB6-βB5. Loop βB3-αB1, which contacts the DNA from the side of the major groove and carries a functionally important arginine in APE1 is almost absent in L1-EN (Figure 5A, 5C). This would allow the prominent hairpin loop βB6-βB5 of L1-EN to push and bend the DNA from the side of the minor groove towards loop βB3-αB1 much more than in the case of APE1. Loop βB6-βB5 is rigidified by multiple internal hydrogen bonds and anchored within the active site cleft. H198, located on the tip of this loop is conserved among mammalian-type L1 endonucleases and points toward the minor groove of the 3' DNA duplex (Figure 2, 5A). The potential of the protein to bend the DNA at the junction of the two A-tracts might be an important parameter for initial DNA recognition.

The direction of the bend towards the major groove would increase the mobility of the adenine downstream

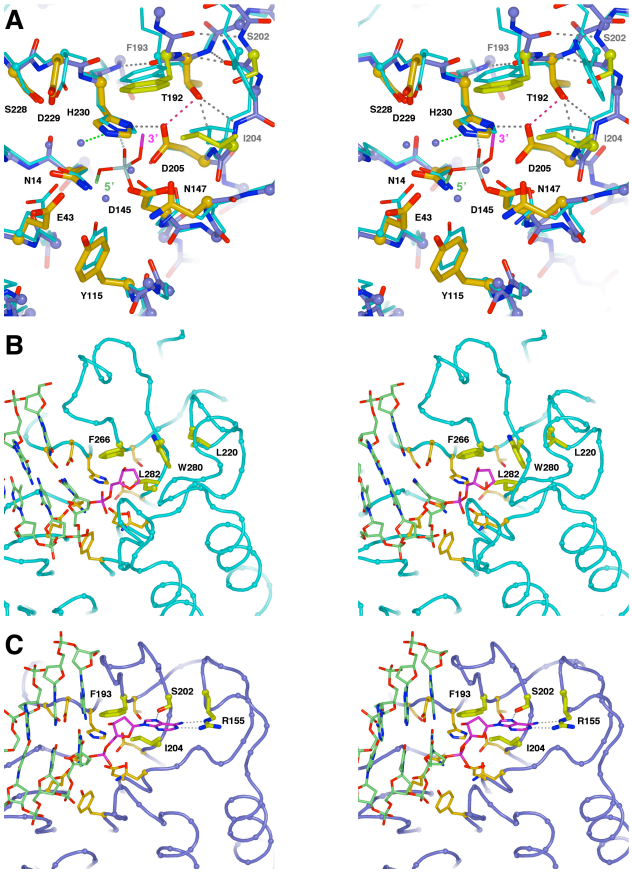


Figure 4. Active site comparison between AP-like retrotransposon-encoded endonucleases (L1-EN, blue) and AP DNA repair endonucleases (APE1, cyan). Selected side-chains are drawn as balls-and-sticks (red, oxygens; blue, nitrogens) and colored mandarine (involved in catalysis) or lemon (recognizing the extra-helical nucleotide). **A.** Superposition of L1-EN and APE1 including water molecules from L1-EN and the scissile phosphate of the APE1 DNA substrate. Upstream (5', lime-green) and downstream (3', magenta) directions are indicated. Identity of catalytic residues between L1-EN and APE1 with respect to chemistry and position indicates a conserved mechanism of phosphodiester hydrolysis. Hydrogen bonds relevant to the newly identified, conserved T192 are drawn as dotted lines. **B**. Recognition of the extra-helical abasic nucleotide by APE1. Upstream (5') DNA is lime-green. Downstream (3') DNA is magenta but omitted for clarity apart from the extra-helical nucleotide. **C.** Model for the accomodation and recognition of an extra-helical adenine by L1-EN. Compared to APE1, space for the base is not restricted and three hydrogen bonds are possible to residues S202 and R155.

(3') of the scissile bond, promoting it to flip and locally unwind the DNA duplex (Figure 5B). Finally, loop βB4-αB2 is also smaller in L1-EN than it is in APE1. The N-terminal half of loop βB4-αB2 is very similar in both proteins and fixed by the same salt bridge (D154-R182 and D219-R254, respectively) (Figure 2). The C-terminal half of loop βB4-αB2 differs significantly between the two proteins, but it might fix bent or unwound downstream (3') DNA also in L1-EN. Most importantly, the smaller size of the loop in L1-EN liberates the space for the downstream (3') adenine to flip.

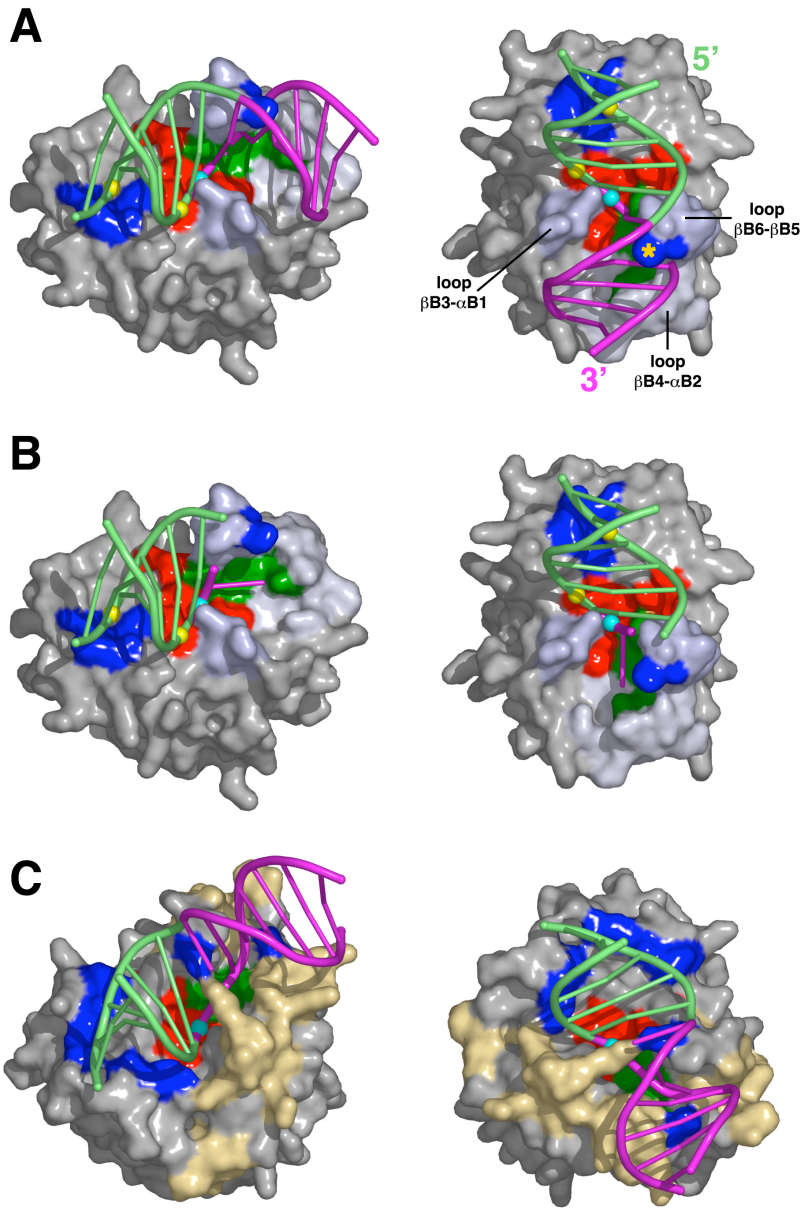


Figure 5. Model for the recognition of A-tract DNA by L1-EN. A. Surface representation of L1-EN (colors as in Figure 2) with a docked NMR model of substrate DNA (Stefl et al., 2004) represented as ribbons. The upstream (5') and downstream (3') duplexes are lime-green and magenta, respectively. Sulfate ions on the surface of L1-EN used to position backbone phosphates of the DNA are yellow, the scissile phosphate in the active site is cyan. Loop Bβ6-Bβ5 with H198 (asterisk) on its tip inserts into the wide minor groove at the TpA step. This likely bends or unwinds downstream (3') DNA, promoting the adenine to flip. Left: view as in Figure 4; Right: view as in Figure 2. **B**. Model including only upstream (5') DNA and the flipped adenine downstream of the scissile bond (views and colors as in A). **C.** APE1 bound to substrate DNA (style, views and colors as in A). Surface patches corresponding to residues that have no equivalent in L1-EN and that occlude the active site cleft are in orange.

The general principle of target DNA binding and bending seems conserved among all retrotransposon-encoded endonucleases (Supplement). The putative DNA-binding loops are quite variable in sequence, however, probably reflecting the sequence variability of the respective DNA targets. Compared to APE1, the three loops on the downstream (3') side of the DNA are smaller in L1 endonucleases, leaving the active site cleft relatively exposed (Figure 3, 5A, 5C). Also this feature may be general for all retrotransposon-encoded endonucleases. It would allow easy access for RNA template and reverse transcriptase in order for TPRT to proceed in a coordinated fashion.

Phylogenetic and functional conclusions for AP-like retrotransposon-encoded endonucleases

With respect to structure, the overall architecture of L1-EN with its relatively accessible active site cleft is generally conserved among all AP-like retrotransposon-encoded endonucleases. The elements of the conserved core as defined by the structural superpositions can be located in most sequences. The biggest variability, where sequences are not alignable or not present at all, is in the area of the molecule formed by the elements including loop βA4-αA2 to loop βA5-βA6 and β-strand βB1. In the most extreme case, the Sam3 element from the CR1 clade, loop βA4-αA2 is linked directly to β-strand βB6, completely deleting the elements in between. A similar situation is present in many sequences of the R1 clade, albeit with slightly longer βA4-βA6 linkers. Another region of interesting architectural variability is loop βB2-αB2, against which the downstream (3') DNA duplex is expected to lean. In most sequences of the L1 clade this loop is fixed by a salt bridge that is otherwise restricted to the subfamily of AP DNA repair endonucleases. In L1-EN, the bridge is formed between D154 in loop βB2-αB2 and R188 in a supporting α–helix that is interrupting strand βB6 (Figure 2). A similar stabilization might occur in the RTE clade, while in all other clades the supporting α-helix

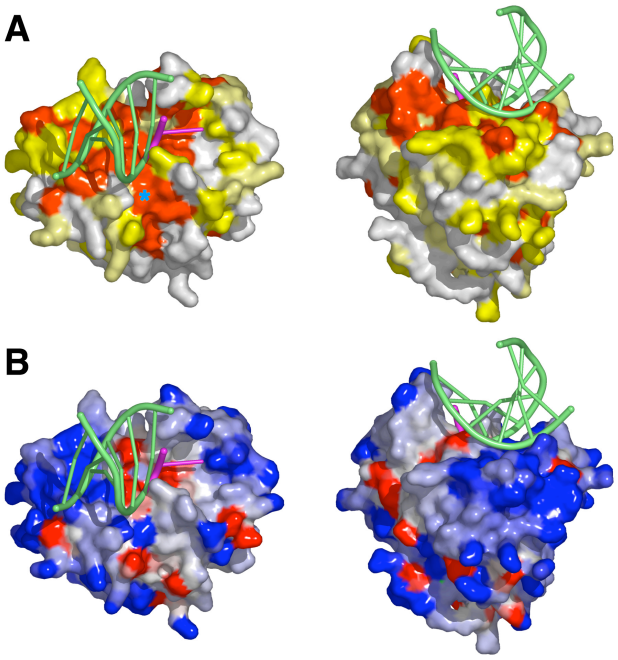


Figure 6. Analysis of conserved surface features among L1 endonucleases. A. Conservation of surface residues. The surface of human L1-EN is color-ramped orange to white according to the conservation of residues among mammalian-type L1 endonucleases. Left: view as in Figure 4 with H96 indicated by an asterisk; Right: view towards the conserved backside of loop Bβ6-Bβ5. **B.** Electrostatic surface potential (-15 kT (red) to +15 kT (blue), GRASP) mapped onto the surface of L1-EN (views as in A). The prominent patch of positive potential is restricted to mammalian-type L1 endonucleases.

seems to be deleted. In those cases loop βB2-αB2 either contains a structure around a conserved tryptophan (currently aligned with L153 of L1-EN) or, like in DNaseI and IPP5, a significant deletion that might even extend far into α–helix αB2. The structure around the conserved tryptophan is restricted to a subset of retrotransposon-encoded endonucleases, which, for reasons of parsimony, are likely to bear a common evolutionary origin (Figure 3, Supplement).

Regarding other retrotransposon- or TPRT-related functions initial clues are provided mainly by the subset of mammalian-type L1 endonucleases. Their alignment allows the identification of some additional conserved and exposed surface residues that are not obviously involved in DNA binding and cleavage (Figure 6A). These are H96 and other residues clustering in and around the small loop βB1-βB2 and a group of aromatic side chains (Y191, F194, Y201) at the backside of loop βB6-βB5. H96 is close to Y115 and might modulate access to the active site cleft. It is conserved as a charged residue with low variability throughout most retrotransposon-encoded endonucleases and many phosphohydrolases. The group of aromatic side chains is restricted to mammalian-type L1 endonucleases. Interestingly, some of the respective ORF2 proteins are thought to hold particularly tightly to their template RNA (a requirement for *cis*-preference in retrotransposition) and aromatic side chain stacking has often been observed in single-stranded RNA binding (Mazza et al., 2002). Additionally, mammalian-type L1 endonucleases share a patch of highly positively charged molecular surface that is formed by the generally variable elements including loop βA4-αA2 to loop βA5-βA6 (Figure 6B). Also this basic patch could possibly be involved in some specialized nucleic acid-binding activity.

Summary and perspectives

The crystal structure of the human L1 endonuclease (L1-EN) is a prototype for AP-like retrotransposon-encoded endonucleases, which nick DNA with variable specificity and are responsible for millions of retrotransposon insertions in eukaryotic genomes. The structure of L1-EN supports an AP-like catalytic mechanism and the recognition of an extra-helical nucleotide. An extensive structure-assisted sequence alignment covers AP-like endonucleases from all known clades of non-LTR retrotransposons and allows new sequences to be compared quickly. The crystal structure of L1-EN together with the present alignment will greatly facilitate attempts to modulate the sequence specificity of any given endonuclease, e.g. in order to convert the respective retrotransposon into a genetic tool with target-site specificity. Furthermore we have started to identify conserved features in L1-EN that might be involved in other TPRT or retrotransposition-related functions. Combined with the existing powerful *in vitro* (Cost et al., 2002) and *in vivo* (Gilbert et al., 2002; Moran et al., 1996; Symer et al., 2002) assay systems the present structure will undoubtedly push the analysis and understanding of non-LTR retrotransposition to a new level.

MATERIALS AND METHODS

Cloning, expression and purification of human L1 endonuclease (L1-EN)

A DNA fragment encoding residues 1-239 of human L1 ORF2p was PCR-amplified using primers *Nco*I-L1O2-N1 (5’ aat ctg gaa acc atg gcg gga tca aat tca cac ata aca ata 3’) and *Xho*I-L1O2_C239 (5’ agc tag ctc gag tta tta aat cct gag ttc tag ttt gat tg 3’) on plasmid pJM130 containing a subcloned, functional L1 element (L1.3, gi:307098, (Sassaman et al., 1997)). The amplified fragment was inserted into the expression plasmid pET-15b (Novagen) using the restriction sites *Nco*I and *Xho*I. L1-EN was overexpressed in *Escherichia coli* BL21(DE3)pLysS (Novagen) after induction with 500 µM isopropylthiogalactoside at an optical density of OD_{600} = 0.7. Cells (2 l) were grown at 37 ºC for 2 h, harvested and lysed by sonication in buffer (20 mM HEPES (pH=7.5), 20 mM EDTA, 10 mM β-mercaptoethanol), supplemented with 300 mM NaCl and 1 mM PMSF. The cleared lysate was applied to a Heparin column (HiTrap HP, 5 ml, Pharmacia), and the protein was eluted at around 850 mM NaCl, diluted in buffer to 250 mM NaCl and loaded onto an ion exchange column (ResourceS, 6 ml, Pharmacia) from where it was eluted at around 500 mM NaCl. After gel filtration chromatography over a Superdex 75 column (HiLoad 26/60, Pharmacia) in buffer containing 300 mM

NaCl but no EDTA the protein was concentrated to 15 mg/ml and stored at 4 ºC.

Characterization and crystallization of human L1-EN

Analytical gel filtration was done on a calibrated Superdex 75 column (PC 3.2/30, SMART system, Pharmacia). Nicking activity was tested as described in (Feng et al., 1996). Crystallization was achieved by vapor diffusion using the hanging drop method. 1 µl of protein solution (15 mg/ml) was mixed with 1 µl of reservoir solution and equilibrated over 500 µl of reservoir solution at 20 ºC. A first crystal form (40 µm x 40 µm x 200 µm, space group $P2_12_12$) was obtained in 2-5 days over a reservoir of 0.2 M $(NH_4)_2SO_4$, 28 % polyethylene glycol 5000 monomethyl ether, 5 mM $MgCl_2$. A second crystal form (80 µm x 80 µm x 80 µm, space group $C222_1$) was obtained over a reservoir of 0.2 M $(NH_4)_2SO_4$, 30% polyethylene glycol 1000, 5 mM $MgCl_2$. Crystals of the first form were transferred to cryo-protectant (reservoir solution mixed 6.5 plus 1.5 with 80% aqueous glycerol) and immediately flash-frozen in liquid nitrogen. Crystals of the second form were flash-frozen directly from the crystallization drop.

Data collection, structure solution and refinement

Data were collected at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) beamline ID-14 and at the European Molecular Biology Laboratory (EMBL, Hamburg, Germany) beamline BW7B. The two different crystallization conditions resulted in crystals that belong to two different but related space groups $P2_12_12$ and $C222_1$. $C222_1$ crystals contain one molecule of L1-EN in the asymmetric unit and diffract X-rays to 2.1 Å, while those of space group $P2_12_12$ contain two molecules per asymmetric unit and their diffraction extends to 1.8 Å. Diffraction data were processed with programs MOSFLM (Leslie, 1992) and SCALA (Evans, 1997). A molecular replacement solution using the manually trimmed structure of the β-sandwich core of APE1 (APE1 has 23 % overall sequence identity) gave a molecular replacement solution using the program MOLREP (Vagin and Teplyakov, 1997), but only in the $C222_1$ space group. After a few cycles of refinement of that model with the program REFMAC5 (Murshudov et al., 1997), the refined model was used to obtain a molecular replacement solution for the $P2_12_12$ space group. That solution was used as input to the program ARP/wARP (Perrakis et al., 2001) for automated model building. The automatically built model was manually adjusted using the program O (Jones et al., 1991) and refined further using REFMAC5 to an R factor of 18.6 % and an R_{free} factor of 22.0 %. Given the high resolution of the diffraction data NCS restraints were not employed at any stage of the refinement. Refinement of the structure in space group $C222_1$ was not pursued due to the poor quality of the data (ice-rings and low completeness), but there are no major structural differences between the two space groups. Figures were generated using the program PYMOL (DeLano, 2002).

Structure-based alignments

For the structure-based alignment of phosphohydrolases the four available structures were first superimposed onto the structure of L1-EN via the central β-sandwich. Residues were classified into three groups by manual inspection. Group1: Residues that can be aligned and superimpose; Group2: Residues that can be aligned but do not superimpose; Group3: Residues that cannot be aligned. Positions in the alignment that only contain residues from group1 are classified as "conserved core".

For the structure-assisted alignment of retrotransposon-encoded endonucleases protein sequences were directly retrieved from the nucleotide database entries or other published sources or, where necessary, manually reconstructed from the raw database entries (selecting those with no or very few frameshifts, extending the N-terminal sequence beyond the first methionine, deriving consensus sequences, etc.). The following sequences were initially (Zingler et al.) aligned in the given order with Clustal W (Thompson et al., 1994) using default parameters except for gap and pair-gap penalties (lowered to 5 and 1, respectively). APE1; L1-clade: L1(L1.3)-Hs, L1-Nc, L1(Tf5)-Mm, L1-Rn, L1-Cf, Sw1-Ol, L1-Dr, Tx1L-Xl, TRE5A-Dd, Zorro3-Ca, Ylli-Yl; RTE1-clade: RTE1-Ce, SjR2-Sj, BovB-LINE-Va; Tad1-clade: Tad1-Nc, Mgr583-Mg, CgT13-Cg; R1-clade: R1-Bm, TRAS1-Bm, SART1-Bm; LOA-clade: LOA-Ds, Lian-Aa, bilbo-Ds; I-clade: I-Dm, MosquI-Aa, You-Dm; Ingi-clade: L1Tc-Tc, IngiTRS-Tb; Jockey-clade: Jockey-Dm, TART-Dm, Juan-Dm; CR1-clade: CR1-Gg, BfCR1-Bf, Q-Ag, Sam3-Ce, Pido-Sj, L2-clade: Maui-Fr. To test the stability of this multiple sequence alignment we then varied the order and number of sequences as well as the gap penalties and classified positions into three groups. Group1: Positions that can always be aligned automatically (black); Group2: Positions that can always be aligned with manual adjustment; Group3: Positions where the alignment is not possible for all sequences. The initial alignment was then adjusted manually, taking into account structural information. Subsequently the following pre-aligned sequences were added manually. Rex1-clade: Rex1-Tn, Rex1-Ol, Rex1-Cp; L2-clade: L2-Ol, L2-Sp, L2-Hs. Basic illustration and secondary structure assignments were done with ESPRIPT (Gouet et al., 1999), the final rendering manually.

Accession numbers and PDB codes

Retrotransposon-encoded endonuclease sequences used for alignments are: L1(L1.3)-Hs, gi:307098; L1-Nc, gi:126296; L1(Tf5)-Mm, gi:3599318; L1-Rn, gi:1791242; L1-Cf, gi:2981630; Sw1-Ol, gi:3746497; Sw1-Cm, gi:3746505; Sw1-Dr, gi:21914808; Tx1L-Xl, gi:214844; TRE5A-Dd, gi:10938; Zorro3-Ca, gi:14286188; Ylli-Yl, gi:20513183; RTE1-Ce, gi:3283066; SjR2-Sj, gi:19067878; BovB-LINE-Va, gi:16076778; Tad1-Nc, gi:409759; Mgr583-Mg, gi:2454620; CgT13-Cg, gi:1237262; R1-Bm, gi:340687; TRAS1-Bm, gi:940388; SART1-Bm, gi:2055274; LOA-Ds, gi:9150; Lian-Aa, gi:2290211; bilbo-Ds, gi:2708264; I-Dm, gi:157749; MosquI-Aa, gi:6635953; You-Dm, gi:11323017; L1Tc-Tc, gi:602092; IngiTRS-Tb, gi:10554; Jockey-Dm, gi:17823; TART-Dm, gi:603662; Juan-Dm, gi:27368147; CR1-Gg, gi:2331057Q; BfCR1-Bf, gi:17529693; Q-Ag, gi:432429;

Sam3-Ce, gi:1166577; Pido-Sj, gi:18091719; Rex1-Tn, (Volff et al., 2000); Rex1-Ol, gi:18157518; Rex1-Cp, 12004981; Maui-Fr, gi:4378023; L2-Ol, gi:12313699; L2-Sp, gi:8289138; L2(MIR)-Hs, (Lovsin et al., 2001). Abbreviations are: Aa, *Aedes aegypti*; Ag, *Anopheles gambiae*; Bf, *Branchiostoma floridae*; Bm, *Bombyx mori*; Ca, *Candida albicans*; Ce, *Caenorhabditis elegans*; Cf, *Canis familiaris*; Cg, *Colletotrichum gloeosporioides*; Cm, *Cyprinodon macularius*; Cp, *Calliactis parasitica*; Dd, *Dictyostelium discoideum*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Ds (LOA), *Drosophila silvestris*; Ds (Bilbo), *Drosophila subobscura*; Fr, *Fugu rubripes*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Mg, *Magnaporthe grisea*; Mm, *Mus musculus*; Nc (L1), *Nycticebus coucang*; Nc (Tad1), *Neurospora crassa*; Ol, *Oryzias latipes*; Rn, *Rattus norvegicus*; Sj, *Schistosoma japonicum*; Sp, *Strongylocentrotus purpuratus*; Tb, *Trypanosoma brucei*; Tc, *Trypanosoma cruzi*; Tn, *Tetraodon nigroviridis*; Va, *Vipera ammodytes*; Xl, *Xenopus laevis*; Yl, *Yarrowia lipolytica*.

PDB accession codes for structures and sequences are: L1-EN (human L1 endonuclease), PDB-ID XXXX; APE1 (human apurinic/apyrimidinic (AP) DNA repair endonuclease), PDB-ID 1dew; ExoIII (bacterial (*Escherichia coli*) AP DNA repair endonuclease), PDB-ID 1ako; DNaseI (bovine DNase I), PDB-ID 1dnk; IPP5 (yeast (*Saccharomyces cerevisiae*) inositol polyphosphate-5-phosphatase), PDB-ID 1i9z; A-tract DNA, PDB-ID 1rvi.

ACKNOWLEDGMENTS

We thank Jos Jonkers, The Netherlands Cancer Institute, Amsterdam, for providing plasmid pJM130. We also thank Gerald Schumann, Paul-Ehrlich-Institute, Langen, Germany, for discussions and a critical reading of the manuscript. O.W. received an EMBO postdoctoral fellowship and is currently funded by a Marie Curie postdoctoral fellowship from the European Union. K. R. is funded by grant NWO-CW OC.01.017 awarded to A.P. and O.W. by the chemistry division of the Dutch National Science Organization. A.P. is an EMBO Young Investigator.

REFERENCES

Barzilay, G., and Hickson, I. D. (1995). Structure and function of apurinic/apyrimidinic endonucleases. Bioessays *17*, 713-719.

Beernink, P. T., Segelke, B. W., Hadi, M. Z., Erzberger, J. P., Wilson, D. M., 3rd, and Rupp, B. (2001). Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape1: implications for the catalytic mechanism. J Mol Biol *307*, 1023-1034.

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

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

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.

Deininger, P. L., Moran, J. V., Batzer, M. A., and Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. Curr Opin Genet Dev *13*, 651-658.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System (http://www.pymol.org), DeLano Scientific LLC, San Carlos, CA, USA.).

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat Genet *35*, 41-48.

Dlakic, M. (2000). Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. Trends Biochem Sci *25*, 272-273.

Dupressoir, A., Morel, A. P., Barbot, W., Loireau, M. P., Corbo, L., and Heidmann, T. (2001). Identification of four families of yCCR4- and Mg2+-dependent endonuclease-related proteins in higher eukaryotes, and characterization of orthologs of yCCR4 with a conserved leucine-rich repeat essential for hCAF1/hPOP2 binding. BMC Genomics *2*, 9.

Eickbush, T. H., and Malik, H. S. (2002). Origins and Evolution of Retrotransposons. In Mobile DNA II (Washington, D.C., USA, ASM Press), pp. 1111 - 1144.

Evans, P. R. (1997). 'SCALA'. Joint CCP4 and ESF-EACBM Newsletter *33*, 22-24.

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

Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315-325.

Goni, F. M., and Alonso, A. (2002). Sphingomyelinases: enzymology and membrane activity. FEBS Lett *531*, 38-46.

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*, 6548-6558.

Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics *15*, 305-308.

Hofmann, K., Tomiuk, S., Wolff, G., and Stoffel, W. (2000). Cloning and characterization of the mammalian brain-specific, Mg2+-dependent neutral sphingomyelinase. Proc Natl Acad Sci U S A *97*, 5895-5900.

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

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J Mol Biol *233*, 123-138.

Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallographica A *47*, 110-119.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A *94*, 1872-1877.

Kazazian, H. H., Jr. (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626-1632.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992). Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science *256*, 1783-1790.

Lara-Tejero, M., and Galan, J. E. (2000). A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. Science *290*, 354-357.

Leslie, A. G. W. (1992). 'MOSFLM'. Joint CCP4 and ESF-EACBM Newsletter *26*.

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*, 2213-2224.

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.

Mack, D. R., Chiu, T. K., and Dickerson, R. E. (2001). Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. J Mol Biol *312*, 1037-1049.

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.

Mazza, C., Segref, A., Mattaj, I. W., and Cusack, S. (2002). Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. EMBO J *21*, 5548-5557.

Mol, C. D., Arvai, A. S., Slupphaug, G., Kavli, B., Alseth, I., Krokan, H. E., and Tainer, J. A. (1995). Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. Cell *80*, 869-878.

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*, 451-456.

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.

Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallographica D *53*, 240-255.

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

Perrakis, A., Harkiolaki, M., Wilson, K. S., and Lamzin, V. S. (2001). ARP/wARP and molecular replacement. Acta Crystallographica D *57*, 1445-1450.

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.

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*, 1417-1428.

Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V., and Feigon, J. (2004). DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. Proc Natl Acad Sci U S A *101*, 1177-1182.

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*, 327-338.

The human genome sequencing consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tsujishita, Y., Guo, S., Stolz, L. E., York, J. D., and Hurley, J. H. (2001). Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. Cell *105*, 379-389.

Vagin, A., and Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. J Appl Cryst *30*, 1022-1025.

Volff, J. N., Korting, C., and Schartl, M. (2000). Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol *17*, 1673-1684.

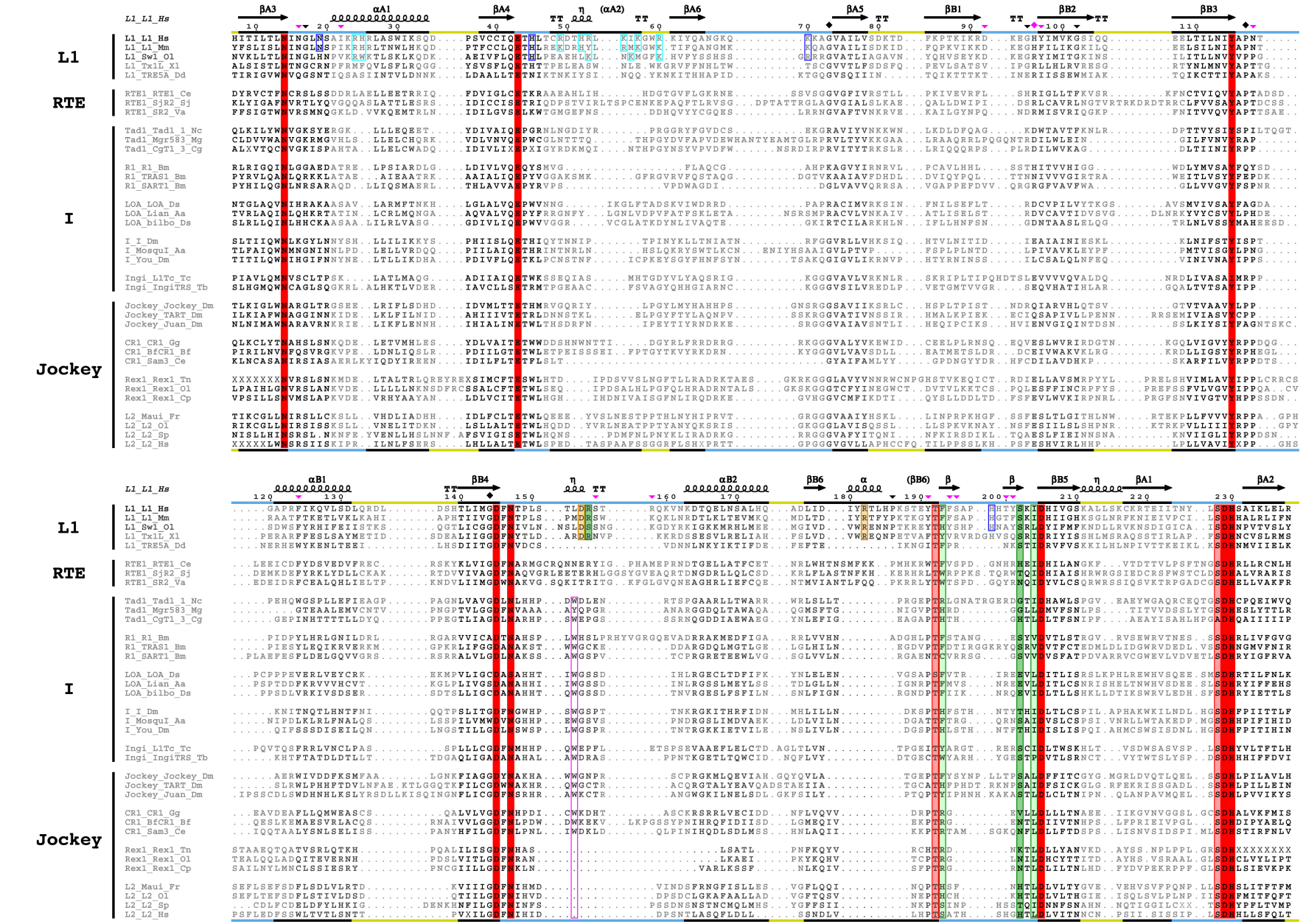
Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., Boeke, J. D., and Moran, J. V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol *21*, 1429-1439.

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*, 167-173.

Weston, S. A., Lahm, A., and Suck, D. (1992). X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. J Mol Biol *226*, 1237-1256.

Whisstock, J. C., Romero, S., Gurung, R., Nandurkar, H., Ooms, L. M., Bottomley, S. P., and Mitchell, C. A. (2000). The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis. J Biol Chem *275*, 37055-37061.

Zingler, N., Weichenrieder, O., and Schumann, G. G. (2004). APE-Type Non-LTR Retrotransposons: Determinants Involved in Target Site Recognition. Cytogenet Genome Res (in press).


L1
RTE
I
Jockey
L1
RTE
I
Jockey

Chapter 3

Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease

Kostas Repanas, Nora Zingler, Liliana E. Layer, Gerald G. Schumann, Anastassis Perrakis and Oliver Weichenrieder

Nucleic Acids Research, In Press

Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease

Kostas Repanas, Nora Zingler, Liliana E. Layer, Gerald G. Schumann, Anastassis Perrakis and Oliver Weichenrieder

ABSTRACT

The human LINE-1 endonuclease (L1-EN) is the targeting endonuclease encoded by the human LINE-1 (L1) retrotransposon. L1-EN guides the genomic integration of new L1 and Alu elements that presently account for ~28 % of the human genome. L1-EN bears considerable technological interest, because its target selectivity may ultimately be engineered to allow the site-specific integration of DNA into defined genomic locations. Based on the crystal structure, we generated L1-EN mutants to analyze and manipulate DNA target site recognition. Crystal structures and their dynamic and functional analysis show entire loop grafts to be feasible, resulting in altered specificity, while individual point mutations do not change the nicking pattern of L1-EN. Structural parameters of the DNA target seem more important for recognition than the nucleotide sequence, and nicking profiles on DNA oligonucleotides *in vitro* are less well defined than the respective integration site consensus *in vivo*. This suggests that additional factors other than the specificity of L1-EN are required for the targeted integration of non-LTR retrotransposons.

INTRODUCTION

In the higher eukaryotes frequently more than 90 % of the DNA does not code for functional proteins or RNA. Much of this DNA has originated from the action of mobile genetic elements, mostly retrotransposons that propagate in a copy-and-paste mechanism via an RNA intermediate. While these elements can be viewed as molecular parasites that are in an evolutionary race with their host genome, they can also be regarded as essential genomic components for slowly reproducing species to adapt to a changing environment. They generate allelic heterogeneity and create new possibilities for genetic recombination, increasing genomic fluidity (Bestor, 2003; Brosius, 2003; Eickbush and Malik, 2002; Han and Boeke, 2005; Kazazian, 2004).

Mobile genetic elements integrate into new genomic locations in two fundamentally different ways. DNA transposons and retrotransposons with long terminal repeats (LTR retrotransposons) use a transposase / integrase to insert a double-stranded DNA copy of the element at the target site. In this case no DNA synthesis takes place at the site of integration. In contrast, non-LTR retrotransposons use a mechanism called target-primed reverse transcription (Eickbush and Malik, 2002). This process is initiated by a targeting endonuclease, which specifically binds to the site of genomic integration. It nicks one strand of the DNA and creates a free 3' hydroxyl end, which is then used as a primer for reverse transcription of the retrotransposon RNA at the site of integration. Endonuclease and reverse transcriptase are two domains of a single retrotransposon-encoded protein. They are thought to rely on the assistance of 'host'-encoded proteins to complete the integration process (Cost et al., 2002; Gasior et al., 2006; Luan et al., 1993).

Most non-LTR retrotransposons are APE-type non-LTR retrotransposons (Zingler et al., 2005). Their targeting endonuclease belongs to a family of metal-dependent phosphohydrolases that includes nucleases like DNaseI (PDB-ID: 1dnk), APE1 (PDB-ID: 1dew), Exo III (PDB-ID: 1ako) and CdtB (PDB-ID: 1sr4) but also sugar phosphatases like I5PP (PDB-ID: 1i9z) and phospholipases like SmcL (PDB-ID: 1zwx) and Bc-SMase (PDB-ID: 2ddt). Members of this family share the same protein scaffold and the same catalytic residues, but a variation of the connecting surface loops has allowed them to develop quite diverse substrate specificities (Dlakic, 2000).

Under the pressure to survive in their respective host species non-LTR retrotransposons have evolved different strategies (Zingler et al., 2005). Stringent elements like R1Bm from *Bombyx mori* (Xiong and Eickbush, 1988) and Tx1L from *Xenopus laevis* (Garrett et al., 1989) encode highly specific targeting endonucleases (Christensen et al., 2000; Feng et al., 1998). They integrate into unique genomic locations (a specific sequence within 28S rDNA for R1Bm or within the apparent DNA transposon Tx1D for Tx1L, respectively) where they do very little or no damage to the host. Promiscuous elements like the human LINE-1 (L1) element (Dombroski et al., 1991) may integrate into several hundred thousand genomic locations. They have a rather short integration-site consensus (5'-TTTT/AA-3' for L1 (Gilbert et al., 2002; Symer et al., 2002; Szak et al., 2002)) that is nicked by the respective targeting endonuclease (Cost and Boeke, 1998; Feng et al., 1996). The host limits the spread of such elements by transcriptional and post-transcriptional silencing mechanisms that reduce activity to tolerable levels (Bogerd et al., 2006; Muckenfuss et al., 2006; Yang and Kazazian, 2006; Yoder et al., 1997).

Clearly, the respective endonucleases play a major role in target site selection (Christensen et al., 2000; Cost and Boeke, 1998; Feng et al., 1998; Takahashi and Fujiwara, 2002). The intriguing question of how different

targeting endonucleases recognize the DNA substrate and how easily new specificities can arise in the course of evolution remains open. There are indications that retrotransposons can evolve back and forth between a stringent and a promiscuous mode-of-action (Kojima and Fujiwara, 2003) and the ability to manipulate and design target specificity would be a crucial step in converting non-LTR retrotransposons into a genetic tool.

Previously, we described the crystal structure of the human L1 endonuclease (L1-EN) (Weichenrieder et al., 2004). Based on structure comparisons and sequence alignments we suggested that the prominent βB6-βB5 hairpin loop may insert into the DNA minor grove and may be particularly important for recognizing the DNA target. Here, we combine a mutational approach (specific point mutants and entire loop grafts) with structural and dynamic analyses. We determine minimal size and structural features of the DNA target and we show that size and flexibility of the βB6-βB5 hairpin loop are crucial for activity. Variation of the loop sequence results in an altered DNA nicking profile including novel sites. This indicates that the engineering of novel specificities may ultimately be feasible.

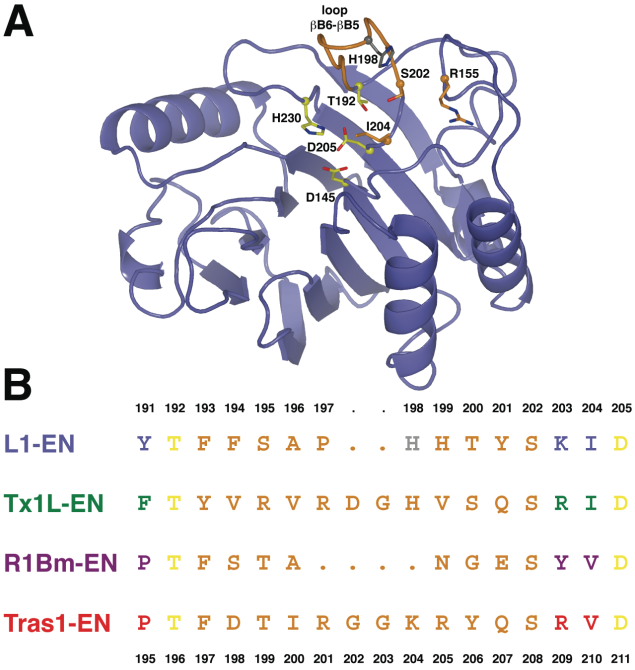


Figure 1: L1-EN point mutants and βB6-βB5 hairpin loop variants. A: Localization of the mutations on the crystal structure of L1-EN. The structure of L1-EN (Weichenrieder et al., 2004) is drawn as ribbons with the backbone of the exchanged loop in orange and with individual point mutants as balls-and-sticks. Yellow: conserved residues, orange: residues potentially contacting DNA, grey: H198. B: Structure-based alignment of the βB5-βB6 hairpin loop. For the chimeric endonucleases LTx and LR1 the respective loop sequences (orange) of Tx1L-EN and R1Bm-EN were grafted onto the L1-EN scaffold between the conserved anchoring residues T192 and S202. For L3G the loop was replaced by three glycines. The loop sequence of TRAS1-EN is shown for comparison. Numbering is from L1-EN (top, PDB-ID: 1vyb) and TRAS1-EN (bottom, PDB-ID: 1wdu) and color-coding of endonucleases is maintained throughout the paper.

RESULTS

The crystal structure of L1-EN suggests elements important for DNA target recognition but not for catalysis

We designed variants of L1-EN that fall into three categories (Figure 1). The first category includes point mutations (D145A, T192V, H230A) of catalytic and structurally important residues that are highly conserved within the entire enzyme family. The second category comprises point mutants (R155A, S202A, I204Y) of moderately conserved non-catalytic surface residues expected to affect the accommodation and recognition of the nucleotide downstream of the scissile bond (Figure 1A, B).

In the third category of L1-EN variants we manipulated the βB6-βB5 hairpin loop, which is positioned to insert into the DNA minor groove with the possibility to read out both sequence and structural parameters (Weichenrieder et al., 2004). It is well suited for a loop-grafting experiment because the anchoring residues T192 and S202 on either side are well conserved among many metal-dependent phosphohydrolases. Therefore, we replaced the entire βB6-βB5 hairpin loop of L1-EN with the corresponding sequences from the R1Bm and Tx1L retrotransposons (Figure 1B). The resulting mutants LR1 and LTx, respectively, were complemented by the loop deletion variant L3G, where we exchanged the entire loop (including S202) for a linker of three glycines.

L1-EN point mutations and loop grafts affect retrotransposition in cell culture

Initially, the L1-EN variants were tested in the context of a functional, tagged L1 element in a well-established cell culture assay (Moran et al., 1996; Wei et al., 2000; Gilbert et al., 2002). We scored successful retrotransposition events by the appearance of neomycin-resistant HeLa cell colonies, subtracting background activity caused by trans complementation or endonuclease-independent retrotransposition.

All variants reduce the frequency of retrotransposition significantly, confirming the relevance of the mutated elements (Table I). The strongest effects are seen with point mutants D145A, T192V, I204Y and H230A and with loop variants LR1 and L3G. To test whether this is directly related to the ability of the enzyme to recognize and nick target DNA we purified the respective L1-EN variants for assays in vitro.

The ability of L1-EN variants to nick plasmid DNA correlates well with the frequency of retrotransposition

Residues T192 and H230 are hydrogen-bonded via D205 (Weichenrieder et al., 2004). These interactions

are apparently essential for the structural integrity of L1-EN as the respective mutants were inherently unstable, degraded easily or precipitated rapidly. From the first category only the D145A mutant could be purified as a negative control for catalytic activity.

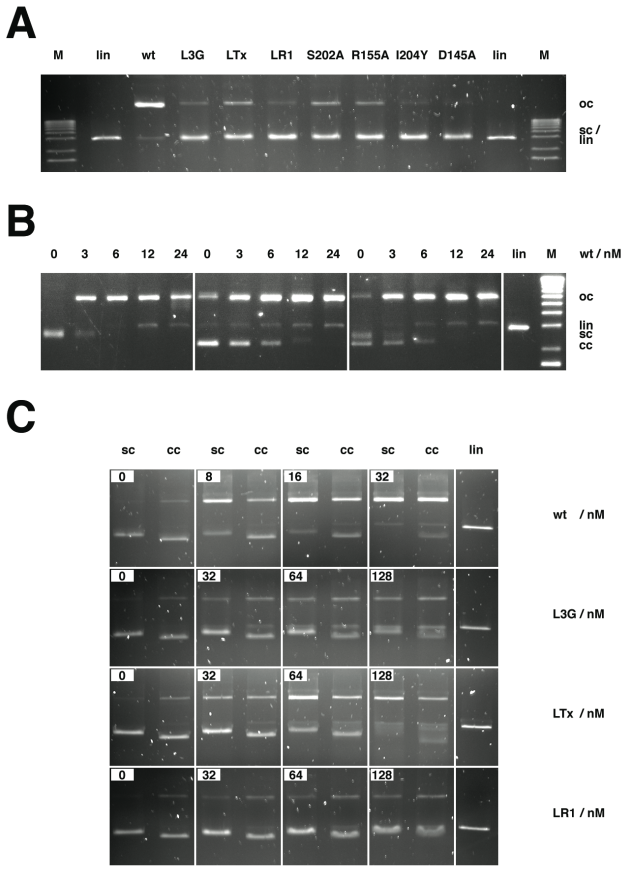


Figure 2: Plasmid nicking activity of L1-EN variants. Experiments were done with wildtype L1-EN (wt), βB5-βB6 hairpin loop variants (L3G, LTx, LR1) and point mutants (S202A, R155A, I204Y, D145A). Supercoiled (sc) plasmid DNA (pBluescript) or relaxed closed circle DNA (cc) was converted into the open circle form (oc) and into linear DNA (lin). Closed circle DNA contains trace amounts of dimer, which runs like open circle DNA both on 1.4 % agarose gels (A, C) and 1.0 % gels (B). M: DNA size marker X (Roche). A: Relative activity of L1-EN mutants (32 nM) on supercoiled plasmid DNA (2 nM). B: Preference of L1-EN for supercoiled target DNA. Supercoiled and closed circle DNA were nicked by increasing concentrations of L1-EN, either separately (2 nM) or in competition (1 nM each). C: Titration of L1-EN hairpin loop variants and selectivity for sc DNA. LR1 and L3G are less active than LTx and show no preference for supercoiled DNA.

The purified L1-EN variants were first analyzed in a plasmid DNA nicking assay (Feng et al., 1996), where supercoiled plasmid is converted into the open circle form that runs considerably slower on an agarose gel (Figure 2). Figure 2A shows a side-by-side comparison of the activities of all L1-EN mutants (32 nM) on 2 nM supercoiled plasmid. Under these conditions wild-type L1-EN converts more than 90 % of plasmid DNA into the open circle form. The three point mutants, R155A, S202A and I204Y, show strongly reduced activity, with S202A being affected the least and I204Y the most. The strong effect of I204Y suggests that L1-EN probably binds double-stranded DNA in an orientation that differs from the one seen in the complex with DNaseI (Suck et al., 1988), because in DNaseI the tyrosine is present and tolerated at this position. This view is supported by the effects of S202A and R155A, which indicate that these moderately conserved amino acids are indeed involved in contacting the nucleotide(s) downstream of the scissile bond, either specifically or non-specifically. For a direct contact with R155A the downstream DNA would have to be distorted or even flipped as in the complex with APE1 (Mol et al., 2000). Among the loop variants, LTx remains most active, at levels similar to the S202A point mutant. In contrast, LR1 and L3G retain little but still detectable activity.

Table I: Comparison of retrotransposition frequencies in vivo and plasmid nicking activities in vitro

L1-EN variant	Retro-transposition frequency [a], %	Plasmid nicking activity [b]
wt	100 ± 17.1	+++
LTx	21 ± 2.4	++
LR1	2 ± 2.3	+
L3G	0 ± 2.2	+
D145A	0 [c]	o
R155A	12 ± 3.3	++
T192V	5 ± 3.0	-
S202A	32 ± 7.8	++
I204Y	1 ± 1.1	+
H230A	0	-

[a] corrected for background activity (≤5 %); for details see supplementary information

[b] scored with respect to L1-EN (wt) according to Figure 2A: (+++) 50-100 %, (++) 10-50 %, (+) ≤10 %, (o) not detectable, (-) not analyzed

[c] as a D145A/N147A double mutant

The structural context of the DNA target is important for its recognition by L1-EN (Cost and Boeke, 1998). When presented with equal amounts of supercoiled and of relaxed, closed circle pBluescript DNA, L1-EN nicks the supercoiled DNA much more efficiently (Figure 2B). Since the βB6-βB5 hairpin loop may well be involved in the recognition of an unusual DNA structure caused by supercoiling, we tested the L1-EN loop variants also in this respect (Figure 2C). While LTx still prefers supercoiled DNA, the very inefficient LR1 shows no detectable preference for supercoiled DNA anymore. The same observation holds true for L3G, where the loop is deleted. This experiment shows that the βB6-βB5 hairpin loop of L1-EN may be particularly important for reading out the structural context of a potential new retrotransposon integration site.

Finally, there is a good correlation between the nicking activities *in vitro* and the retrotransposition frequencies *in vivo*, indicating that the activity of the endonuclease is limiting over a considerable range (Table I). Consequently, alterations in the nicking specificity of the endonuclease should lead to changes in integration specificity. To distinguish whether our mutations simply impair catalysis or indeed alter target recognition we veri-

fied *in vitro* if and how nicking specificities were affected.

Efficient DNA nicking by L1-EN requires a minimum of five base-pairs upstream and three base-pairs downstream of the target site

Genomic L1 pre-integration sites have been analyzed statistically and a consensus sequence has been reconstructed. In the 5' to 3' direction the substrate strand consists of an upstream tract of four to five strongly conserved thymidines (T-tract) followed downstream by two more moderately conserved adenines, with the integration occurring at the poly(T)-A junction (Gilbert et al., 2002; Symer et al., 2002; Szak et al., 2002). In contrast to previous approaches (Cost and Boeke, 1998) we chose this type of asymmetric target for a DNA oligonucleotide nicking assay (Figure 3).

We designed a DNA duplex consisting of 14 T-A pairs, followed by two A-T pairs and a single clamp of four C-G pairs (Figure 3A (Cwt)). We find the 5' labeled substrate strand (the bottom strand in all figures) to be nicked throughout the entire T-tract with very similar relative frequencies and only the first five thymidines are spared. Nicking at the poly(T)-A junction is enhanced not more than 4 to 5 fold (Fig. 3B (wt)). As shown previously (Cost and Boeke, 1998), the observed nicking patterns result from multiple independent endonucleolytic nicking events and not from a cryptic 3' to 5' exonuclease activity of L1-EN.

For a closer analysis of the DNA structural parameters required for efficient nicking we manipulated the complementary DNA strand (upper strand in all figures). A mismatched adenine (A:C) in position (+1) immediately downstream of the target site diminishes the preference for the poly(T)-A junction, reducing it to the levels observed for nicking within the T-tract (Figure 3A, C (Cim)). This suggests, that at least during the initial step of recognition of a poly(T)-A junction by L1-EN, this nucleotide position needs to be base-paired properly with an unobstructed minor groove. Mismatching the complete remainder of downstream DNA in addition to position (+1) does not cause any further reduction of nicking efficiency at the poly(T)-A junction (Figure 3A, C (C56m)). Next, we tested to which degree the complementary strand is required downstream of the target site by deleting an increasing number of nucleotides from the 5' end. The results show that the complementary strand needs to extend downstream by at least one nucleotide. However, for nicking at the poly(T)-A junction to be preferred over the adjacent T-tract, at least three downstream base-pairs are required (Figure 3A, C (C53-, C54-, C55- and C56-)). Upstream of the target site, L1-EN prefers at least five nucleotides to be base-paired. If this is not the case nicking is significantly reduced (Figure 3A, C (C35-)). In summary (Figure 3D), our data suggest that preferential recognition of a poly(T)-A junction by L1-EN requires five nucleotides upstream that should be base-paired at least close to the target site and three base-pairs downstream which are just sufficient to form a short independent stem that does not need to stack on the upstream duplex for stability. Thus, the minor groove at the poly(T)-A junction would be flexible and could easily be widened by external strain on the DNA or simply by the insertion of the βB6-βB5 hairpin loop, pushing the downstream DNA into a position to be contacted by S202 and R155.

The protruding hairpin loop of L1-EN is crucial for recognition of the DNA target structure

The relative enzymatic activities of the three L1-EN loop variants are similar in the plasmid DNA nicking and duplex DNA nicking assays with LTx being the most active and LR1 being the least active (Figure 2C, 3B). LTx still nicks T-tract DNA, but the specificity for the poly(T)-A junction has disappeared. We conclude that the βB6-βB5 loop of LTx is less well suited to recognize a poly(T)-A junction, although it does functionally replace the βB6-βB5 loop of L1-EN to a large degree. In sharp contrast, LR1 and L3G do not show any significant endonucleolytic activity, even at the highest concentrations (Figure 3B).

To extend the analysis of the respective nicking profiles we designed long DNA oligonucleotides (Dwt and Dhy) with more sequence variation (Figure 4). Dwt contains the genomic target sequences of human L1-EN, Tx1L-EN and R1Bm-EN on a single DNA duplex. This design assures that the potential target sites are present at equal concentrations and compete for the respective endonuclease under identical conditions. Dhy is identical, except that the upstream sequences of the respective target sites have been replaced by T-tracts (Figure 4A). In addition to L1-EN, we also used wild-type Tx1L-EN (wTx) as a positive control in this assay (Christensen et al., 2000). Figure 4B demonstrates the difference in nicking specificity between the sequence-specific Tx1L-EN and the promiscuous L1-EN. Tx1L-EN nicks almost exclusively at the expected target site, after nucleotide 29 on Dwt and to a lesser extent on the corresponding hybrid site on Dhy. L1-EN nicks preferentially at the poly(T)-A junctions on Dwt or Dhy, but also within extended T-tracts and non-canonical sequences like after nucleotides 19 and 20. This gives rise to characteristic and reproducible nicking profiles (Figure 4C).

The nicking profile of the LTx chimera is different from both L1-EN and Tx1L-EN. LTx does not nick after nucleotide 29 on Dwt, the preferred site for Tx1L-EN. On Dhy there is no specific nicking at this position either, despite the upstream T-tract that was introduced and expected to fit the L1-EN scaffold (Figure 4B). This suggests that one cannot simply combine and exchange upstream (L1-EN scaffold) and downstream (βB6-βB5 loop) recognition elements in a modular fashion to generate a desired target specificity.

Nevertheless there are novel nicking sites for LTx. The most prominent of those is after nucleotide 23 on Dwt, a site that is nicked neither by L1-EN nor by Tx1L-EN (Figure 4 A, B). The downstream sequence (5' AGCT 3') of this novel site is very similar to the sequence (5' AGTT 3') downstream of nucleotide 29, the site nicked by Tx1L-EN.

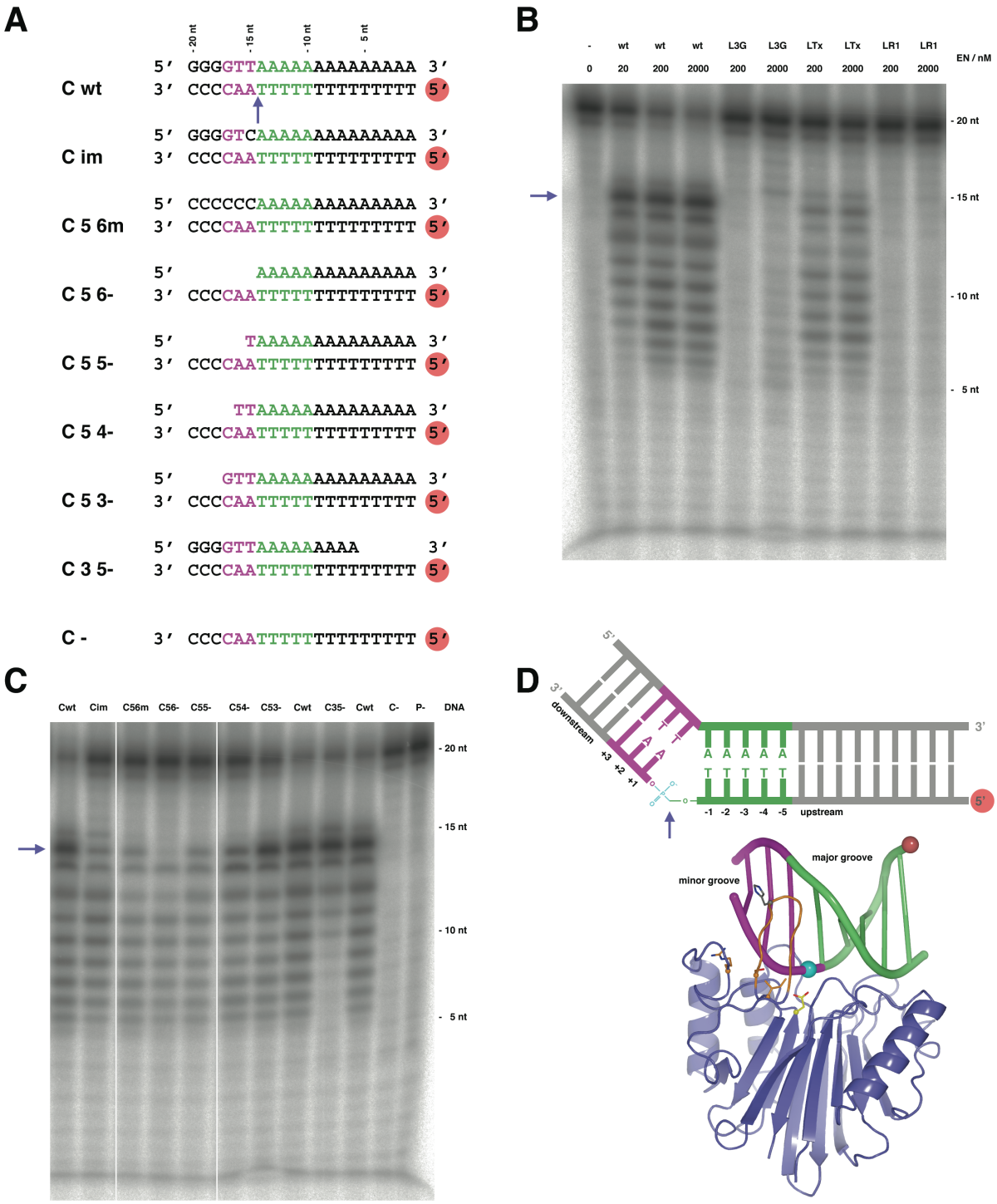


Figure 3: Characterization of the DNA target of L1-EN. A: DNA substrate duplexes containing mismatches or single-strand deletions of the complementary (top) strand. Three base-pairs (magenta) downstream of the poly(T)-A junction (blue arrow) and five base-pairs (lime) upstream are highlighted. Red Circle: 5' end labeled. B: Activity and target recognition of L1-EN hairpin loop variants. DNA duplexes (Cwt, 180 nM) were titrated with increasing concentrations of L1-EN and L1-EN hairpin loop variants. Products were analyzed on autoradiographs of denaturing polyacrylamide gels. Blue arrow: poly(T)-A junction. C: Substrate requirements of L1-EN. L1-EN (160 nM) was used to nick substrates from (A). Lane (P-): Cwt without L1-EN protein. D: Model for DNA target recognition by L1-EN. Top: Scheme of target DNA including the consensus L1 integration sequence. Bottom: Three-dimensional model adapted from (Weichenrieder et al., 2004) with L1-EN represented as in Figure 1. The upstream DNA duplex (T-tract geometry, lime) is thought to be contacted by the L1-EN protein scaffold, while the orientation and flexibility of the downstream DNA duplex (magenta) is probed by the insertion of the βB5-βB6 hairpin loop (orange) into the widened minor groove at the poly(T)-A junction.

The structural context upstream of nucleotide 23, however, is different and appears to be suited much better to support recognition and nicking by LTx. On Dhy the sequence downstream of nucleotide 23 is exchanged (5' TTTT 3') and nicking by LTx is negligible. Apparently, the LTx βB6-βB5 hairpin loop plays a dominant role for the recognition of downstream DNA. This interpretation is supported by the general nicking profile of LTx (Figure 4C), suggesting that novel target sites can be engineered indeed.

In clear contrast to LTx the βB6-βB5 loop of LR1 cannot functionally replace the βB6-βB5 loop of L1-EN, as replacement results in a low nicking activity. With respect to specificity, LR1 rather seems to avoid T-tracts and produces a very distinct nicking pattern that is quite similar to the one from the loop deletion variant L3G (Figure 4B, Supplementary Figure 1). The prominent nick of LR1 on Dwt after nucleotide 14 (the preferred nicking site for R1Bm-EN (Feng et al., 1998)) does not seem to be specific, since it is present also with L3G (not shown), L1-EN (Figure 4B) and other variants (Supplementary Figure 1).

Loop grafting has thus produced chimeric endonucleases with altered and novel nicking preferences. In contrast, all analyzed point mutants display nicking patterns that are identical to those of L1-EN (Supplementary Figure 1). They lose activity to various degrees, but maintain specificity. Contrary to the hairpin loop these residues seem to play a rather passive role in contacting an unusual or bendable DNA structure and they might need to be replaced simultaneously to cause any significant effect on nicking specificity.

Requirements for genomic integration of L1 elements are more stringent than requirements to nick target DNA

To test whether the altered nicking specificity of the LTx endonuclease is reflected by an altered integration site preference of the respective L1 variant we determined the genomic pre-integration sequences from several neomycin-resistant HeLa cell clones obtained in the cell culture assay (Gilbert et al., 2002). Comparison of the *in vitro* nicking profiles to the integration site consensus sequences confirms that for the wildtype L1 element, the nicking specificity of the endonuclease and integration site selection match. However, in the case of LTx, they differ significantly. Like L1, the chimeric LTx element prefers to integrate into locations with a T-tract upstream of the nicking site and only a subset of nicking sites appears to be used for integration (Figure 4C, D). This points to additional requirements for targeting. These might be the rigidity of the T-tract that could play a more important role for target recognition or subsequent integration steps *in vivo* or rather also other constraints such as the need for base-pairing between the 3' ends of retrotransposon RNA and target site DNA.

Loop grafting results in beta-hairpin loops of similar orientation and does not perturb the rest of the L1-EN structure

The distinct effects of the exchanged βB6-βB5 hairpin loop sequences on DNA target recognition and hence nicking specificity are intriguing and may largely relate to the respective structures (Figure 5). We therefore determined the crystal structures of LTx (Figure 5C, D) and LR1 (Figure 5E, F) at 2.3 Å and 1.8 Å resolution, respectively (Table II) and compared them to the existing structure of L1-EN (Figure 5A, B) (Weichenrieder et al., 2004). According to an analysis with the program ESCET (Schneider, 2000), the common scaffold and catalytic center of the three enzyme variants are essentially unchanged (see also Figure 5G, H), despite some variance in the crystal packing. The exchanged βB6-βB5 loop sequences are well-ordered in both variants, forming protruding beta-hairpins as in wildtype L1-EN, and their orientation is similar.

The backbone of the LR1 hairpin loop superimposes well onto the backbone of the L1-EN hairpin loop (Figure 5G, H). Since the βB6-βB5 hairpin loop of LR1 is two amino acids shorter it lacks the tip (P197 and H198 of L1-EN) that bends towards the minor groove of a putative DNA substrate (Figure 3D). Furthermore, residue T200 of L1-EN is replaced by a glycine in LR1, eliminating an additional possibility of LR1 to interact with the substrate and finally, the LR1 hairpin loop lacks the positive charges of the L1-EN and LTx hairpin loops that might mediate initial contacts with the negatively charged DNA backbone (Figure 5F). The backbone of the LTx hairpin loop is twisted slightly with respect to the βB6-βB5 hairpin loop of L1-EN, especially at the distal end (Figure 5G, H). There, the RDGH sequence of Tx1L-EN (Figure 1B) replaces P197 and H198 of L1-EN, forming a more extended tip with side chains that could all make favorable DNA contacts (Figure 5D).

The structure of the catalytic center is not perturbed by the exchange of the loop sequence, suggesting that the mechanism of phosphodiester hydrolysis is not affected directly. It therefore seems likely that certain properties of the βB6-βB5 hairpin loop itself are causing the observed differences in activity and specificity. As the largest structural differences between the three loop variants are at the tip of the loop, we created a point mutation (H198A) in this region of L1-EN. However, the mutation only reduces activity, but does not change target specificity (Supplementary figure 1). Together with the observation that the chimeric LTx still nicks T-tract DNA despite an entirely different loop sequence this argues against the requirement of sequence-specific protein-DNA contacts. The initial affinity between L1-EN and its target may therefore be based on passive, non-specific contacts resulting from simple complementarity between the shapes of the βB6-βB5 hairpin loop of the endonuclease and the minor groove of the DNA. According to this model, the LR1 hairpin loop is just too short to reach the minor groove properly, explaining why the nicking pattern resembles that of L3G, where the loop is missing entirely.

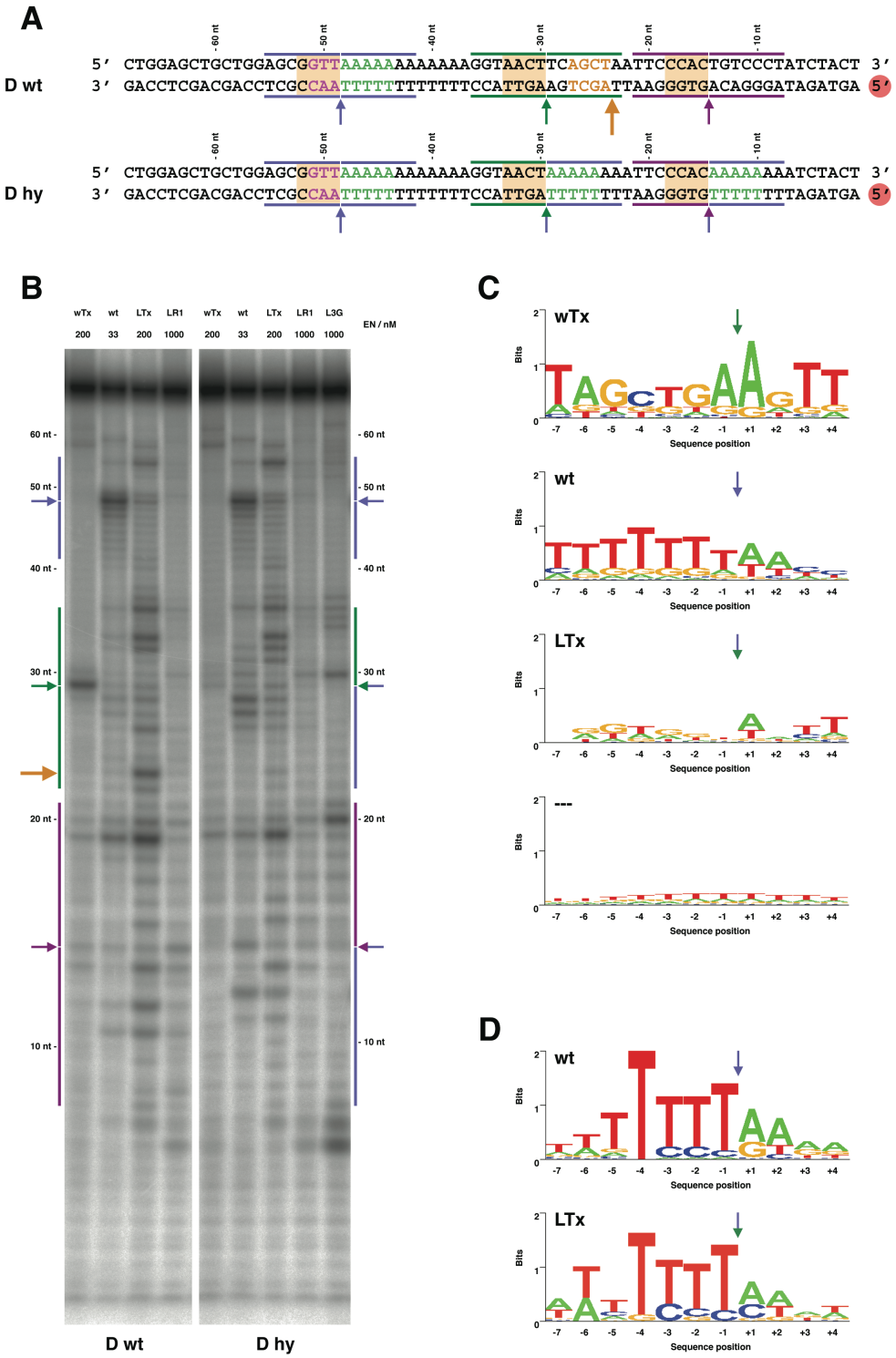


Figure 4: Target specificity of L1-EN mutants. A: DNA multi-substrate duplexes. Dwt contains wild-type target sites (arrows with seven flanking nucleotides marked by horizontal lines) for L1-EN (blue), Tx1L-EN (dark green) and R1Bm-EN (purple). Dhy contains hybrid target sites designed for nicking by LTx (dark green/blue) and LR1 (purple/blue), where seven upstream base pairs of the ideal target sites of Tx1L-EN and R1Bm-EN are replaced by a T-tract. Up/downstream base pairs important for recognition by scaffold and βB5-βB6 hairpin loop of L1-EN are lime and magenta, respectively. Nucleotides on the marked target sites thought to be in the reach of the various βB5-βB6 hairpin loops are on an orange background. The major novel target site of LTx on Dwt is marked by an orange arrow with the downstream nucleotides highlighted in orange. Red Circle: 5' end labeled. B: Specificity of L1-EN βB5-βB6 hairpin loop variants. DNA duplexes (180 nM) were nicked by the indicated amounts of endonuclease. Products were analyzed on autoradiographs of denaturing polyacrylamide gels. Colors and symbols as in (A). C: Sequence logos representing nicking profiles. (---), hypothetical logo obtained by assuming random nicking of Dwt. D: Sequence logos representing genomic pre-integration site consensus sequences. Top, n = 35, from (Gilbert et al., 2002). Bottom, n = 14. For details see Supplementary Table I.

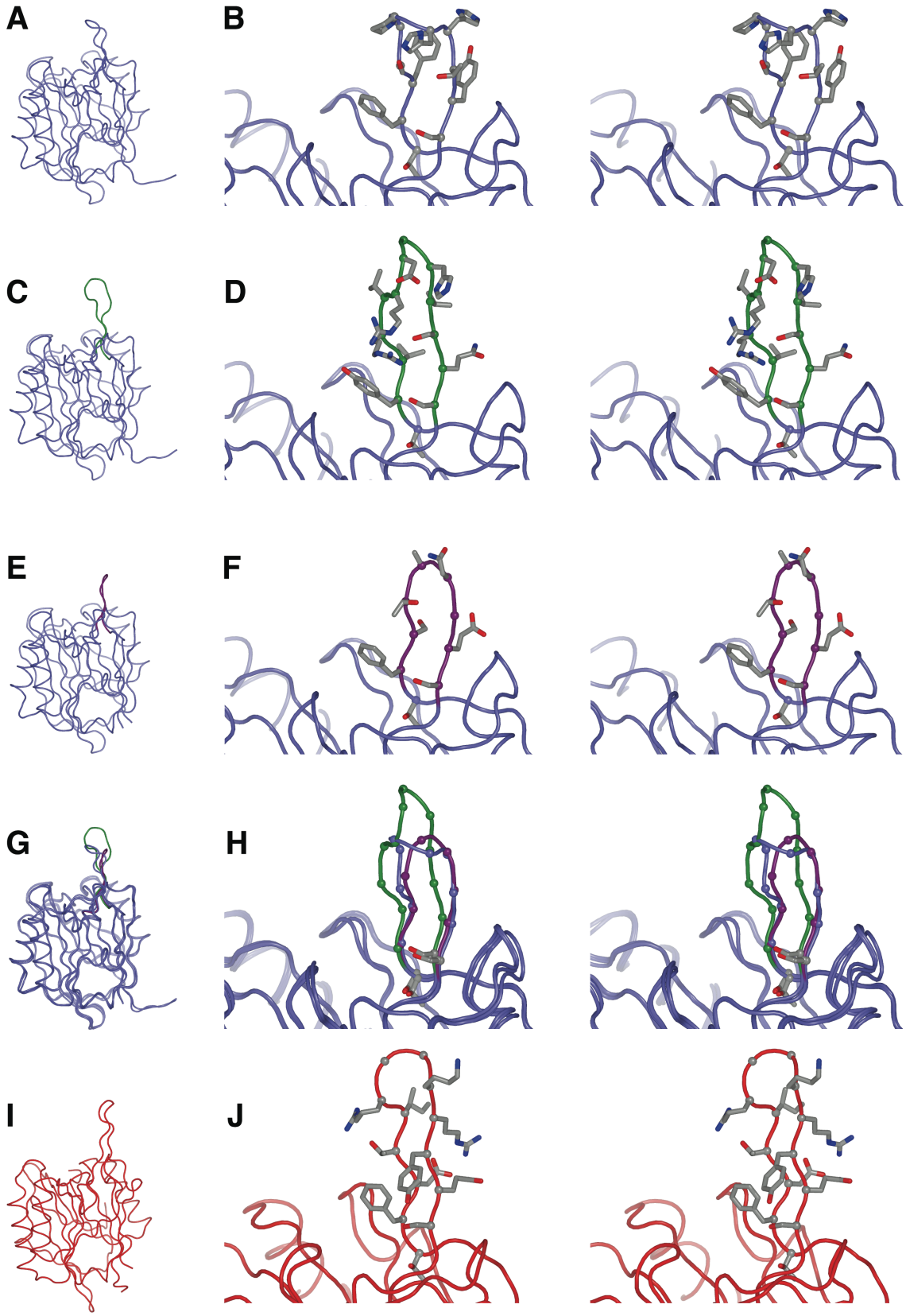


Figure 5. Crystal structures of L1-EN βB5-βB6 hairpin loop variants. The structures of L1-EN and of the two chimeras LTx and LR1 are compared to each other and to the structure of TRAS1-EN. A, B: L1-EN (blue). C, D: LTx (dark green/blue). E, F: LR1 (purple/blue). G, H: Superposition of (A), (B) and (C) illustrating differences in size and orientation of the βB5-βB6 hairpin loop. I, J: TRAS1-EN (red). Structures are represented as tubes and seen from the side (A, C, E, G, I) or from the front (stereo) zooming in on the loop region (B, D, F, H, J). Side chains of the hairpin loops are shown as balls-and-sticks with carbons in grey, oxygens in red and nitrogens in blue. In (H) side-chains are omitted apart from T192 and S202.

An additional property of the βB6-βB5 hairpin loop that may be relevant for target selectivity and that would not become obvious from a static crystal structure is its dynamic behavior in the course of the catalytic nicking cycle. In a crucial initiation step a flexible βB6-βB5 hairpin loop may be needed to probe the dynamics of the minor groove at the junction of the two non-stacking DNA stems.

Normal mode analysis indicates different flexibilities of the grafted hairpin loops

Normal Mode Analysis (NMA) is a powerful molecular modeling approach that is particularly suited for calculating slow, large-scale movements within proteins, which would be too expensive computationally for full-scale molecular dynamics simulations. We used the web-based server WEBnm@ (Hollup et al., 2005) to analyze the C-alpha chains of L1-EN, LTx and LR1. As an additional reference we included TRAS1-EN, which is encoded by the telomere-specific APE-type retrotransposon TRAS1 from *Bombyx mori*. Its structure (Figure 5I, J) is characterized by a βB6-βB5 beta-hairpin loop that, like Tx1L-EN, contains eleven residues (Maita et al., 2004). We calculated the respective average deformation energies of the lowest vibrational mode and also plotted the normalized squared atomic displacements along the sequence of each protein (Supplementary Figure 2).

Low deformation energies indicate that large regions of the protein, possibly domains, can be displaced. For the relatively inactive LR1 we obtain the highest deformation energy (4345), which decreases with increasing loop size via L1-EN (1290) to LTx (684) and TRAS1-EN (510). Furthermore, we clearly identify the βB6-βB5 hairpin loop as the most flexible region in each protein, with a big difference in the extent of the atomic displacement between LR1 and the other three proteins.

Taken together, these calculations suggest that an additional reason for the low activity and altered specificity of LR1 in our assays is the missing flexibility of the hairpin loop that is potentially required during the catalytic cycle to lock the DNA target in a suitable position for effective binding and subsequent hydrolysis of the phosphodiester bond.

Table II: Data collection and refinement statistics

Data Collection	**LTx**	**LR1**
Resolution, Å	2.3	1.8
Space group	$P2_12_12_1$	$C222_1$
Cell dimensions, Å	a=54.7, b=70.1, c=130.2	a=58.6, b=67.6, c=128.3
R_{merge}, % [a]	11.2 (48.8)	7.8 (44.2)
Completeness, % [a]	99.8 (100.0)	96.4 (98.6)
I/σ(I) [a]	8.8 (2.3)	11.9 (2.7)
No. of reflections		
unique observed	22873	23062
total measured	79345	88132

Refinement	**LTx**	**LR1**
R_{cryst}, %	21.6	18.5
R_{free}, %	26.8	22.2
Number of		
molecules in asymmetric unit	2	1
atoms	3948	2039
ions	6	3
glycerol molecules	-	1
water molecules	159	185
Ramachandran plot		
Most favored regions, %	88.5	91
Allowed regions, %	10.2	8.5
Generously allowed regions, %	1.4	0.5
R.m.s.d. from ideal geometry		
Bond lengths, Å	0.018	0.013
Bond angles, °	1.81	1.4

[a] Values in parentheses correspond to those in the outer resolution shell (1.89-1.8 Å and 2.4-2.3 Å for LR1 and LTx, respectively)

DISCUSSION

The structural context of the DNA target is highly important for efficient nicking by L1-EN

DNA target specificity of L1-EN has been studied before with plasmid DNA (Feng et al., 1996) and with special DNA duplexes that contained a symmetric junction of two T-tracts (Cost and Boeke, 1998). The present study confirms such junctions to be ideal nicking substrates for L1-EN and corroborates the importance of the DNA structure for molecular recognition. We extend the previous analyses to asymmetric DNA targets and determine minimal substrate requirements for the flanking upstream and downstream sequences. Furthermore, we look at the nicking specificity of L1-EN on more general DNA substrates and compare it to the integration specificity of L1 elements *in vivo*.

We find that with unstrained duplex DNA, L1-EN requires a minimum of five base-pairs upstream and three base-pairs downstream of the target site for efficient target recognition. On the upstream duplex L1-EN recognizes mainly the T-tract (A-tract) geometry (see also (Cost and Boeke, 1998)) that is primarily characterized by its very narrow minor groove (Stefl et al., 2004). Downstream, the three base-pairs are just enough to form an independent stem. In the case of a T-A junction following the T-tract (poly(T)-A junction), the downstream adenine is not stacked on the upstream thymidine (Stefl et al., 2004) and thus, the downstream stem can more easily be bent away with an associated widening of the minor groove. Most likely, this local flexibility is a feature that is recognized by L1-EN in addition to the narrow minor groove of the T-tract, leading to the enhanced nicking efficiency observed at the junction. On a strained substrate such as supercoiled plasmid DNA, the difference between cleaving T-tract DNA and a poly(T)-A junction would probably be even more pronounced. The torsional strain might widen the minor groove at the junction even further and facilitate the structural recognition of the DNA target.

Although the structure of L1-EN would allow the accommodation of a flipped nucleotide at position (+1) downstream of the scissile bond (Weichenrieder et al., 2004), we do not find any evidence for the base-specific recognition of such a nucleotide. At least for the initial target recognition the nucleotide needs to be part of a downstream stem. However, this does not rule out the possibility that the flexibility (or 'flippability') of the nucleotide is required in consecutive steps of the integration process.

L1 integration specificity is influenced by additional factors

In conclusion, L1-EN recognizes structural features of the DNA target rather than specific nucleotides in the sequence. The 5' TTTT/AA 3' integration site consensus sequence may fulfill these structural requirements in an ideal way, but many alternative sequences seem to have similar structural features and are nicked *in vitro*. The requirements for integration seem stricter than the requirements for nicking, indicating that the endonuclease may not be the only component determining integration site selection (Zingler et al., 2005). Additional specificity factors could influence the choice of nicking site in the first place (co-targeting factors) or select among already nicked sites the ones that are suitable for integration (post-nicking factors). The latter possibility is favored by reports of endonuclease-independent retrotransposition (Morrish et al., 2002) and L1-induced chromosomal breaks (Gasior et al., 2006).

Structure and dynamics of the βB6-βB5 beta-hairpin loop are more important for activity and specificity of L1-EN than sequence

During DNA target site recognition, the conformational space available to the downstream DNA duplex is probed by the insertion of the βB6-βB5 beta-hairpin loop of L1-EN into the minor groove at a poly(T)-A junction, according to the presented model (Figure 3D). The presence of the loop is important for nicking activity and both nicking activity and target specificity are very sensitive to structural changes of the loop, especially at its tip. Similar to the situation in TRAS1-EN (Maita et al., 2004) a deletion of the tip (LR1) or of the entire loop (L3G) results in an altered specificity and much reduced activity. To examine the importance of the amino acid sequence we exchanged residue H198 in the tip of the loop, which had no impact on the nicking pattern. Even the substitution of the entire loop with a different sequence and an extended reverse turn (LTx) was tolerated rather well. This suggests that the conformational flexibility of the beta-hairpin loop probing the DNA minor groove may be much more important than its sequence, especially if target recognition proceeds via the structural flexibility of the DNA at the poly(T)-A junction. This hypothesis is supported by the presented normal mode analysis. The βB6-βB5 hairpin loop of LTx may be able to functionally replace the βB6-βB5 hairpin loop of L1-EN because it is flexible enough to insert partially into the minor groove of many L1-EN targets to probe the conformational space of the downstream duplex. The βB6-βB5 hairpin loop of LR1 may be too rigid for this function. In its natural context on R1Bm-EN (Feng et al., 1998) it may only be required as a counter bearing for the target DNA, which would then be probed sequence-specifically from the side of the major groove by an extension of surface loop βB4-αB2, that is unique to R1Bm-EN (see (Weichenrieder et al., 2004) for alignment).

Can novel integration specificities be engineered ?

The L1 retrotransposon bears considerable potential as a genetic tool (Ostertag and Kazazian, 2001). It can be delivered to cells by an adenovirus vector (Soifer and Kasahara, 2004) and its suitability for *in vivo* mutagenesis has recently been demonstrated with a synthetic, highly active mouse L1 element called ORF*eus* (An et al., 2006). The application of similar L1 retrotransposons for gene delivery into defined genomic locations requires engineering of the endonuclease target specificity as one of the most crucial steps. This appears feasible since

there are many natural APE-type non-LTR retrotransposon endonucleases with distinct target specificities that all share the same protein scaffold and the same catalytic site (Weichenrieder et al., 2004; Zingler et al., 2005).

Loop grafting experiments have been shown to mimic evolutionary processes (Aharoni et al., 2005), allowing novel specificities to be engineered (Jones et al., 1986; Park et al., 2006). The analysis of the presented L1-EN βB6-βB5 hairpin loop variants shows that the respective grafting experiments worked successfully from a structural point of view and that other surface loops may be manipulated in a similar way in the future. From a functional point of view, we could show that the DNA nicking profile of L1-EN is quite sensitive to structural changes of the studied loop and that novel specificities can indeed be acquired. For further improvements high-resolution structures of retrotransposon endonucleases in complex with their respective DNA targets would be of great help.

Finally, the apparent existence of additional targeting factors poses a further challenge for the engineering of novel integration specificities. One such factor may be the need for complementary bases between the 3' end of retrotransposon RNA and the 3' end of nicked genomic DNA. Tools like the LTx variant will allow us to investigate such requirements in the future.

MATERIALS AND METHODS

Preparation and purification of L1-EN variants

Point mutants and loop variants of L1-EN were generated in the context of the retrotransposition reporter plasmid pCEP4/L1.3/ColE1/*mneoI*$_{400}$ (Gilbert et al., 2002) as described in the supplementary information. For the overexpression in *Escherichia coli* Rosetta II cells (Novagen) of mutated L1-EN domains with N-terminal poly-histidine tags DNA corresponding to residues 1 – 239 of wild-type L1-EN was PCR-amplified (Weichenrieder et al., 2004) and inserted into the *Nco*I/*Xho*I cloning sites of expression plasmid pETM11 (Zou et al., 2003). Proteins were purified over Ni-chelating chromatography and heparin affinity columns and quantified spectroscopically or on denaturing SDS polyacrylamide gels. For Tx1L-EN protein (residues 1-239) DNA was amplified from plasmid pE1EN (Christensen et al., 2000) using primers Tx1L-EN-N1 and Tx1L-EN-C239. For crystallization the respective proteins were expressed and purified without tag as described in (Weichenrieder et al., 2004). Purified protein (> 1 mg/ml) was stored frozen at -80 ºC at NaCl concentrations above 300 mM.

Retrotransposition reporter assay

Retrotransposition frequencies of wild-type and mutant L1 constructs were determined by applying the rapid, quantitative transient L1 retrotransposition assay described previously (Wei et al., 2000). HeLa cells (2 x 10^5) were plated in each well of a six-well dish and grown to 50-80% confluency in DMEM. The following day, triplicate dishes were transfected using 6 µl Fugene-6 transfection reagent (Roche) and 2 µg of a Qiagen plasmid 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. After 10-14 days, G418R colonies were stained with Giemsa solution and counted. The recovery of integrated L1 elements for sequencing is described in (Gilbert et al., 2002).

Plasmid nicking

Supercoiled pBluescript plasmid DNA was prepared from *E. coli* DH5α cells. Closed circle plasmid DNA was obtained by simultaneous digestion and religation of supercoiled DNA (15 µg/ml) with 5 U/ml *Hin*d III and 900 U/ml T4 DNA ligase resulting in only trace amounts of dimeric product. DNA was quantified after linearization on agarose gels containing ethidium bromide. Nicking reactions (10 or 60 µl) were done in single tubes or 96-well trays in 20 mM Na-HEPES (pH = 7.5), 100 mM NaCl, 10 mM $MgCl_2$, 0.1 mg/ml bovine serum albumin and 4 mM dithiothreitol. Final concentrations were 2 nM DNA (3.6 µg/ml) and 2 - 128 nM protein, which had been diluted in protein buffer (20 mM Na-HEPES (pH = 7.5), 300 mM NaCl, 10 mM $MgCl_2$, 0.3 mg/ml bovine serum albumin and 10 mM dithiothreitol) before. After 30 minutes at 37 ºC reactions were stopped by the addition of DNA loading buffer containing EDTA (17 mM final). Reaction products were separated on 1.0 or 1.4 % agarose gels (0.5 x TBE) containing 0.5 µg/ml ethidium bromide and visualized by fluorescence.

Oligonucleotide nicking

Gel-purified synthetic oligonucleotides were labeled at the 5' end with radioactive phosphate (^{32}P) using [γ-^{32}P]ATP and T4 polynucleotide kinase and were re-purified on gel. Equimolar amounts (450 nM) of unlabeled complementary and substrate strands were mixed with a trace amount of labeled substrate. The mixture was annealed in 5 mM Na-HEPES (pH = 7.5) by heating to 90 ºC and slow-cooled to room temperature. After testing various pH and salt conditions nicking reactions (50 µl) were done in 50 mM Na-HEPES (pH = 6.5), 150 mM NaCl, 10 mM $MgCl_2$, 0.1 mg/ml bovine serum albumin and 1 mM dithiothreitol. Final concentrations were 180 nM DNA (0.5-7.5 µg/ml) and 20 - 2000 nM protein, which had been diluted in protein buffer (5 mM Na-HEPES (pH = 7.5), 300 mM NaCl, 10 mM $MgCl_2$, 0.5 mg/ml bovine serum albumin and 5 mM dithiothreitol) before. After 30 minutes at 37 ºC reactions were stopped by the addition of 175 µl of 380 mM Na-acetate (pH = 7.5), phenol extraction and ethanol precipitation. Reaction products were separated on 10 % denaturing polyacrylamide gels and quantified in a phosphoimager to produce sequence logos.

Crystallization

Untagged LTx (20 mM Na-HEPES (pH = 7.0), 200 mM NaCl) was concentrated to 15 mg/ml. Sitting drops (200 nl protein plus 200 nl reservoir solution) were set up at room temperature using a Mosquito robot. Single crystals appeared over night from a reservoir (75 µl) containing 160 mM $MgCl_2$, 370 mM $(NH_4)_2SO_4$ and 33.8 % PEG 6000. Untagged LR1 (20 mM Na-HEPES (pH = 7.0), 200 mM NaCl) was concentrated to 10 mg/ml. Hanging drops (2 µl protein plus 2 µl reservoir solution)

were set up manually at 4°C. Crystals appeared after several days over a reservoir (500 µl) containing 10 mM $MnSO_4$, 200 mM $(NH_4)_2SO_4$ and 31% PEG 1000. Hair-seeding improved reproducibility significantly. In both cases crystals were transferred to a cryo-solution containing 15 % glycerol (mixing reservoir and 80 % glycerol stock solution) and flash-frozen in liquid nitrogen.

Data collection and structure solution

Diffraction data were collected at beamline ID23-1 at the European Synchrotron Radiation Facility in Grenoble, France. Diffraction images were processed by MOSFLM (Leslie, 1992) and SCALA (Evans, 1997). The structures were solved by molecular replacement using MOLREP (Vagin and Teplyakov, 1997) with L1-EN (PDB ID: 1vyb) as search model. Automatic model building was done with ARP/wARP (Cohen et al., 2004) to a completeness of 90 % for LTx and 98 % of LR1. Models were completed manually and structures were refined using REFMAC (Murshudov et al., 1997) and COOT (Emsley and Cowtan, 2004) iteratively.

Normal Mode Analysis

For normal mode analysis the PDB files of L1-EN, LTx, LR1 and TRAS1-EN were provided to the web-based server WEBnm@ following the standard protocol to calculate and analyze the first six vibrational modes (Hollup et al., 2005:
http://www.bioinfo.no/tools/normalmodes).

ACKNOWLEDGMENTS

Retrotransposition frequencies and genomic pre-integration sites were determined in the Schumann lab at the Paul-Ehrlich-Institute, biochemical and structural analyses were done at the Netherlands Cancer Institute. We thank Drs. J. V. Moran, and D. Carroll for providing plasmids pCEP4/L1.3/ColE1/*mneoI*$_{400}$ and pE1EN. Special thanks go to Jef D. Boeke, Greg Cost and Titia K. Sixma for helpful discussions. We also thank beam line scientists at the ESRF for assistance with data collection. This research was supported by a personal VIDI grant (NWO-CW 700.54.427) to O.W. by the Dutch National Science Organization (NWO) and in part by grants Schu1014/2-1, Schu1014/2-2, Schu1014/2-3 and Schu1014/2-4 of the *Deutsche Forschungsgemeinschaft* (DFG) to G.G.S.. K.R. was funded by grant NWO-CW 700.51.012 awarded to A.P. and O.W.. Coordinates and structure factors for LTx and LR1 have been deposited with the PDB.

REFERENCES

Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q.G.S., Roodveldt, C. and Tawfik, D.S. (2005) The 'evolvability' of promiscuous protein functions. *Nat Genet*, **37**, 73-76.

An, W., Han, J.S., Wheelan, S.J., Davis, E.S., Coombes, C.E., Ye, P., Triplett, C. and Boeke, J.D. (2006) Active retrotransposition by a synthetic L1 element in mice. *Proc Natl Acad Sci U S A*, **103**, 18662-18667.

Bestor, T.H. (2003) Cytosine methylation mediates sexual conflict. *Trends Genet*, **19**, 185-190.

Bogerd, H.P., Wiegand, H.L., Hulme, A.E., Garcia-Perez, J.L., O'Shea, K.S., Moran, J.V. and Cullen, B.R. (2006) Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc Natl Acad Sci U S A*, **103**, 8780-8785.

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

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.

Cohen, S.X., Morris, R.J., Fernandez, F.J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V.S., Kleywegt, G.J. and Perrakis, A. (2004) Towards complete validated models in the next generation of ARP/wARP. *Acta Crystallogr D Biol Crystallogr*, **60**, 2222-2229.

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.

Dlakic, M. (2000) Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. *Trends Biochem Sci*, **25**, 272-273.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F. and Kazazian, H.H., Jr. (1991) Isolation of an active human transposable element. *Science*, **254**, 1805-1808.

Eickbush, T.H. and Malik, H.S. (2002) Origins and Evolution of Retrotransposons. In *Mobile DNA II*. ASM Press, Washington, D.C., USA, Vol. 49, pp. 1111 - 1144.

Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*, **60**, 2126-2132.

Evans, P.R. (1997) 'SCALA'. *Joint CCP4 and ESF-EACBM Newsletter*, **33**, 22-24.

Feng, Q., Moran, J.V., 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 U S A*, **95**, 2083-2088.

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

Gasior, S.L., Wakeman, T.P., Xu, B. and Deininger, P.L. (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol*, **357**, 1383-1393.

Gilbert, N., Lutz-Prigge, S. and Moran, J.V. (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell*, **110**, 315-325.

Han, J.S. and Boeke, J.D. (2005) LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *Bioessays*, **27**, 775-784.

Hollup, S.M., Salensminde, G. and Reuter, N. (2005) WEBnm@: a web application for normal mode analyses of proteins. *BMC Bioinformatics*, **6**, 52.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S. and Winter, G. (1986) Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, **321**, 522-525.

Kazazian, H.H., Jr. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626-1632.

Kojima, K.K. and Fujiwara, H. (2003) Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol Biol Evol*, **20**, 351-361.

Leslie, A.G.W. (1992) 'MOSFLM'. *Joint CCP4 and ESF-EACBM Newsletter*, **26**.

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.

Maita, N., Anzai, T., Aoyagi, H., Mizuno, H. and Fujiwara, H. (2004) Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J Biol Chem*, **279**, 41067-41076.

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**, 451-456.

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.

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. *Nat Genet*, **31**, 159-165.

Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Lower, J., Cichutek, K., Flory, E., Schumann, G.G. and Munk, C. (2006) APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J Biol Chem*, **281**, 22161-22172.

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica D*, **53**, 240-255.

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

Park, H.S., Nam, S.H., Lee, J.K., Yoon, C.N., Mannervik, B., Benkovic, S.J. and Kim, H.S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, **311**, 535-538.

Schneider, T.R. (2000) Objective comparison of protein structures: error-scaled difference distance matrices. *Acta Crystallogr D Biol Crystallogr*, **56**, 714-721.

Soifer, H.S. and Kasahara, N. (2004) Retrotransposon-adenovirus hybrid vectors: efficient delivery and stable integration of transgenes via a two-stage mechanism. *Curr Gene Ther*, **4**, 373-384.

Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V. and Feigon, J. (2004) DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. *Proc Natl Acad Sci U S A*, **101**, 1177-1182.

Suck, D., Lahm, A. and Oefner, C. (1988) Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I. *Nature*, **332**, 464-468.

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**, 327-338.

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. (2002) Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *Embo J*, **21**, 408-417.

Vagin, A. and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J Appl Cryst*, **30**, 1022-1025.

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

Weichenrieder, O., Repanas, K. and Perrakis, A. (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure*, **12**, 975-986.

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.

Yang, N. and Kazazian, H.H., Jr. (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol*, **13**, 763-771.

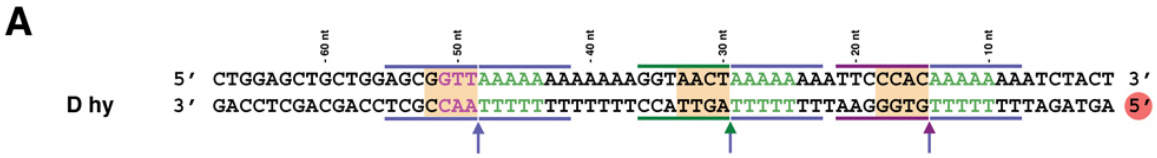
Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, **13**, 335-340.

Zingler, N., Weichenrieder, O. and Schumann, G.G. (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res*, **110**, 250-268.

Zou, P., Gautel, M., Geerlof, A., Wilmanns, M., Koch, M.H. and Svergun, D.I. (2003) Solution scattering suggests cross-linking function of telethonin in the complex with titin. *J Biol Chem*, **278**, 2636-2644.

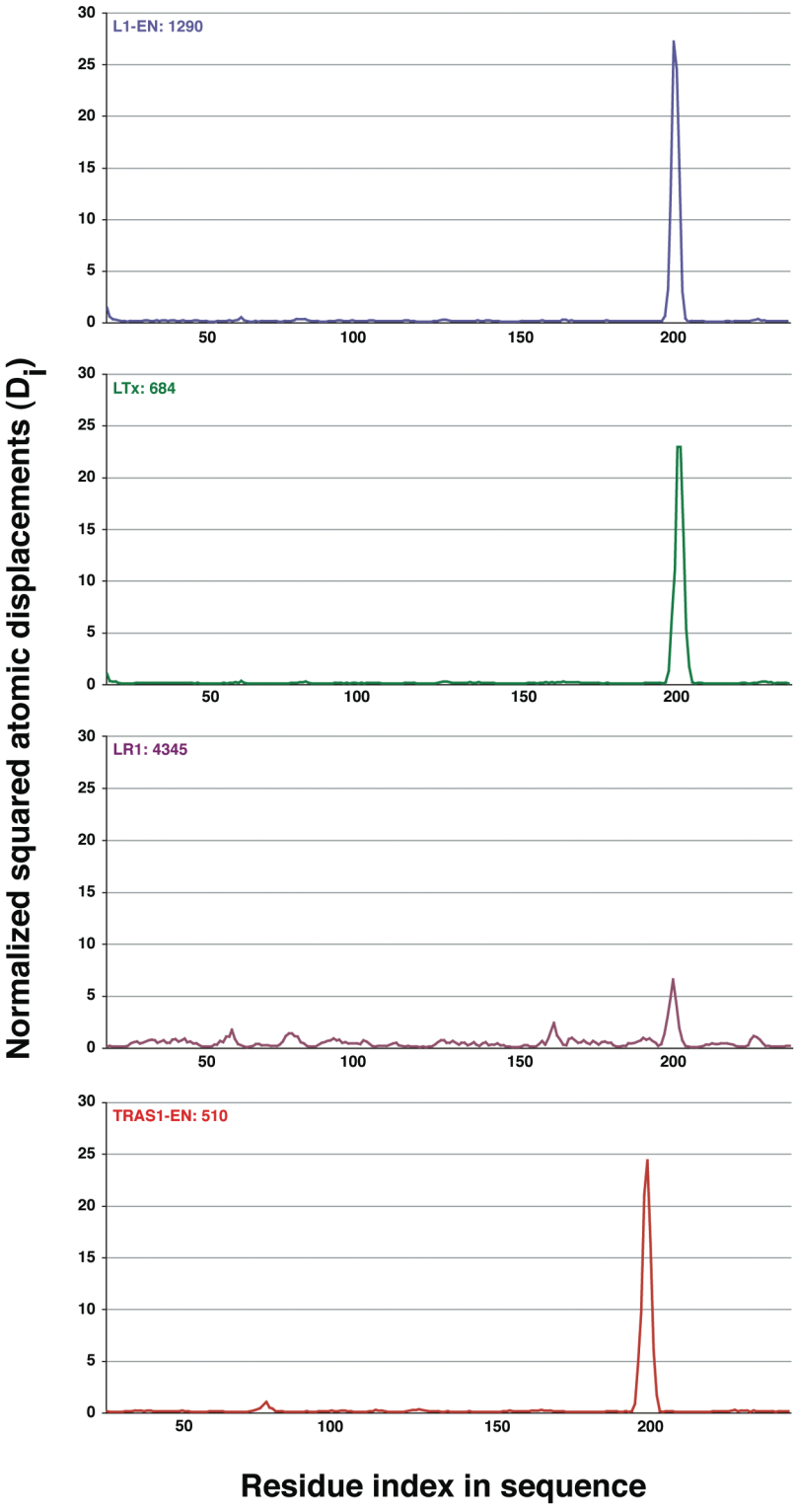
Supplementary Information

Supplementary Figure 1



Supplementary figure 1: Specificity of L1-EN βB5-βB6 hairpin loop variants and point mutants. A: DNA multi-substrate duplex (Dhy). Dhy contains a wild-type target site (blue arrow with seven flanking nucleotides marked by horizontal lines) that is ideal for L1-EN. Hybrid target sites are designed for nicking by LTx (dark green / blue) and LR1 (purple / blue), where the seven upstream base-pairs of the ideal target sites of Tx1L-EN and R1Bm-EN, respectively, are replaced by a T-tract. Upstream and downstream base-pairs important for recognition by the scaffold and the βB5-βB6 hairpin loop of L1-EN are colored lime and magenta, respectively. Nucleotides on the marked target sites that are thought to be in the reach of the various βB5-βB6 hairpin loops are on an orange background. Red Circle: 5' end (labeled) of the substrate strand. B: DNA duplexes (Dhy, 180 nM) were nicked by the indicated amounts of endonuclease. Products were analyzed on autoradiographs of denaturing polyacrylamide gels. Colors and symbols are as in (A).

Supplementary Figure 2



Supplementary figure 2: Normal mode analysis. C-alpha chains of L1-EN (blue), LTx (dark green), LR1 (purple) and TRAS1-EN (red) were analyzed on the WEBnm@ server (Hollup et al., 2005). Values for the lowest vibrational mode (mode 7) are shown. Normalized squared atomic displacements are plotted against residue number and average deformation energies are indicated in the top left corner.

Supplementary table I: Nicking sites and genomic pre-integration sites

Nicking sites and frequencies on Dwt					Genomic pre-integration sites	
Position on Dwt	Sequence (5' - 3') ------- ++++ 7654321 1234	Frequency, % wTx [a]	wt [b]	LTx	Sequence (5' - 3') wt [b,c] ------- ++++ 7654321 1234	Sequence (5' - 3') LTx ------- ++++ 7654321 1234
13	TAGGGAC AGTG	4	0	4	TACTTTT ATGA	TTTGTTT AAAA
14	AGGGACA GTGG	4	1	1	ATTTTAT AAAG	GAATTTC AAAT
15	GGGACAG TGGG	0	0	2	GACTTCT AAAA	ATGTCTT CTTT
16	GGACAGT GGGA	0	0	1	TATTTTT ATGT	AATTTTT AAGG
17	GACAGTG GGAA	0	0	2	GTTTCTT ATAC	TTTTTCT AAGT
18	ACAGTGG GAAT	0	2	2	TTTTCTT ACTG	TTTTCTT AGGT
19	CAGTGGG AATT	7	10	22	TTTTTTC ATAT	AATTCTT CTAA
20	AGTGGGA ATTA	4	3	5	TATTTCT AATG	ATTTTTT ATTT
21	GTGGGAA TTAG	0	1	1	GGTTTTC AAAT	TTCTCTT CTTA
22	TGGGAAT TAGC	0	1	2	TATTCTT ACTA	TTCTTTC AATC
23	GGGAATT AGCT	0	1	8	TTTTTTA ATTA	CATTTGT CACA
24	GGAATTA GCTG	0	1	1	CATTTCT GCAT	ATTTTCT GTGT
25	GAATTAG CTGA	0	0	1	AATTCTT AAGA	AACTTTT AGAG
26	AATTAGC TGAA	0	1	3	GTTTTCT AAGA	AAATTTT CAAA
27	ATTAGCT GAAG	2	2	1	ATGTCTT GTAA	
28	TTAGCTG AAGT	9	3	3	TGATTTT AAAA	
29	TAGCTGA AGTT	57	2	2	AAATTTT GAGG	
30	AGCTGAA GTTA	8	1	1	CCTTTCT AAAA	
31	GCTGAAG TTAC	0	0	1	TTTTTTT AACA	
32	CTGAAGT TACC	0	1	5	CTTTTTC AAGA	
33	TGAAGTT ACCT	0	2	9	TTTTTTT CACT	
34	GAAGTTA CCTT	0	0	1	TCTTTTT GAGA	
35	AAGTTAC CTTT	0	1	2	TTTTTTT GAGG	
36	AGTTACC TTTT	0	2	9	TTTTTTT AAGA	
37	GTTACCT TTTT	0	0	2	TTTTTTT GTTT	
38	TTACCTT TTTT	0	0	1	ATTTTTT AAAA	
39	TACCTTT TTTT	0	0	0	TATTTCT GTAT	
40	ACCTTTT TTTT	0	1	1	TTTTTTT AAAA	
41	CCTTTTT TTTT	0	1	1	ATTTCTT GCGG	
42	CTTTTTT TTTT	0	1	0	GGATTTT GAAA	
43	TTTTTTT TTTT	0	2	0	CAGTTTT AAAG	
44	TTTTTTT TTTT	0	2	0	TTTTATT GAAA	
45	TTTTTTT TTTA	0	3	0	CAGTTTT AAGG	
46	TTTTTTT TTAA	0	6	1	TTTTTTT AAAC	
47	TTTTTTT TAAC	0	12	1	TTTTTTT GAGA	
48	TTTTTTT AACC	0	29	2		
49	TTTTTTA ACCG	5	8	2		
50	TTTTTAA CCGC	0	0	0		

[a] Tx1L-EN
[b] L1-EN
[c] (Gilbert et al., 2002)

Supplementary table II: DNA primers

Primer name	Nucleotide sequence (5' – 3')
GS73	GGAAACCCATCTCACGTG
GS263	GTGTCGAGGAATGTATCC
GS313	GTCAATTTTGCCTCCTCCGGTATATTCTGTTGATTTG
GS314	ACAGAATATACCGGAGGAGGCAAAATTGACCACATAG
GS315	GAGAAACATGGCCATCTCTCACCCTGACATAGGTATATTCTGTTGATTTG
GS316	ATGTCAGGGTGAGAGATGGCCATGTTTCTCAATCCAAAATTGACCACATAG
GS317	TGGATTCTCCGTTCGCCGTACTGAAGGTATATTCTGTTGATTTG
GS318	ATACCTTCAGTACGGCGAACGGAGAATCCAAAATTGACCACATAG
GS323	AATAATGGGCGCCTTTGCCACCCCACTGTCAACATTAG
GS324	CAGTGGGGTGGCAAAGGCGCCCATTATTAATGTGTGG
GS334	CCACACCACACCTATGCCAAAATTGACCACATAG
GS335	GTGGTCAATTTTGGCATAGGTGTGGTGTGGTGC
GS336	GTCAACATTAGACGCATCAACGAGACAGAAAGTC
GS337	TTCTGTCTCGTTGATGCGTCTAATGTTGACAGTGG
GS338	ATCAACAGAATATGTCTTTTTTTCAGCACCACAC
GS339	GGTGCTGAAAAAAAGACATATTCTGTTGATTTGGG
GS340	CACACCTATTCCAAATATGACCACATAGTTGGAAG
GS341	CAACTATGTGGTCATATTTGGAATAGGTGTGGTG
Tx1L-EN-N1	AATCTGGAAACCATGGCCTTGAGTATAAGCACACTTAATACTAATGGCTG
Tx1L-EN-C239	AGCTAGGGATCCTTATTAGATTGACATTCTCAGGGATACACAATTGTGGT

Oligonucleotides and retrotransposition reporter plasmids

In order to enable easy modification of the L1-EN domain in the context of retrotransposition reporter plasmids, the 3.7 kb *Not*I/*Bcl*I-fragment of pJM101/L1.3 (Moran et al., 1999) was subcloned into pBluescript KS+ (Stratagene) to create plasmid pNZ01. pNZ01 was then used as a template for site-overlap extension (SOE-) PCR to generate the desired mutations. Inner primers introducing the mutations were used as indicated below. Outer primers (GS73 and GS263) included the restriction sites *Pml*I and *Xba*I, which were used to reinsert the mutated SOE-PCR products back into pNZ01. From the resulting subclones, the mutated sequences were transferred into the plasmid rescue vector pCEP4/L1.3/ColE1/*mneoI*$_{400}$ (Gilbert et al., 2002) using the restriction sites *Not*I and *Bcl*I. This resulted in the following reporter plasmids: pNZ74 (L3G) using primers GS313 and GS314, pNZ75 (LTx) using GS315 and GS316, pNZ76 (LR1) using GS317 and GS318, pColE1-S202A using GS334 and GS335, pColE1-R155A using GS336 and GS337, pColE1-T192V using GS338 and GS339 and pColE1-I204Y using GS340 and GS341. For the negative control, pColE1-H230A, the respective *Not*I/*Bcl*I fragment of pTAMH230A/L1.3 (Morrish et al., 2002) was directly transferred to pCEP4/L1.3/ColE1/*mneoI*$_{400}$. For the βB6-βB5 hairpin loop mutants pNZ74 (L3G), pNZ75 (LTx) and pNZ76 (LR1) additional control plasmids (pNZ83, pNZ84 and pNZ85, respectively) were prepared by SOE-PCR as described above using the respective subclones as PCR templates and inner primers GS323 and GS324 to introduce a D145A/N147A double mutation in the active site. Primer sequences are listed in the supplementary table I.

Chapter 4

To flip or not to flip: *insight into the DNA cleavage mechanism of human LINE-1 retrotranposon endonuclease*

Kostas Repanas, Gloria Fuentes, Serge X. Cohen, Alexandre M.J.J. Bonvin, Oliver Weichenrieder and Anastassis Perrakis

Submitted for publication

To flip or not to flip: *insight into the DNA cleavage mechanism of human LINE-1 retrotranposon endonuclease*

Kostas Repanas, Gloria Fuentes, Serge X. Cohen, Alexandre M.J.J. Bonvin, Oliver Weichenrieder and Anastassis Perrakis

ABSTRACT

The human LINE-1 endonuclease (L1-EN) contributes in defining the genomic integration sites of L1 and Alu elements that colonize more than a quarter of the human genome. Understanding how L1 recognizes DNA and facilitates nicking is the first step towards utilization of L1-EN retrotransposons as vehicles for gene delivery. We demonstrate that mutation of the catalytic aspartate (D145A) residue and the conserved isoleucine (I204Y), arginine (R155A) and serine (S202A) residues that may be responsible to accommodate a flipped out base during catalysis, all negatively affect *in vitro* activity. The total loss of activity of the aspartate mutant supports its role in generating the nucleophile. The crystal structures of three mutants indicate a very robust catalytic scaffold with minor structural rearrangements. Based on crystal structures and biochemical observations we constructed two computational models for DNA recognition: one involving a flipped-out nucleotide downstream the scissile phosphodiester; and one not. Although both models are feasible, the dramatically reduced activity of the isoleucine mutant can only be explained based on the flipped out nucleotide model; the activity of the arginine and serine mutants are also compatible with this model. Comparative molecular dynamics simulations however show that the arginine mutation is likely to destabilize the general DNA binding area and this could also explain the reduced activity of this variant, without assuming the flipped-out base mechanism. Collectively the dynamics studies and the modeling suggest that DNA backbone cleavage likely but not necessarily proceeds through a flipped out nucleotide base. L1-EN has a robust scaffold where the major flexible part is a protruding loop, that we previously showed to be crucial for inferring sequence specificity, and which together with additional surface elements facilitate DNA binding.

INTRODUCTION

L1 retrotransposons are mobile long interspersed nuclear elements (LINE-1) that lack retroviral-like long terminal repeats and are able to copy and paste themselves from one genomic location to the other with the help of an endonuclease-reverse transcriptase self-encoded activity. They propagate via an RNA intermediate and have created more than 17% of the mass of the human genome (Lander et al, 2001; Han & Boeke, 2005; Babushok & Kazazian, 2007). The Alu elements that likely propagate by hijacking the L1 retrotransposition mechanism, make up an additional 11% of the genome. L1 elements have a short loose genomic integration site consensus sequence 5'-TTTT/AA-3' (Gilbert et al, 2002; Symer et al, 2002; Szak et al, 2002) that is nicked by their targeting endonuclease (Feng et al, 1996; Cost & Boeke, 1998), to free a 3' OH group and prime reverse transcription (Luan et al, 1993). Previously, we described the crystal structure of the human L1 endonuclease (L1-EN) (Weichenrieder et al, 2004), a member of the family of metal-dependent phosphohydrolases (Dlakić, 2000). Based on structure comparisons and sequence alignments we suggested that L1-EN may accommodate a flipped-out nucleotide, in a DNA cleavage intermediate similar to the one characterized for the human DNA repair enzyme apurinic/apyrimidinic endonuclease 1 (APE1) (Mol et al, 2000).

The human LINE-1s are semi-specific for the target they recognize and integrate into, which differentiates them from elements of other organisms that exhibit higher specificity. Mobile elements like R1Bm from *Bombyx mori* (Xiong & Eickbush, 1988) and Tx1L from *Xenopus laevis* (Garrett et al, 1989) encode highly specific targeting endonucleases (Feng et al, 1998; Christensen et al, 2000). Each of these elements integrate into unique genomic locations, a specific sequence within 28S rDNA for R1Bm and within DNA transposon Tx1D for Tx1L (Zingler et al, 2005). The respective endonuclease largely determines the integration site. We previously exchanged the surface exposed βB6-βB5 hairpin loop of L1-EN for the ones of R1Bm and Tx1L endonucleases and determined the corresponding crystal structures (LR1 & LTx). Biochemical and structural analysis of the loop chimeras showed that the loop is indeed a major specificity element. We thus proposed that the prominent βB6-βB5 hairpin loop probably inserts into the DNA minor groove and may be a particularly important surface element involved in DNA target recognition (Repanas et al, 2007).

In this manuscript we aim to promote our understanding of how L1-EN is recognizing target DNA. In the enduring absence of an experimental L1-EN:DNA complex

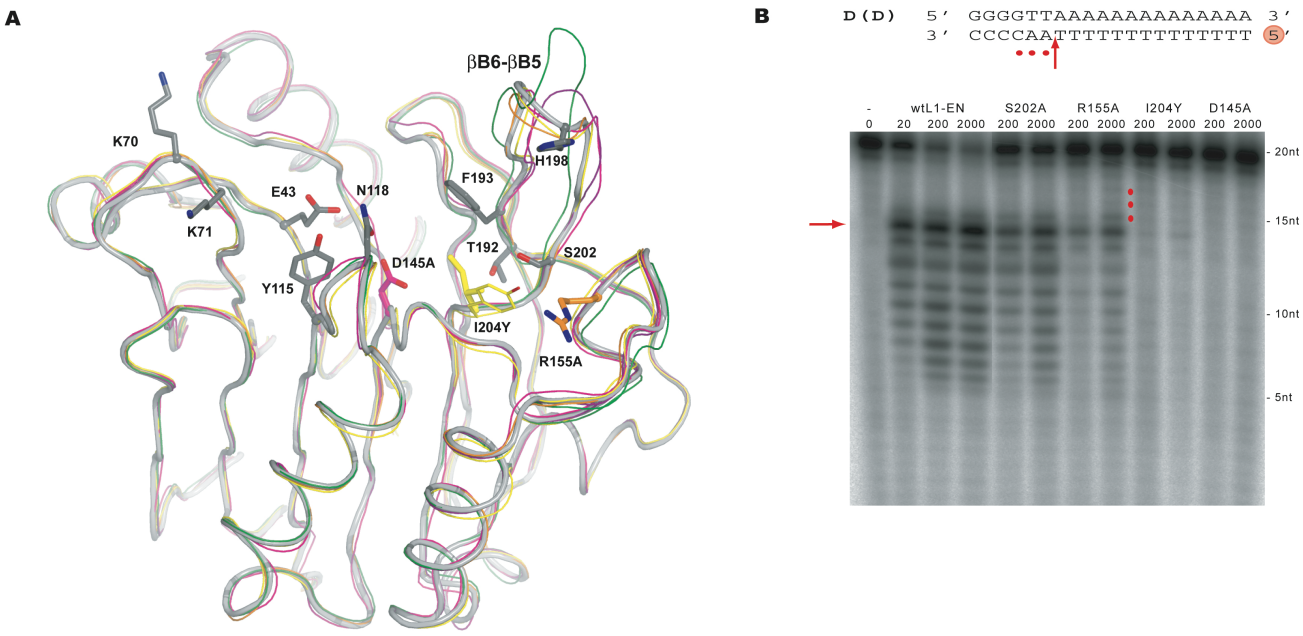


Figure 1. A: Ribbon representation of L1 endonuclease with key residues discussed in the text shown as ball and stick models. I204, R155 and S202 are thought to form a pocket for the accommodation of a flipped-out nucleotide. Residues that have been mutated are color coded: D145A-magenta, R155A-orange, and I204Y-yellow. Same colored ribbons of the respective variant structures are superposed on wtL1-EN. Ribbons of loop exchange variants LTx (green) and LR1 (purple) are also superposed for comparison. T192 and S202 are the anchoring residues of loop βB6-βB5 that also served as reference points in the loop exchange mutants (Table 2). Oxygen: red, Nitrogen: blue. **B:** Activity and target recognition of L1-EN point mutants. 20mer DNA substrate duplex containing the poly(T)-A junction. Red Circle: 5' end (labeled) of the substrate strand. DNA duplex (180 nM) was titrated with increasing concentrations of L1-EN and L1-EN variants. Products were analyzed on autoradiographs of denaturing polyacrylamide gels. Red arrow: poly(T)-A junction. Red dots: longer products downstream of the main nicking site.

structure, we employ the computational docking program HADDOCK (High Ambiguity Driven DOCKing) (Dominguez et al, 2003). HADDOCK can incorporate available information like experimental and bioinformatics data, in order to drive the docking process (van Dijk et al, 2005). Its use in a variety of cases predicting protein:DNA or protein:RNA complexes has been previously reported (Kalodimos et al, 2004; Kopke Salinas et al, 2005; Kamphuis et al, 2005; van Dijk et al, 2006). HADDOCK further allows for the introduction of protein flexibility, firstly on side-chains of the interface and then for both backbone and side-chains.

To focus our understanding of how the L1-EN structure infers its functional properties and the role of site-mutants that affect the actual catalysis step, we opted to also study the robustness and plasticity of the L1-EN catalytic scaffold. Flexibility in protein systems can constitute a major component in creating their specificity and catalytic power, hence leading to the creation of a functional entity. Molecular dynamics simulations can reveal the role of protein flexibility in structure and mechanism, especially when carried out at the atomic level. Catalytic processes in enzymes exploit the energy available from local and global molecular motions just as much as from the particular chemical environment (Dodson & Verma, 2006) Many different theoretical methods exist for the description of protein flexibility, but molecular dynamics (MD) is probably the most powerful (McCammon et al, 1977; van Gunsteren et al, 1983). Since the very first applications to proteins in the late 1970s, MD has been successfully used to study the dynamic behavior of various proteins (Karplus & McCammon, 2002; Karplus & Kuriyan, 2005). It is generally believed that protein behavior during MD simulations most likely reflects physically meaningful processes (Rueda et al, 2007).

We follow a mutational approach combined with structural, docking, MD and biochemical studies. Specifically we aim to test the hypothesis of the occurrence of a flipped out base as an intermediate to DNA cleavage similar to the APE1 mechanism (Mol et al, 2000) and to provide further insight to the dynamic properties of the DNA binding surface.

The structure of wild type L1-EN (Weichenrieder et al, 2004) comprises a rigid core that sculpts the surface of the likely DNA binding site, and a prominent protruding loop (βB6-βB5, Figure 1A). We have speculated that catalysis may proceed with an APE1-like mechanism with a flipped out nucleotide. This was illustrated by a model constructed using the positions of two sulphate ions and the scissile phosphate position as anchor points for the placement of substrate DNA. This model together with sequence conservation indicated that R155, S202 and I204 are forming a pocket that accommodates a flipped out adenine, and was used as a guide for the mutagenesis experiments presented in this manuscript. In contrast to DNA repair enzymes APE-1 in humans (Mol et al, 2000) and Exonuclease III in Escherichia coli (Mol et al, 1995), L1-EN can provide sufficient space for a

flipped out nucleotide. The turned ribose could rest on I204 as in APE-1, while the semi-conserved R155 and S202 could do base specific hydrogen bonds with the nucleotide. We have also shown two 'loop graft' mutants, LTx and LR1, where the βB6-βB5 loop was exchanged for that of the specific R1Bm and Tx1L retrotransposons (Figure 1A, magenta and green ribbons). These indicated that loop grafting can alter the sequence specificity of L1-EN *in vitro* and of L1 integration specificity in cell cultures (Repanas et al, 2007). The readout of the sequence appears to be 'indirect' and L1 most likely recognizes structural features inferred by the sequence and not the sequence itself.

Here we report the *in vitro* function and crystal structures of point mutants of EN-L1: D145A, the main catalytic residue that renders the protein inactive; I204Y that has implications to the accommodation of a speculated flipped-out DNA base during cleavage; and that of R155A and S202A that are in a position to fix the flipped out nucleotide. By comparing these structures and the conserved position of bound small molecule ligands among them, we construct two computational models for 'normal' and 'flipped-out' DNA complexes and validate them using our mutant structures. Finally, by studying the dynamic properties of various L1-EN structures via MD simulations (of wild type L1-EN, two L1-EN loop graft mutants LTx & LR1, the R155A mutant and that of retrotransposon-encoded TRAS1-EN (Maita et al, 2004) we propose how the L1-EN activity is likely to be modulated by dynamic structural properties.

Table 1. Summary of mutants

Name	Mutation type	New residues	3D structure	Activity	Affects specificity
D145	Point	A	+	-	n/a
R155	Point	A	+	**	-
S202	Point	A	-	***	-
I204	Point	Y	+	*	-
LTx	Loop-exchange	YVRVRDGH VSQ	+	**	+
LR1	Loop-exchange	FSTANGE	+	*	+
L1-EN	Wild-type	n/a	+	*****	n/a

RESULTS AND DISCUSSION

Active site and peripheral L1-EN mutants affect cleavage activity but not specificity or structure

A summary of functional and structural data for all mutants used in this study is available as Table 1. Structural comparisons are summarized in Supplementary Table 1 (Theobald & Wuttke, 2006).

Aspartate 145 is a totally conserved residue in the family of metal-dependent phosphohydrolases. This mutation totally abolishes activity (Figure 1B) while no structural differences occur (Figure 1A and 2; purple model) despite being situated at the heart of the L1 endonuclease. These observations confirm that D145 is absolutely required for the actual catalytic event (Mol et al, 2000).

The isoleucine at position 204 was mutated to tyrosine to mimic the L1 homolog bovine pancreatic deoxyribonuclease I (DNaseI) (Weston et al, 1992). DNAseI-like enzymes form a small group of metal-dependent phosphohydrolases that display conservation in this tyrosine, which is usually a small hydrophobic residue among other family members. DNAseI binds DNA in a non-specific manner and has not been shown to involve recognition of an extra-helical nucleotide for catalysis to proceed, like APE1 (Mol et al, 2000). The isoleucine at position 204 provides enough space for accommodating the nucleotide adjacent to the scissile phosphate in a possible flip-out mechanism, similar to the model for APE1 catalysis. Should the longer and bulkier tyrosine moiety assume a relative position similar to the one in DNAseI active site, it should obscure activity. Indeed the I204Y mutation has a dramatic effect in catalysis although residual activity can still be detected, compared to the totally inactive D145 mutation (Figure 1B). However, the overall structure is only moderately affected and the bulkier and more hydrophilic tyrosine is accommodated rather well in the L1-EN scaffold, (Figures 1A and 2; yellow model). Since I204 is rather far from the actual catalytic residues and an isoleucine is totally unlikely to participate in actual catalysis, the only reasonable explanation is that indeed the tyrosine is not tolerated due to steric hindrance upon DNA binding. Since wt L1-EN is only moderately specific demonstrating the characteristic laddering pattern in Figure 1B, it is totally unlikely that the I204Y mutation results in complete loss of specificity in our assay and subsequent loss of only apparent activity; that is further supported by cleavage of supercoiled plasmid DNA (data not shown).

Arginine 155 could hydrogen bond to an extra-helical nucleotide, should L1-EN follow the APE1 model and recognize a flipped out base. The structure of the R155A mutant remains virtually unaffected (Figures 1A and 2; orange model), while activity is reduced (Figure 1B). The R155A is significantly more active than the I204Y mutant: that is compatible with the hypothesis that isoleucine is absolutely crucial for allowing space for the flipped out nucleotide, while the arginine would only be needed to stabilize the base for more efficient catalysis.

That also explains why sequences with nucleotides other than adenine downstream of the scissile bond can be accommodated during cleavage at that position, although less efficiently (Repanas et al, 2007); the lesser efficiency is likely due to less efficient fixation of the base by R155. Finally, the R155A mutation affects only nicking activity but has little apparent effect on specificity; if anything the observation of the slightly relatively more pronounced longer cleavage products outside the A-tract are compatible with our hypothesis for its role. The activity of the S202A mutant is less affected (Figure 1B); that is consistent with a lesser role in fixing the flipped out nucleotide: the hydrogen bond possibly contributed by S202 to fixing the flipped-out nucleotide can be sufficient even if less efficient than that of R155 for fixing a flipped out base. This also can explain why R155 is important but not strictly essential for catalysis to proceed.

Table 2. Data collection and refinement statistics.

Data collection	I204Y	R155A	D145A
Resolution, Å	2.2	2.0	2.4
Cell a,b,c (Å) α, β, γ (°)	42.6, 93.8, 126.6 90, 90, 90	43.1, 93.6, 125.9 90, 90, 90	43.2, 93.5, 125.7 90, 90, 90
Space group	$P2_1$	$P2_1$	$P2_1$
R_{merge}, % [a]	7.6 (48.2)	8.0 (55.3)	9.2 (43.7)
Completeness, % [a]	86.2 (88.6)	97.3 (96.5)	84.1 (87.0)
I/σ(I) [a]	6.8 (1.5)	12.1 (2.5)	9.3 (2.0)
No. of reflections			
Unique observed	46416	65720	37199
Total measured	62476	248082	78085
Twinning			
Operator	-h, -k, l	-h, -k, l	-h, -k, l
Twin fraction	0.45	0.1	0.29
Refinement			
R_{cryst}, %	14.2	17.5	17.3
R_{free}, %	20.8	24.7	25.6
Number of :			
molecules in AU	4	4	4
atoms	8198	8082	7930
ions	6	9	4
water molecules	435	389	430
Ramachandran plot			
Most favored regions, %	89.6	91.6	82.1
Allowed regions, %	10.4	8.2	17.1
Generously allowed, %	-	0.2	0.8
Rmsd.from ideal geometry			
Bond lengths, Å	0.003	0.006	0.007
Bond angles, °	0.55	0.77	0.86

Structures of L1-EN mutants provide a structurally robust scaffold with conserved binding features

Crystals of R155A diffracted X-rays to 2.0 Å, those of I204Y to 2.2 Å and the ones of D145A to 2.3 Å. All three crystals belong to space group $P2_1$, contain four copies of each protein in the asymmetric unit and are twinned (twin law -h,-k,l). Details for data collection, treatment for twinning and refinement statistics are summarized in Table 2. The structures were solved by molecular replacement using L1-EN as a search model; the mutated residues could be clearly confirmed and identified in the electron density maps. The three new structures presented here, combined with the wild type L1-EN structure and those of the two hairpin variants LTx and LR1 we have previously reported (Repanas et al, 2007), enable us to take a much more detailed look in and around the active site of L1-EN.

Looking at the active site regions of all structures collectively (Figure 2), most of the induced differences apart from the mutations themselves seem to concentrate on E43 and H230. The glutamate has been shown to coordinate a Mg^{+2} ion essential for catalysis, in many metal-dependent phosphohydrolases. The histidine is likely to create the attacking nucleophile as show for APE-1 (Mol et al, 2000). Comparing the overall positioning with respect to the wild type L1-EN structure, the structure of the I204Y variant is the one exhibiting the largest changes, besides the prominent substitution of isoleucine to tyrosine. It is not straightforward to explain these differences since the tyrosine that replaces the isoleucine points away from the active site.

It is notable that Mg^{+2} ions were necessary for crystallization to proceed in all cases; but Mg^{+2} could not be identified in any of these structures, with the exception of LR1 (Repanas et al, 2007). In the LR1 structure we confirmed the metal position by substituting Mg^{+2} for Mn^{+2} and collecting anomalous X-ray diffraction data at 1.7 Å wavelength. The presence of a metal position was clearly confirmed, but despite extensive effort we were unable to model it accurately, due to dynamic or static disorder, that manifested itself as a 'double' anomalous peak. Due to this lack of conclusive data we are not discussing this specific structure and possible implications of the metal positioning in catalysis. Collectively, the superposition of the structures, with the various positions that residues can adopt, illustrates the plasticity of this region and hints to breathing motions while scanning and adapting the surface depending on the DNA substrate.

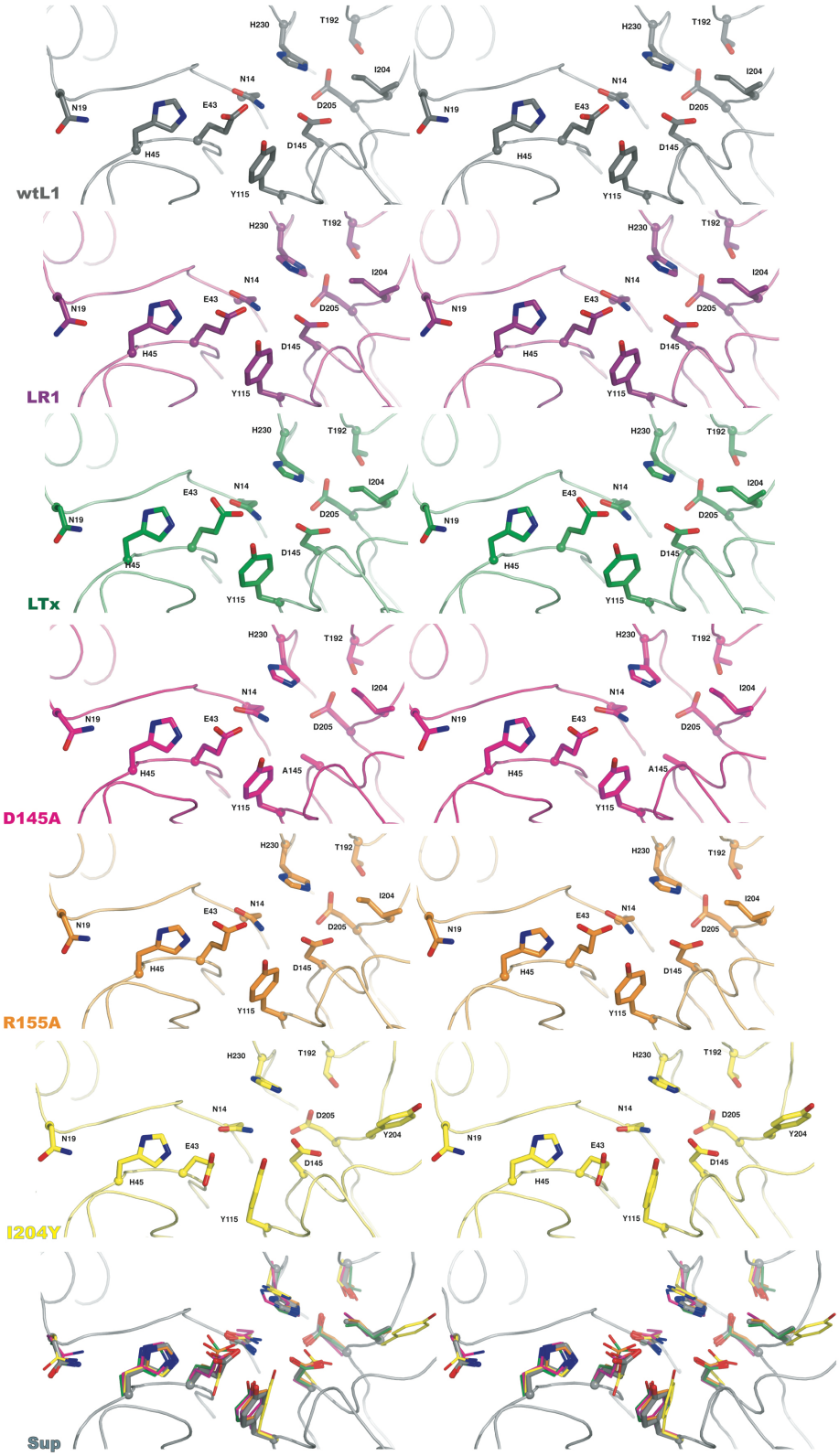


Figure 2. Stereo views of the L1-EN active site and of the five variants discussed in the text, as well as superposition of all six. Grey: L1-EN, purple: LR1, green: LTx, magenta: D145A, orange: R155A and yellow: I204Y. Color coding follows throughout the paper. Oxygen: red, Nitrogen: blue.

All crystallization conditions contained sulfate ions (ammonium sulphate was used together with Polyethylene glycol as a precipitant) and the cryo-protection medium contained glycerol. In the three mutant structures we present here, as well as in the previously solved structures of wtL1-EN, and the hairpin exchange variants LTx and LR1, both sulfate ions and glycerol molecules could be clearly located. After superimposition of all available structures (Figure 3), it can be seen that many of these ligands occupy common sites. The most frequently occupied position for a sulfate ion is between His45 and Asn19; all six available structures bind an ion at this position. Three structures (LTx, R155A and L1-EN) have a sulfate ion between the 'other' side of His45 and Tyr115. Other binding sites are clear but not consistent between structures. In LTx, Arg155 binds a sulfate ion that could indicate a stabilizing role contacting a DNA phosphate, or the extra-helical nucleotide in the flipped-out base model, that are both discussed later. In LTx we also find a sulfate ion in the active site of the protein, coordinated by His230, Asp145, Glu43 and Tyr193 (that replaces the wild type Phe193 present in L1-EN). LR1 also binds a glycerol moiety, which is located straight in the active site; the glycerol hydrogen bonds to His230, Tyr115, Asp145, Asn147; it is also in contact with Glu43 via the bound - albeit not clearly modeled - metal ion. The glycerol moiety superimposes well onto the DNA backbone phosphate of the APE1 complex and the LTx sulfate ion (Figures 2 and 3).

Collectively the analysis of all biochemical assays and structures - argue that the active site mutants presented here are in sharp contrast with the activity profile of the loop exchange variants LTx and LR1 we have previously described, and which affect the specificity of cleavage (Table 1). The activity of the point mutants and their structures are compatible with the initial speculation that DNA cleavage might proceed via an extra-helical nucleotide. We further our analysis, by testing the hypothesis of the flip-out mechanism using computational modeling methods.

Computational docking of an A-tract DNA substrate on L1-EN driven by mutational data.

The L1 endonuclease could follow two main modes for binding and cleaving its DNA substrate. The first is that of DNaseI that appears to cleave DNA without the need to flip a nucleotide. The second is that of APE1 which in order to repair an abasic site most likely flips out a nucleotide. We have previously constructed such models manually based on the crystal structure of wild type L1 (Weichenrieder et al, 2004). In the absence of an experimentally determined structure of an L1-EN:DNA complex, we decided to extend our docking studies and model the complex computationally, using the HADDOCK software (Dominguez et al, 2003) and suitable DNA substrates (Stefl et al, 2004). Together with the mutant structures we attempt to extend our understanding and distinguish computationally, which of the two aforementioned binding modes is the most favorable.

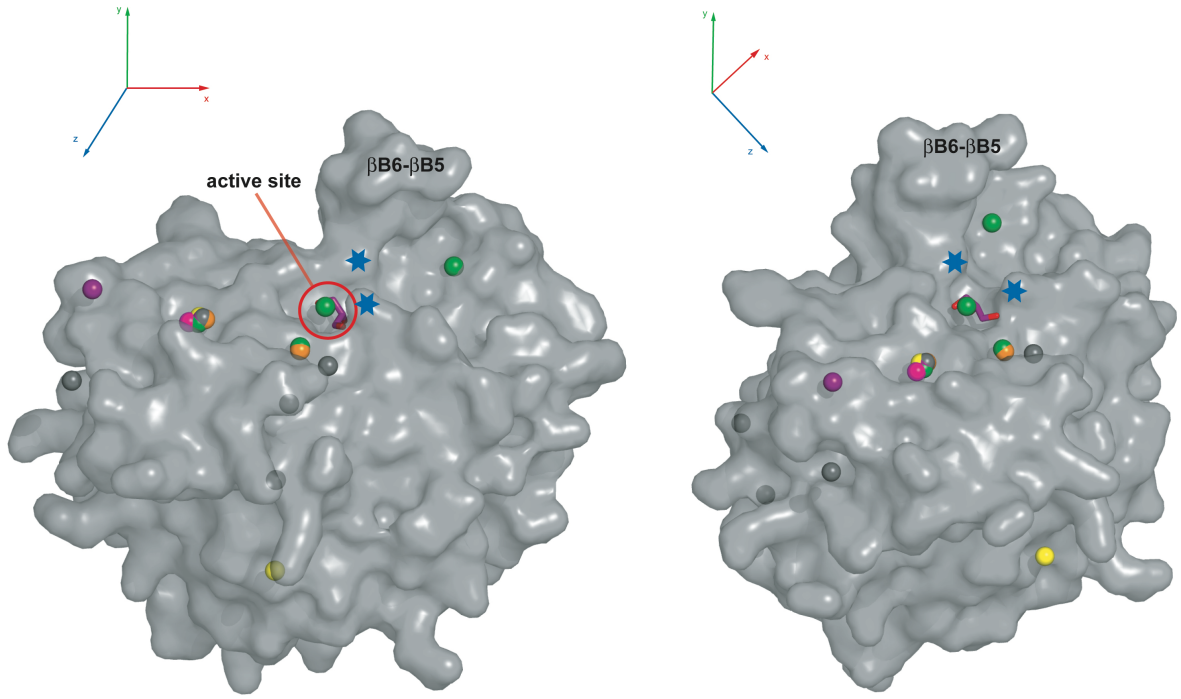


Figure 3. Surface representation of L1-EN with all sulfate ions (as spheres) and glycerol (as sticks) from the six available crystal structures superimposed. The active site is marked with a red circle. Blue asterisks indicate the narrow "channel" formed by N118 and F193 above the active site. The βB6-βB5 hairpin is labeled to aid orientation and front and side views of the protein are shown.

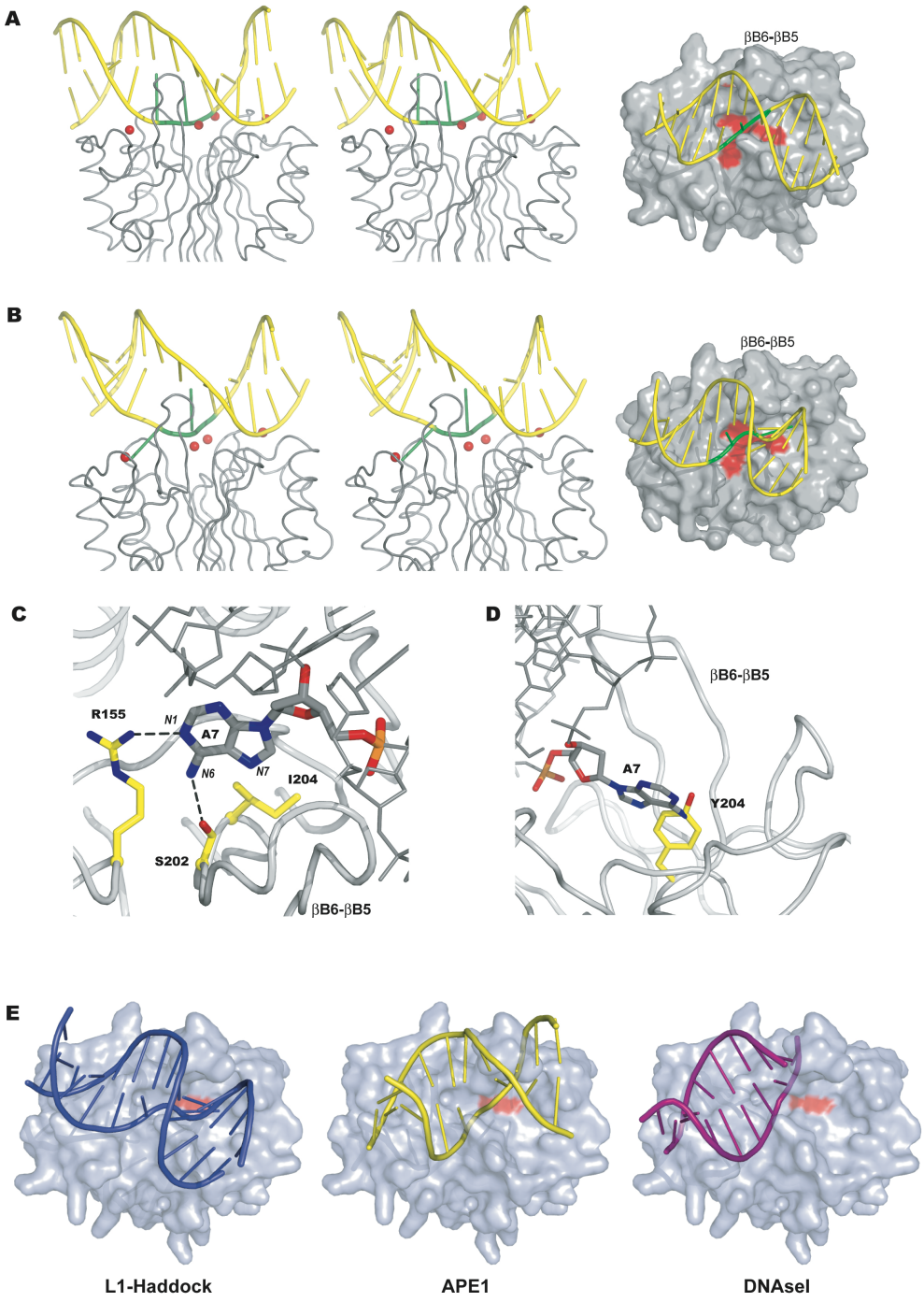


Figure 4. A: Stereo view of "non-flipped" DNA (wild-type) docked to L1-EN with HADDOCK. The model depicted is the representative of the best cluster. The βB6-βB5 hairpin is seen inserting in the DNA minor groove. Red spheres represent sulfate ions from the LTx crystal structure. DNA cartoon is shown in yellow, with the bases at the TpA step targeted for cleavage in green, Gray: L1-EN ribbon /Alternative representation: L1-EN is drawn as a grey surface and is viewed from the top, so that the βB6-βB5 hairpin is pointing towards the reader. The double-stranded A-tract DNA substrate is drawn as a yellow cartoon, with the TpA step that is the supposed target for cleavage highlighted in green. A red colored surface indicates the position of active site residues of L1-EN involved in catalysis. **B:** Same as in A but with "flipped-out" DNA as substrate. **C:** Close-up of the "flipped-out" HADDOCK model to illustrate the accommodation of adenine 7 by R155, S202 and I204. L1-EN is depicted as light-gray ribbon with the hairpin labelled and the DNA as dark-gray sticks. Oxygen: red, Nitrogen: blue, Phosphorus: orange. **D:** Alternative view of same model, illustrating the clash between I204Y mutant (superposed in yellow) and adenine 7 of the TpA step, which are shown as ball and stick models. Coloring as in C. **E:** Three surface representations of L1-EN with superposed DNA cartoons from "flipped-out" HADDOCK (blue), APE1 (yellow, pdb code: 1DE8) and DNAse I (magenta, pdb code: 1DNK). The position of the I204Y mutation is marked as red surface

The first DNA target, dT_4A_4 (pdb id: 1RVI) resembles the preferred target substrate of the L1-EN, having a unique geometry that includes a narrow minor groove that widens at the TpA step, making it a good candidate for the docking study. To initiate the docking process we used our mutational data as well as residues shown to coordinate bound ions in the six crystal structures discussed earlier. Hence, in the HADDOCK formalism, certain residues were defined as 'active', depending on the predicted level of involvement in the DNA binding interface, and their neighbors as 'passive' (Table 3). The docking partner, dT_4A_4, was annotated in a similar fashion, defining as active the adenine and thymine at the TpA step and the rest of the molecule as passive. Complexes were scored as described in the Supplementary Information sections and for simplicity a representative model obtained with HADDOCK is illustrated in Figure 4A. The substrate DNA is docked on the predicted interface and the βB6-βB5 hairpin can be seen penetrating the wide minor groove. Major groove contacts are mediated by Asn118 and neighboring residues on the same small loop. Residues Asn118 and Phe193 form a narrow passage of less than 6.5 Å, where the DNA can be locked in position while catalysis of the phosphodiester bond takes place (Figure 3, blue asterisks). Four of the sulfate ions we previously described, coincide with backbone phosphates of the DNA. The docked substrate is placed on an orientation suitable for cleavage, with the target phosphodiester bond between Thy6 and Ade7 in the proximity of the active site.

The second DNA model contains an extra-helical adenine at the T-A junction. To construct it we chose to use angles from two known structures that contain flipped out nucleotides: that of APE1 bound to abasic DNA (pdb code: 1DE9); and that of β-glucosyltransferase bound to a 13-mer DNA (pdb code: 1SXP). The model of L1-EN complex with this *in silico* "manipulated" DNA was created following exactly the same protocol and using identical restraints as above, to enable unbiased comparison of the two models. A number of complexes were selected for further inspection and the best resulting complex is shown in Figure 4B. The overall binding of this DNA appears similar to that of the wild type (illustrated in Figure 4A,B) but this time the extra-helical adenine at position 7 is in hydrogen bonding distance from R155, S202 and close to I204 that create a pocket for accommodating the flipped-out nucleotide (Figure 4C). Arginine 155 makes a single hydrogen bond to N1 of the adenine base in this model; however one can easily imagine that a slight re-arrangement can result to R155 contacting both N1 and N6, while the hydroxyl of S202 (now hydrogen bonding N6) contacts N7 instead, as it was actually modeled manually (Weichenrieder et al, 2004). This rearrangement would imply a more important role of R155 than S202 in catalysis, consistent with the activity experiments (Figure 1B). Similar to the non-flipped DNA docking, the bond to be cleaved at the TpA step is positioned near the active site, where direct or water mediated contacts could take place. In the case of a nucleotide flipping-out, its Watson-Crick partner would need to be stabilized during catalysis in a way similar to that of APE1, which provides two loops contacting the orphan base from both minor and major DNA grooves. These contacts in L1-EN could be achieved by F193 of the main hairpin βB6-βB5 in the minor groove and the βB3-αB1 loop containing N118 in the major groove, stabilizing the orphan thymidine in the HADDOCK model.

When the I204Y variant structure is superposed on the "flipped-out" complex, there is a very obvious clash between the extra-helical adenine and Y204 that would be present in DNaseI (Figure 4D). That would argue in favor of the base-flipping mechanism, illustrating that the bulky tyrosine instead of an isoleucine is obstructing the recognition of the extra-helical nucleotide and thus rendering the EN inactive. On the contrary, superposition of the I204Y variant structure on the "non-flipped" EN:DNA complex does not seem to obstruct DNA binding in a major way. With the currently available models the behavior of this mutant is favoring the recognition of an extra-helical nucleotide during cleavage. Of course, given the available data, we cannot rule out the possibility that the tyrosine can rotate and point to different directions which would render any EN:DNA complex unstable or inactive, even in the case that base-flipping would not be required.

Still, it is not evident from the docking results alone, which mode of binding is preferred. DNA positioning appears similar independent of the base flipping. The non-flipped substrate is docked marginally better, and therefore placed 1-2Å closer to the active site and ions that could take the place of backbone phosphates. Both runs yield complexes that resemble to a greater extent the way DNA is bound on APE1 structures, rather than how DNaseI binds its target (Figure 4E); in the sense that a larger binding interface is required with the DNA sitting on both sides of the hairpin loop. This is in agreement with our mutational data. If DNA was bound as in DNAseI, we would expect no reduction in EN activity when I204 is mutated to tyrosine: our combined biochemical results indicate that the tyrosine in the place of the isoleucine is not tolerated at all; only residual activity can be detected in three different assays (Figure 1B and Repanas et al, 2007).

Superposition of the inactive D145A variant structure on either of the docked complexes confirms that no major changes occur on the protein surface and that the DNA could still bind as before, although catalysis cannot proceed without the aspartate in place. One residue that is in favorable position to contact target DNA in all the HADDOCK models examined is Y115. This is in agreement with observations in the crystal structures, since it could be involved in ion co-ordination, and it is situated in the vicinity of the active site. The above also agree with recent studies in DNA repair enzyme APE1, where this conserved tyrosine is thought to be directly involved in catalysis and binding/recognition of DNA (Melo et al, 2007).

When we superpose the structures of the previously reported LTx and LR1 variants on this new docking models, there is a striking difference on the way the two exchanged hairpins could interact with the DNA substrate. The morphology and positioning of the loops is quite different: the Tx1L hairpin penetrates and pushes the minor groove; the shorter R1Bm hairpin barely reaches the DNA backbone. These structural differences could explain the reduced activity and altered specificity these variants exhibit as we initially suggested based on the apo- structures and biochemical data (Repanas et al, 2007).

Table 3. Definition of the ambiguous interaction restraints (AIRs) for the two HADDOCK runs

	Residue number
L1-EN	
Active	R155, F193, S202, I204, N19, H45, K70, H198
Passive	N147, Q159, T157, H199, A196, R53, K71, T119, N118, P197, D229, N16, S20
DNA	
Active	T6, A7
Passive	1-5, 8-24

Apart from the structural characteristics of the loops that penetrate the minor groove, the charge properties of the three different loop variants display some differences that further support and explain the modes of DNA binding. The active wt L1-EN protein has two histidines at the top of the loop that are positively charged at least at lower

pH; the less active LTx loop mutant harbors one positively charged histidine and a negatively charged glutamate at the tip of the loop, while it offers two additional positive charges with two arginines positioned at the loop base; the very little active LR1 loop graft mutant offers no positive charges at the tip of the loop, and a negatively charged aspartate in the middle. The endonuclease from insect retrotransposon TRAS1 (Maita et al, 2004), TRAS1-EN, has a similar prominent loop. This exhibits three positively charged residues close to the tip of the loop, two arginines and a lysine. Since the insertion of the loop to the minor groove according to both our models is essential, we conclude that positive charges mediate better placement of the DNA and subsequent cleavage. In such a model where the loop probes structural features of the DNA, while charges definitely have a role, the dynamic flexibility of the loop and the DNA binding interface in general should be crucial for target recognition, DNA binding and efficient phosphodiester bond cleavage. Dynamic processes seem to play a significant role both for initial binding and subsequent nicking of the DNA substrate, which might proceed via a base flipping mechanism. Here we use a molecular dynamics approach to analyze further this flexibility hypothesis, which we initially demonstrated with a normal mode analysis (Repanas et al, 2007).

Molecular dynamics simulations confirm a connection between flexibility and functionality

To further understand the functional role of loop plasticity and flexibility regarding its importance for target DNA recognition, binding and cleavage, and at the same time confirm that the L1-EN scaffold is rigid as suggested by the mutant structures, we performed classical molecular dynamics simulations using GROMACS to approximate the obscure experimental representation of protein elasticity. A total of five 10 nanosecond simulations were performed on: wtL1-EN, LTx, LR1, R155A and finally for comparison on the L1-EN homolog TRAS1-EN, the sequence-specific APE-type non-LTR retrotransposon endonuclease that cleaves telomeric repeats of insects.

Standard GROMACS analysis was performed for all trajectories for factors such as temperature, pressure and energy, to ensure that once the system reaches equilibrium after the first couple of nanoseconds, it remains stable along the trajectory. The average backbone root mean square fluctuation (RMSF) values during simulation (a measure of the overall flexibility of the system) were calculated during the last 6 ns of the simulations. The RMSF plots of the atomic positions in the trajectories of LR1, R155A, LTx and L1-EN (Figure 6A) shows that certain regions are more flexible than others and that these regions coincide across the different structures. Focusing on the region of the β-hairpin (βB6-βB5), which lies roughly between residues 190-200, we observe that for the case of the relatively inactive LR1 this area is much more rigid compared to the flexible hairpins on L1-EN, LTx and R155A. For L1-EN and R155A the plots practically superimpose as expected; the residues in question are identical in both cases. The detailed simulations collectively confirm and quantify our hypothesis for the importance of loop flexibility for efficient cleavage of substrate DNA.

Our simulations reveal additional regions that have an elevated degree of flexibility (Figure 5A and 5B). The area around βA4-αA2 near E43 that is involved in ion coordination, the area around K70-K71 (βA6-βA5) that possibly contacts DNA, and finally the small loop (βB3-αB1) where N118 is situated that could be involved in major groove interactions.

A clear variation in the different trajectories is in the R155A mutant trajectory: the loop (βB6-αB2) around the mutation is comparably more flexible. The salt-bridge between D154 and R182, adjacent to the R155A mutation, is conserved and restricted to AP DNA repair endonucleases and mammalian-type L1 endonucleases (Weichenrieder et al, 2004). It could serve as an extra fixing point for this loop and is present for instance in L1-EN, APE1 and Tx1L-EN, but missing from DNaseI, R1Bm-EN and TRAS1-EN. This salt bridge is disrupted in the R155A simulation (Figure 5C), along the timeline of the last five nanoseconds, whilst it is stable in L1-EN where the neighboring arginine is present. From that we conclude that the R155A mutation destabilizes the surrounding region and results in the salt bridge breaking. This possibly explains why the whole loop appears to be more flexible than the one in the wild type protein. If the role of this loop is to contact the DNA and fix it in a way for cleavage to occur, the absence of the arginine and the increased mobility could explain the reduced activity this mutant exhibits, likely important during DNA binding in general, even if it is not directly responsible for binding the flipped out DNA base.

The region around R155 in the L1-EN homolog TRAS1-EN, is also of interest. R155 corresponds to K168 in TRAS1-EN. It is known that a mutation of K168 to alanine or even to an also positively charged arginine, which mimics human L1-EN, drastically reduces the activity of this highly specific endonuclease (Maita et al, 2004). This loop is in a region of high mobility also in TRAS1-EN, looking at the rainbow matrix representation of root mean square deviation (RMSD) between distances of atoms during the trajectory (Figure 5D). The equivalent of the D154 - R182 salt-bridge that could provide extra stability to this loop is missing in the case of TRAS1-EN. This further suggest that this peripheral loop, almost 17 Å away from the protein's active site, has an important role in aiding function and modulates activity, while having no apparent effect on protein specificity, as seen in both L1 and TRAS1 endonucleases. A small loop that includes D130 (N118 in L1-EN) also shows increased flexibility; this aspartate is known to retain activity but lose specificity for telomeric targets when mutated to alanine, possibly pointing towards major groove interactions (Maita et al, 2004).

Principal component analysis (PCA) reveals that in the trajectory of the R155A mutant, the βB3-αB1 loop of N118 also shows flexibility (Suppl. Figure 1). Since motions of the βB6-βB5 hairpin loop coincide on the same eigenvector with the βB3-αB1 loop, this could indicate that such motions are paired to each other. It would be an attractive idea to consider these regions moving in an anti-correlated fashion upon DNA binding and regulating the width of the "channel" they seem to create above the

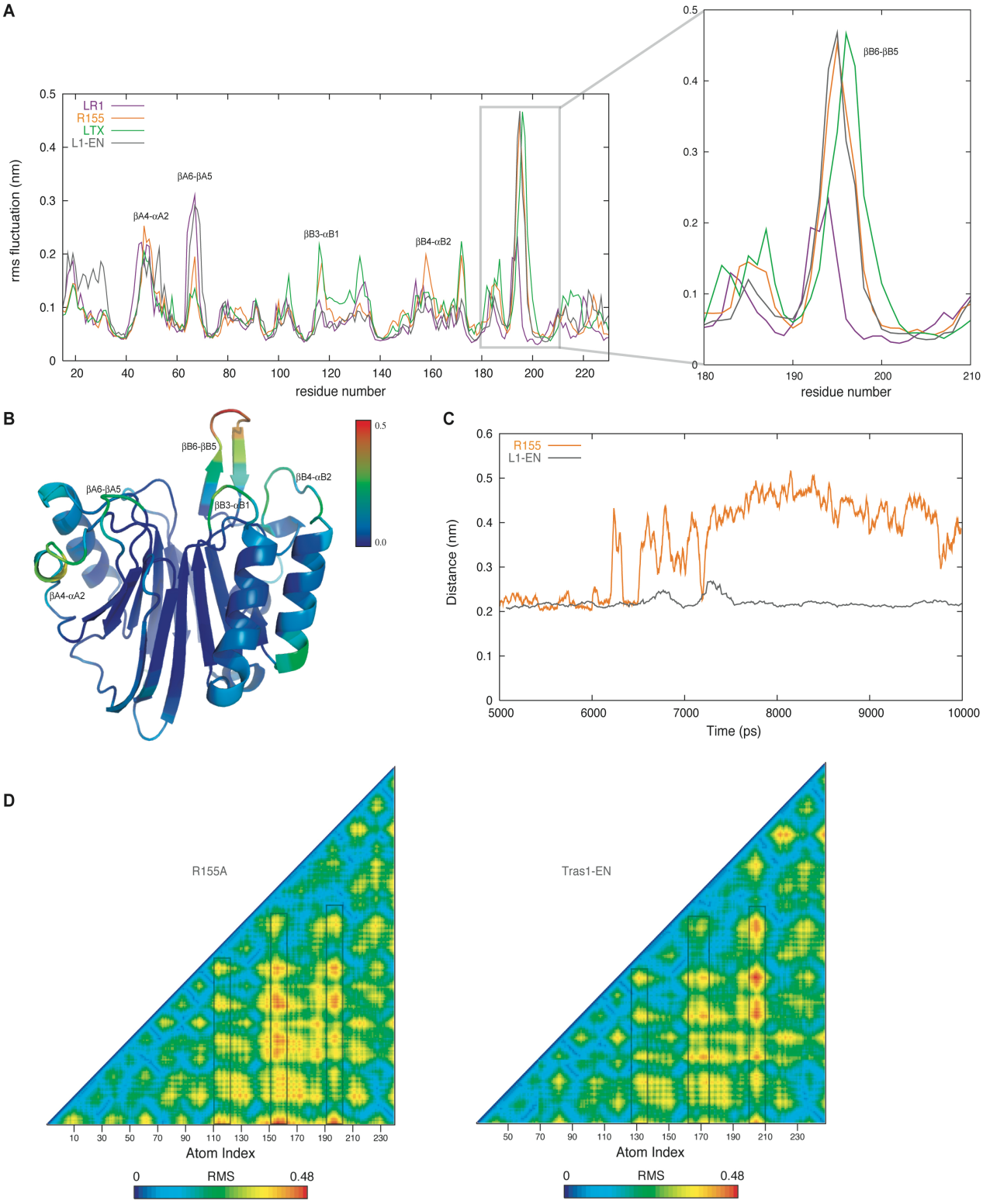


Figure 5. A: RMS fluctuations calculated over the last 6 ns of the 10 ns MD simulations of L1-EN, LTx, LR1 and R155A. On the right, blow-up of the βB6-βB5 hairpin region. **B:** Cartoon representation of L1-EN colored according to RMSF (0.0-0.5) values of R155A trajectory as an example. Flexible areas are named and correspond to the ones in A. **C:** Behavior of the D154-R182 salt-bridge during the last 5 ns of the trajectory, in the L1-EN and R155A simulations. **D:** Root mean square Cα-Cα distance fluctuation matrices for the R155A and TRAS1-EN trajectories. Flexible areas that coincide in the two proteins are boxed.

active site of L1-EN (Figure 3,blue asterisks). This would also agree with all docking models, where the hairpin inserts into the minor groove of the DNA while the N118 loop interacts from the other side with the DNA's major groove. The two loops appear to lock the DNA in position, which also agrees with our docking results.

In the LTx trajectory the motions of the exchanged hairpin are indeed so strong that they dominate the first, as well as the second eigenvector, almost at the same levels. This observation most likely indicates that this loop has enhanced plasticity, which is probably necessary for specifically sensing the target DNA geometry and positioning the substrate for subsequent cleavage, in the context of the highly specific wt Tx1L endonuclease. The opposite is true for LR1. This exchanged hairpin is much less flexible and even in the second eigenvector, where the big motion of the K70-K71 loop is removed, it shows up but in no case reaching the fluctuation levels of L1-EN, LTx or R155A. This way, we rule out the possibility that we "miss" the motion of LR1 hairpin due to another, more dominant motion and we can conclude that a combination of structural architecture and plasticity, or rather the lack of it, can explain why the LR1 variant is almost inactive in our biochemical assays.

Summary and Perspectives

The three new crystal structures of L1 endonuclease point-mutants presented here help us better understand the catalytic mechanism of DNA cleavage and the roles crucial residues play in it. Two DNA binding models are constructed by computationally docking a "wt" and a "flipped-out" double-stranded nucleic acid on the surface of L1-EN. Both models point towards a DNA orientation similar to that seen in the APE1:DNA crystal structures, possibly requiring the recognition of an extra-helical nucleotide in a pocket formed by R155, S202 and I204. All these residues reduce EN activity but not affect specificity when mutated to A, A and Y respectively. Furthermore we performed MD simulations to pinpoint surface loops that together with the prominent hairpin βB6-βB5 are flexible and involved in DNA recognition/binding, in wtL1-EN, in L1-EN loop grafts LTx and LR1, and in the related highly specific TRAS1-EN. The presence of such loops on the DNA binding interface and their degree of plasticity could be a major factor in conferring EN cleavage specificity and thus drive the integration targeting of the respective retrotransposon. A clear understanding of the above factors would greatly aid efforts to engineer/manipulate specificity of L1 and related endonucleases, hence enhancing our knowledge of retrotranspositional mechanisms and how retrotransposons could form the next generation of gene delivery vectors.

Experimental Procedures

Preparation and purification of L1-EN variants

L1-EN variants were cloned, expressed and purified as described previous (Weichenrieder et al, 2004; Repanas et al, 2007).

Crystallization

The D145A mutant was concentrated to 15mg/ml and crystallized from 22% PEG4000, 0.2 M Ammonium Sulphate (AS) and 20mM $MgCl_2$. The I204Y mutant was concentrated to 8mg/ml and crystallized from 25% PEG4000, 0.2 M Ammonium Sulphate (AS) and 20mM $MgCl_2$. The R155A mutant was concentrated to 11mg/ml and crystallized from 34% PEG4000, 0.2 M Ammonium Sulphate (AS) and 20mM $MgCl_2$.

Structure solution and refinement

Diffraction data were collected at beamline ID23-1 at the European Synchrotron Radiation Facility in Grenoble, France. Diffraction images were integrated by MOSFLM (Leslie, 2006) and scaled in SCALA (Evans, 2006). The structures were solved by molecular replacement using MOLREP (Vagin & Teplyakov, 2000) with L1-EN (PDB ID: 1VYB) as a search model. Automatic model building was done with ARP/wARP (Cohen et al, 2004) and initial refinement was done using REFMAC5 (Murshudov et al, 1997) and COOT (Emsley & Cowtan, 2004) iteratively. Although electron density maps were of good quality, refinement statistics failed to converge, with free-R ranging between 32.5% for the I204Y mutant, 30.9 % for the D145A mutant, and 27.3% for the R155A mutant. Examining the diffraction data with the program X-triage (Zwart et al, 2006) we detected significant indications for twining; possible in this apparently monoclinic but metrically close to orthorhombic space group. We used the twin operator (-h, -k, l) to refine the twin fraction in the refinement program of the phenix package (Afonine et al, 2005). The twin fraction refined to values ranging between 0.45, 0.29, 0.1 (for I204Y, D145A, R155A respectively) and in all cases resulted to considerably improved free-R values (20.8%, 25.6%, 25.2% for the I204Y, D145A and R155A mutants respectively) and improved electron density maps. All de-twinned maps were inspected and models were finalized and validated using the Molprobity server (Lovell et al, 2003).

Oligonucleotide nicking

Gel-purified synthetic oligonucleotides of a 20 base pair long A-T rich oligonucleotide that contains the preferred consensus L1-EN cleavage site 5'TTTT/AA 3' were labeled at the 5' end with radioactive phosphate (^{32}P) using [g-^{32}P]ATP and T4 polynucleotide kinase and were re-purified on gel. Equimolar amounts (450 nM) of unlabeled complementary and substrate strands were mixed with a trace amount of labeled substrate. The mixture was annealed in 5 mM Na-HEPES (pH = 7.5) by heating to 90 ºC and slow-cooled to room temperature. Nicking reactions (50 µl) were done in 50 mM Na-HEPES (pH = 6.5), 150 mM NaCl, 10 mM $MgCl_2$, 0.1 mg/ml bovine serum albumin and 1 mM dithiothreitol. Final concentrations were 180 nM DNA (0.5-7.5 µg/ml) and 20 - 2000

nM protein, which had been diluted in protein buffer (5 mM Na-HEPES (pH = 7.5), 300 mM NaCl, 10 mM $MgCl_2$, 0.5 mg/ml bovine serum albumin and 5 mM dithiothreitol) before. After 30 minutes at 37 ºC reactions were stopped by the addition of 175 µl of 380 mM Na-acetate (pH = 7.5), phenol extraction and ethanol precipitation. Reaction products were separated on 10 % denaturing polyacrylamide gels.

Docking Protocol: HADDOCK

As starting structures for the protein partner we used our L1-EN crystal structure (pdb code: 1VYB), together with two available additional structures of alternative space-group, in order to account for certain degree of flexibility within the protein. For the DNA partner we used an NMR ensemble consisting of nine structures (pdb code: 1RVI). Ambiguous Interaction Restraints (AIRs) are shown in Table 3. As active residues, we choose those residues involved in the interaction interface using experimental information extracted from sequence, structural and mutational data; for all of them a relative solvent accessibility of >50% was required, defined with NACCESS (Hubbard & Thornton, 1993). Neighbors of the active residues that could also be part of the interface are defined as passive residues and the same solvent accessibility criteria apply.

The docking protocol consists of three steps: *1*.rigid-body docking, *2*.semi-flexible simulated annealing and *3*.refinement in explicit solvent. During the initial rigid-body docking stage all combinations of different starting structures are used to create 1000 protein:DNA complexes. The best 200 of these in terms of intermolecular energy according to HADDOCK scores were selected for the simulated annealing step and the final water refinement (as described in Dominguez et al, 2003). Default HADDOCK (version 2.0) parameters were used and the top scoring complexes were further visually inspected for contacts between DNA bases and residues in the protein involved in both catalysis and DNA stabilization, as indicated in the Supplementary Information section. The HADDOCK package is freely available to academic users

(http://www.nmr.chem.uu.nl/haddock).

Molecular Dynamics Simulations

A total of five molecular dynamics (MD) simulations in explicit solvent were performed using the GROMACS 3.0 package (Lindahl et al, 2001; Van Der Spoel et al, 2005) and the GROMOS96 force field (Daura et al, 1998). L1-EN (pdb code: 1VYB), LTx, LR1, R155A and TRAS1-EN (PDB CODE: 1WDU) crystal structures were used as starting models for the MD simulations. Structures were solvated in a cubic box of SPC water (Berendsen et al, 1981) using a minimum distance of 14 Å between the protein and the box edges. After a steepest decent energy minimization step with positional restraints on the solute, negatively (Cl^-) or positively (Na^+) charged ions were introduced depending on the run, to obtain an electro-neutralized system. A second energy minimization was performed, followed by five successive 20 ps MD equilibration runs. During these, the position restraints force constant on the solute Kposre was decreased progressively (1000, 1000, 100, 10, 0 kJ mol^{-1} nm^{-2}). After these equilibration stages, a 10 ns production run was performed.

Solute, solvent and counterions were weakly coupled independently to reference temperature baths at 300 K (τ = 0.1 ps) (Berendsen et al, 1984). The pressure was maintained by coupling the system weakly to an external pressure bath at one atmosphere (1atm=101,325 Pa). The LINCS algorithm was used to constrain bond lengths, allowing an integration time step of 2 fs to be used (Hess et al, 1997). The non-bonded interactions were calculated with a twin-range cut-off of 0.8 and 1.4 nm (van Gunsteren & Berendsen, 1990). The long-range electrostatic interactions beyond the 1.4 nm cut-off were treated with the generalized reaction field model, using a dielectric constant of 54 (Tironi et al, 1995). Trajectory coordinates and energies were stored at 0.5 ps intervals. The analysis was performed routinely for the last 6 ns of each simulation, to allow for the system to start equilibrating, using the set of programs within GROMACS. Statistics for energy terms and RMSD are available in Supplementary Information.

To identify smaller scale motions that were possibly overshadowed by more dominant ones and examine possible correlations, we analyzed the projections of each trajectory on the eigenvectors of its covariance matrix. The fast, free and flexible GROMACS package is available @

http://www.gromacs.org/ .

Hardware

The GROMACS computations were performed on a cluster of 5 Apple Xserve dual G5 processors at 2.3 GHz having 1GB of RAM memory each and using Lam-MPI over gigabit ethernet as a inter-node communication implementation.

HADDOCK docking runs were performed on a Transtec (Tubingen, Germany) computer cluster operating with 32, 2.0 GHz, 64 bit Opteron processors.

REFERENCES

Afonine PV, Grosse-Kunstleve RW, Adams PD (2005) The Phenix refinement framework. *CCP4 Newsletter July, Contribution 8* .

Babushok DV, Kazazian HH (2007) Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* .

Berendsen HJC, Postma JPM, van Gunsteren WF, Di Nola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **81**: 3684-3690.

Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, Pullman B (eds) pp 331-342. Dordrecht: Reidel Publishing Company

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

Cohen SX, Morris RJ, Fernandez FJ, Ben Jelloul M, Kakaris M, Parthasarathy V, Lamzin VS, Kleywegt GJ, Perrakis A (2004) Towards complete validated models in the next generation of ARP/wARP. *Acta Crystallogr D Biol Crystallogr* **60**: 2222-2229.

Cost GJ, Boeke JD (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.

Daura X, Mark AE, van Gunsteren WF (1998) Parametrization of aliphatic CHn united atoms of GROMOS96 firce field. *Journal Computational Chemistry* **19**: 535-547.

Dlakić M (2000) Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. *Trends Biochem Sci* **25**: 272-273.

Dodson G, Verma CS (2006) Protein flexibility: its role in structure and mechanism revealed by molecular simulations. *Cell Mol Life Sci* **63**: 207-219.

Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**: 1731-1737.

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**: 2126-2132.

Evans P (2006) Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr* **62**: 72-82.

Feng Q, Moran JV, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905-916.

Feng Q, Schumann G, Boeke JD (1998) Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proc Natl Acad Sci U S A* **95**: 2083-2088.

Garrett JE, Knutzon DS, Carroll D (1989) Composite transposable elements in the Xenopus laevis genome. *Mol Cell Biol* **9**: 3018-3027.

Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**: 315-325.

Han JS, Boeke JD (2005) LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *Bioessays* **27**: 775-784.

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A Linear Constraint Solver for Molecular Simulations. *Journal of Computational Chemistry* **18**: 1463-1472.

Hubbard SJ, Thornton JM (1993) NACCESS. In . London:Department of Biochemistry and Molecular Biology, University College:

Kalodimos CG, Biris N, Bonvin AM, Levandoski MM, Guennuegues M, Boelens R, Kaptein R (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305**: 386-389.

Kamphuis MB, Bonvin AM, Monti MC, Lemonnier M, Muñoz-Gómez A, van den Heuvel RH, Díaz-Orejas R, Boelens R (2005) Model for RNA Binding and the Catalytic Site of the RNase Kid of the Bacterial parD Toxin-Antitoxin System. *J Mol Biol* **357**: 115-126.

Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* **102**: 6679-6685.

Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**: 646-652.

Kopke Salinas R, Folkers GE, Bonvin AM, Das D, Boelens R, Kaptein R (2005) Altered Specificity in DNA Binding by the lac Repressor: A Mutant lac Headpiece that Mimics the gal Repressor. *Chembiochem* **6**: 1628-1637.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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 JP, 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, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Rein-

hardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Leslie AG (2006) The integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr* **62**: 48-57.

Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analisys. *Journal Molecular Modeling* **7**: 306-317.

Lovell SC, Davis IW, Arendall WB, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**: 437-450.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (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.

Maita N, Anzai T, Aoyagi H, Mizuno H, Fujiwara H (2004) Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J Biol Chem* **279**: 41067-41076.

McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* **267**: 585-590.

Melo LF, Mundle ST, Fattal MH, O'Regan NE, Strauss PR (2007) Role of active site tyrosines in dynamic aspects of DNA binding by AP endonuclease. *DNA Repair (Amst)* **6**: 374-382.

Mol CD, Izumi T, Mitra S, Tainer JA (2000) DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected]. *Nature* **403**: 451-456.

Mol CD, Kuo CF, Thayer MM, Cunningham RP, Tainer JA (1995) Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* **374**: 381-386.

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**: 240-255.

Repanas K, Zingler N, Layer LE, Schumann GG, Perrakis A, Weichenrieder O (2007) Determinants for target selectivity of the human LINE-1 retrotransposon endonuclease . *Submitted* .

Rueda M, Ferrer-Costa C, Meyer T, Pérez A, Camps J, Hospital A, Gelpí JL, Orozco M (2007) A consensus view of protein dynamics. *Proc Natl Acad Sci U S A* **104**: 796-801.

Stefl R, Wu H, Ravindranathan S, Sklenár V, Feigon J (2004) DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. *Proc Natl Acad Sci U S A* **101**: 1177-1182.

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD (2002) Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**: 327-338.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3**: research0052.

Theobald DL, Wuttke DS (2006) THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* **22**: 2171-2172.

Tironi IG, Sperb R, Smith PE, Vangunsteren WF (1995) A Generalized Reaction Field Method for Molecular-Dynamics Simulations. *Journal of Chemical Physics* **102**: 5451-5459.

Vagin A, Teplyakov A (2000) An approach to multi-copy search in molecular replacement. *Acta Crystallogr D Biol Crystallogr* **56**: 1622-1624.

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comput Chem* **26**: 1701-1718.

van Dijk AD, Fushman D, Bonvin AM (2005) Various strategies of using residual dipolar couplings in NMR-driven protein docking: Application to Lys48-linked di-ubiquitin and validation against (15)N-relaxation data. *Proteins* **60**: 367-381.

van Dijk M, van Dijk AD, Hsu V, Boelens R, Bonvin AM (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res* **34**: 3317-3325.

van Gunsteren WF, Berendsen HJ, Hermans J, Hol WG, Postma JP (1983) Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proc Natl Acad Sci U S A* **80**: 4315-4319.

van Gunsteren WF, Berendsen HJC (1990) Computer-Simulation of Molecular-Dynamics - Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie-International Edition in English* **29**: 992-1023.

Weichenrieder O, Repanas K, Perrakis A (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure (Camb)* **12**: 975-986.

Weston SA, Lahm A, Suck D (1992) X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J Mol Biol* **226**: 1237-1256.

Xiong Y, Eickbush TH (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.

Zingler N, Weichenrieder O, Schumann GG (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res* **110**: 250-268.

Zwart PH, Grosse-Kunstleve RW, Adams PD (2006) Characterization of X-ray data sets. *In Preparation* .

Supplementary Information

Supplementary Table 1: RMSD (in Å) of all structures after pairwise maximum likelihood superposition

		L1-EN A	L1_EN B	LTx A	LTx B	LR1 A	R155 A	R155 B	R155 C	R155 D	D145 A	D145 B	D145 C	D145 D	I204 A	I204 B	I204 C	I204 D
L1-EN	A		0.31	0.70	0.71	0.36	0.45	0.32	0.29	0.49	0.47	0.44	0.47	0.46	0.43	0.50	0.37	0.45
L1-EN	B	0.31		0.62	0.63	0.37	0.40	0.33	0.32	0.45	0.50	0.47	0.52	0.44	0.45	0.49	0.44	0.43
LTx	A	0.70	0.62		0.17	0.62	0.61	0.70	0.69	0.70	0.78	0.77	0.73	0.69	0.76	0.79	0.74	0.75
LTx	B	0.71	0.63	0.17		0.64	0.62	0.70	0.70	0.71	0.78	0.77	0.75	0.71	0.76	0.78	0.74	0.76
LR1	A	0.36	0.37	0.62	0.64		0.45	0.40	0.38	0.46	0.46	0.50	0.46	0.46	0.45	0.49	0.47	0.43
R155	A	0.45	0.40	0.61	0.62	0.45		0.38	0.36	0.31	0.53	0.53	0.53	0.46	0.53	0.50	0.48	0.51
R155	B	0.32	0.33	0.70	0.70	0.40	0.38		0.25	0.42	0.38	0.41	0.51	0.44	0.41	0.48	0.41	0.43
R155	C	0.29	0.32	0.69	0.70	0.38	0.36	0.25		0.41	0.40	0.41	0.49	0.41	0.40	0.47	0.40	0.41
R155	D	0.49	0.45	0.70	0.71	0.46	0.31	0.42	0.41		0.53	0.52	0.59	0.52	0.56	0.52	0.55	0.55
D145	A	0.47	0.50	0.78	0.78	0.46	0.53	0.38	0.40	0.53		0.46	0.47	0.49	0.52	0.61	0.48	0.53
D145	B	0.44	0.47	0.77	0.77	0.50	0.53	0.41	0.41	0.52	0.46		0.44	0.48	0.50	0.58	0.52	0.54
D145	C	0.47	0.52	0.73	0.75	0.46	0.53	0.51	0.49	0.59	0.47	0.44		0.43	0.57	0.59	0.58	0.53
D145	D	0.46	0.44	0.69	0.71	0.46	0.46	0.44	0.41	0.52	0.49	0.48	0.43		0.57	0.57	0.55	0.52
I204	A	0.43	0.45	0.76	0.76	0.45	0.53	0.41	0.40	0.56	0.52	0.50	0.57	0.57		0.43	0.41	0.44
I204	B	0.50	0.49	0.79	0.78	0.49	0.50	0.48	0.47	0.52	0.61	0.58	0.59	0.57	0.43		0.44	0.37
I204	C	0.37	0.44	0.74	0.74	0.47	0.48	0.41	0.40	0.55	0.48	0.52	0.58	0.55	0.41	0.44		0.41
I204	D	0.45	0.43	0.75	0.76	0.43	0.51	0.43	0.41	0.55	0.53	0.54	0.53	0.52	0.44	0.37	0.41	

Supplementary Table 2: HADDOCK clusters characteristics[a]

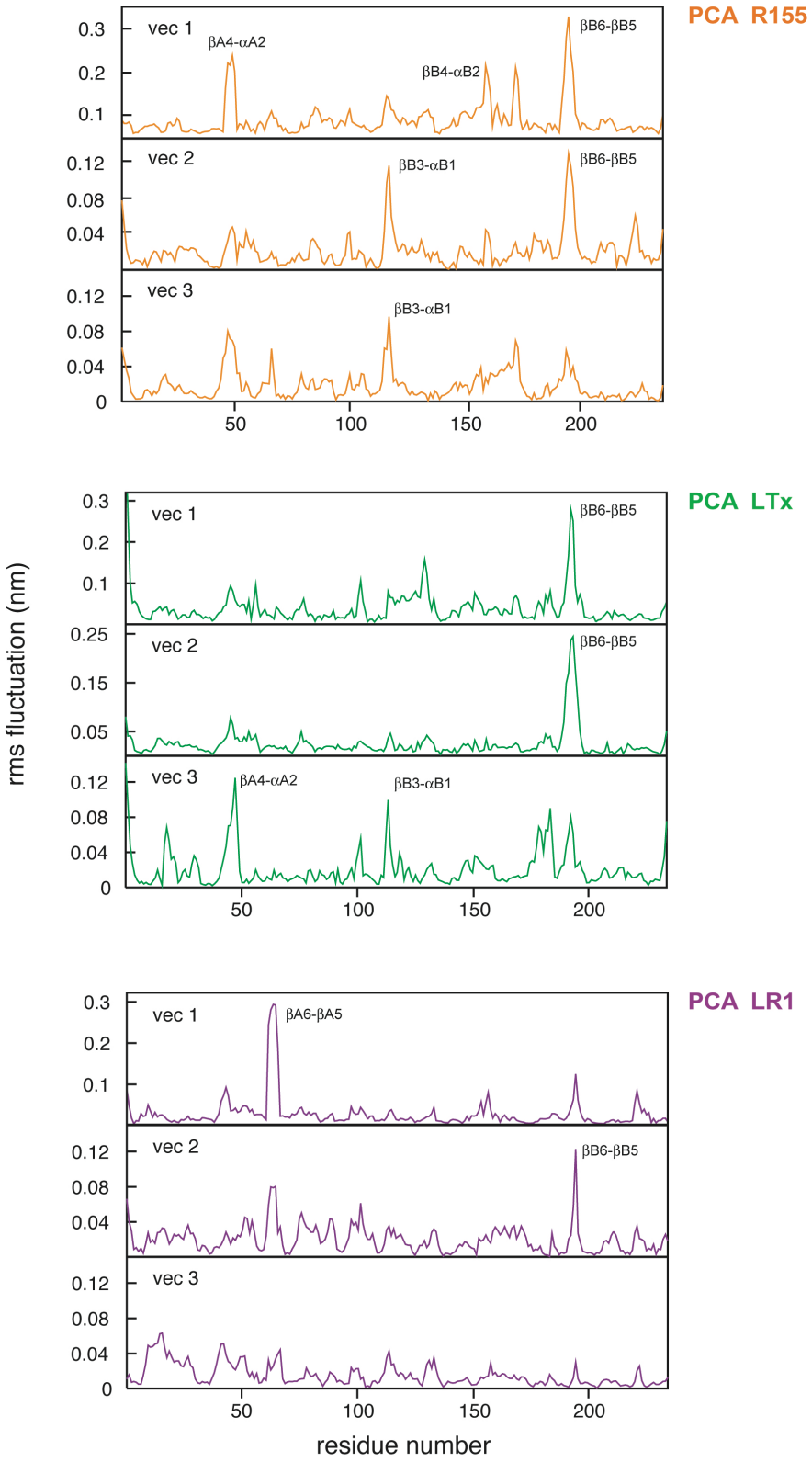
Run-type	Cluster	N_{str}	RMSD	Score	BSA [Å^2]	E_{air} [kcal/mol]
Flipped	**1**	**35**	**0.6 ± 0.1**	**-12.8 ± 6.2**	**1256 ± 67**	**269 ± 7**
	2	31	0.7 ± 0.1	-38.7 ± 7.6	1767 ± 158	359 ± 24
	3	24	0.8 ± 0.1	-56.3 ± 3.4	1863 ± 74	392 ± 19
Non-flipped	1	35	0.5 ± 0.2	-58.4 ± 6.0	1916 ± 65	268 ± 5
	2	**33**	**0.7 ± 0.1**	**-49.0 ± 6.0**	**1934 ± 107**	**269 ± 10**
	3	10	0.8 ± 0.1	-34.3 ± 9.2	1802 ± 177	285 ± 18

[a]: N_{str} , number of structures in cluster; RMSD, average root mean square positional deviation from the best score structure within each cluster; BSA , buried surface area; Score, HADDOCK score; **Bold** designates clusters selected as best, as explained in the text.

The supplementary table depicts various characteristics and energy terms for the three best clusters (concerning number of structures) obtained after the two different HADDOCK runs with “flipped” and “non-flipped” (wild-type) DNA substrates. The docking solutions were clustered based on positional RMSD using a 4 Å cut-off; only clusters with at least 4 members were analyzed, but not depicted in this table. Averages were calculated over the best 10 structures of each cluster to remove differences originating from cluster size.

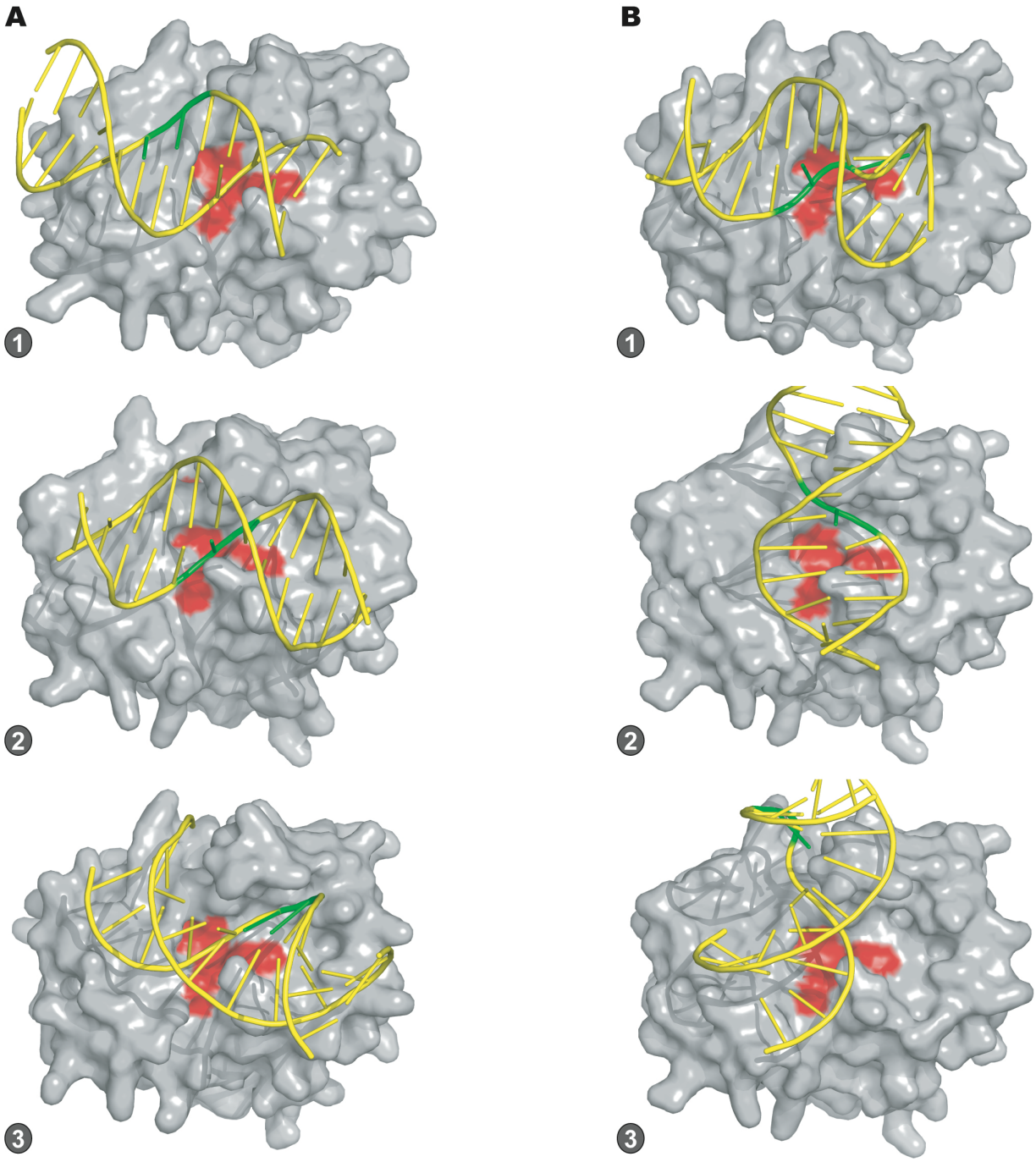
Final selection of the best cluster was based on a combination of factors such as: Haddock score, highly populated clusters and geometrical positioning of DNA substrate to facilitate catalysis (as illustrated in supplementary figure 2). The geometrical requirements are getting additional weight in the final selection, in order to filter out solutions that score well but do not favor DNA nicking.

Supplementary Figure 1



Supplementary Figure 1: Principal components analysis results for R155A, LTx and LR1 variants. Projection of the Cα atom fluctuation along the first three eigenvectors are plotted for each trajectory. Loop names as in Figure 5.

Supplementary Figure 2



Supplementary Figure 2. Visualization of best HADDOCK resulting models. L1-EN is drawn as a grey surface representation and is viewed from the top, so that the βB6-βB5 hairpin is pointing towards the reader. The double-stranded A-tract DNA substrate is drawn as a yellow cartoon, with the TpA step that is the supposed target for nicking highlighted in green. A red surface indicates the position of active site residues of L1-EN involved in catalysis. Bear in mind that these catalytic residues were not used in the definition of the ambiguous interaction restraints, in order to avoid bias of the docking solutions. **A:** Run using "non-flipped" (wild-type) DNA substrate. Models depicted are the representatives of cluster 1 (1), cluster 2 (2) and cluster 3 (3), as numbered in Supplementary Table 2. The model from cluster 2 better satisfies the requirements for catalysis, since the targeted scissile phosphodiester bond (green), is favorably placed in the active site of L1-EN (red). **B:** Run using "flipped" DNA substrate. Models depicted are the representatives of cluster 1 (1), cluster 2 (2) and cluster 3 (3). The model from cluster 1 better satisfies the same requirements for catalysis as explained in **A**.

Supplementary Table 3: MD simulations statistics

Structural and energy[a] statistics for the five MD simulations (4-10 ns).

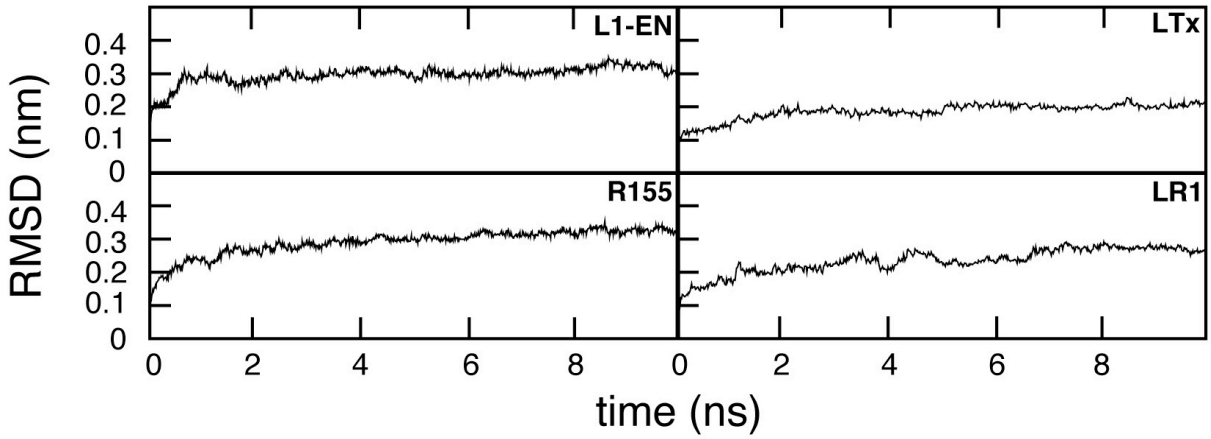
	Average heavy atoms RMSD[b] with respect to the starting structure (nm)	Coulomb's electrostatic energy (kJ/mol)			Lennard-Jones (van der Waals) energy (kJ/mol)		
		intra-protein	protein-solvent	Total	intra-protein	protein-solvent	Total
L1-EN	0.31 (0.01)	-29072 (1240)	-22531 (1814)	-51603 (2197)	-8735 (88)	-1486 (124)	-10222 (152)
R155A	0.31 (0.01)	-27090 (1563)	-25787 (2688)	-52878 (3109)	-8669 (88)	-1527 (120)	-10196 (149)
LR1	0.19 (0.02)	-25913 (2170)	-29923 (3479)	-55836 (4100)	-8486 (81)	-1390 (114)	-9877 (140)
LTx	0.24 (0.02)	-26908 (1702)	-25421 (2564)	-52330 (3077)	-8782 (88)	-1516 (120)	-10298 (149)
TRAS1-EN (resi 68-245)*	0.26 (0.03)*	-22885 (1653)	-27581 (2749)	-50465 (3208)	-7698 (87)	-1454 (113)	-9152 (143)

[a]The non-bonded energies were calculated with the GROMOS96 force field using a twin range cutoff of 0.8 and 1.4 nm with a reaction field correction. The energies are the sum of short- (SR) and long-range (LR) terms; 1-4 terms have not been included. Standard deviations are indicated in parentheses.

[b]The average heavy atom positional RMSD values were calculated with respect to the starting frame for each simulation after superposition on the backbone atoms.

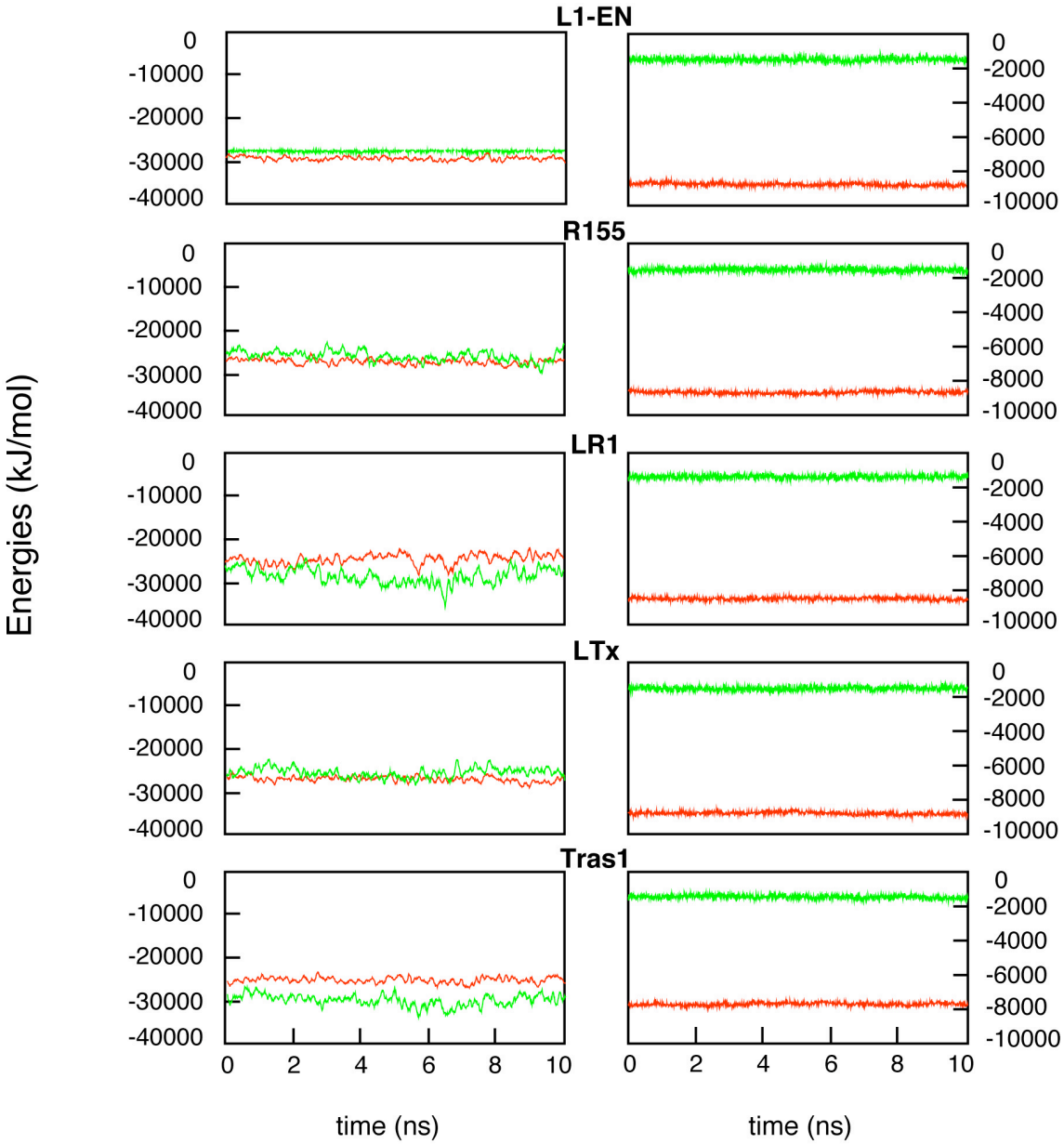
*Only residues 68-245 were included in the analysis of the Tras1-EN trajectory to avoid bias due to a missing region in the original PDB file between residues 59-68.

Supplementary Figure 3



Supplementary Figure 3. Time evolution of the RMSD from the starting structure during the 10 ns simulations of L1-EN, R155A, LTx and LR1.

Supplementary Figure 4



Supplementary Figure 4. Evolution of the electrostatic (left panel) and Van der Waals (right panel) energy components during the five 10 ns simulations. Red curves indicate intra-protein energies and green curves indicate protein-solvent energies as listed in the Supplementary Table 3.

Chapter 5

LINE-1 endonuclease friends and family:
Structural and functional connections in a family of metal-dependent phosphohydrolases

Kostas Repanas, Oliver Weichenrieder and Anastassis Perrakis

To be submitted for publication

LINE-1 endonuclease friends and family: Structural and functional connections in a family of metal-dependent phosphohydrolases

Kostas Repanas, Oliver Weichenrieder and Anastassis Perrakis

Introduction

In this mini-review we discuss structural and functional relationships among proteins that belong to the extended enzyme family of metal-dependent phosphohydrolases. This is the family to which the endonuclease of the L1 retrotransposon belongs, alongside many other diverse and functionally unrelated proteins. The focus is on ten family members for which a three-dimensional structure is available.

This sundry group includes "founding" member bovine pancreatic deoxyribonuclease I (DNAseI), DNA repair apurinic/apyrimidinic (AP) endonucleases, sphingomyelinases and retrotransposon-encoded AP-like endonucleases, alongside an inositol phosphatase, a bacterial genotoxin and a nitrophorin from the saliva of blood-sucking insects. Of course, the family encompasses many more proteins for which no structural information is available; for instance the yeast carbon catabolite repressor protein (Ccr4p) and the vertebrate circadian-clock-regulated protein nocturnin. It is quite striking that in general, although members of the family are responsible for truly unrelated functions and share less than 20% sequence identity, they still use the same set of catalytic residues (Figure 1) (Dlakić, 2000).

All members of this extended family share a common DNaseI-like fold, which consists of a four-layered α/β sandwich. A well-maintained core of opposing β-sheets is flanked by α-helices and topped on the substrate binding interface with surface loops, which direct protein function by varying in number, length and flexibility (Figure 2). Many of the key residues in and around the active site are conserved; still the enzymes cleave a variety of phosphoester substrates, or not, like for instance in the case of nitrophorin where it is actually not clear if the protein makes use of its preformed active site.

Here we address all members initially individually, reporting the status of research for each protein thus far and later proceed to make comparisons regarding their structural fold, the existence of a conserved active site within a not-so-well-conserved protein sequence and the importance of metal ion(s) for protein function. A question that without doubt arises is whether the members of this family all emerge from different ancestors that at a certain point in evolutionary time adopted a similar fold, or whether there once existed a common ancestor and due to evolutionary pressure certain differences started accumulating, to finally give rise to a variety of proteins with unique individual attributes (Dlakić, 2000; Ofran & Margalit, 2006).

Deoxyribonuclease I

DNAseI is the founding member of this family and one of the first crystal structures to have been determined. This glycoprotein from bovine pancreas is an endonuclease that hydrolyzes the P-O3' bond in various DNA substrates. The enzyme exhibits strong preference for double-stranded substrates and cleaves DNA with varying cutting rates, which indicates that it recognizes sequence-dependent structural discrepancies on the DNA. The cutting frequency of DNase I depends on both local and global helix parameters (Suck et al, 1984; Suck & Oefner, 1986).

A number of co-crystal structures with substrate have illustrated that the enzyme binds to the DNA in the minor groove and the sugar-phosphate backbones of both strands, thus forcing the minor groove to widen around 3Å. The combination of biochemical and structural data confirmed that flexibility and minor groove geometry are crucial factors in determining DNase I activity. Contacts of the enzyme with the DNA duplex extend over a total of six base-pairs and enzyme activity is optimal in the presence of Ca^{2+}, Mg^{2+} or Mn^{2+}. There is speculation that a second active site exists, situated about 15Å away from the one initially identified. DNAseI has been studied extensively and now constitutes an established tool in the molecular biology laboratory (Suck et al, 1988; Kabsch et al, 1990; Weston et al, 1992; Suck, 1994).

Base excision repair enzymes: *APE1 and ExonucleaseIII*

The most common structural abnormality in cellular DNA is the breaking of the N-glycosylic bond that consequently results in production of AP sites. Such damaged sites are created both spontaneously or by damage-specific DNA glycosylases, but can be repaired

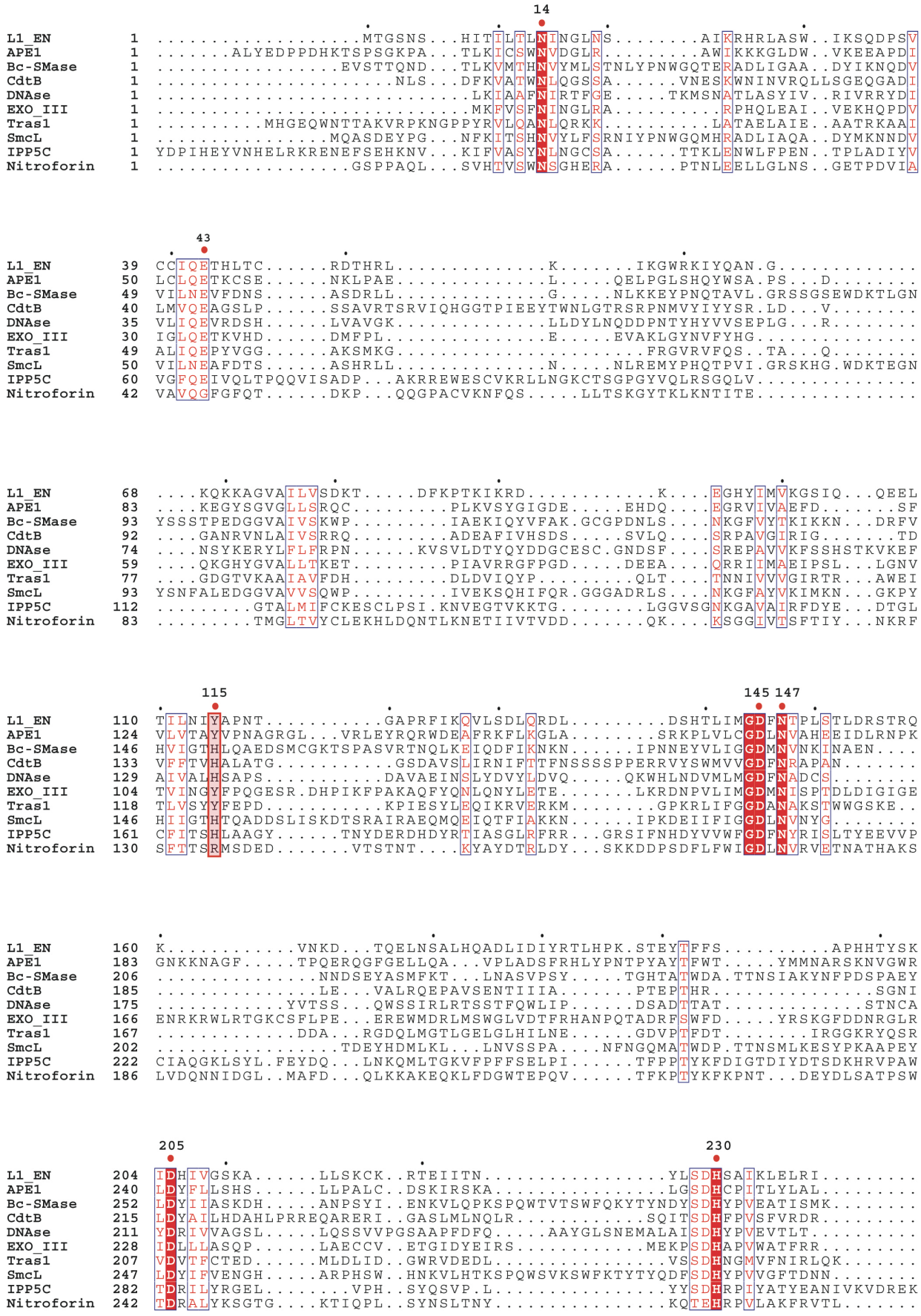


Figure 1. Structure based sequence alignment. The ten proteins were aligned using both sequence and structure information in the program T-Coffee (Expresso). Absolutely conserved residues are white and boxed with red shading. Semi-conserved residues are colored red and boxed blue. Red box and magenta shading indicates the Y/H, part of the active site. All residues of the conserved active site are numbered as in L1-EN and dotted red.

through the base excision repair (BER) pathway. Two well-studied enzymes, critical for initiation of BER, are the multifunctional exonuclease III (Exo III) in *Escherichia coli* and the major human AP endonuclease 1 (APE1) (Barzilay & Hickson, 1995).

These enzymes detect, recognize, and cleave the DNA phosphodiester backbone 5' of AP sites to create a free 3'-OH end for DNA polymerase repair synthesis. In humans, AP sites are processed by APE1, which is homologous to the *E.coli* enzyme Exo III. Exo III additionally contains 3'-repair diesterase, 3'to 5' exonuclease, 3'-phosphomonoesterase and ribonuclease activities (Demple & Harrison, 1994). APE1 is the best studied enzyme of the two and a number of high resolution crystal structures were reported of the protein in the apo- form and also in complex with DNA, both of a catalytic intermediate and of a product complex. APE1 binds abasic DNA in a pre-formed surface, inserts two loops in the minor and major grooves and in this way severely kinks the DNA helix in order to flip-out the AP site, lock the substrate in position, cleave the sugar-phosphate backbone and prime DNA repair synthesis (Mol et al, 2000).

Currently three different mechanisms are proposed as to how the nucleophilic attack might occur: first that H309 is the one generating the attacking nucleophile, facilitated by a metal ion (M1, Figure 3C) (Mol et al, 1995; Gorman et al, 1997), then the complex of APE1 bound to abasic DNA suggests that D210 might actually play the histidine part, also aided by a single metal ion (Mol et al, 2000), while a third mechanism dictates that a second metal ion (M2) coordinated by D210, N212 and H309 indeed creates the attacking nucleophile (Beernink et al, 2001). Recently, a fourth possibility was added to the above, proposing that the phenolate of Y115 could be the nucleophile attacking the scissile phosphate (Mundle et al, 2004; Melo et al, 2007).

AP-like retrotransposon-encoded endonucleases: *L1-Endonuclease and TRAS1-Endonuclease*

The human L1 retrotransposon encodes ORF2p which contains both reverse-transcriptase and endonuclease activities. It has been shown that the L1 endonuclease (L1-EN) has nuclease activity but is not an AP enzyme like APE1 or ExoIII. Furthermore, L1-EN is essential for retrotransposition in cultured cells and prefers the 5' TTAAAA 3' consensus target sequence that corresponds to sites of *de novo* L1 insertions *in vivo* (Feng et al, 1996; Cost & Boeke, 1998; Cost et al, 2002). We have previously solved the crystal structure of L1-EN, that illustrates how a prominent β-hairpin on the DNA binding interface could sense the wide minor groove at the TpA step in T_nA_n sequences and enable the enzyme to orient and nick the target duplex possibly by accommodating an extra-helical nucleotide in a surface pocket. By mutating key active site or peripheral residues and exchanging prominent loops for others of related but highly specific endonucleases, we were able to identify ways to manipulate and direct the specificity of the L1 endonuclease and possibly of the whole retro-element, although additional host factors might also play a role (Weichenrieder et al, 2004; Repanas submitted). By determining crystal structures of these mutants and performing computational studies we further showed that flexible loops on the surface of the endonuclease could be major players in DNA recognition and nicking, conferring stronger or looser specificity in different retrotransposon-encoded endonucleases and possibly reflecting a pattern encountered throughout the whole family of phosphohydrolases (Repanas submitted).

In the silkworm *Bombyx mori* we find many telomeric repeat-specific retrotransposons, like TRAS1, TRAS3, TRAS4, TRASW and SART1. TRAS1 elements are abundantly transcribed and actively retrotransposed into TTAGG telomeric repeats in a highly sequence-specific manner (Okazaki et al, 1995). To ensure correct propagation, TRAS1 retrotransposon encodes an ultra specific endonuclease (TRAS1-EN) that is responsible for recognizing and digesting the target telomeric DNA. Biochemical experiments have confirmed that it is mainly the endonuclease domain that determines the target-site specificity of that retrotransposon, recognizing less than 10 base pairs around the initial cleavage site. Similar to L1-EN, the TRAS1 endonuclease creates a cleavage first at the bottom strand of the target, which is then followed by a subsequent nick at the top strand according to the current target primed reverse transcription (TPRT) model (Anzai et al, 2001; Maita et al, 2004; Fujiwara et al, 2005). Both of these proteins share low sequence identity but high structural similarity with APE1 and ExoIII.

Neutral Sphingomyelinases: *SmcL and Bc-SMase*

Members of the family of neutral sphingomyelinases (N-SMases) are thought to be the major mediators for the stress-induced production of ceramide. Although a common link exists, that of the need for phosphodiester bond cleavage, SMases are unique compared to other DNaseI family members. Instead of targeting nucleic acids, they are responsible for the hydrolysis of the membrane phospholipid sphingomyelin (SM) to ceramide and phosphocholine. The reaction catalyzed by N-SMases resembles that of the phospholipases, where a phosphodiester bond is hydrolyzed, while removing a soluble molecule from the insoluble lipid. Apart from phospholipase C activity they also exhibit hemolytic activity (Matsuo et al, 1996).

The three-dimensional crystal structures of sphingomyelinase C from *Listeria ivanovii* (SmcL) and from *Bacillus cereus* (Bc-SMase), confirm previous predictions suggesting that they both belong to the DNaseI superfamily of metal-dependent phosphohydrolases. Unfortunately, neither of the two structures contains the SM substrate bound to the protein, so details regarding catalysis are still vague. An SM moiety though could be modeled in the active site, using information from a bound phosphate ion on SmcL and extrapolating from the

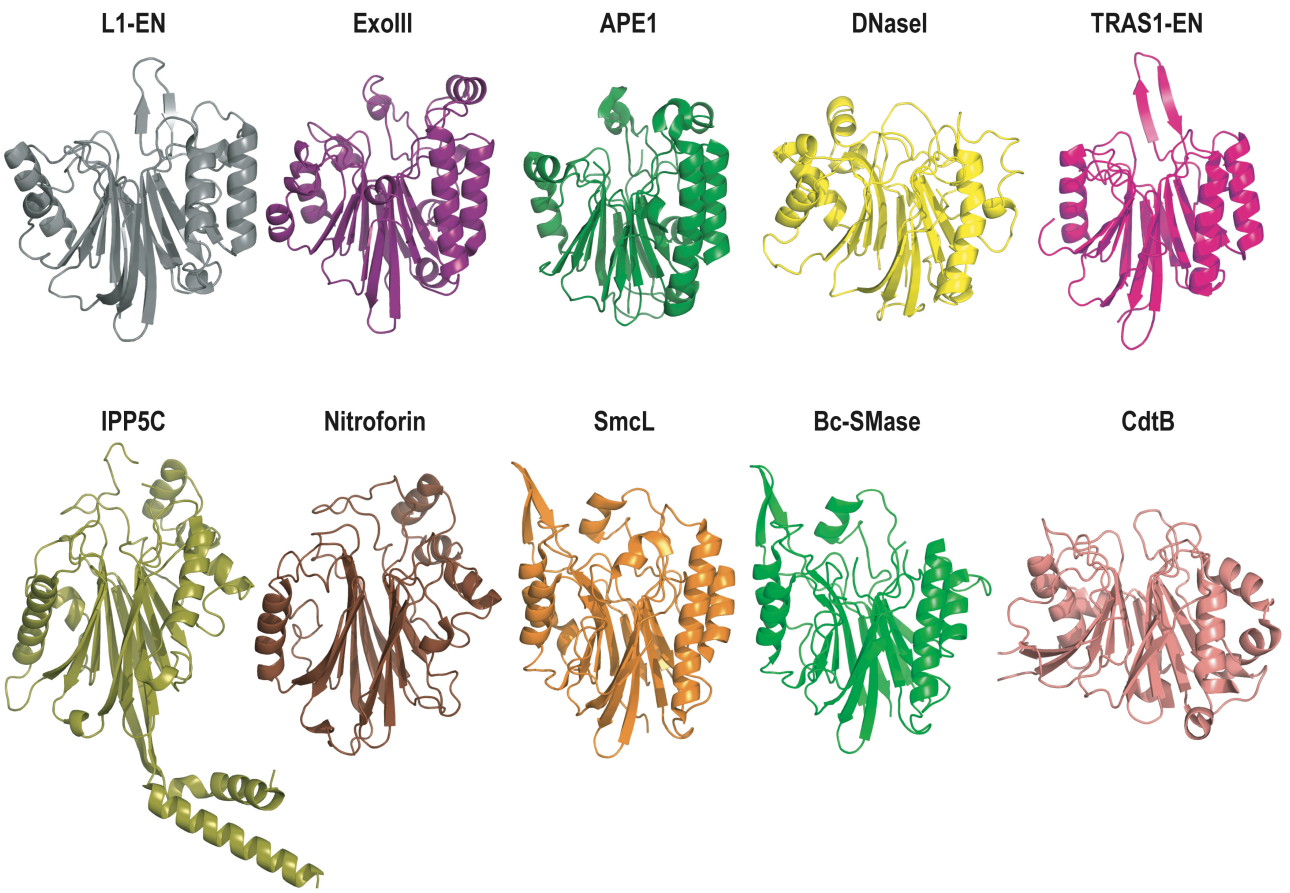


Figure 2. Cartoon drawings showing secondary structure elements in the ten members of the metal-dependent phosphohydrolases family presented here. Same orientation for all-substrate binding on top. *Grey:* L1-EN, *Deeppurple:* ExoIII, *Dark green:* APE1, *Yellow:* DNAseI, *Magenta:* TRAS1-EN, *Olive:* IPP5C, *Chocolate:* Nitroforin, *Orange:* SmcL, *Green:* Bc-SMase, *Salmon:* CdtB. Color coding follows throughout the paper.

APE1:DNA complex. In this way, several key residues could be identified as catalytically essential and it was further shown biochemically that Mg^{2+} and Co^{2+} are favoring activity, while Ca^{2+} is acting as an inhibitor. Another distinct feature of bacterial SMases is the presence of an elongated and highly hydrophobic β-hairpin, which together with other hydrophobic surface loops mediates docking to the cell membrane as well as properly orienting SM in the active site cleft (Matsuo et al, 1996; Openshaw et al, 2005; Ago et al, 2006; Clarke et al, 2006).

Inositol polyphosphate 5-phosphatase (IPP5C)

Inositol polyphosphate 5-phosphatases (IP5P) hydrolyze the phosphate at position number five in the inositol ring of several signaling molecules derived from phosphatidylinositol. IP5Ps are central in the regulation of phosphoinositide signaling and in the pathogenesis of human diseases. Moreover, these proteins are involved in regulating cell growth, apoptosis, cytoskeletal organization, intracellular calcium signaling, and post-synaptic vesicular trafficking (Majerus, 1992).

Four different groups of inositol polyphosphate 5-phosphatases exist depending on substrate preference; the available crystal structure of *Schizosaccharomyces pombe* synaptojanin (SPsynaptojanin) belongs to group II, catalyzing both soluble and lipid inositol phosphates and participating in synaptic vesicle trafficking. This type of enzyme has an N-terminal Sac1-like domain encoding polyphosphoinositide phosphatase activity, a central 5-phosphatase domain recognizing mainly inositol (4,5)-biphosphate targets, and finally a proline rich region at the C-terminus. The two structures crystallized are those of the apo- form of the protein, as well as that complexed with the reaction product inositol (1,4)-biphosphate (Majerus et al, 1999; Whisstock et al, 2000).

The crystal structure confirms previous predictions that the fold and catalytic mechanism is conserved compared to DNAseI, APE1, ExoIII and other family members. It further illustrates how the conserved DNAseI catalytic core has expanded to create a subfamily of inositol polyphosphate enzymes. A great puzzle regarding inositol phosphatases is their site selectivity, the question being why for instance 5-phosphatases only dephosphorylate the 5 position of a substrate that might be phosphorylated at positions 3, 4, 5 and combinations. In the product complex, the 4-phosphate of $Ins(1,4)P_2$ is misoriented by 4.6Å compared to the optimal reactive geometry observed in the APE1, explaining the dephosphorylation site selectivity of the 5-phosphatases, i.e.

why 5-phosphatases are not 4-phosphatases. A recent study illustrates that SPsynaptojanin might encompass much broader substrate specificity than previously appreciated, suggesting that it most likely performs multiple roles in cell signaling and may regulate distinct pathways (Tsujishita et al, 2001; Chi et al, 2004).

Bacterial genotoxin (CdtB)

The cytolethal distending toxin (CDT) is a tripartite bacterial toxin that initiates a eukaryotic cell cycle block at the G2 stage prior to mitosis and is widely distributed among gram-negative bacteria including *Escherichia coli, Campylobacter spp., enterohepatic Helicobacter spp., Actinobacillus actinomycetemcomitans* and *Haemophilus ducreyi*. Subunits CdtA and CdtC associate with nuclease CdtB forming a holotoxin, which enables delivery of CdtB into the host cell where the enzyme creates DNA lesions acting as a genotoxin (Lara-Tejero & Galán, 2001; Ceelen et al, 2006).

The crystal structure of CDT revealed that CdtB is a member of the DNaseI family and is bound to two ricin-like lectin domains, thus forming a ternary complex with three independent molecular interfaces. CdtA and CdtC create a deeply grooved, highly aromatic surface critical for toxicity. CdtB is the active subunit of the holotoxin, adopting the DNaseI-like fold and binding to the DNA. Once delivered inside the cell, CdtB enters the nucleus and exhibits endonuclease activity that results in DNA double-strand breaks. Moreover it has been shown that CdtB demonstrates nuclease activity *in vitro* and *in vivo*. What the three-dimensional structure further revealed was that the N terminus of subunit CdtC produced a steric block at the active site of CdtB, inhibiting its DNase activity. This could be a self-regulatory mechanism for the holotoxin, possibly necessary to preclude random DNase activity (Lara-Tejero & Galán, 2002; Nesić et al, 2004; Hu et al, 2006).

Nitrophorin from bloodsucking insects (cNP)

Nitrophorins are proteins found in the saliva of blood-feeding insects and are responsible for nitric oxide (NO) delivery, transport and storage. Delivery of the reactive molecule NO while feeding is known to induce vasodilation and inhibit blood coagulation. The *Cimex lectularius* (the bedbug) nitrophorin (cNP) stores and releases NO in a pH-dependent manner by using a ferric heme protein but does so with a protein evolutionarily unrelated to the well studied *Rhodnious prolixus* (the kissing bug) nitrophorins (Weichsel et al, 1998; Weichsel et al, 2000). Unlike the Rhodnius nitrophorins, the Cimex nitrophorin does not bind histamine, normally secreted by the victim as a response to the bite, but rather binds two molecules of NO reversibly, one to the heme and the other to the cysteine thiolate, which binds the heme when NO is absent (Walker, 2005).

The crystal structure of cNP determined to 1.75Å resolution confirmed the distant relationship between these two types of nitrophorins and further indicated the high level of structural similarity with the DNaseI family of functionally unrelated proteins. The protein fold is an extensive β sandwich motif, with the bound heme inserted into a distal hydrophobic pocket above one face of the sandwich. The structure of this protein also illustrates that the active site of nitrophorin, although similar to that of other members, does not participate to the reversible metal-assisted S-nitroso formation by reduction of the heme iron. Instead, the hydrophobic "active" pocket that binds the ferric heme is around 17Å away from the literally "inactive" site (Weichsel et al, 2005).

Structural comparison

All ten protein structures of family members discussed here, form a similar four-layered α/β sandwich, with the two β-sheets situated in the inside and flanked by a number of α-helices on the outside. The protein core formed by the two anti-parallel β-sheets is maintained among the different proteins, although the number and length of the β-strands might vary. The biggest differences however, are observed regarding loop regions found at the substrate-binding surface, which is always seen on the top in the current view to facilitate comparison (Figure 2).

The overall fold forms a long groove at one end of the α/β sandwich similar to DNaseI, but it is worth noting the differences on the surface loops, when different substrates need to be targeted. For example, the biggest difference in ExoIII, APE1 and DNAseI is that in the case of the unspecific DNAseI the surface loops surrounding the groove are missing. Proteins like the L1 and TRAS1 endonuclease on the other hand, need to target a similar T-A rich DNA substrate and seem to have developed a prominent hairpin in the middle of the top surface (Figure 2). This type of hairpin loop, together with smaller surface loops appear to be responsible for the specificity of the endonucleolytic cleavage. Specific contacts, but also overall structure and flexibility of such loops can direct the specificity of the respective endonuclease.

It is also noteworthy that when necessary, surface loops have evolved in a different way in order to assist function. A similar, prominent hydrophobic loop is found on the surface of bacterial SMases. This loop is responsible for mediating binding to the cell membrane and is not found in other family members. It could be though similar to the loop of IPP5C that also needs to interact with the membrane, but this loop is not fully modeled due to disorder in the crystal structure. These similar enzymes have probably evolved on a common, suitably functional scaffold, while the variety of surrounding loops direct the specificity of substrate selectivity, binding and catalysis.

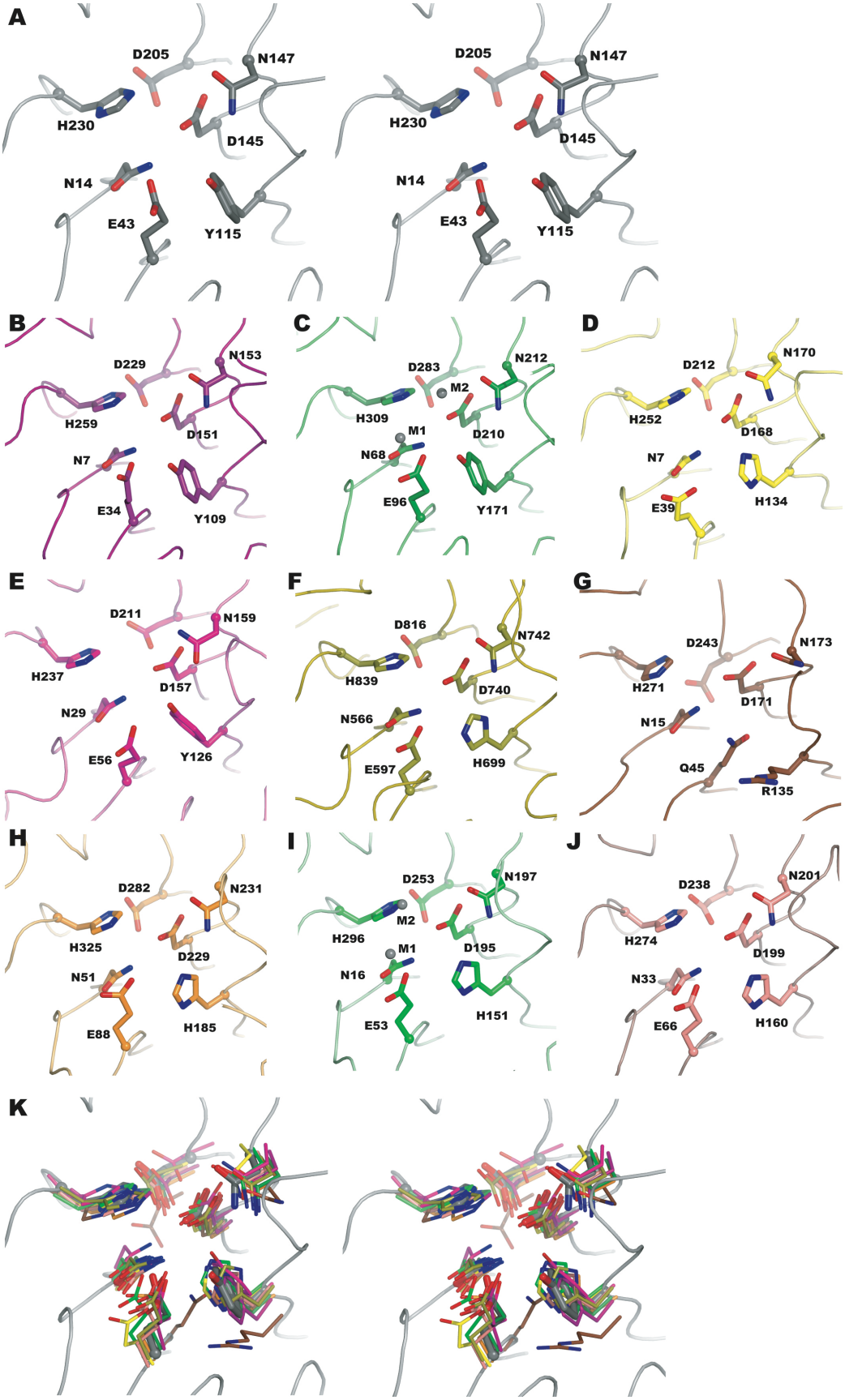


Figure 3. A: Stereo view of the active site of L1-EN. Single views of the active sites of **(B)** ExoIII, **(C)** APE1, **(D)** DNAseI, **(E)** TRAS1-EN, **(F)** IPP5C, **(G)** Nitroforin, **(H)** SmcL, **(I)** Bc-SMase, and **(J)** CdtB. **K:** Stereo superposition of nine active sites on that of L1-EN. The same seven corresponding residues are always displayed as ball and stick models in the respective color. Oxygen red. Nitrogen blue.

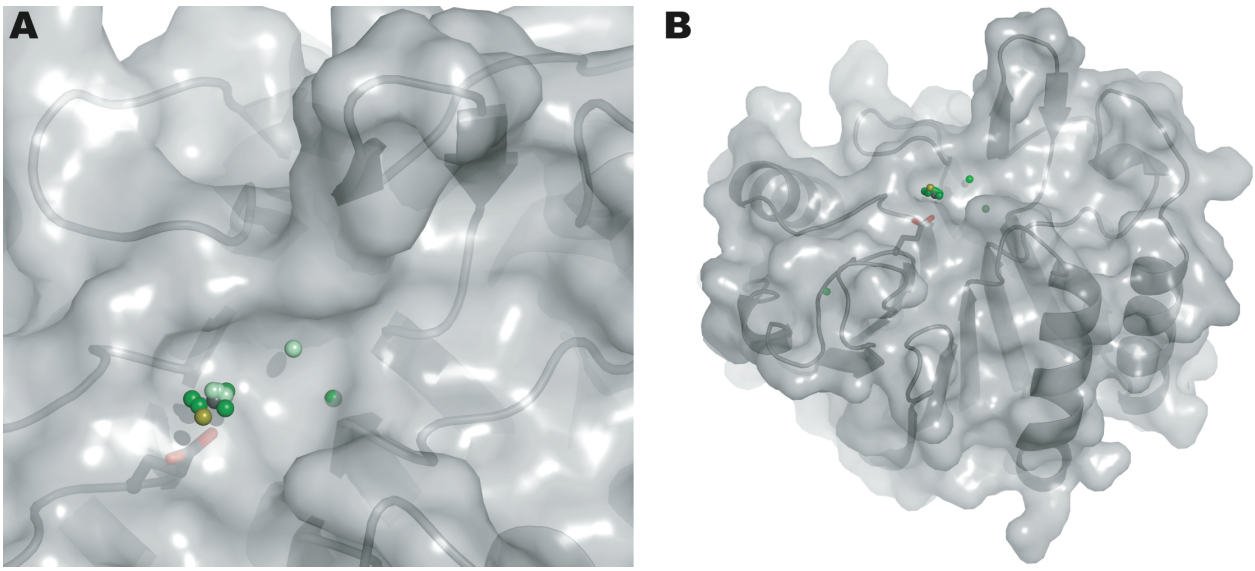


Figure 4. **A:** Positions of metal ions bound on the different structures are mapped on zoomed-in surface of L1-EN. Colored spheres represent the metals from the respective crystal structure. *Grey:* L1-EN, *Dark green:* APE1, *Pale green:* Bc-SMase, *Olive:* IPP5C. The conserved glutamate responsible for metal ion co-ordination is shown as a ball and stick model under the surface, together with a cartoon of L1-EN. **B:** Same as A but full view with orientation as in Figure 2.

Residue conservation in the active site

It is quite interesting to observe that in such a divergent group of functionally unrelated proteins, a number of key residues forming the active site cleft are conserved through different subfamilies and evolutionary stages. These highly conserved residues point towards a common or at least a very similar catalytic mechanism (Figure 3). L1-EN residues depicted in figure 3A include: N14, E43, Y115, D145, N147, D205, H230, that are totally conserved apart from few exceptions (Figure 1). The tyrosine at position 115 for instance is maintained among the AP-like endonucleases (3A, B, C, E), but is otherwise substituted for a histidine (3D, F, H, I, J) or an arginine (3G) in the rest of the family members.

Most of these residues are conserved even in the case of nitrophorin where the active site does not serve a purpose for the proteins' function that is heme/NO binding. Nitrophorin is the only of these ten proteins where two of the conserved active site residues are exchanged for others, but still of similar properties. The residues in question are Y115, which is an arginine in nitrophorin and E43 that is missing in nitrophorin (G45, Figure 1). Although Q42 of L1-EN (Figure 1) is replaced in sequence for Q45 in nitrophorin (Figure 1 and 3G), in the structural superposition this glutamine is situated much closer to E43 of L1-EN, pointing towards the protein surface whilst Q42 of L1-EN is always buried between the two opposing β-sheets. The above portray an example of evolutionary relationships connecting sequence, structure and function, illustrating how residues that are normally conserved within a family, might be substituted for others when they are not anymore essential for a specific function.

In general it is quite remarkable that the seven residues forming the active site are so well conserved among this diverse group of proteins that target a variety of substrates. The superposition (Figure 3K) of all nine active sites on that of L1-EN nicely illustrates the structural similarities within the family and enables us to extend speculations for those family members, for which a crystal structure is not yet available.

Importance of metal ion binding for the catalytic mechanism

A characteristic of this family of proteins is undoubtedly the necessity for metal ions, in order for the enzymes to properly function. It has been shown experimentally that activity decreases or is abolished in the absence of Mg^{+2} or other metals. A number of such metal ions have been found bound in different crystal structures, either in the presence or in the absence of substrate.

The different positions of metal ions bound in the crystal structures, can be seen projected on the surface of L1-EN (Figure 4). The main metal coordinating residue is always a glutamate (E43 in L1-EN), in all the different structures. This residue is totally conserved, but as mentioned previously not in the case of nitrophorin where it is replaced for a glycine since the protein does not appear to need the presence of a metal ion at that position for its function.

Apart from two structures, that of APE1 crystallized at pH 7.5 and of Bc-SMase co-crystallized with cobalt (pH 6.5), binding two metal ions, in general only one metal ion is found bound in the active site of all proteins. It is quite striking however that although the second metal site in Bc-SMase and in APE1 is found in the proximity of the active site, it is not at the same position

in the two structures. APE1 residues that are coordinating the second metal ion are D210, N212, H309 and water mediated via Y171 (Figure 3C). On the other hand, the only direct hydrogen bond with the second metal in SMase is by H296 and two further water-mediated contacts by D195 and N197 (Figure 3I).

For Bc-SMase, three different crystal structures are available, each one with different type and number of metal ions bound: Bc-SMase + single Mg^{2+}, Bc-SMase + single Ca^{2+} and Bc-SMase + double Co^{2+}. It appears that differences in the binding mode of Mg^{2+}, Co^{2+}, Ca^{2+} in the active site of Bc-SMase are reflected in differences in its activity *in vitro*. This led to the belief that a water-bridged double divalent ion in the active site of Bc-SMase is the necessary architecture for activity. However, a distinction is yet to be made between one or two metal ion mechanisms for catalysis in the different enzymes.

What is not clear is if the presence of the second metal ion in APE1 and Bc-SMase crystal structures can have a crucial role in the catalytic mechanism. It is worth noting that in both cases this second ion is bound in the absence of target substrate, DNA for APE1 or sphingomyelin for Bc-SMase. One factor that can be crucial for metal binding and physiological relevance is the pH level during the crystallization process. The second metal in APE1 appears only at pH 7.5 that is the optimal for protein function, but there is no available DNA complex at this pH, and all other complexes contain only one metal. For Bc-SMase, all crystallization was performed at pH 6.5 and only co-crystals with cobalt bind two metals, while all activity assays for this protein are carried out at pH 7.5.

With the currently available information it is not possible to conclude whether one or two metal ions are necessary for catalysis and one can only speculate about the existence of a common underlying catalytic mechanism for the whole family of enzymes. However, the necessity for metal ions like magnesium that have strict geometrical requirements, could enhance the specificity of the respective protein and help in distinct functions. The accumulation of data and the combination with existing knowledge about metal ion assisted catalysis in DNA and RNA polymerases (Yang et al, 2006) could enable us to draw firmer conclusions in the future.

Concluding remarks

In the family of metal-dependent phosphohydrolases, although the sequence identity is quite low, the essential active site residues are very well maintained among the family members. This was confirmed by several crystal structures that indeed showed a conserved architecture in the active site, which is situated in a common protein core.

It is believed that instead of adopting a similar fold after having originated from different ancestors, members of this family most likely did possess a common phosphoesterase fold. Such a fold was probably the type of scaffold these functionally diverse proteins used in order to conform to their catalytic mechanism. This common fold and structural scaffold, has most likely been subjected to many rounds of evolution in order to yield enzymes that serve a variety of functions and fulfill different requirements in the cell.

Through evolution, the need to recognize and bind to an increasing number of varied substrates, brought about significant changes regarding the number, length and positioning of surface loops, while maintaining the same protein core. Hence, those differences observed in the surface loops seems to be what gives to each protein its unique characteristics and specific function. Adaptation through evolution depending on the type of substrate and the requirement for strict distinction between targets could indeed be the force behind such dissimilarity; for instance the need of APE1 to scan a DNA duplex looking for abasic sites to repair or the necessity of Bc-SMase to interact with the membrane.

More crystal structures of substrate complexes need to be determined, in order to enable us to conclude with certainty about common ways of catalysis and metal ion requirements, within this extended family of enzymes.

References

Ago H, Oda M, Takahashi M, Tsuge H, Ochi S, Katunuma N, Miyano M, Sakurai J (2006) Structural basis of the sphingomyelin phosphodiesterase activity in neutral sphingomyelinase from Bacillus cereus. *J Biol Chem* **281**: 16157-16167.

Anzai T, Takahashi H, 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**: 100-108.

Barzilay G, Hickson ID (1995) Structure and function of apurinic/apyrimidinic endonucleases. *Bioessays* **17**: 713-719.

Beernink PT, Segelke BW, Hadi MZ, Erzberger JP, Wilson DM, Rupp B (2001) Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape1: implications for the catalytic mechanism. *J Mol Biol* **307**: 1023-1034.

Ceelen LM, Decostere A, Ducatelle R, Haesebrouck F (2006) Cytolethal distending toxin generates cell death by inducing a bottleneck in the cell cycle. *Microbiol Res* **161**: 109-120.

Chi Y, Zhou B, Wang WQ, Chung SK, Kwon YU, Ahn YH, Chang YT, Tsujishita Y, Hurley JH, Zhang ZY (2004) Comparative mechanistic and substrate specificity study of inositol polyphosphate 5-phosphatase Schizosaccharomyces pombe Synaptojanin and SHIP2. *J Biol Chem* **279**: 44987-44995.

Clarke CJ, Snook CF, Tani M, Matmati N, Marchesini N, Hannun YA (2006) The extended family of neutral sphingomyelinases. *Biochemistry* **45**: 11247-11256.

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

Demple B, Harrison L (1994) Repair of oxidative damage to DNA: enzymology and biology. *Annu Rev Biochem* **63**: 915-948.

Dlakić M (2000) Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. *Trends Biochem Sci* **25**: 272-273.

Feng Q, Moran JV, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905-916.

Fujiwara H, Osanai M, Matsumoto T, Kojima KK (2005) Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. *Chromosome Res* **13**: 455-467.

Gorman MA, Morera S, Rothwell DG, de La Fortelle E, Mol CD, Tainer JA, Hickson ID, Freemont PS (1997) The crystal structure of the human DNA repair endonuclease HAP1 suggests the recognition of extrahelical deoxyribose at DNA abasic sites. *EMBO J* **16**: 6548-6558.

Hu X, Nesic D, Stebbins CE (2006) Comparative structure-function analysis of cytolethal distending toxins. *Proteins* **62**: 421-434.

Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC (1990) Atomic structure of the actin:DNase I complex. *Nature* **347**: 37-44.

Lara-Tejero M, Galán JE (2001) CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect Immun* **69**: 4358-4365.

Lara-Tejero M, Galán JE (2002) Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions. *Trends Microbiol* **10**: 147-152.

Maita N, Anzai T, Aoyagi H, Mizuno H, Fujiwara H (2004) Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J Biol Chem* **279**: 41067-41076.

Majerus PW (1992) Inositol phosphate biochemistry. *Annu Rev Biochem* **61**: 225-250.

Majerus PW, Kisseleva MV, Norris FA (1999) The role of phosphatases in inositol signaling reactions. *J Biol Chem* **274**: 10669-10672.

Matsuo Y, Yamada A, Tsukamoto K, Tamura H, Ikezawa H, Nakamura H, Nishikawa K (1996) A distant evolutionary relationship between bacterial sphingomyelinase and mammalian DNase I. *Protein Sci* **5**: 2459-2467.

Melo LF, Mundle ST, Fattal MH, O'Regan NE, Strauss PR (2007) Role of active site tyrosines in dynamic aspects of DNA binding by AP endonuclease. *DNA Repair (Amst)* **6**: 374-382.

Mol CD, Izumi T, Mitra S, Tainer JA (2000) DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected]. *Nature* **403**: 451-456.

Mol CD, Kuo CF, Thayer MM, Cunningham RP, Tainer JA (1995) Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* **374**: 381-386.

Mundle ST, Fattal MH, Melo LF, Coriolan JD, O'Regan NE, Strauss PR (2004) Novel role of tyrosine in catalysis by human AP endonuclease 1. *DNA Repair (Amst)* **3**: 1447-1455.

Nesić D, Hsu Y, Stebbins CE (2004) Assembly and function of a bacterial genotoxin. *Nature* **429**: 429-433.

Ofran Y, Margalit H (2006) Proteins of the same fold and unrelated sequences have similar amino acid composition. *Proteins* **64**: 275-279.

Okazaki S, Ishikawa H, Fujiwara H (1995) Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. *Mol Cell Biol* **15**: 4545-4552.

Openshaw AE, Race PR, Monzó HJ, Vázquez-Boland JA, Banfield MJ (2005) Crystal structure of SmcL, a bacterial neutral sphingomyelinase C from Listeria. *J Biol Chem* **280**: 35011-35017.

Suck D (1994) DNA-protein interactions. Flip out and modify. *Curr Biol* **4**: 252-255.

Suck D, Lahm A, Oefner C (1988) Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I. *Nature* **332**: 464-468.

Suck D, Oefner C (1986) Structure of DNase I at 2.0 A resolution suggests a mechanism for binding to and cutting DNA. *Nature* **321**: 620-625.

Suck D, Oefner C, Kabsch W (1984) Three-dimensional structure of bovine pancreatic DNase I at 2.5 A resolution. *EMBO J* **3**: 2423-2430.

Tsujishita Y, Guo S, Stolz LE, York JD, Hurley JH (2001) Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. *Cell* **105**: 379-389.

Walker FA (2005) Nitric oxide interaction with insect nitrophorins and thoughts on the electron configuration of the {FeNO}6 complex. *J Inorg Biochem* **99**: 216-236.

Weichenrieder O, Repanas K, Perrakis A (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure (Camb)* **12**: 975-986.

Weichsel A, Andersen JF, Champagne DE, Walker FA, Montfort WR (1998) Crystal structures of a nitric oxide transport protein from a blood-sucking insect. *Nat Struct Biol* **5**: 304-309.

Weichsel A, Andersen JF, Roberts SA, Montfort WR (2000) Nitric oxide binding to nitrophorin 4 induces complete distal pocket burial. *Nat Struct Biol* **7**: 551-554.

Weichsel A, Maes EM, Andersen JF, Valenzuela JG, Shokhireva TKh, Walker FA, Montfort WR (2005) Heme-assisted S-nitrosation of a proximal thiolate in a nitric oxide transport protein. *Proc Natl Acad Sci U S A* **102**: 594-599.

Weston SA, Lahm A, Suck D (1992) X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J Mol Biol* **226**: 1237-1256.

Whisstock JC, Romero S, Gurung R, Nandurkar H, Ooms LM, Bottomley SP, Mitchell CA (2000) The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis. *J Biol Chem* **275**: 37055-37061.

Yang W, Lee JY, Nowotny M (2006) Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. *Mol Cell* **22**: 5-13.

SUMMARY

The work described in this thesis is an attempt to enrich our knowledge regarding the biology of mobile genetic elements and in particular retrotransposons, that until recently were dismissively viewed as genetic parasites or "junk" DNA. The accumulating data however, of ongoing research and of many genome sequencing projects, are slowly but steadily changing the aforementioned views. Transposable elements are increasingly accepted as modulators of gene expression and drivers of genome evolution.

The first chapter serves as a general introduction into the different types of mobile genetic elements, while putting the focus on the LINE-1 (L1) retrotransposons and the endonuclease they encode. The human L1 endonuclease (L1-EN), which is the main focus of this work, is the targeting endonuclease encoded by the L1 retrotransposon. L1's are responsible for more than 1.5 million retrotransposition events in the history of the human genome, contributing more than a quarter to human genomic DNA (L1 and Alu elements). An overview is given on the existing knowledge of L1 biology, highlighting unexplored areas and taking a look at how retrotransposons might affect our own and other genomes, in both negative -e.g disease- and positive -e.g gene expression regulation- ways.

In Chapter 2 we present the crystal structure of human L1 endonuclease. This was the first structure of a retrotransposon-encoded protein at the time of publication and still constitutes a prototype for retrotransposon-encoded endonucleases involved in target-primed reverse transcription. L1-EN is related to the well-studied human DNA repair endonuclease APE1 and its nicking specificity is a major determinant for retrotransposon integration site selection. Structure-based endonuclease alignments reveal a set of key conserved residues and suggest that DNA recognition may proceed via the accommodation of an extra-helical nucleotide within a pocket on the surface of the endonuclease. The present analysis will help to refine phylogenetic and functional relationships among metal-dependent phosphohydrolases and provides a basis for manipulating non-LTR retrotransposon integration site selection.

The main question addressed in Chapter 3 is what determines the target selectivity of L1-EN. This is especially intriguing since similar retrotransposable elements encode endonucleases that have different levels of specificity; a fact that might subsequently be reflected in the integration specificity of the retrotransposon. Based on our crystal structure, we designed and generated L1-EN variants in order to analyze and manipulate DNA target site recognition. We determined two additional crystal structures, which confirm the fact that it is indeed possible to exchange the prominent hairpin of L1-EN with those of highly specific retrotransposon endonucleases. Biochemical analysis shows that the hairpin grafts are functional *in vitro*, resulting in altered specificity, while individual point mutations do not change the nicking pattern of L1-EN. Structural parameters of the DNA target seem more important for recognition than the nucleotide sequence, and nicking profiles on DNA oligonucleotides *in vitro* are less well defined than the respective integration site consensus *in vivo*. This could be an indication that additional factors other than the specificity of L1-EN, or the respective endonuclease, are required for the targeted integration of non-LTR retrotransposons.

Chapter 4 focuses in better understanding how L1-EN recognizes DNA and facilitates subsequent nicking. This is made possible by following a multidisciplinary approach that involves mutagenesis, X-ray crystallography, data driven computational docking and molecular dynamics simulations. We show that mutation of catalytic and peripheral residues that may be responsible to accommodate a flipped out base during catalysis, all greatly reduce *in vitro* endonuclease activity. Three new crystal structures of point mutants indicate a very robust catalytic scaffold with minor structural rearrangements. Combining all the available structures with our mutational data, we construct computational models that explore different modes of DNA recognition, asking the question weather L1-EN flips out a nucleotide or not. Although we cannot exclude either mechanism, the recognition of an extra-helical nucleotide is a strong possibility. Comparative molecular dynamics simulations show that flexibility of a prominent hairpin, strategically positioned over the active site, is likely a necessity for DNA recognition, positioning and nicking. The idea that additional surface loops are also flexible and possibly assist in DNA binding, both in L1-EN and in a related but highly specific insect retrotransposon endonuclease, is also explored. It is likely that differences in surface loops can confer different degrees of target DNA nicking specificity.

Finally, in Chapter 5 we set out to make a structural and functional comparison of all those members in the family of metal-dependent phosphohydrolases, for which a crystal structure is available. By determining the structure of the L1-EN we confirmed previous suggestions that indeed this enzyme is a member of this extended family, the founding member of which is bovine pancreatic De-

oxyribonuclease I. This chapter illustrates how all members discussed conform to the characteristic DNAse I fold and that various enzymes responsible for diverse functions in the cell use a common, pre-formed active site located at the same position on the surface of each protein. It is likely that members originate from a common ancestor with the same fold that under evolutionary pressure diverged to result in similar proteins, but with unique roles. The core of each protein is highly similar, but large differences occur regarding surface loops located both in the proximity of the active site and further spread over an extensive substrate binding area. Exactly those differences could be what gives to each of these proteins a characteristic function. The requirement for metal ions is also discussed, together with the possible existence of a common catalytic mechanism.

In conclusion, the present work makes a contribution to our thus far incomplete understanding of L1 retrotransposon biology and self-replication mechanism, and could be the first step towards utilization of L1 retrotransposons as vehicles for gene delivery. The L1 endonuclease bears considerable technological interest, because its target selectivity may ultimately be engineered to allow the site-specific integration of DNA into defined genomic locations. Of course further research is needed in order to clarify issues such as the importance of additional host factors in the specificity of retrotransposon integration, as well as the possible synergistic relationship between endonuclease and reverse transcriptase activities that are encoded as one full-length protein. To take our current knowledge to the next level, it would be of great significance to obtain structural information for both an endonuclease:DNA complex and a putative endonuclease:reverse transcriptase complex.

SAMENVATTING

Het werk beschreven in dit proefschrift is een poging om onze kennis te verrijken betreffend de biologie van de mobiele genetische elementen en specifiek de retrotransposons, die tot dusver negatief worden gezien als genetische parasieten of "junk" DNA. De accumulerende resultaten echter, van huidig onderzoek en van vele genoom sequentie projecten, zijn langzaam maar veranderen gestaag de bovengenoemde inzichten. Verplaatsbare elementen worden snel geaccepteerd als modulatoren van genexpressie en het drijvend wiel van genoom evolutie.

Het eerste hoofdstuk dient als een algemene inleiding voor de verschillende types van mobiele genetische elementen, terwijl de focus gezet wordt op de LINE-1 (L1) retrotransposons en de endonucleases die zij coderen. De humane L1 endonuclease (L1-EN), die het centrale onderwerp is van dit werk, is de aanleggende endonuclease gecodeerd door de L1 retrotransposon. L1's zijn verantwoordelijk voor meer dan 1.5 miljoen retrotranspositie gebeurtenissen in de geschiedenis van het humane genoom, en daartoe dragen zij bij aan meer dan een kwart van het humane genomische DNA (L1 en Alu elementen). Een overzicht is gegeven van de bestaande kennis over L1 biologie, nadruk leggend op niet-onderzochte gebieden en kijkend op hoe retrotransposons ons eigen en andere genomen in zowel negatieve – e.g. ziekte- als in positieve –e.g. genexpressie regulatie- manieren, zouden kunnen beïnvloeden.

In Hoofdstuk 2 presenteren wij de kristalstructuur van de humane L1 endonuclease. Dit was de eerste structuur van een retrotransposon-gecodeerd eiwit op het moment van publicatie en representeert nog steeds als een prototype voor retrotransposon-gecodeerde endonucleases, die betrokken zijn bij het doelwit geïnitieerde omgekeerde transcriptie. L1-EN is familie van de goed bestudeerde humane DNA herstel endonuclease APE1 en zijn inkerving specificiteit is een beslissende factor voor de plaats selectie van retrotransposon integratie. Op structuur gebaseerde endonuclease schikkingen onthullen een groep belangrijke geconserveerde residuen en suggereren dat DNA herkenning voort kan gaan door de beschikbaarheid van een extra spiraalvormig nucleotide binnenin een buidel op de oppervlakte van de endonuclease. De huidige analyse zal de phylogenetische en functionele relaties tussen metaal-afhankelijke phosphohydrolases helpen verfijnen en een basis leveren voor het manipuleren van non-LTR plaats selectie van retrotransposon integratie.

De hoofdvraag gericht in Hoofdstuk 3 is wat bepaalt de selectiviteit voor het mikpunt van L1-EN. Dit is vooral fascinerend aangezien soortgelijke retrotransposeerbare elementen coderen voor endonucleases die verschillende niveaus van specificiteit hebben; een feit die misschien vervolgens gereflecteerd kan worden op de integratie specificiteit van de retrotransposon. Gebaseerd op onze kristal structuur, hebben we L1-EN varianten ontworpen en gemaakt om de plaats herkenning van DNA mikpunt te analyseren en te manipuleren. We hebben twee aanvullende kristal structuren bepaald, die het feit bevestigen dat het inderdaad mogelijk is om de prominente haarspeld van L1-EN uit te wisselen met die van zeer specifieke retrotransposon endonucleases. Biochemische analyse laat zien dat de haarspeld transplantaties functioneel zijn *in vitro*, resulterend in een verandering in specificiteit, terwijl individuele punt mutaties de inkerving patroon van L1-EN niet veranderen. Structurele parameters van het DNA doelwit lijkt belangrijker te zijn voor de herkenning dan de nucleotide sequentie, en inkerving profielen op DNA oligonucleotiden *in vitro* zijn in mindere mate gedefinieerd dan de respectieve integraties plaats overeenstemmend *in vivo*. Dit kan een indicatie zijn dat andere aanvullende factoren dan de specificiteit van L1-EN of de respectieve endonuclease, nodig zijn voor de doelgerichte integratie van non-LTR retrotransposons.

Hoofdstuk 4 richt zich op het beter begrijpen van hoe L1-EN DNA herkent en de daaropvolgende inkerving bevordert. Dit is mogelijk gemaakt door een multidisciplinaire benadering te volgen bestaande uit mutagenese, röntgenstraal kristallografie, met behulp van data gestuurde computer koppeling en moleculaire dynamische simulaties. We laten zien dat een mutatie in de katalytische en omliggende residuen, die verantwoordelijk kunnen zijn voor het toelaten van een uitgestoken base tijdens katalyse, allen *in vitro* de endonuclease activiteit in hoge mate verminderd. Drie nieuwe kristal structuren van punt mutanten duiden een erg robuust platform met kleine structurele veranderingen aan. Gecombineerd met alle beschikbare structuren met onze mutatie data, bouwen we computer modellen die de verschillende wijze van DNA herkenning, afvragend of L1-EN een nucleotide uitsteekt of niet. Alhoewel wij beide mechanismen niet kunnen uitsluiten, is de herkenning van een extra spiraalvormige nucleotide een grote mogelijkheid. Vergelijkende simulaties van moleculaire dynamiek laten zien dat flexibiliteit van een prominente haarspeld, strategisch gepositioneerd over de actieve plaats, waarschijnlijk een noodzaak is voor het herkennen van DNA, het positioneren en het inkerven. Het idee dat extra lussen aan de oppervlakte ook flexibel zijn en misschien assisteren met het binden aan DNA, zowel in L1-EN als in een gerela-

teerde maar uiterst specifiek insecten retrotransposon endonuclease, wordt ook onderzocht. Het is waarschijnlijk dat verschillen in lussen aan de oppervlakte verschillende niveaus van specificiteit voor DNA doelwit inkerving.

Tenslotte in Hoofdstuk 5 zijn we begonnen met het maken van een structurele en functionele vergelijking van alle familieleden van metaal-afhankelijke phosphohydrolases, waarvan een kristal structuur beschikbaar is. Door de structuur van L1-EN te bepalen, bevestigen wij de eerdere suggesties dat dit enzym inderdaad lid is van deze uitgebreide familie, wiens eerste lid de Deoxyribonucleas I in de runderen alvleesklier is. Dit hoofdstuk illustreert hoe alle besproken leden zich aanpassen aan de karakteristieke vouw van DNAse I en dat verschillende enzymen die verantwoordelijk zijn voor de diverse functies in de cel, een algemene van tevoren gevormde actieve plaats gebruiken die zich op dezelfde positie aan de oppervlakte van elke eiwit bevindt. Het is waarschijnlijk dat leden van een gemeenschappelijke voorouder afstammen met dezelfde vouw die door druk van evolutie afweek resulteren tot gelijke eiwitten met unieke functies. De kern van elk eiwit lijkt heel erg op elkaar, maar grote verschillen komen voor wat betreft de lussen aan de oppervlakte gelegen in zowel in de buurt van de actieve plaats als verder verspreid over een uitgebreide substraat bindingsgebied. Precies die verschillen kunnen verantwoordelijk zijn dat elk van deze eiwitten een karakteristieke functie heeft. De behoefte aan metaal ionen wordt ook besproken, samen met het mogelijke bestaan van een algemene katalytische mechanisme.

Tenslotte, het huidige werk levert een bijdrage aan ons tot dusver incomplete begrip van L1 retrotransposon biologie en zelfreplicerend mechanisme, en het zou een eerste stap kunnen zijn richting het gebruik maken van L1 retrotransposons als transportmiddel voor gen overdracht. De L1 endonuclease levert een aanzienlijk technologische interesse op, omdat zijn doelwit selectiviteit uiteindelijk omgebouwd zou kunnen worden om plaats specifieke integratie van DNA in bepaalde genomische locaties toe te staan. Natuurlijk is verder onderzoek nodig om zaken zoals de belang van extra gastheer factoren in de specificiteit van retrotransposon integratie en ook zoals de mogelijke overeenkomende relatie tussen endonuclease en de omgekeerde transcriptase activiteiten die als een compleet eiwit gecodeerd zijn, op te helderen. Om ons huidige kennis naar de volgende niveau te tillen, zou het van groot belang zijn om structurele informatie van zowel een endonuclease:DNA complex als van een mogelijke endonuclease:omgekeerde transcriptase complex te verkrijgen.

ACKNOWLEDGEMENTS

When Tassos told me to stay for a PhD in his lab, my first reaction was to look behind me and make sure there was nobody there, then think …hmmm sounds more interesting than the Greek army, but what really did it was that at least he was not an Olympiakos supporter. On a slightly more serious note now, many thanks for giving me the opportunity to work in your lab and NKI, it was a good collaboration and I learned a lot. Great scientist, sports analyst and philosopher, and also very patient guy, endlessly waiting for me to sort out my complicated list of priorities. Also it was great that he was available online practically 24/7 and replying "real time" during the write-up. Thanks and all the best for the future….and present, very important as well.

Oli, thanks for letting me join you on the L1-EN adventure and also thanks for spending time and effort with me especially in the beginning. Great mind and passion for the retro-elements, but sometimes 2-3 experiments ahead for my own poor standards. Ok, you and Tassos did not always agree on project issues and it has to be said that the 2-bosses-thing was tough at times, but it was also fruitful, giving rise to many interesting discussions and ideas. Best of luck to you as well with the new step as group-leader. As far as the crystals are concerned, honestly I didn't plan to get them in the first year!! Imagine, I even had to cancel my tickets for Sonar in Barcelona (Rafa can confirm) and go for data collection instead.

Titia, thank you for being my promotor (…well, we didn't really give you a choice, did we?). It is great to have someone like you around the lab, with extensive knowledge on biology but also with understanding towards students and with good will to discuss and suggest new ideas. Respect and I wish you all the best.

So, the lab then: Nuno for being there when it all started-best of luck man-and sorry for the Euro2004, Valeria Christodoulou and Vangelis de Marco it was great having you in the lab-always willing to help/assist/guide with the work but also discuss Italian politics and Greek Football, Suzan-Angelina-Cristiane for "gracing" us with their presence and balancing the men/women equation in the lab when most needed and seriously now for all their help and collaboration and good will to make the lab a better place to work in, Dave for putting a lot of effort to start us up with the robots and make our life easier, Hilge for the NMR stuff and for discussing all the latest in tennis-he is a Federer fan but nobody's perfect..…and no it was not accidental, Mobien-the man-the legend-all the best for you and think again seriously about supporting PAOK,ok? And now that I said PAOK….Nikolaaakiiiiiiii, όχι στα βαθειά παιδί μου!Take care. Serge for helping out numerous times and for spending time and time again to get the cluster running (open a mac-helpline and be millionaire man!), Mattheos-"how will this improve my daily life"-Kakaris good friend and good cook, Marouane-our one and only terrorist/fundamentalist/extremist videogamer, Diederick, Patrick-who I never managed to beat in squash, Krista: take care people. Dene, all the best man. Koen, Rebecca, and of course executive-perakislab-member: Roelof, who was so happy when I finally left because he would stop the nomadic-bench-life.

Many thanks go to Maria for making the connections, Jos for starting us up with the retrotransposons, Gerald for the Tx1L-R1Bm collaboration, Alexandre for the computational stuff, Geerten for the NMR ideas, and also all the people that constantly help at the Grenoble & Hamburg synchrotrons.
Of course survival in the NKI wouldn't be as easy for the ethnic minority of structural biologists if Titias' people were not around: Meindert, Mark-thanks for everything, Puck, Ganesh-Bollywood is still waiting for you man, Valerie, Joyce, Chris, Gretel, Sasha-the "Pelevin" of biacore, Sari, guys thanks for all the discussions-help-borels-fun it was good to work together, Alex-Annete-Francesca all the best for you guys…and yes! Pim + Herrie for invaluable help and for answering the same questions over and over again somehow managing to keep a straight face and pretend they never heard this before!!

Marlijn + Suellen: for all the great help that might range from sending crystals safely to the synchrotron to finding houses and convincing bank managers that we are not wanted in 3rd world countries like Greece. Pete I'll make it to Barca one fine day, Mike, Sebastian, Menno, Luis, Annette, Roderick, Thijn, Katrien, Marielle, Ron + his micro-array team and everybody on the second floor. Vaso and Rafa for friendship/concerts/parties/cigarette breaks, Helen, Dieuwke, the H4-guys, Zong-the bravest of us all, Iman, Paul-Andre, Andrej, Scott, Carla, France who forgot all about science the moment Lou arrived.

Who else….who else…..ah!.....Gloria, the woman I love most in the planet (….yes mom…you too) and who I admire for still putting up with me and all the weirdness & stupidity that at times seem to surround me, thanks for that and also for actively helping with the computational part of the L1-EN and for introducing me to all the nice people at the Bijvoet in Utrecht…….btw Danny we're still waiting for you in Madrid, and you too Viktor "the-2-meter-chinese" Hsu :)!

Onno for showing me the real "boring" Amsterdam-flying in from Warsaw for European Sounds so that all 3 listeners are happy and giving me the opportunity to play weird music at places like the Leiden Cathedral and Schiphol hangars, Helene who finally made it to the big "alive" poets society, Aude-if only there was a bar in Nieuw Sloten ee?, Pavlina, Anne-Sophie, Paulo the brazilian squash champion, Petro (μάθε τένις ρεεεεε), Julia, Rafita, Barbara+Mark+???, Axelman, Tom, Paul.

And then there are all these Greek idiots that even if I don't see them as often as I would like to, when we meet είναι σαν να μην πέρασε μια μέρα, πλατεία Μαβίλλη 4 παρά …και τα γνωστά: stefanos, xionia, katerina+bill, adamou-gkouvas-gkelias-dougalis, tzeni, stefanos+maria, akos, markos+sofia, george, sakis, yannis, george-g, gewrgia, tzeni-viky-mitso and I stop, the rest can fill in the official complaint form for not seeing their name here. Οικογένεια Ρεπανά…..κανείς δεν το περίμενε.

Very special thanks to Takis Papistas and Tassos Economou……after all it's partly their fault as well and they know it!

And finally (….ooufff he said finally)….εννοείται οτι δεν ξέχασα να ευχαριστήσω τους γονείς και την αδερφή μου….άντε και τον Μήτσο, για την στήριξη, την υπομονή και την αγάπη τους όλα αυτά τα χρόνια. Ναί μαμα, τρώω καλά :) . Δεν θα γράψω περισσότερα γιατί υπάρχει κίνδυνος να το γυρίσουμε σε Βραζιλιάνικη σαπουνόπερα (sorry Cristiane).

That's that….and don't forget to support Open Access, it's the future.

List of Publications

Kostas Repanas, Nora Zingler, Liliana Layer, Gerald Schumann, Anastassis Perrakis and Oliver Weichenrieder. (2007) ***Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease***, *Nucleic Acids Research, in press.*

Kostas Repanas, Oliver Weichenrieder and Anastassis Perrakis. (2007) **LINE-1 endonuclease friends and family: *Structural and functional connections in a family of metal-dependent phosphohydrolases*,** *mini-review in preparation*.

Kostas Repanas, Gloria Fuentes, Serge Cohen, Alexandre Bonvin, Oliver Weichenrieder and Anastassis Perrakis. (2007) **To flip or not to flip: *insight into the DNA cleavage mechanism of human LINE-1 retrotranposon endonuclease,*** *submitted*.

Kostas Repanas, Oliver Weichenrieder and Anastassis Perrakis. (2004) ***Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon***, *Structure,* 12(6):975-86.

Georgios Sianidis, Spyridoula Karamanou, Eleftheria Vrontou, Kostantinos Boulias, **Kostantinos Repanas**, Nikos Kyrpides, Anastasia Politou and Anastassios Economou (2001) ***Cross-talk between catalytic and regulatory elements in a DEAD motor domain is essential for SecA function***, *EMBO Journal,* 20: 961-970

Propositions

1. Mobile Genetic Elements are drivers of genome evolution. Increasing interest and extensive research is slowly but steadily changing public opinion about the elements formerly known as “junk”. (Petsko 2003, Kazazian 2004, Han & Boeke 2005)

2. The crystal structure of human L1 endonuclease (L1-EN) is the first of a retrotransposon-encoded protein and a prototype for retrotransposon-encoded endonucleases involved in target-primed reverse transcription.

3. It is possible to manipulate the L1-EN target selectivity and in this way direct DNA integration into novel genomic locations.

4. L1-EN turned out to be more promiscuous than we thought, but still cautious enough not to be caught for a snapshot with her DNA partners.

5. Knowledge regarding L1-EN and related proteins could pave the way for utilising retrotransposons as vehicles for gene delivery in novel genetic therapies yet to come.

6. As a researcher, every day you learn something new. All this accumulating knowledge should help you realise that you know nothing. (Freely adapted from Hellenic Philosophy)

7.“You can’t always get what you want" ... also in Experimental Science. (Rolling Stones, Let it Bleed, 1969)

8. The same scientific result can be interpreted in different ways by different people.

9. The end of crystallography as we know it is coming with the free electron laser being built in Hamburg, while I am writing these lines.

10. 99% of Biological Research in the whole world is restricted to 3% of the human genome sequence.

11. Since Prof. R. Plasterk - renounced for his contribution in the field of mobile genetic elements - is Minister of 'Education, Science and Culture' in the Netherlands, I feel the way is open for me to get that Ministry in Greece - as soon as it evolves - either by means of natural selection or intelligent design - from Ministry of 'Educational and Religious affairs' to something sensible.