Strukturelle und funktionelle Analyse des ORF1p Proteins des humanen LINE-1 Retrotransposons

Structural and functional analysis of the ORF1p protein of the human LINE-1 retrotransposon

DISSERTATION

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Elena Khazina

aus Novosibirsk, Russland

Tübingen

2010

Tag der mündlichen Qualifikation:	03. Dezember 2010
Dekan:	Professor Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Professor Dr. Thilo Stehle
2. Berichterstatter:	Professor Dr. Andrei Lupas

This thesis describes work carried out in the laboratory of Dr. Oliver Weichenrieder, in the department of Dr. Elisa Izaurralde at the Max Planck Institute for Developmental Biology, Tübingen, Germany, from February 2007 until September 2010. The work was supervised by Prof. Dr. Andrei Lupas at the Max Planck Institute for Developmental Biology, Tübingen, Germany and by Prof. Dr. Thilo Stehle at the Eberhard-Karls-University, Tübingen, Germany. I declare that this thesis is the product of my own work. Wherever parts of the work have been published, wherever other sources have been used as well as wherever parts of the work were done by colleagues of mine, this has been indicated accordingly.

Acknowledgements

First and above all, I would like to thank Dr. Oliver Weichenrieder for giving me the opportunity to work on a very interesting thesis topic. I am also grateful to him for the excellent supervision and teaching during my thesis, for his great support in all cases, and for very helpful and inspiring discussions.

Also, I would like to thank Dr. Elisa Izaurralde for her scientific guidance and support throughout my thesis.

I would like to express my gratitude to Prof. Dr. Thilo Stehle for supervising this thesis at the Eberhard Karls University, Tübingen, and for giving me the opportunity to attend his excellent crystallographic courses. I would also like to thank Prof. Dr. Andrei Lupas for the supervision of this PhD project at the Max Planck Institute for Developmental Biology, Tübingen.

I thank all past and present members of the Department of Biochemistry at the Max Planck Institute for Developmental Biology, whom I had a pleasure to work with, for the nice working atmosphere, helpful discussions and introducing me to the new laboratory techniques.

I am very grateful to my mother and all my family for the everlasting support and encouragement, without which this thesis would not be possible.

I also thank my friends for their support, patience, understanding and help.

Finally, I would like to thank my boyfriend, Marcus, for his kindly support, care and encouragement, as well as for the helpful scientific discussions.

Table of contents

Zusammenfassung

Retrotransposons sind mobile genetische Elemente, die sich über einen "copy-and-paste" Mechanismus replizieren, bei dem eine neue Kopie ihrer DNA im Genom erzeugt wird. Sie spielten eine wichtige Rolle in der Evolution von eukaryotischen Genomen. Im humanen Genom ist LINE-1 (L1) das am häufigsten vorkommende Retrotransposon, wo es 17% der gesamten genomischen DNA ausmacht. L1 hatte somit einen großen Einfluss auf die Evolution des humanen Genoms und ist auch heute noch aktiv. Es wird als wichtigste Quelle für humane interindividuelle genomische Variation betrachtet. Des Weiteren wird es in Zusammenhang mit verschiedenen menschlichen Krankheiten gebracht. Während die Datenfülle zur Bedeutung des L1 Elements über die letzten Jahre ständig anwuchs, blieben die mechanistischen Grundlagen des Retrotranspositionsprozesses weiterhin weitgehend unverstanden.

Die Retrotransposition von L1 verläuft über "target-primed reverse transcription" (TPRT). Hierbei sind die reverse Transkription eines RNA-Intermediates des mobilen Elementes und die Integration der entstehenden Kopie in das Genom direkt gekoppelt. Dies ist der typische Mechanismus für non-LTR Retrotransposons, d.h. für Retrotransposons, die keine langen terminalen Sequenzwiederholungen aufweisen ("long terminal repeats", LTRs). Die beiden von L1 kodierten Proteine, L1ORF1p und L1ORF2p, sind beide essentiell für die Retrotransposition. L1ORF2p enthält eine Endonukleasedomäne und eine Domäne für die reverse Transkriptase. Die Rolle von L1ORF1p hingegen war zu Beginn dieser Arbeit wesentlich unklarer, da keine Sequenzhomologie zu Proteinen bekannter Funktion erkennbar war. Es war lediglich bekannt, dass die N-terminale Hälfte von L1ORF1p eine "coiled coil" enthält, die für die Oligomerisierung zuständig ist, und dass die C-terminale Hälfte positiv geladen ist und Nukleinsäuren bindet. Um die Funktionsweise und Phylogenie des L1ORF1p zu verstehen, beabsichtigten wir deshalb dessen Struktur mittels Röntgenkristallographie zu ermitteln.

Im ersten Teil dieser Arbeit identifizierten wir mittels bioinformatischer Methoden eine nicht-kanonische RRM (RNA recognition motif) Domäne sowohl im humanen L1ORF1p Protein als auch in vielen phylogenetisch unverwandten ORF1p Proteinen anderer non-LTR Retrotransposons. Damit konnten wir zeigen, dass die ORF1p Proteine von non-LTR Retrotransposons trotz der häufigen Präsenz von

Gag-ähnlichen CCHC Zinkfingern nicht mit dem retroviralen protein Gag verwandt sind. Insgesamt besteht das humane L1ORF1p Protein somit aus einer N-terminalen "coiled coil", einer zentralen RRM Domäne und einer zusätzlichen C-terminalen Domäne (CTD). Wir bestimmten die Grenzen der drei genannten Domänen auf experimentelle Weise und zeigten, dass die "coiled coil" sowohl notwendig als auch hinreichend für eine Trimerisierung des Proteins ist. Am wichtigsten war dabei die Kristallstruktur, die wir für die isolierte humane RRM Domäne bestimmen konnten. Diese zeigt konservierte Salzbrücken, die verlängerte Schleifen der L1ORF1p RRM Domäne stabilisieren und ein besonderes Charakteristikum der L1ORF1p RRM Domäne darstellen. Des Weiteren zeigten wir, dass für die Nukleinsäurebindung die RRM- und die CTD Domäne auf einer gemeinsamen Polypeptidkette liegen müssen. Es werden vorrangig einzelsträngige Substrate gebunden, und wir bestimmten für die Bindung notwendige Aminosäuren (Khazina und Weichenrieder, 2009). Anhand einer NMR Struktur, die wir für ein RRM-CTD Konstrukt bestimmt haben, sieht man allerdings, dass zwischen den beiden Domänen in der Abwesenheit der N-terminalen keine spezifischen Bindungen ausgebildet werden. Somit blieb weiterhin unklar, wie die RRM und CTD Domänen bei der Nukleinsäurebindung kooperieren, welche Rolle die "coiled coil" dabei spielt, und warum einzelsträngige Substrate bevorzugt werden.

Diese Fragen wurden im zweiten Teil dieser Arbeit beantwortet. Hierzu lösten wir die Strukturen dreier Kristallformen des humanen L1ORF1p Trimers, aus denen ersichtlich wird, dass die "coiled coil" als zentrales Gerüst für die flexible Verankerung der RRM und CTD Domänen fungiert. Der Aufbau erinnert an die trimeren "coiled coils" von viralen Fibern und Membranfusionsproteinen, auch auf Grund von Ionen, die von polaren Resten im Kern der "coiled coil" koordiniert werden. Des Weiteren bieten die Strukturen einen grundsätzlichen Einblick in die Flexibilität eines RNA Bindeproteines, das aus mehreren Domänen besteht. Es wird ersichtlich, wie die RRM und CTD Domänen zusammen Nukleinsäuren binden können, ohne dabei gegenseitig spezifische Bindungen auszubilden. Das Oberflächenpotential des Proteins legt nahe, dass einzelsträngige Nukleinsäuren um das Molekül herum, in den besonders stark positiv geladenen Spalten zwischen den RRM und CTD Domänen, binden. Da die Abmessungen und Verkrümmungen dieser Spalten eine Bindung doppelsträngiger Substrate nicht zulassen würde, ergibt sich aus diesem Modell auch die Präferenz für einsträngige Substrate. Um diesen vorgeschlagenen Substratbindemodus zu verifizieren machten wir Bindeexperimente

in vitro, wobei wir Präferenzen für bestimmte Nukleinsäuresubstrate feststellen konnten. Abschließend führten wir *in vivo* eine Mutationsanalyse durch, die nicht nur bestätigt, dass die filigrane Architektur von L1ORF1p höchst relevant für die Retrotransposition ist, sondern auch auf weitere mögliche Funktionen von L1ORF1p hinweist, die über RNA-Bindung hinausgehen. (Khazina et al., 2011).

Die Resultate dieser Arbeit bieten einen wichtigen Einblick sowohl in die Funktion und molekularen Mechanismen des L1ORF1p Proteins, als auch in seine evolutionäre Geschichte. Sie werden die Forschung in vielen Feldern wie Zellbiologie, Strukturbioloie, Virologie, Genomevolution und Medizin erheblich vorantreiben.

Summary

Retrotransposons are mobile genetic elements, which replicate via a “copy-and-paste” mechanism, thereby creating a new copy of their DNA in the genome. Retrotransposons have played an important role in the evolution of eukaryotic genomes. LINE-1 (L1) is the most abundant retrotransposon in the human genome, directly accounting for 17% of the genomic DNA. L1 has shaped the human genome in many ways during evolution and is still active nowadays. It is considered as the major source of human interindividual genetic variation, and has been implicated in several human diseases as well. While data on the impact and significance of the L1 element has been accumulating in recent years, the understanding of the underlying molecular mechanism of retrotransposition has been lagging behind.

L1 retrotransposes via target-primed reverse transcription (TPRT), where the reverse transcription of an RNA intermediate is coupled to the integration of the new copy of mobile element into the genome. This mechanism is typical for all retrotransposons that lack long terminal repeats (non-LTR retrotransposons). L1 encodes two proteins called L1ORF1p and L1ORF2p. Both proteins are essential for retrotransposition. L1ORF2p contains a nicking endonuclease and a reverse transcriptase domain. The role of L1ORF1p was much more elusive in the beginning of this PhD thesis, because it lacks sequence homology with any protein of known function. It only was known that the N-terminal half of the L1ORF1p contains a coiled coil, which mediates multimerization of the protein, and that the C-terminal half of the protein is positively charged and binds nucleic acids. To understand the molecular mechanics and phylogeny of L1ORF1p we decided to obtain a high resolution structure of the protein.

In the first part of my thesis we used bioinformatics to identify a non-canonical RRM (RNA recognition motif) domain within human L1ORF1p, as well as in the ORF1p proteins from many phylogenetically unrelated non-LTR retrotransposons. This showed that the ORF1p proteins from non-LTR retrotransposons are not related to the retroviral protein Gag, despite the presence of Gag-like CCHC zinc knuckles in many of them.

In addition to the central RRM domain of the L1ORF1p, we experimentally determined the domain boundaries of the coiled coil and of a C-terminal domain (CTD), and we showed that coiled coil is necessary and sufficient for a trimerization

of human L1ORF1p. Most importantly, we crystallized the human RRM domain and determined a high resolution crystal structure, revealing extended loops stabilized by conserved salt bridges as a characteristic feature of the L1ORF1p RRM domain. Furthermore, we found that nucleic acid binding requires both RRM and CTD domains on the same polypeptide chain, and we identified residues required for that function (Khazina and Weichenrieder, 2009). However, an NMR structure that we determined for the RRM-CTD construct did not reveal any contact between the RRM and CTD domains in the absence of the coiled coil. As a consequence, it still was not clear how RRM and CTD domains cooperate in nucleic acid binding, why single-stranded substrates would be preferred and what the role of the coiled coil was.

These questions were answered in the second part of my thesis. For that, we solved three crystal forms of the human L1ORF1p trimer, which show that the coiled coil serves as a scaffold for the flexible attachment of the RRM and CTD domains. The assembly is similar to the trimeric coiled coils of viral fibres and membrane fusion proteins, also because of ions that are coordinated by polar residues in the core of the coiled coil. Furthermore, the structures provide an insight into the flexibility of a multidomain RNA binding protein and suggest how the RRM and CTD domains can cooperate in nucleic acid binding without directly contacting each other. The surface potential of the protein indicates single-stranded nucleic acids to bind around the molecule in the highly positively charged cleft between RRM and CTD domains. The depth and curvature of this cleft do not allow accommodation of a double-stranded nucleic acid substrate, explaining the preference for single strands. To validate the putative nucleic acid binding surfaces, we did *in vitro* binding experiments that also revealed preferences for certain nucleic acid substrates. Finally, we did a mutational analysis *in vivo* showing that the delicate architecture of L1ORF1p is highly relevant for retrotransposition and indicating functions of L1ORF1p that go beyond RNA binding (Khazina et al., 2011).

Together, the results obtained during my PhD studies provide an important insight into the function and molecular mechanics of the L1ORF1p, as well as its evolutionary history. These data will advance research in many fields including cell and structural biology, virology, genome evolution and medicine.

1 Introduction

1.1 Different types of mobile elements in the human genome

The completion of the first human genome sequence revealed that nearly half of our genome is derived from transposable elements (also known as 'jumping genes'), discrete pieces of DNA that can move within the genome. This fact is especially striking because the protein-coding regions comprise only 1.5% of the human genomic DNA (Lander et al., 2001) (**Fig.1**).

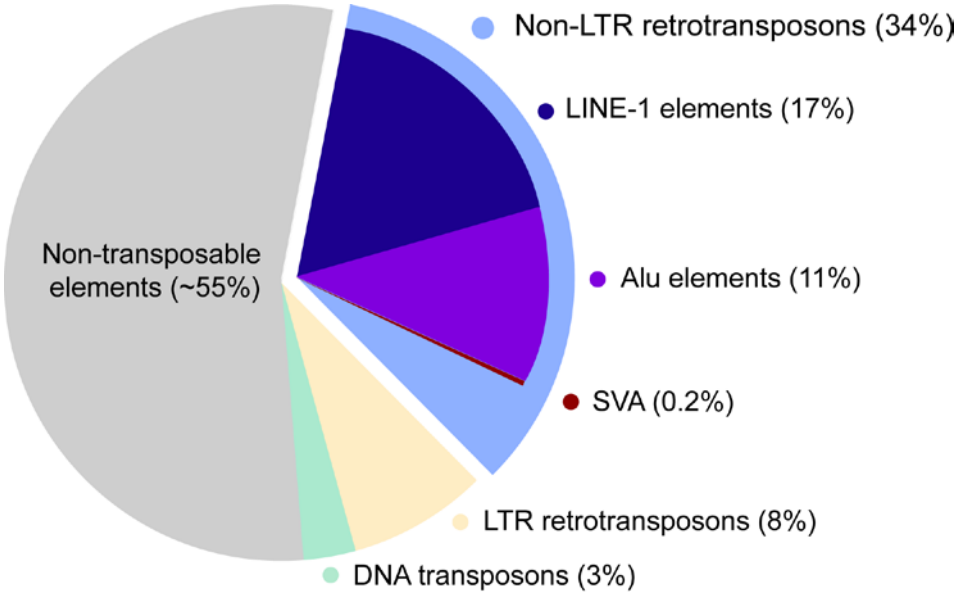

Fig. 1. Composition of the human genome.

Based on the mechanism used for transposition, mobile elements can be divided into two major classes: DNA transposons and retrotransposons. DNA transposons can excise themselves from the genome, move as DNA and insert themselves into new genomic sites (Craig, 2002). This is called 'cut-and-paste' mechanism. Nowadays, DNA transposons are not active in the human genome, but they were active during early primate evolution until ~ 37 million years (Myr) ago (Pace and Feschotte, 2007).

Retrotransposons replicate via an RNA intermediate that is reverse transcribed and inserted into a new genomic location (Craig, 2002) (**Fig.2**). This way of transposition is also known as 'copy-and-paste' mechanism.

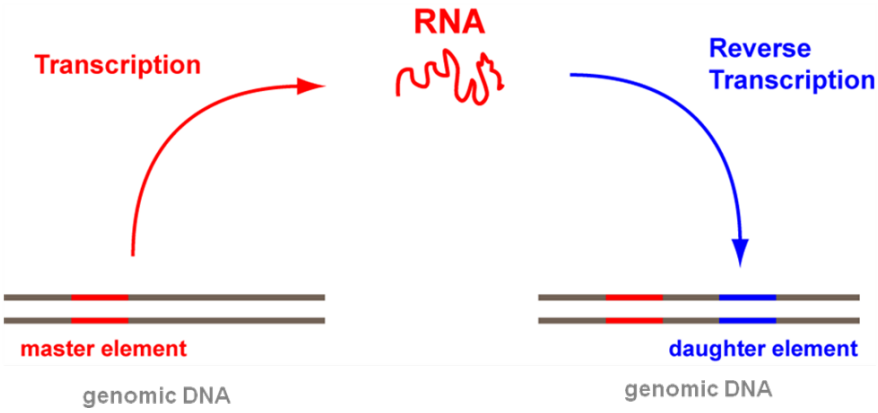

Fig. 2. Propagation of retrotransposons.

Retrotransposons can be subdivided into two groups distinguished by the presence or absence of long terminal repeats (LTRs). LTRs are repetitive sequences of 300 - 1000 bp in length that are found directly at 5' and 3' ends of long terminal repeat (LTR) retrotransposons and retroviruses. The replication mechanism of LTR retrotransposons is very similar to the one used by retroviruses. An example for human LTR elements is endogenous retroviruses (HERVs), which along with related elements account for ~ 8% of the genomic DNA (**Fig.1**). Most HERVs inserted in the human genome more than 25 Myr ago, and presently are virtually inactive (Lander et al., 2001) (Mills et al., 2007).

Non-LTR retrotransposons (NLRs, lacking LTRs) replicate via the target-primed reverse transcription (TPRT), which is fundamentally different from the replication mechanism used by LTR retrotransposons and retroviruses. LINE-1 (L1) is the most abundant non-LTR element in the human genome, directly accounting for 17% of genomic DNA, and is still active nowadays. The insertion rate is estimated to be 1 insertion in every 100 newborns (Huang et al., 2010). Alu and SVA elements comprise another 11% of the genome and are also active through the use of the L1 retrotransposition machinery (Lander et al., 2001) (Belancio et al., 2008) (Dewannieux et al., 2003).

1.2 Non-LTR retrotransposons in the human genome

Non-LTR retrotransposons are currently the only active elements in the human genome, as mentioned above. Therefore it is interesting to look more in detail at the structure of these elements and mechanism of transposition.

1.2.1 The L1 element – gene structure and retrotransposition cycle

There are more than 500.000 copies of L1 in the human genome, but only ~ 100 of them are functional full-length elements (Brouha et al., 2003). The full-length L1 is ~ 6 kb long and consists of a 5' UTR containing an internal RNA polymerase II promoter (shown as a black arrow in **Fig.3**), two open reading frames (L1ORF1p and L1ORF2p) and a 3' UTR ending with an oligo(A)-rich tail of variable length (see for a recent review Cordaux and Batzer, 2009) (**Fig.3**).

Fig. 3. A full-length L1 element.

L1ORF1p encodes a 40 kDa protein, which binds single-stranded nucleic acid and is necessary for RNP formation (see below). L1ORF2p encodes a 150 kDa multidomain protein containing an N-terminal endonuclease (EN), a reverse transcriptase (RT), and a cysteine-rich C-terminal CCHC zinc-knuckle motif (Z) of unknown function. The crystal structure of the endonuclease domain shows that it is closely related to the human apurinic/apyrimidinic DNA repair endonuclease (APE1) and its nicking specifity together with other parameters is important for integration site selection. (Weichenrieder et al., 2004) (Repanas et al., 2007) (**Fig 4**).

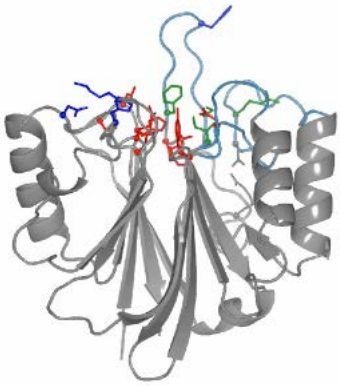

Fig. 4. The crystal structure of L1 endonuclease.

The reverse transcriptase seems most closely related to telomerase RT, both mechanistically and on the sequence level (A.M. Schneider and O. Weichenrieder, unpublished data).

The molecular machinery encoded by L1ORF1p and L1ORF2p is required for L1 retrotransposition. First, the transcribed L1 RNA is exported into the cytoplasm, where it is translated. Both encoded proteins, L1ORF1p and L1ORF2p associate with the L1 RNA molecule, from which they have been translated (**Fig. 5**) (Wei et al., 2001).

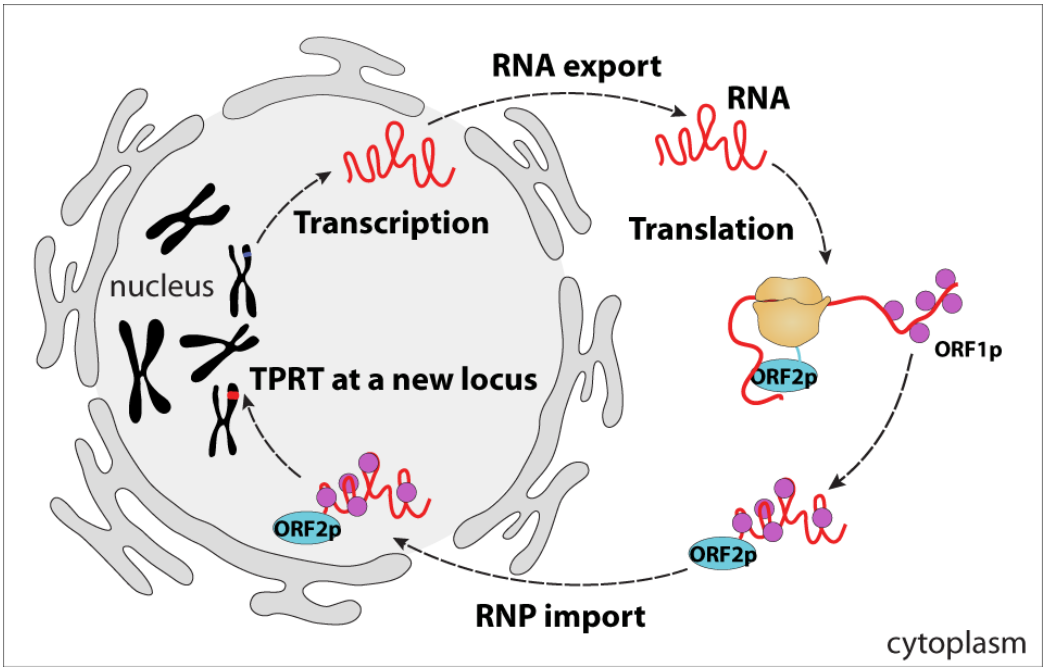

Fig. 5. Retrotransposition cycle of L1.

The mechanism for this apparent *cis*-preference is not yet understood, but it assures that only functional L1 RNA is incorporated into L1 RNPs (Wei et al., 2001). After RNPs are imported back into the nucleus, the endonuclease domain of L1ORF2p nicks chromosomal DNA at a preferential site. This frees a 3' hydroxyl group on the nicked DNA strand that serves as a substrate for the reverse transcriptase domain of L1ORF2p to carry out the reverse transcription. It is unclear how the second DNA strand is nicked and how the integration of the element is finished. This process of the L1 integration into a genome is also known as target-primed reverse transcription (TPRT) (Luan et al., 1993) (Cost et al., 2002).

Both L1ORF1p and L1ORF2p are essential for the L1 retrotransposition. The fact that L1 encodes its own reverse transcriptase makes it the only autonomous

element in the human genome. Other non-LTR retrotransposons use the L1 machinery for their propagation, and therefore are called non-autonomous elements.

1.2.2 The Alu element

The most abundant non-autonomous non-LTR retrotransposons in the human genome are Alu elements. With more than 1 million copies they are the most successful elements in the genome in terms of copy number (Lander et al., 2001). A full-length Alu element is ~ 300 bp long and has a dimeric structure formed by the fusion of two monomers derived from the 7SL RNA gene (a component of the signal recognition particle) (Ullu and Tschudi, 1984) (**Fig.6**).

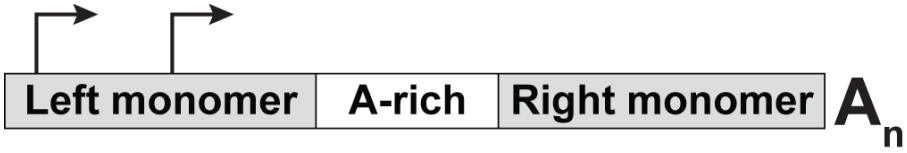

Fig. 6. Alu element.

The monomers are separated by an A-rich linker region. The 5'-region contains an internal RNA polymerase III promoter (A and B boxes, shown as arrows in **Fig.6**) and the element ends with an oligo(A)-rich tail of variable length (Batzer and Deininger, 2002). Alu elements do not encode proteins, but, likely, use the L1 retrotransposition machinery and transpose via the TPRT mechanism (Dewannieux et al., 2003). It is not clear how they recruit the L1ORF2p protein so efficiently, because, as mentioned above, L1 encoded proteins show a strong *cis*-preference for L1 RNA. After transcription Alu RNA is exported to the cytoplasm, where it forms RNPs with SRP9 and SRP14 proteins (Chang et al., 1996). One likely hypothesis is that the Alu RNPs interact with ribosomes positioning Alu RNA in proximity to the nascent L1ORF2p (Boeke, 1997). Another possibility is that Alu RNPs recruit L1ORF2p in the nucleus and then directly proceed with TPRT (GarciaPerez et al., 2007). L1ORF1p seems not required, but can enhance the Alu retrotransposition (Dewannieux et al., 2003) (Wallace et al., 2008).

1.2.3 The SVA element

Another example of non-autonomous non-LTR rerotransposons are SVA elements. There are ~ 3000 copies of SVA element in the human genome. A full-length SVA element is ~ 2 kb long and consists of a hexamer repeat region, an Alu-like region, a region composed of a variable number of tandem repeats (VNTR), a HERV-K10-like region and an oligo(A)-rich tail of variable length (Ostertag et al., 2003) (Wang et al., 2005) (**Fig.7**).

Fig. 7. SVA element.

SVA elements lack an internal promoter, so they have to rely on promoter activity in flanking regions. It was suggested that SVA elements are transcribed by RNA polymerase II. SVA elements, like Alu elements, do not encode any proteins and are mobilized by the L1 retrotransposition machinery (Ostertag et al., 2003) (Wang et al., 2005). It is still unclear which determinants in an SVA element allow it to gain access to the L1ORF2p protein and at which stage of retrotransposition this happens (Hancks and Kazazian, 2010).

1.3 Impact of non-LTR retrotransposons on the human genome evolution

Although the first mobile elements were discovered in the 1940s by Barbara McClintock (McClintock, 1956), their functional significance has been long underestimated. Transposable elements were even called selfish DNA or “junk DNA”. But over the past decades the knowledge about how mobile elements affect our genome has accumulated, revealing that non-LTR retrotransposons have a huge impact on both the structure and function of the genome.

1.3.1 Impact of non-LTR retrotransposons on human genome structure

There are 65 cases known where non-LTR retrotransposons insertions caused heritable human diseases, such as haemophilia, cystic fibrosis, neurofibromatosis and others (reviewed in Han and Boeke, 2005). Insertional mutagenesis that occurred in these cases is only one of the many ways in which retrotransposons can generate genomic instability (**Fig. 8**).

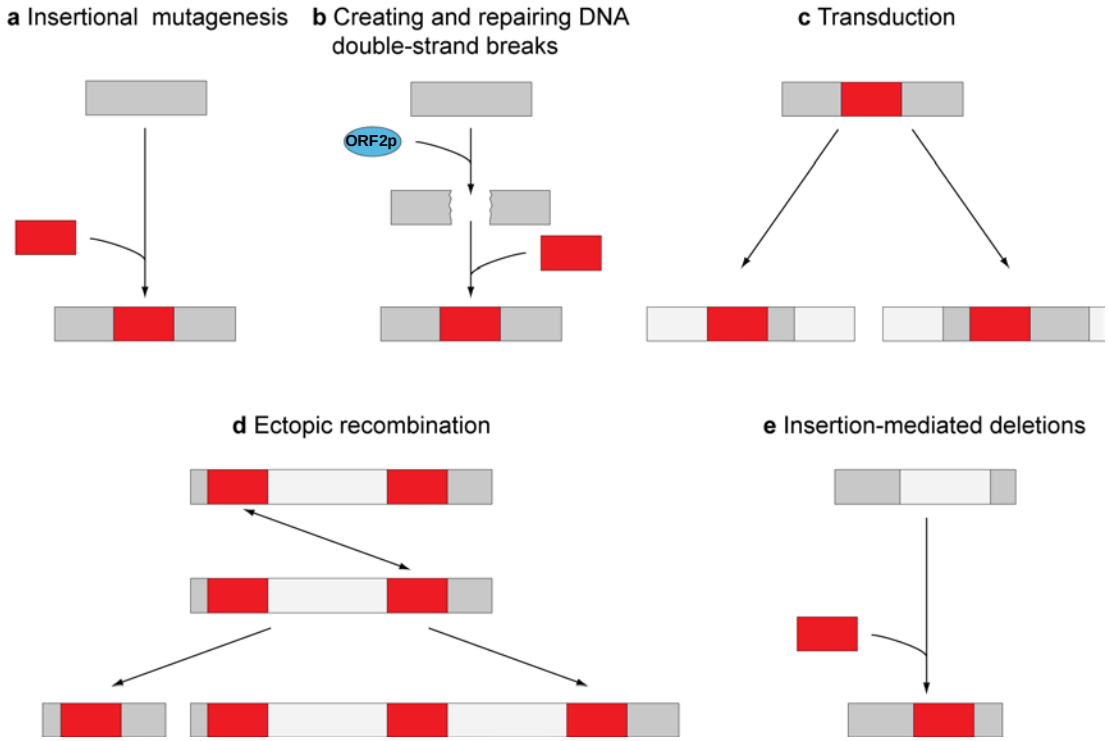

Fig. 8. Impact of retrotransposons on human genome structure (adopted from Cordaux and Batzer, 2009).

L1 can integrate into DNA with double-strand breaks (DSBs) in an endonuclease-independent way, thereby repairing the DSBs (Morrish et al., 2002). L1ORF2p can also generate DSBs itself, many of which are not associated with the L1 insertions and are prone to recombination (Gasior et al., 2006) (**Fig.8b**). Disfunctional telomeres also can serve as a substrate for the endonuclease-independent L1 insertions. This lead to the hypothesis by Morrish et al. that the endonuclease-indepent retrotransposition is an ancient mechanism of RNA-mediated DNA repair, which existed before L1 acquired an endonuclease activity (Morrish et al., 2007).

L1 and SVA elements can sometimes carry upstream or downstream flanking sequences, which also become integrated in the genome along with the mobile element sequence. This process is called 5' or 3' transduction, respectively (**Fig. 8c**). 5' transduction occurs when transcription starts from the promoter located upstream and continues into the retrotransposon sequence. In case of 3' transduction, the RNA transcription does not terminate at the weak polyadenylation signal in the mobile element, but at an alternative signal located further downstream (Wang et al., 2005) (Moran et al., 1999). Transduction can lead to the exon shuffling and creation of new genes. One example is the SVA-mediated transduction of the AMAC1 (acyl-malonyl condensing enzyme 1) gene, which resulted in the formation of a new gene family in the human genome (Xing et al., 2006).

Due to the high copy number of L1 and Alu elements in the genomic DNA, recombination between non-allelic copies of these elements can occur (**Fig. 8d**). Such recombination can lead to deletions, duplications or inversions of the genome fragments. In particular, Alu elements are known to be significantly enriched at the boarders of the segmental duplications in the human genome. Segmental duplications are quite large stretches of DNA (>10 kb) with more than 90% sequence identity. The role of Alus in creating segmental duplications might be important since about 5% of the human genome has been duplicated in the past 40 Myr (Bailey et al., 2003). Deletions mediated by the recombination between retrotransposons are also an important source of genetic variation, and in several cases such deletions have been reported to cause diseases (Deininger and Batzer, 1999) (Han et al., 2008).

Deletions of the integration site can be also occasionally generated by L1 and Alu elements in the retrotransposition process (**Fig. 8e**). One of the examples is the L1 mediated deletion of 46 kb DNA from the gene encoding pyruvate dehydrogenase complex, component X (*PDHX*), which caused the pyruvate dehydrogenase complex deficiency (Mine et al., 2007).

1.3.2 Impact of non-LTR retrotransposons on human gene expression

Apart from changing the human genome structure in the ways described above, retrotransposons also have a considerable effect on the gene expression, since about 80% of human genes contain a L1 insertion.

One way how L1 and Alu elements affect the gene expression is providing alternative splicing sites and promoting exonization (**Fig. 9a**). Another, more subtle way is an attenuation of the gene expression. Transcription levels of the genes containing L1 insertions can decrease because polymerase is known to stall in the L1 sequence. Polyadenylation signals in the mobile element can also lead to a premature termination of the gene transcript (Han et al., 2004) (**Fig. 9b**).

The third way, in which L1 can modulate gene expression, is generation of transcripts in both directions from its sense and antisense promoters. Transcription from the antisense promoter may regulate the expression of some genes (**Fig. 9c**) (Faulkner et al., 2009).

Finally, L1 and Alu insertions are often silenced through the DNA methylation, which is one of the mechanisms used by cells to protect the genome. Silencing could also spread to genes nearby the insertion and repress their transcription (Hata and Sakaki, 1997) (Rubin et al., 1994) (**Fig. 9d**). L1 induced silencing is important for X chromosome inactivation in eutherians (Chow et al., 2010).

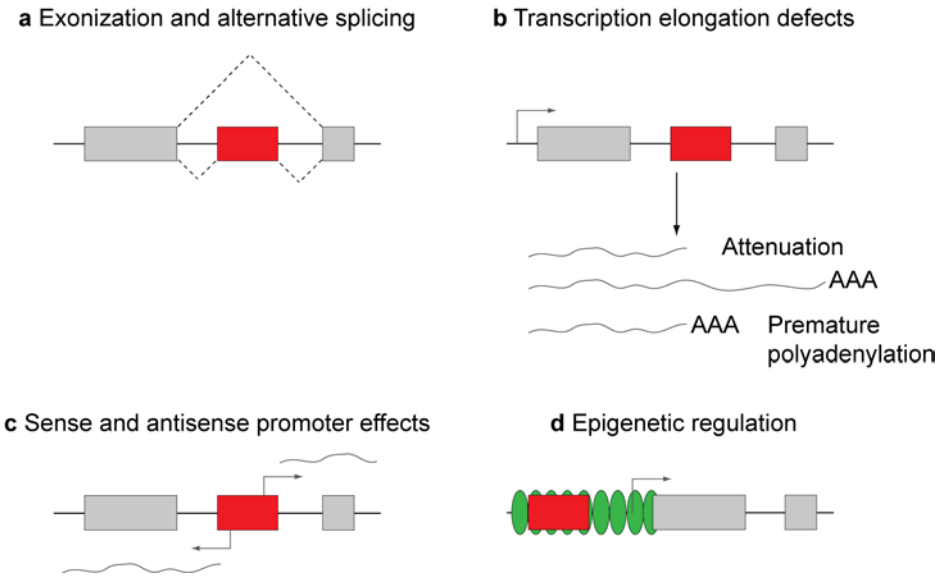

Fig. 9. Impact of retrotransposons on gene expression (adopted from Cordaux and Batzer, 2009).

1.3.3 Mechanisms used by the cell to control non-LTR retrotransposition

With so many ways of how retrotransposons can affect the genome structure and function, the cell has developed several mechanisms in order to protect the integrity of the genome.

One of these ways is methylation of retrotransposon sequences, as mentioned above. It is especially important in the germline cells. In mice, a knockout of the *Dnmt3L* gene, which encodes a *de novo* methyltransferase, leads to the activation of L1 and other retrotransposons in the sperm cells precursors, and, consequently, to a failure of the male germline (Bourchis and Bestor, 2004).

Recently, another type of chromatin modification, histone deacetylation, has been involved to play a role in the silencing of freshly delivered engineered L1 insertions in human embryonic carcinoma cells (GarciaPerez et al., 2010). This silencing requires the insert to be integrated via TPRT.

It is suggested that RNA-based silencing mechanisms are used for the retrotransposon inhibition as well. The activity of the L1 antisense promoter leads to production of small antisense RNAs, and it has been shown that depletion of Dicer in the cells transfected with L1 element results in decreased levels of retrotransposition (Yang and Kazazian, 2006). Several studies have shown piRNAs to be involved in the repression of retrotransposition during the germline development. Mutations in Piwi proteins in mice lead to the increase in activity of L1 and other elements in spermatocytes (Aravin et al., 2007). However, this does not seem to be a consequence of RNA slicing activity of the Piwi proteins, but rather an indirect effect of the piRNAs on DNA methylation patterns. This view is corroborated by recent studies showing that Tudor domain-containing (TDRD) proteins are also involved in piRNA mediated repression of retrotransposons. MILI and MIWI2 proteins, which are responsible for piRNA biogenesis in mice, interact with TDRD1 and TDRD9, respectively. MILI/TDRD1 complex localizes to pi-bodies, which are the hypothetical sites for processing transposon transcripts into piRNAs (Aravin et al., 2009) (Reuter et al., 2009). Activity of this complex is crucial for *de novo* methylation of retrotransposons in the mouse germ line (Aravin et al., 2008). TDRD9 is essential for male fertility. *Tdrd9* mutant germ cells lack L1-derived piRNAs and are also defective in *de novo* methylation of L1 retrotransposons (Shoji et al., 2009). MIWI2 and TDRD9 localize to cytoplasmic granules, distinct from pi-bodies. These granules contain components typical for P-bodies, enriched in translationally repressed mRNAs, such as GW182, DCP1a and DDX6 (Shoji et al., 2009). The existence of such chimeric piP-bodies suggests cooperation between the piRNA and RNA silencing pathways.

Another line of retrotransposition control is carried out by the APOBEC proteins. There are seven different APOBEC3 protein forms encoded in the human

genome. They were first identified due to their role in the HIV replication inhibition. APOBEC3A inhibits L1 retrotransposition, with APOBEC3B and 3C involved in this process to a lesser extent. APOBEC3G inhibits Alu activity. The mechanism of retrotransposon inhibition by APOBEC proteins is not clear (see Schumann, 2007 for a review).

1.3.4 L1 retrotransposition activity in the germline and somatic cells

Despite different control mechanisms, in some cells non-LTR retrotransposons are active. It would be beneficial for a mobile element to retrotranspose in germline cells or their precursors in the embryo, so that the new insertion is passed to future generations, and this is really the case: L1 insertions and higher levels of L1ORF1p have been observed in mouse germ cells, theca cells of adult ovaries and in embryonic testis (Ostertag et al., 2002) (Trelogan and Martin, 1995).

L1ORF1p and full-length RNA transcripts have been detected also in several transformed cell lines (Garcia-Perez et al., 2007). In a recent study Iskow et al. detected new L1 insertions occuring at high frequency in human lung cancer genomes. Genome-wide analyses of methylation status suggest that altered DNA methylation may be responsible for the high levels of L1 retrotransposition observed in these tumors (Iskow et al., 2010). In general, the loss of methylation associated with cancer could lead to derepression of mobile elements, while new insertions would further promote the genetic instability in cancer cells (Schulz, 2006).

However, in some cases L1 retrotransposition has been observed in somatic cells as well. First, it was shown that L1 can retrotranspose in the brain stem cells in transgenic mice (Muotri et al., 2005). Then, human neural progenitor cells, derived either from fetal brain or from embryonic stem cells, were found to support L1 retrotransposition as well. A highly sensitive quantitative PCR method was used to show that retrotransposition occurs in neural progenitor cells also in vivo (Coufal et al., 2009). These results suggest that L1 plays a role in creating interindividual differences through its activity in brain cells.

The role of L1 retrotransposition in creating interindividual genetic variation is supported also by recent genome-wide assays, which draw a much more dynamic portrait of our genome than previously appreciated (Beck et al., 2010) (Ewing and Kazazian, 2010) (Huang et al., 2010) (Iskow et al., 2010). Remarkably, the bulk of L1

retrotransposition activity results from only ~80 to 100 highly active elements in each individual human genome, which generally are very dimorphic (allele frequencies < 5%) and hence not systematically explored yet (Beck et al., 2010) (Huang et al., 2010).

1.4 Origin and function of the L1ORF1p

Many non-LTR retrotransposons encode a protein (ORF1p) upstream of the reverse transcriptase and endonuclease (ORF2p). For L1ORF1p and some other ORF1p proteins it has been shown that the protein binds nucleic acids and forms RNPs with the retrotransposon RNA. This triggered comparisons to the retroviral RNA packaging protein Gag that is also found in the context of LTR retrotransposons, and also to RNA packaging proteins from RNA viruses. In the following section I will therefore summarize the functional data on L1ORF1p, which was available at the beginning of my thesis, and then briefly describe the RNA binding proteins of LTR retrotransposons, retroviruses and RNA viruses.

1.4.1 L1ORF1p

Both proteins encoded by L1 are necessary for retrotransposition, as mentioned above. In case of L1ORF2p it is known that its endonuclease and reverse transcriptase enzymatical activities are needed for the propagation of L1. The role of the L1ORF1p in the retrotransposition remained much more elusive for a long time, because the amino acid sequence of the protein lacked homology with any protein of a known function (Hohjoh and Singer, 1996). L1ORF1p is a 40 kDa protein in humans. Its size varies across species as the N-terminal region is not conserved regarding both sequence and length (Furano et al., 2004) (Martin, 2006). Sedimentation studies and atomic force microscopy indicate that purified murine L1ORF1p forms unusual, dumbbell-shaped trimers that are held together by a coiled coil formed between sequences in the N-terminal halves of the monomers (Basame et al., 2006) (Martin et al., 2003). The other, well-conserved half of murine L1ORF1p is highly basic and binds nucleic acids. It is known that L1ORF1p binds single-stranded RNA and DNA with high affinity (Hohjoh and Singer, 1997) (Kolosha and Martin, 1997). At the beginning of this thesis no classical sequence motifs were identified in the L1ORF1p, which would help to understand the mode of nucleic acid binding. Later on, the NMR structure of the murine L1ORF1p C-terminal domain

(CTD) became available (Janusczyk et al., 2007). The structure shows a rare αββββαα fold (no futher examples in the PDB) and does not relate L1ORF1p to any protein of known function. The CTD domain alone has very low affinity to nucleic acids, and therefore cannot explain the properties of the trimeric form (**Fig.10**).

Fig. 10. NMR structure of the CTD from mouse L1ORF1p.

Other facts, which were known about L1ORF1p in the beginning of the thesis are the following. L1ORF1p is thought to be a nucleic acid chaperone with annealing and displacement activities. (Kulpa and Moran, 2005) (Martin, 2006). Like L1ORF2p, L1ORF1p shows a remarkable *cis*-preference, that is, it associates preferentially with its encoding transcript (Wei et al., 2001) (Kulpa and Moran, 2005). L1ORF1p can be localized in the cytoplasm (in putative stress granules) as well as in the nucleus (Goodier et al., 2007) (Kirilyuk et al., 2008), and it can also be identified in large L1 RNPs fractionated from cytoplasmic extracts (Kulpa and Moran, 2005) (Martin, 1991) (Hohjoh and Singer, 1996).

1.4.2 The Gag proteins of retroviruses and LTR retrotransposons

RNA packaging proteins of non-LTR retrotransposons on the sequence level are unrelated to their counterparts from LTR elements, which resemble RNA binding proteins of retroviruses in their architecture and post-translational processing. Proteins responsible for RNA binding in LTR retrotransposons and retroviruses are called nucleocapsids. A nucleocapsid protein is translated initially as a part of multidomain precursor protein called Gag. After viral particle assembly and maturation, Gag is proteolytically processed by the retroelement- or virus-encoded protease into several fragments, including the nucleocapsid protein. Zinc finger motifs are a characteristic feature of nucleocapsid proteins. It has been shown also that

nucleocapsids possess nucleic acid chaperone activity, that is, they facilitate arrangement of nucleic acids into thermodynamically most stable form (see Wu et al., 2010 for review).

1.4.3 RNA packaging proteins from RNA viruses

There are myriads of different RNA viruses, but all of them have to package their genome into virus particles. The two main ways for protecting the viral genome are:

(1) the process known as encapsidation, which is used by most positive-sense RNA and double-stranded RNA viruses. In this case, the viral genome is placed into a protein shell, called capsid. Encapsidation can take different functional and structural forms.

(2) coating the length of genomic RNA with a nucleocapsid protein, which is common for negative-sense RNA viruses (Raymond et al., 2010).

The latter mechanism, coating, is more similar to the NLR RNA packaging than the encapsidation. Most of the negative-sense RNA viruses, such as human respiratory syncytial virus (RSV) (Tawar et al., 2009),vesicular stomatitis virus (VSV) (Ge et al., 2010) and others, have symmetrical ring-like or helix-like RNPs, however. A regular helical architecture has not been described for non-LTR retrotransposons RNPs so far. But, interestingly, a recent structure of Rift Valley fever virus (RVFV) RNP determined by EM showed string-like appearance, without any helical symmetry (Raymond et al., 2010). A similar organization was described for the reconstituted complex of L1ORF1p with RNA visualized by AFM (Basame et al., 2006).

Structures of some nucleoproteins were solved, including influenza A virus (Ye et al., 2006), rabies virus (Albertini et al., 2006), HRSV (Tawar et al., 2009), VSV (Green et al., 2006), and Borna disease virus (BDV) (Rudolph et al., 2003). These proteins have a common bilobal architecture, and the connection between the lobes is suggested to be a hinge point providing flexibility for the protein upon interaction with the viral polymerase or other factors (Green et al., 2006) (Tawar et al., 2009).

The interior between two lobes is often positively charged and serves as the RNA binding site, contacting primarily the backbone and leaving the bases exposed to different extent. Both lobes can have highly flexible protrusions, which are responsible for the interaction with the neighbouring subunits in the RNP (**Fig.11**) (Albertini et al., 2006) (Green et al., 2006) (Tawar et al., 2009).

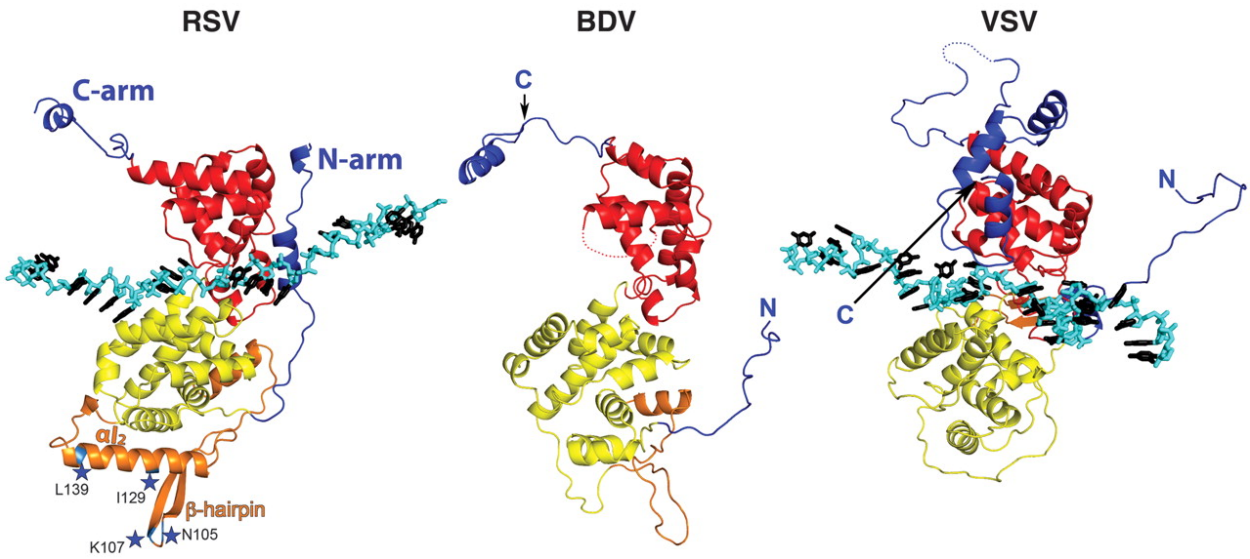

Fig. 11. 3D fold of RSV, BDV and VSV N. (Left) The variable region is in orange, with the β hairpin highlighted and labeled with the RSV604 resistance mutation sites. The BDV (center) and VSV (right) N structures, respectively, oriented and colored identically to RSV N. The conserved domains are red and yellow. The RNA is colored in blue and black (adopted from Tawar et al., 2009).

Nucleoproteins are essential for the viral RNA replication, because naked, not coated, RNA cannot serve as a template for efficient synthesis of a new genome copy. It was suggested that due to the likely flexibility of nucleoproteins the local change in the RNP structure is possible, which would result in the transient opening of the RNA binding groove during the readout by polymerase (Tawar et al., 2009) (Albertini et al., 2006) (Green et al., 2006).

Some mutations in nucleoproteins show defects in the viral RNA synthesis, but do not affect the RNA binding properties. Together with other lines of evidence this points to the fact that nucleoproteins are involved in interactions with viral and cellular factors necessary for RNA transcription and replication (Ortin, 2003).

Despite the common architecture and RNA binding principles, nucleoproteins from different viruses are very divergent on the protein sequence level, which makes it not possible to predict their common ancestor and rather suggests that they might be the result of convergent evolution (Albertini et al., 2008).

1.5 Aims of the work

Despite a huge impact of L1 mediated retrotransposition and its consequences on the human genome, the molecular details of the L1 propagation were poorly

understood at the beginning of this work. The role of the L1ORF1p in retrotransposition was particularly obscure. The most important questions necessary for understanding the L1ORF1p function were:

- what is the domain structure of the protein
- how it binds nucleic acids
- what is the evolutionary origin of L1ORF1p and L1 retrotransposon.

To answer these questions, we first identified distinct domains within L1ORF1p computationally, and then verified the predictions experimentally. Then, we used X-ray crystallography and NMR to determine the structures of individual domains, as well as of the trimeric L1ORF1p. From these studies L1ORF1p emerges as a highly sophisticated RNA binding protein that shows an unprecedented flexibility in the arrangement of its individual domains. These structures together with *in vitro* and *in vivo* experiments also suggest how single-stranded nucleic acids are bound by the trimer.

Phylogenetic analyses of a mammalian L1ORF1p suggest an ancient origin of the RRM domain and support the modular evolution of non-LTR retrotransposons. Finally, the trimeric structure of L1ORF1p is interesting, when compared to viral fibres and membrane fusion proteins, which form similar trimers. This provides a new twist in the discussion about the origin of non-LTR retrotransposons.

2 Materials and methods

The methods used in the experimental part of this thesis were carried out as described in standard laboratory manuals. In this section, only modified methods are described in detail.

2.1 Materials

2.1.1 Chemicals

All chemicals were purchased at analytical grade from the following companies unless stated otherwise: Invitrogen, Serva, Sigma, Sigma Aldrich, PeqLab, Applichem, Alfa Aesar, Roche, Aldrich, Applied Biosystems, Qiagen, Merck, Roth, Gibco, Riedel-de Haen, Acros Organics, Fluka, Bio-Rad, New England Biolabs, Hampton, Promega.

2.1.2 Enzymes

DNA modifying enzymes, e.g. restriction endonucleases, T4 ligase, DNase I and *Pfu* polymerase were obtained from Stratagene, New England Biolabs, and Roche. Reactions were carried out according to manufacturer's directions.

2.1.3 Buffers and solutions

Coomassie staining solution	45% Methanol
	10% Acetic acid
	1 g/l Brilliant Blue R-250 Coomassie
Destaining solution	25% Isopropanol
	10% Acetic acid
2x protein sample buffer	100 mM Tris/HCl pH 6.8
	4% SDS
	0.05% bromphenol blue
	10% β-mercaptoethanol

1x Laemmli buffer	0.1% SDS
	190 mM Glycine
	24.8 mM Tris base
10x M9 salts	420 mM Na_2HPO_4
	220 mM KH_2PO_4
	86 mM NaCl
	0.13 mM NH_4Cl
	pH 7.2
5x DNA Agarose dye	20% Ficoll 400
	1 mM EDTA
	0.1% SDS
	0.05% Bromphenol blue
1x TBE buffer	89 mM Tris, pH 8.3
	89 mM Boric acid
	2.5 mM EDTA
1000x Vitamin mix	2.86 mM Choline chloride
	1.13 mM Folic acid
	2.1 mM Panthotenic acid
	5.55 mM Myo-Inositol
	2.02 mM Pyridoxal phosphate
	1.48 mM Thymidine HCl
	0.13 mM Riboflavin
	4.09 mM Biotin

2.1.4 Media

LB medium	300 mM NaCl, pH 7.0
	1% Peptone
	0.5% Yeast extract
LB agar	300 mM NaCl, pH 7.0
	1% Peptone
	0.5% Yeast extract
	2% Bacto agar
Minimal medium for	
Se-methionine cultures	1x M9 salts
	1 mM $MgSO_4$
	0.1 mM $CaCl_2$
	0.2% Glucose
	1x Vitamin mix
Minimal medium for	
^{15}N-, ^{13}C- cultures	1x M9 salts containing 128 mM $^{15}NH_4Cl$
	2 mM $MgSO_4$
	0.1 mM $CaCl_2$
	0.2% Glucose
	1x Vitamin mix
DMEM culture medium	Dulbecco's Modified Eagle Medium with
	25 mM D-Glucose
	1 mM Sodium Pyruvate
	no L-Glutamine

2.1.5 Bacterial strains

Escherichia coli XL1-blue (K12) (cloning strain, Invitrogen)

Escherichia coli Rosetta 2 (DE3) (protein expression strain, Novagen)

2.2 Methods

2.2.1 Bioinformatics

Individual NLRs and their ORF1p sequences were identified by tBLASTn searches using queries from the literature or from RepBase (Jurka et al., 2005). ORF1p sequences were analyzed for similarity to known domains using profile hidden Markov models as implemented in HHpred (Söding et al., 2005).

2.2.2 Cloning

DNA sequences corresponding to the different human L1ORF1p constructs were PCR amplified from a plasmid (pJM103) encoding a functional human L1 element (Moran et al., 1999). PCR products were purified using QIAquick PCR Purification Kit (Qiagen).

Restriction digest of the purified PCR products and respective vectors was done according to the enzyme supplier's instructions. Digestion products were separated on a preparative agarose gel and then isolated using the Qiagen Gel Extraction Kit and then ligated.

For standard cloning steps chemically competent *E.coli* XL1-blue cells were used for transformation of plasmid DNA.

To generate plasmids containing mutations in the protein of interest site-directed PCR mutagenesis was used. A typical reaction mixture for mutagenesis PCR contains the following components:

10x Pfu Ultra buffer	5 µl
Pfu Ultra DNA Polymerase	1 µl
primer forward (2 µM)	1 µl
primer reverse (2 µM)	1 µl
dNTPs (10 mM each)	1 µl

template DNA	200 ng
sterile H_2O	filled up to a total volume of 50 µl

The performed PCR program was the following:

step	temperature	time	
denaturation	94°C	5 min	
denaturation	94°C	1 min	35 cycles
primer annealing	56°C	1 min	
elongation	68°C	2 min/kb template	
final elongation	68°C	10 min	

After PCR 20 U of DpnI restriction enzyme were added to the reaction to digest (Dam)-methylated, non-mutated template DNA. After incubation for 2h at 37°C, 5 µl of the digested reaction mixture were used for transformation of chemically competent *E.coli* XL1-blue cells, or in case of large plasmids (~18 kb) for electroporation of electrocompetent *E.coli* XL1-blue cells.

M121A/M125I/M128I triple mutation corresponding to the murine sequence had been introduced into all trimeric L1ORF1p constructs to avoid aberrant initiation of bacterial translation.

Plasmid DNA was isolated using the Qiagen Mini Prep Kit and the sequences of cloning or mutagenesis products were verified by DNA sequencing.

2.2.3 Protein expression and purification

We have used several vectors for protein expression: pETM11 (derived from pET24d (Novagen), with a cleavable His-tag), pET15b ((Novagen), with a short uncleavable His_6-tag), pETM60 ((derived from pET24d (Novagen), with a cleavable NusA-tag) and pGEX6p1 ((GE Healthcare), with a cleavable GST-tag).

pETM11 was used for cloning and expression of hL1ORF1p-C (tagMVS^{254}-R^{328}), pET15b for hL1ORF1p-ΔN/1 (MAS^{106}-Q^{330}HHHHHH) and hL1ORF1p-ΔN/2 (MAS^{106}-N^{326}HHHHHH), hL1ORF1p-MC^{H6} (MGN^{157}-Q^{330}HHHHHH), pETM60 for

hL1ORF1p-M (GAMGN^{157}-D^{252}), and pGEX6p1 for hL1ORF1p-MC (GPLGSN^{157}-Q^{330}). Proteins were expressed in the *E. coli* strain Rosetta 2(DE3) (Novagen) at 20°C overnight. To uniformly label hL1ORF1p-MC^{H6}, hL1ORF1p-M, hL1ORF1p-C with $^{15}N/^{13}C$ or ^{15}N, cells were grown in M9 minimal medium supplemented with $^{15}NH_4Cl$ with or without $^{13}C_6$-glucose. Selenomethionine containing minimal medium was used to produce Se-Met substituted proteins.

Proteins were purified from cleared cell lysates by Ni^{2+} - or glutathione affinity steps. The proteins with uncleavable His-tag were directly subjected to gel filtration. In other cases the affinity tags were removed by proteolyitc cleavage (hL1ORF1p-MC, hL1ORF1p-M, hL1ORF1p-C) and proteins were further purified by heparin-affinity chromatography (hL1ORF1p) and gel filtration. Dithiothreitol was added to the samples of the pure proteins for nuclear magnetic resonance (NMR) spectroscopy to the final concentration of 1 mM.

2.2.4 Crystal growth and optimization

hL1ORF1p-M

Crystalline clusters of hL1ORF1p-M (9.7 mg/ml in 5 mM Tris/Cl, pH 8.0, and 300 mM NaCl) were obtained by vapor diffusion at 18°C mixing 0.8 µl of protein solution with 0.8 µl of reservoir (2.2M Na-malonate, pH 7.0) over 500 µl reservoir. Crystals were optimized by hair-seeding (1.7-1.9 M Malonate) and flash frozen in liquid nitrogen without additional cryoprotection.

hL1ORF1p-ΔN/1 and hL1ORF1p-ΔN/2

Initial crystals of hL1ORF1p-ΔN/1 (18 mg/ml in 5 mM Tris/Cl, pH8.0, and 300 mM NaCl, 2.0 mM β-ME) were obtained in many conditions by vapor diffusion (18°C) mixing 0.2 µl of protein solution with 0.2 µl of reservoir over an 80 µl reservoir. Crystals were optimized by manual screening around several initial conditions and flash frozen in liquid nitrogen with additional cryoprotection.

The best-diffracting crystal (2.1 Å resolution) belonged to cfI and was obtained with protein (hL1ORF1p-ΔN/1) containing seleno-methionine. It was obtained in sitting drops by mixing 0.3 µl sample (18 mg/ml in 5 mM Tris/Cl, pH8.0, and 300 mM NaCl, 2.0 mM β-ME) and 0.3 µl reservoir (100 mM Na-Hepes (pH=7.0), 1.1 M Na-

malonate), suspended over a reservoir of 80 µl. Cryoprotection was achieved by a final concentration of 2.0 M Na-malonate.

Crystals of cfII were grown from native protein (hL1ORF1p-ΔN/2, 24 mg/ml in 5 mM Tris/Cl, pH8.0, and 300 mM NaCl, 2.0 mM β-ME) over a reservoir of 200 mM K-citrate, 20% PEG 3350 and were cryoprotected with 10% PEG 400. Crystals of cfIII were grown from selenomethionine-containing protein (hL1ORF1p-ΔN/2, 16 mg/ml in 5 mM Tris/Cl, pH8.0, and 300 mM NaCl, 2.0 mM β-ME) over a reservoir of 100 mM Tris-Cl (pH=8.5), 100 mM Mg-acetate, 12% PEG 8000 and were cryoprotected with 25% PEG 400.

2.2.5 Data collection, structure determination and refinement

hL1ORF1p-M

Crystals containing seleno-methionine diffracted better than the initial crystals from the native protein. Diffraction data for the selenomethionine derivative were collected at a single wavelength (0.97154 Å) on beamline PXII of the Swiss Light Source. Images were processed by XDS (Kabsch, 2010). The structure was solved by single anomalous dispersion (SAD). We used autoSHARP (Vonrhein et al., 2007) to search for three selenium sites per molecule. Assignment of the correct hand and solvent flattening (optimum contrast at 51.6%) was done automatically. In the resulting map, ARP/wARP (Cohen et al., 2008) was able to trace 92% of the final model and built 43% of the side chains. The model was completed manually in COOT (Emsley et al., 2010), including alternative conformations. Refinement was done in REFMAC (Murshudov et al., 1997) and COOT iteratively, using anisotropic B-factors.

hL1ORF1p-ΔN/1 and hL1ORF1p-ΔN/2

Diffraction data were collected on a Mar 225 CCD detector on beamline PXII (X10SA) of the Swiss Light Source (Villigen, Switzerland) at a temperature of 90 K (see **Table 3** for Data Collection and Refinement Statistics). Data were recorded at multiple wavelengths; at the selenium absorption peak (0.97868 Å, K-edge) and at a high energy remote point (0.97100 Å). Images were processed with XDS (Kabsch, 2010).

The structure of cfI was solved by a combination of single wavelength anomalous dispersion (SAD) and molecular replacement (MR). In a first step a

polyalanine model (derived from PDB-ID: 2wpq) of the coiled coil could be placed by molecular replacement using Phaser MR (McCoy et al., 2007) from within the CCP package (CCP4, 1994) and the dataset collected at the peak wavelength. In a second step, using the resulting phase information, Phaser EP (McCoy et al., 2007) identified potential seleno-methionine sites, twelve of which had previously been found independently by SHELX C/D/E (Sheldrick, 2008). Six of these sites allowed us to manually place three copies of the RRM-domain high resolution crystal structure that we had determined previously (PDB-ID: 2w7a); the other six sites helped to build the CTD domains in the subsequent steps.

The structure containing the coiled coil and the RRM domains was used to start automated model building using ARP/wARP (Cohen et al., 2008) and BUCCANEER (Cowtan, 2006). This resulted in models containing large parts of the coiled coil and of the RRM domains that were subsequently refined in REFMAC (Murshudov et al., 1997) using the dataset collected at the remote wavelength. The failure to auto-build the CTD domains probably results from the weaker electron density for this part of the structure. Most likely this is due to static disorder in the crystal, considering the flexibility within the structure and the fact that different domain orientations are found with almost identical unit cell parameters (see below, cfII and cfIII). The structure was completed by iterative cycles of BUCCANEER runs and manual building in COOT (Emsley et al., 2010) combined with REFMAC refinement of the partial model after each of such cycles.

Chloride ions were assigned based on the similarity to other coiled coils, on suitable coordination distances and on crystallographic evidence (i.e. refinement as chlorides leads to temperature factors that are similar to the surrounding residues and to a good agreement with the crystallographic data).

Two other crystal forms (cfII and cfIII) were identified by molecular replacement (PHASER MR) using different CTD truncations of cfI as input models.

The missing CTD domains were placed manually in the electron density obtained from the respective molecular replacement solution, and the structures were rebuilt and completed manually in COOT.

Final refinement rounds for all three structures were done in PHENIX (Adams et al., 2010), refining TLS parameters for the individual domains between the hinges in addition to individual B-factors. Stereochemical properties were analyzed with MOLPROBITY (Davis et al., 2007) and WHATCHECK (Hooft et al., 1996).

2.2.6 Figures and homology modelling

Figures were generated in Pymol (http://pymol.org/) using the APBS plugin to visualize electrostatic surface potentials (Baker et al., 2001). The homology model for the ZfL2 esterase was created via HHPred (Söding et al., 2005) using MODELLER (Sali et al., 1995).

2.2.7 Nucleic acid binding experiments

Analytical size-exclusion chromatography was done on an AKTA™ Purifier-10 equipped with a Superdex200 10/300GL or Superdex75 10/300GL column (GE Healthcare), monitoring optical density (OD) simultaneously at 230 nm, 260 nm, and 280 nm. Protein concentrations were estimated from the theoretical molar extinction coefficients ε_{280} at 280 nm. Nucleic acid concentrations were estimated from ε_{260} as provided by the manufacturers. The relative contributions of nucleic acid and protein to the total absorption at each wavelength were calculated assuming constant ratios of $\varepsilon_{230}/\varepsilon_{280}$ for each substance (Müller et al., 2006).

$$c(P)=(OD_{230}(tot)-Q(R)*OD_{280}(tot)) / (\varepsilon_{280}(P)*d*(Q(P)-Q(R))$$

$$c(R)=(OD_{230}(tot)-Q(P)*OD_{280}(tot)) / (\varepsilon_{280}(R)*d*(Q(R)-Q(P)),$$

where $Q(P)=OD_{230}(P) / OD_{280}(P)$ and $Q(R)=OD_{230}(R) / OD_{280}(R)$, P=protein, R=RNA.

Components were mixed in chromatography buffer (20 mM Tris/HCl, pH 8.0, 200-300 mM NaCl, and 0-10 mM $MgCl_2$) using starting concentrations between 20 and 100 µM. After 5 min at 18°C or 20°C, 100 µl were injected on the column (18°C) at a flow rate of 0.5 ml/min.

2.2.8 Cell culture

Mutants of the L1 reporter construct were generated by site-directed mutagenesis as described above using the plasmid pJM101/L1.3 as a template (**Fig.12**). DNA sequencing was used to verify that no other mutations were introduced in the L1 reporter construct. Plasmid DNA for transfection was isolated using the Qiagen Midi Prep Kit.

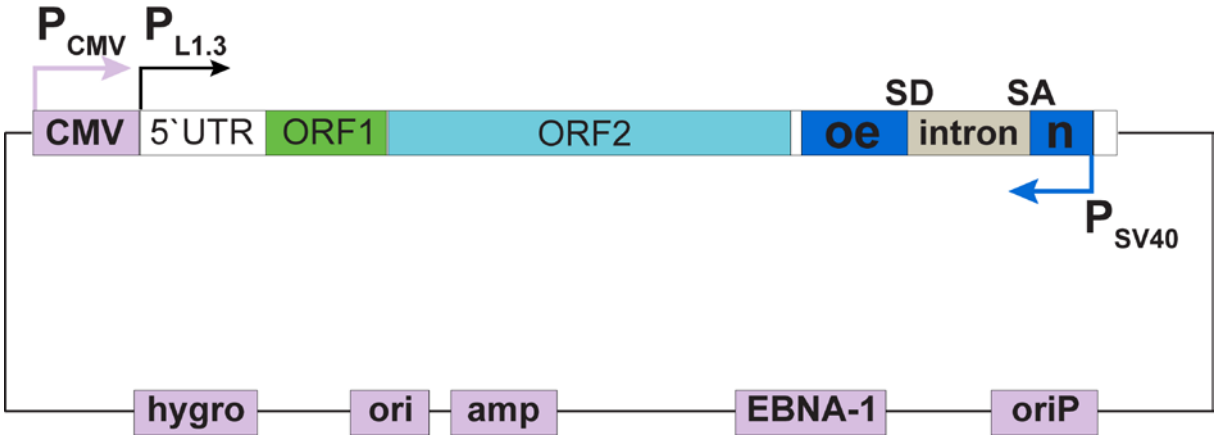

Fig. 12. Schematic representation of pJM101/L1.3. Regulatory elements of the plasmid backbone are shown in light violet color, the parts of the L1 reporter construct are shown as following: L1 5'UTR in white and the L1 promoter is indicated as a black arrow; L1ORF1 in green, L1ORF2 in light blue; the neomycin resistance gene in dark blue; the intron in gray, with splicing sites indicated above and SV40 promoter as a black arrow.

The retrotransposition cell culture assay was modified from reference (Moran et al., 1996) (**Fig.14**). Briefly, wild type L1 reporter construct and mutant variants were transfected in HeLa cells (with confluency ~90-100%) in a six-well plate. For monitoring the transfection efficiency, luciferase reporter vector (modified from pCINeo-Rluc, Promega, **Fig.13**) was co-transfected with each L1 construct.

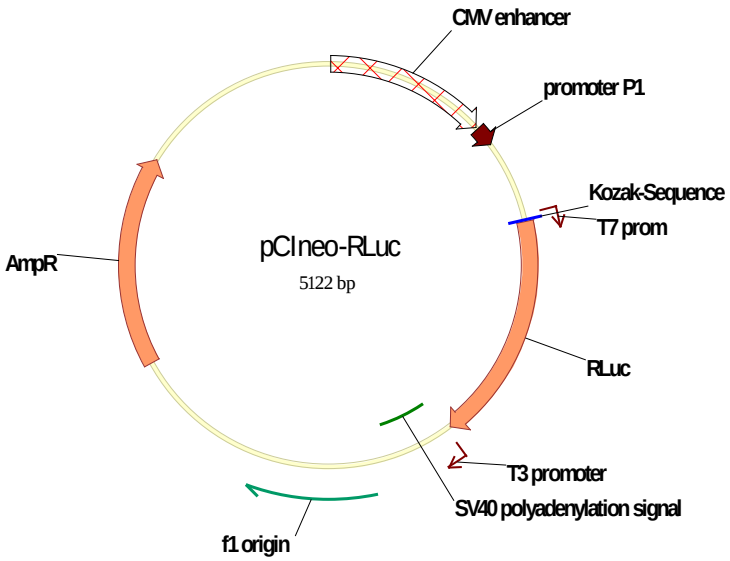

Fig. 13. Plasmid map of the modified pCINeo-Rluc.

Cells in each well were split 48h posttransfection into two wells. One half of the cells was used to measure the luciferase activity levels on day 3 posttransfection. The other half of the cells was grown for 12-13 days in DMEM containing G418. The

G418^r cells were fixed and stained with Giemsa, colony numbers were scored, and the retrotransposition frequency was determined as the number of G418^r colonies per number of transfected cells.

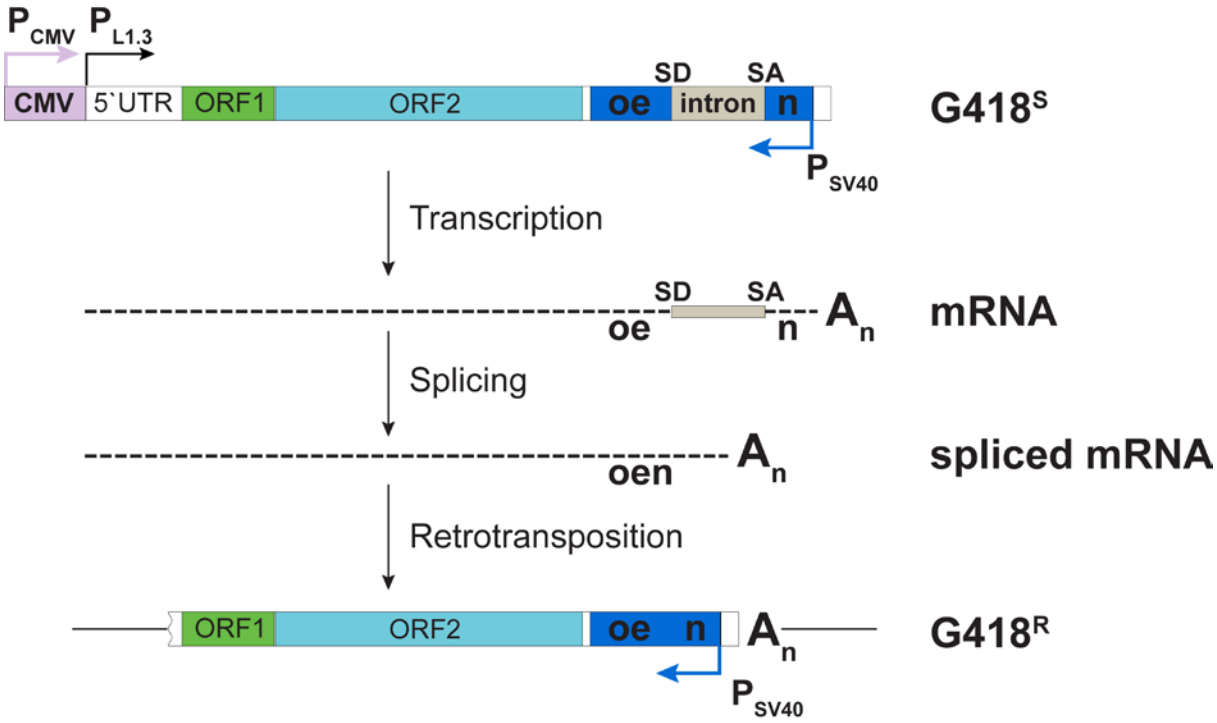

Fig. 14. Schematic representation of L1 retrotransposition assay. In the transfected plasmid, transcription of the L1 reporter construct is driven from CMV promotor. Transcription of the neomycin resistance gene from SV40 promotor cannot yield a functional protein product, because there is an intron in the gene in the orientation opposite to the transcription direction. After transcription of the L1 construct from CMV promotor, the intron can be spliced out. If retrotransposition of the reporter construct occurs, the neomycin resistance gene can be transcribed from the SV40 promotor and also can be translated due to the absence of the intron. This confers the resistance of the cells to neomycin, and allows the colony growth.

2.2.9 Luciferase assay

To measure the levels of luciferase activity, the cells were washed with ice-cold PBS on the next day after splitting, then incubated with 500 µl of the Passive Lysis Buffer (Promega) per well for 1h at room temperature. After the incubation the plates were frozen at -20°C to improve the lysis, then thawed and the lysed cells in each well were resuspended manually. 4 µl of lysed cells from each well in the 6-well plate were pipetted into 96-well plate, this was done in triplicate to improve the measurement accuracy. The luciferase activity was measured using the reagent for *Renilla* luciferase from the Dual-Luciferase Reporter Assay system (Promega).

3 Results

3.1 The L1ORF1p encodes an RRM domain

3.1.1 Identification of three distinct domains in the human L1ORF1p

To identify structured domains in the human L1ORF1p, we subjected the sequence of the protein to a search for remote protein domain homologues using the HHpred server (Söding et al., 2005). This is a very sensitive and reliable tool for remote homology detection, which is based on pair-wised comparison of profile hidden Markov models (HMMs). In contrast to most conventional search methods HHpred searches not sequence, but alignment databases, like Pfam or SMART. These alignments are used then for calculating profile HMMs. HMMs are similar to simple sequence profiles, but in addition to the amino acid frequencies in the columns of a multiple sequence alignment they contain information about the frequency of inserts and deletions at each column, which helps to further improve the sensitivity (Söding, 2005b).

The HHpred search revealed a potential RRM (RNA recognition motif) domain in the L1ORF1p, which is known to be the most common eukaryotic RNA-binding domain. The RRM (M) domain is followed by the CTD (C) domain (Januczyk et al., 2007), which has no structural homologues in other proteins and is unique to vertebrate L1ORF1p proteins.

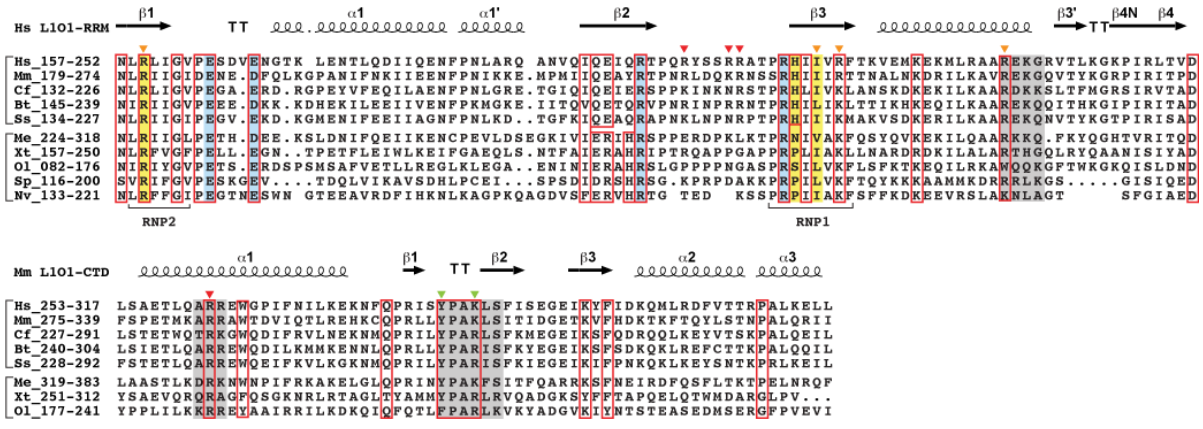

Fig. 15. Structure-based sequence alignments of the RRM and CTD domains. (L1O1-RRM, top) and CTD (L1O1-CTD, bottom) domains show highly conserved

residues boxed in red. Surface residues only conserved in placental mammals (group I) or only outside of placental mammals (group II) are boxed separately. Residues forming the conserved salt bridges are shaded in blue. Residues providing aromatic, RNA-binding side-chains in canonical RRMs are shaded in yellow. Triangles mark residues mutated in this study with a strong (red), moderate (orange) or negligible (green) effect on RNA-binding. Additional motifs mutated in a previous study (Moran1996) are shaded in gray. The C-terminal sequences of Sp and Nv cannot be confidently aligned to the mammalian-type CTD domain. Gene identifiers: Hs, *Homo sapiens* (gi:307098); Mm, *Mus musculus* (gi:198644); Cf, *Canis familiaris* (gi:116175029); Bt, *Bos taurus* (gi:66734172); Ss, *Sus scrofa* (gi:148645275); Me, *Macropus eugenii* (gi:151302550); Xt, *Xenopus tropicalis* (gi:85740540); Ol, *Oryzias latipes* (gi:3746501), Sp, *Strongylocentrotus purpuratus* (gi:111740418); Nv, *Nematostella vectensis* (gi:149338150).

The N-terminal half of the protein is predicted to contain 14 repeats of the coiled coil, which leads to trimerization of the L1ORF1p in case of the murine homologue (Martin et al., 2003) (Basame et al., 2006) (**Fig.15, 16**).

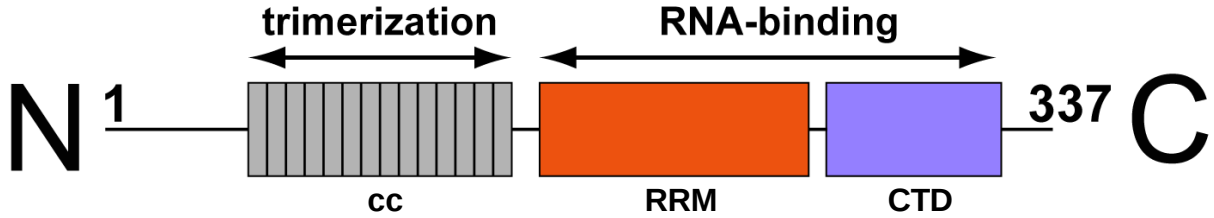

Fig. 16. Domain organisation of the human L1ORF1p. The hL1ORF1p contains the coiled coil domian (cc, shown in gray) in the N-terminal half of the protein, which was known to be responsible for trimerization of the murine homologue, and RRM- and CTD-domains in the C-terminal half (shown in red and blue, respectively), which is responsible for nucleic acid binding.

We mapped the preliminary domain boundaries more precisely by the deletion analysis of the protein. To find out where the coiled coil C-terminal boundary is, we have designed a series of L1ORF1p constructs containing only the C-terminal half of the protein. After expression and purification of these recombinant constructs we did size exlusion chromatography followed online by multiangle static laser-ligtht scattering (MALLS). This allows do determine both hydrodynamic radius and molecular weight of the protein. The largest monomeric fragment (L1ORF1p-MC) that we could identify comprises both RRM- and CTD-domains and is rather globular as indicated by an r_H of approximately 20Å (**Fig. 17**).

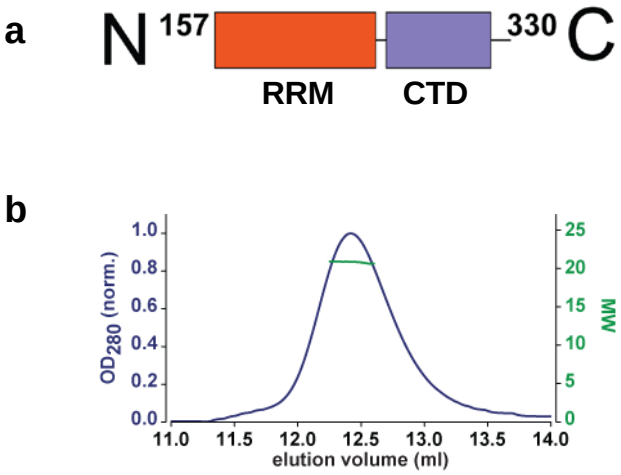

Fig. 17. The RRM-CTD construct is monomeric in solution. (a) Schematic representation of the RRM-CTD construct. **(b)** Size-exclusion chromatography and MALLS support a globular, monomeric state.

This represents a significantly larger portion of the protein as compared to previous studies with the murine protein (Basame et al., 2006) (Januczyk et al., 2007) (Martin et al., 2000). Furthermore, we can show that the predicted RRM- and CTD-domains (hL1ORF1p-M and hL1ORF1p-C, respectively) are soluble independently from each other and remain monomeric at concentrations up to 100 µM (**Fig 18a, b**). When mixed at these concentrations, they also do not detectably interact with each other (**Fig. 18c**). Such conclusion can be made because the elution volume of the domain mixture is the same as of the domains alone, whereas the formation of higher molecular weight complex would lead to an earlier elution in gel filtration.

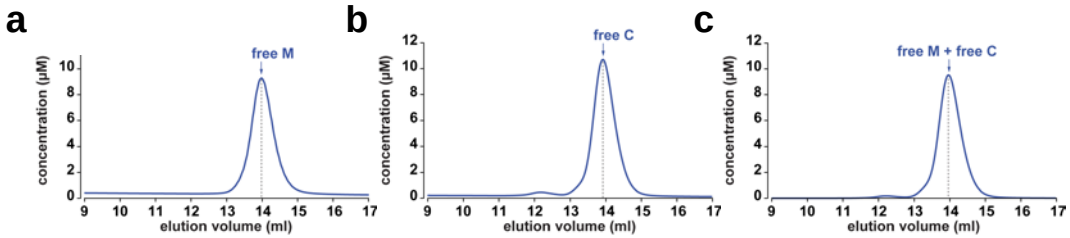

Fig. 18. The RRM and CTD domains do not interact in solution. Size –exclusion chromatography indicates that RRM domain is monomeric in solution **(a)**, the CTD domain is monomeric as well **(b)**, and there is no interaction between RRM- and CTD domains in solution **(c)**, as can be concluded from the elution volume of the mixture.

3.1.2 The crystal structure of the RRM domain in L1ORF1p shows extended loops and noncanonical RNP motifs

To verify the presence of the predicted RRM domain in the L1ORF1p and to reveal its molecular details we determined the crystal structure of hL1ORF1p-M.

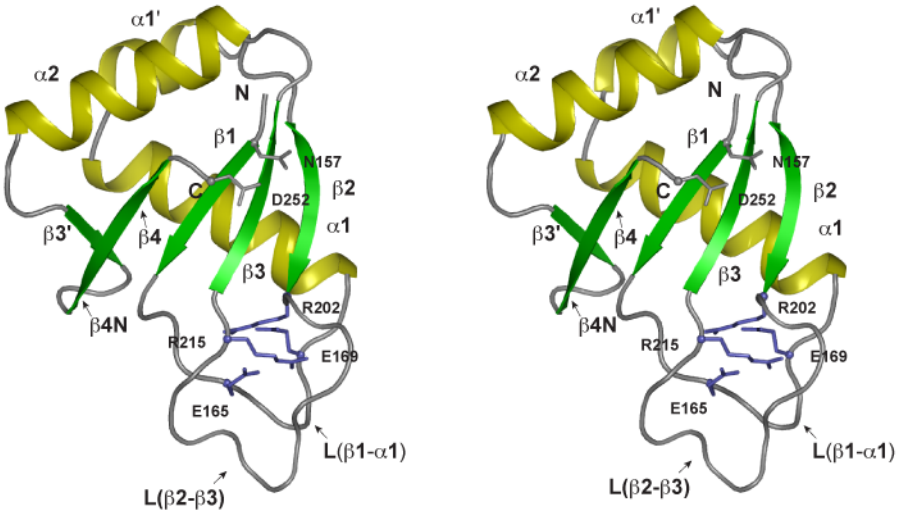

Fig. 19. Crystal structure of the RRM domain of human L1ORF1p (stereo). Ribbons representation with α-helices in yellow and β-strands in green. Blue sticks, side chains forming the conserved salt-bridges; gray sticks, side chains of the N-terminal N157 and of the C-terminal D252.

The structure was solved by single-wavelength anomalous dispersion (SAD) from seleno-methionine substituted protein and was refined at 1.4 Å resolution to an R_{free} of 18.5 %. (**Fig. 19, Table 1**).

The protein shows a classical RRM fold with the typical βαββαβ topology, where the two α-helices are packed against one surface of the four-stranded, anti-parallel β-sheet (2.8 Å r.m.s.d. over 76 C_α positions compared with the classical U1A-RRM; PDB-ID: 1oia; see **Fig. 20**). In the present case there is an additional small β-hairpin (β3'/β4-N) that is located between helix α2 and strand β4, and an extra α-helix α1' within the loop L(α1-β2). Whereas the β-hairpin is occasionally observed in other RRM domains, the helix α1' has not been seen before. The two salt bridges (E165-R215 and E169-R202) that are formed between loop L(β1-α1) and the extended loop L(β2-β3) are another unique feature of the hL1ORF1p RRM domain. These salt bridges stabilize the structures of the loops and fix their relative orientations (**Fig. 15, 19**). They likely are of functional importance, since a single point mutation (E165G) results in a strong nucleolar localization of the protein (Goodier et al., 2007). Interestingly, the unique parts of the RRM-domain interact with each other in the crystal; helix α1' fits nicely into the cleft between the loops L(β1-α1) and L(β2-β3), but there is no evidence so far that this might be physiologically relevant.

Table 1: Data collection, phasing and refinement statistics for hL1ORF1p-M

Data Collection	**SeMet data (SAD)**
Wavelength, Å	0.97154
Resolution range, Å	56 - 1.4
Space group	$P2_1$
Unit cell	
dimensions (a / b / c), Å	32.4 / 54.7 / 57.8
angles (α / β / γ), °	90 / 103.0 / 90
R_{merge}, % [a]	5.4 (53.3)
Completeness, % [a]	99.4 (96.6)
Completeness (anomalous), % [a]	97.9 (94.7)
Mean I/σ(I) [a]	13.9 (2.8)
Number of unique reflections [a]	38608 (2754)
Multiplicity [a]	3.7 (3.5)
Multiplicity (anomalous) [a]	1.9 (1.8)
Phasing	
R_{cullis}	0.922
Phasing power	0.60
Mean figure of merit	0.20
Refinement	
R_{cryst}, %	14.2
R_{free}, %	18.6
Number of	
molecules per asymmetric unit	
protein molecules	2
malonate ions	2
atoms (excluding water)	1717
water molecules	298
Average B-factor (anisotropic), $Å^2$	14.1
Ramachandran plot	
most favored regions, %	97.1
allowed regions, %	2.9
R.m.s.d. from ideal geometry	
bond lengths, Å	0.015
bond angles, °	1.58

[a] Values in parentheses correspond to those in the outer resolution shell (1.40-1.44 Å)

Canonical RRM domains are characterized by two conserved sequence signatures, RNP1 ([RK]-[G]-[**FY**]-[GA]-[**FY**]-[ILV]-[X]-[FY]) located on β-strand β3, and RNP2 ([ILV]-[**FY**]-[ILV]-X-N-L) located on β-strand β1 (**Fig. 20**). These strands provide aromatic side chains on the surface of the β-sheet (positions 3, 5 in RNP1 and position 2 in RNP2) that are frequently involved in base-stacking or in hydrophobic interactions with nucleic acid substrates (Maris et al., 2005). In the human L1ORF1p RRM domain the RNP1 (P-R-**H**-I-**I**-V-R-F) and RNP2 (L-**R**-L-I-G-V) sequences deviate significantly from the consensus signature (**Fig. 15, 20**). This may explain why this RRM domain was not identified earlier and raises the question if and how the β-sheet surface of this domain is involved in nucleic acid binding.

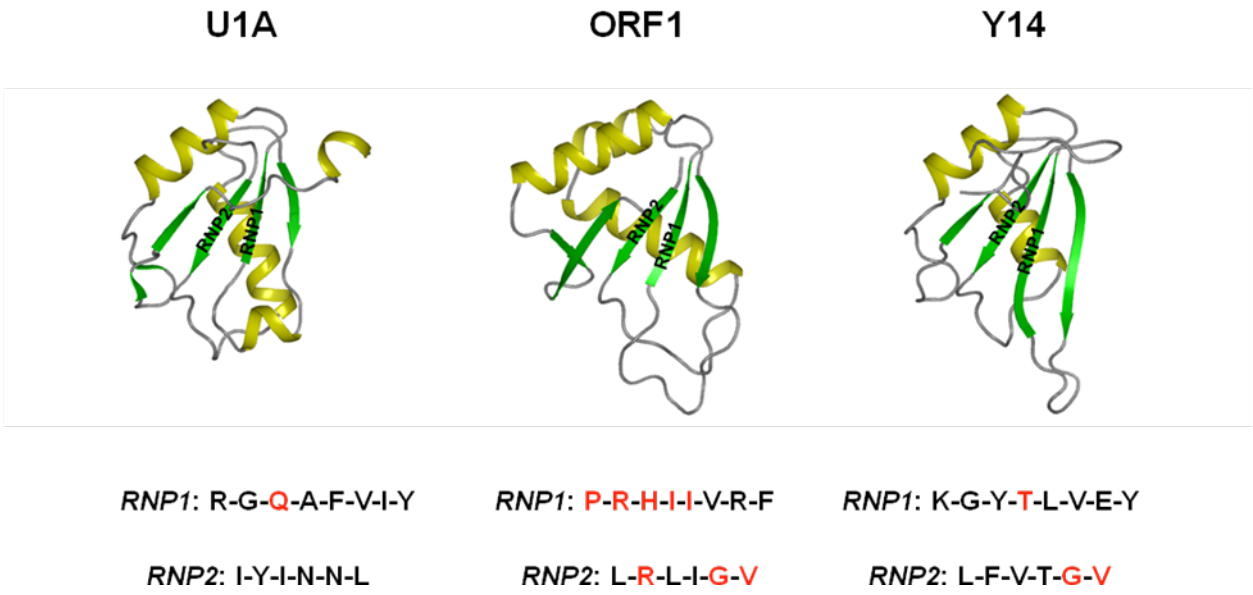

Fig. 21. Structural comparison of different RRM-domains. Ribbons representations with α-helices in yellow and β-strands in green. Canonical residues in the RNP motifs are shown in black, non-canonical ones in red.

3.1.3 Sequence conservation and the distribution of surface charge indicate the interface involved in nucleic acid binding

The C-terminal half of L1ORF1 is highly positively charged but the isolated CTD domain is not sufficient to mediate strong nucleic acid binding (Janusczyk et al., 2007). As a classical single-strand specific nucleic acid binding domain the presently identified RRM domain may therefore play a major role. The structure shows a highly asymmetric distribution of charges with a strongly basic surface that includes the

canonical β-sheet but also the adjacent surface of the extended loop L(β2-β3) that is unique to the present RRM domain (**Fig. 21**).

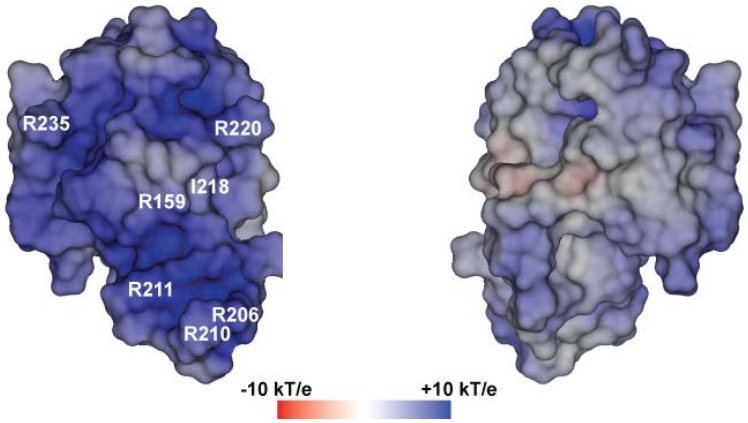

Fig. 21. Electrostatic potential mapped on the molecular surface of the RRM domain (pI = 10.6). Potentials are contoured from -10 kT/e (red) to +10 kT/e (blue). Left: view as in **Fig. 19**, onto the surface of the β-sheet and the adjacent loop L(β2-β3). Right: backside view, 180° from left.

Furthermore, we analyzed the sequence conservation of the RRM domain among five placental mammals and found that the most highly conserved surface side chains cluster on and around the basic β-sheet surface (**Fig. 15, 22**). These side-chains include N157 and D252 that link the N-and C-termini of the RRM domain (**Fig. 19**), R159 on strand β1, H216, I218 and R220 on strand β3 and K227, E228 and R235 on helix α2.

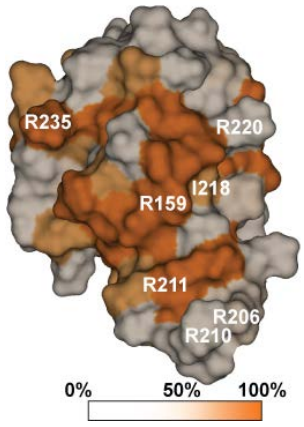

Fig. 22. **Sequence conservation of the RRM surface.** Sequence similarity among placental mammals (**Fig. 15**, group I) is color-ramped: white (50% or less) to orange (100%).

Many of those residues do not fulfill any obvious structural roles and are likely conserved for functional reasons. To test if they are important for nucleic acid binding we constructed a series of point mutants (see below).

3.1.4 Efficient nucleic acid binding requires the cooperation of the RRM and CTD domains

To test for stable nucleic acid binding under constant buffer conditions we used analytical size exclusion chromatography, monitored by triple wavelength UV absorption spectroscopy. We estimated the concentrations of the individual protein and RNA components as they eluted from the column, providing insight into the stoichiometry of the complexes.

No interaction was detected between the isolated RRM domain (hL1ORF1p-M) and a 27-mer poly(U) RNA substrate (27U RNA), at concentrations up to 75 µM. Similarly, we did not detect any interaction with the isolated CTD domain (hL1ORF1p-C) or with a protein sample where the individual RRM and CTD domains were pre-mixed at equimolar concentrations (**Fig. 23**).

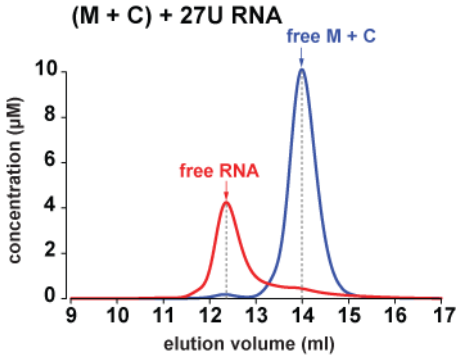

Fig. 23. **The mixture of isolated RRM- and CTD-domains does not bind RNA.** Size exclusion chromatography was done with 27U RNA (red line) and the mixture of RRM- and CTD-domains (blue line). Elution volumes of the components are indicated by arrows and dashed gray lines.

With both RRM and CTD domains on a single polypeptide chain (hL1ORF1p-MCH6), however, the RNA substrate was bound quantitatively. The majority of the RNA was found in an equimolar complex with the human L1ORF1p-MCH6 fragment. Even with an excess of protein only a small fraction of the RNA bound additional protein molecules (probably up to three) (**Fig. 24**).

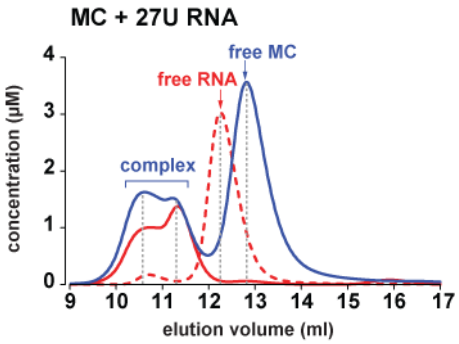

Fig. 24. The RRM- and CTD-domains expressed on one polypeptide chain bind RNA. Size exclusion chromatography was done with 27U RNA (red solid line) and the RRM-CTD construct (blue line). Chromatography run with the RNA only is shown as red dashed line. Elution volumes of the components are indicated by arrows and dashed gray lines.

The enhanced RNA affinity of the RRM-CTD fragment over the mixture of the individual domains can be explained by their cooperation and by the extremely short linker sequence that probably constrains the relative positions of the two domains (Shamoo et al., 1995).

For the more detailed analysis of the RRM-CTD binding properties we have done a number of point mutations in both RRM and CTD domains. We have selected conserved surface residues, which do not have an obvious structural role, and some of them have been shown previously to be important for the L1 retrotransposition in the cell culture assay (Moran et al., 1996).

The mutational analysis confirms that both domains participate in RNA-binding (27U RNA). As a result, in size exclusion chromatography, the RNA no longer co-elutes with the mutated protein (strong effect) or elutes significantly later than in the complex with the wild-type RRM-CTD fragment (intermediate effect) (**Fig. 25**).

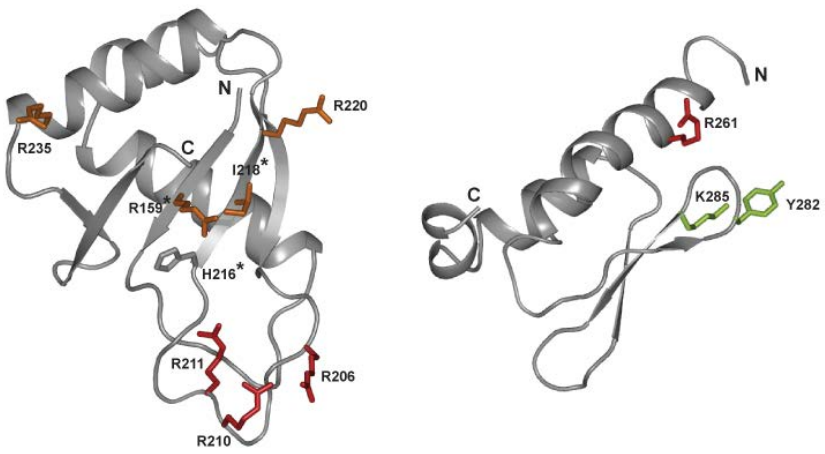

Fig. 25. Localization of mutated side-chains. The RRM (left) and CTD (right, PDB-ID 2jrb (Janusczyk et al., 2007)) domains are shown as ribbons with selected side chains as sticks (for colors see triangles in **Fig. 15**; asterisks: aromatic side chain in canonical RRMs).

The most severe effects are shown by the R206A/R210A/R211A triple mutant on the extended loop L(β2-β3) of the RRM domain and by the R261A mutant on helix α1 of the CTD domain. The single R220A, R159A, I218Y and R235A mutants on the RRM domain have an intermediate effect, while the Y282A/K285A mutant on the loop L(β1-β2) of the CTD domain behaves quasi identically to the wild-type protein (**Fig. 15, 25, 26**).

Although none of the mutants abolished RNA binding completely, the results confirm the importance of the basic protein surface of the RRM domain for RNA binding and show that cooperation with the CTD domain is essential. Many of the exchanged arginine residues may solely make contacts to the phosphate-ribose backbone of the RNA, thereby fixing its conformation. R159 and R261, however, appear particularly important, as they are invariant in sequence alignments and are functionally required at several steps in retrotransposition (Moran et al., 1996) (Kulpa and Moran, 2005) (Goodier et al., 2007) (Martin et al., 2005). They may be involved in multiple contacts, possibly stacking on bases or locking the relative orientations of the RRM and CTD domains on the RNA. Furthermore, the shallow surface cavity centered over the hydrophobic I218 seems essential, since the tyrosine substitution frequently found in canonical RRM domains (position 3 in the RNP1 motif) reduces RNA binding. The negligible effect of the Y282A/K285A mutation on RNA binding indicates that the Y^{282}PAKLS motif in the CTD domain probably does not interact directly with RNA. It rather plays a structural role and the original alanine substitution of the entire motif is likely to affect the structural integrity of the CTD (Moran et al., 1996) (Janusczyk et al, 2007). A similar effect can be expected for the original alanine substitution of the R^{235}EKG motif (Moran et al., 1996) on the RRM domain, although we see an RNA-binding defect for the single R235A substitution alone.

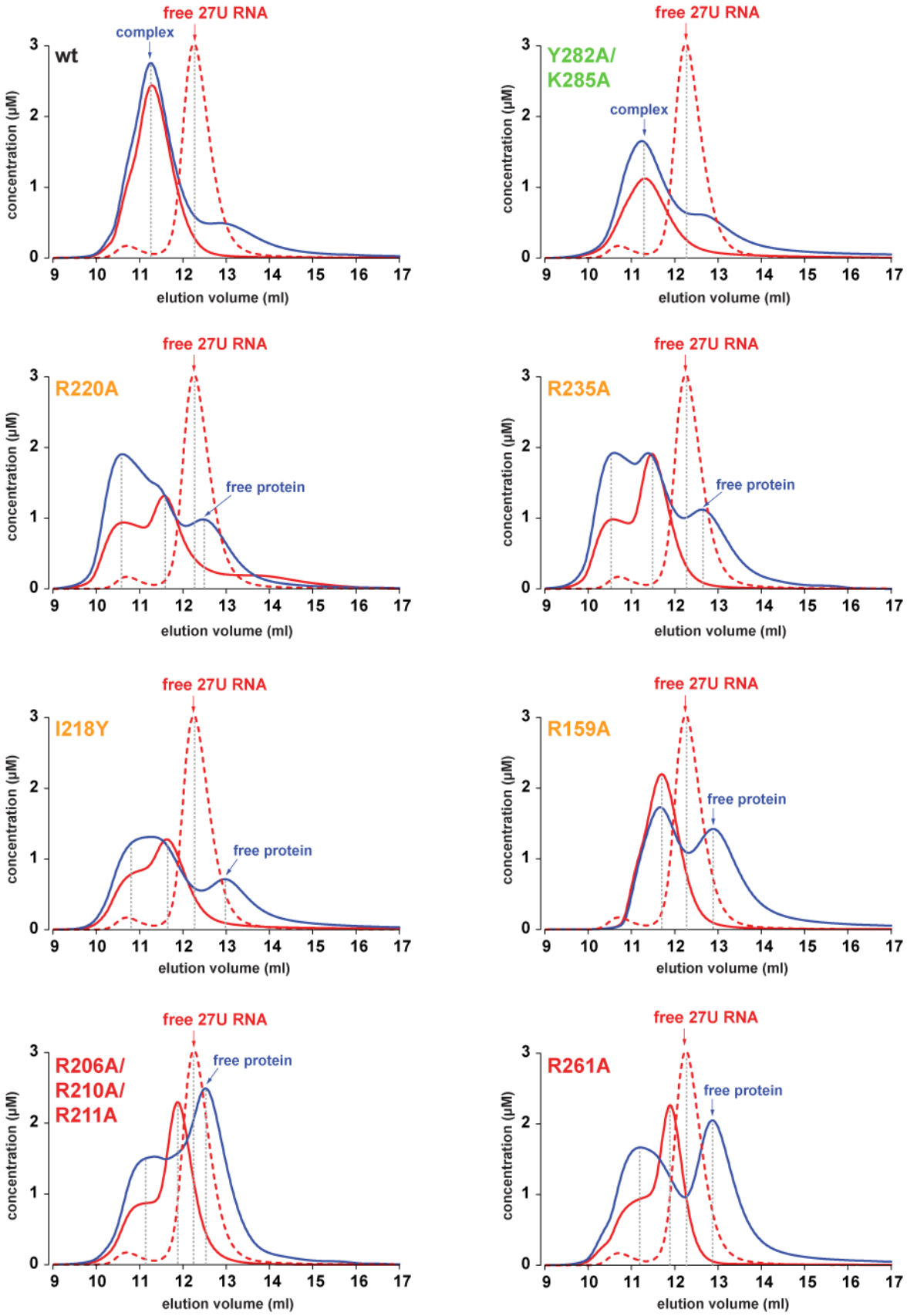

Fig. 26. Mutational analysis of the RNA-binding properties of the human L1ORF1p RRM-CTD double domain. Size exclusion chromatography was done with 27-mer poly(U) RNA (27U RNA, 40 µM at start, red lines) in the absence (dashed lines) or in the presence (solid lines) of RRM-CTD (hL1ORF1p-MC^{H6})

protein variants (50 μM at start, blue solid lines). Elution volumes of the free components and of the complexes are indicated by arrows and dashed gray lines, while apparent concentrations are calculated from the relative absorption properties of the components. The respective mutations are indicated. Colors signal strong (red), moderate (orange) and negligible (green) effects on RNA-binding.

3.1.5 The RRM-CTD fragment binds single-stranded nucleic acid and competes with the formation of base-paired structures

To exclude that complex formation simply results from electrostatic attraction of the negatively charged RNA backbone by the positively charged protein surface, we tested highly structured Alu RNA (SA86) (Weichenrieder et al., 2000) as a substrate in size exclusion chromatography. Most of the phosphate-ribose backbone of this 86-mer RNA is conformationally fixed and most of its nucleotides are involved in base-pair interactions. In the gel filtration assay it did not bind to the RRM-CTD fragment (**Fig. 27**).

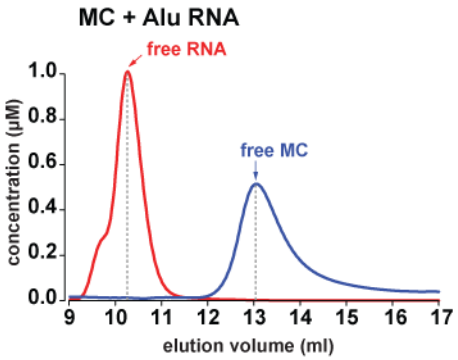

Fig. 27. The RRM-CTD construct expressed does not bind Alu RNA. Size exclusion chromatography was done with Alu RNA, SA86 (red solid line) and the RRM-CTD construct (blue line). Elution volumes of the components are indicated by arrows and dashed gray lines.

To test whether weak secondary structures or the nucleotide composition of the RNA substrate affect the interaction with the RRM-CTD fragment, we selected an alternative 27-mer (5' UAACAAUAUUAACUUUAAAUAUAAAUG 3') derived from the human L1 RNA (27L1 RNA). It corresponds to the 3'-terminal nucleotides of a longer 41-mer that specifically copurifies with endogenous human L1ORF1p (Hohjoh and Singer, 1996). In size exclusion chromatography 27L1 RNA is delayed with respect to 27U RNA, indicating that it folds into a more compact stem-loop structure. Nevertheless, 27L1 RNA also binds quantitatively to hL1ORF1p-MC^{H6}. In contrast to 27U RNA, each 27L1 RNA molecule recruits at least two or even three protein monomers (**Fig. 28**).

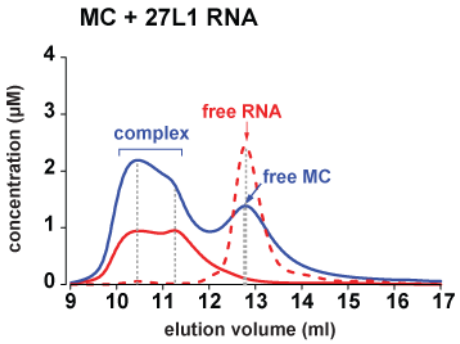

Fig. 28. The RRM-CTD construct binds 27L1 RNA. Size exclusion chromatography was done with 27L1 RNA (red solid line) and the RRM-CTD construct (blue line). Chromatography run with the RNA only is shown as red dashed line. Elution volumes of the components are indicated by arrows and dashed gray lines.

This shows that the RRM-CTD fragment can distinguish between RNA sequences and will consequently have preferential binding sites on longer RNA substrates. Apparently, hL1ORF1p-MC^{H6} is able to melt the 27L1 RNA stem-loop and stabilizes the unfolded conformation of the RNA substrate with one protein monomer occupying around nine nucleotides.

To investigate whether binding to hL1ORF1p-MC^{H6} is limited to RNA we also tested a 29-mer DNA of mixed sequence (29 DNA) as well as its reverse complement (29c DNA) (Martin and Bushman, 2001). In the absence of protein, each sample elutes as a single peak at the same position as the other, indicating an extended conformation without secondary structure. In the presence of a slight molar excess of hL1ORF1p-MC^{H6} 29 DNA is bound with equimolar stoichiometry (**Fig. 29**).

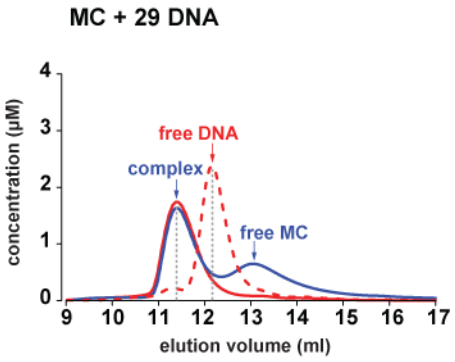

Fig. 29. The RRM-CTD construct binds single stranded DNA. Size exclusion chromatography was done with 29-mer DNA (red solid line) and the RRM-CTD construct (blue line). Chromatography run with the DNA only is shown as red dashed line. Elution volumes of the components are indicated by arrows and dashed gray lines.

The same is true for 29c DNA (data not shown). When both complexes are mixed together, the DNA strands readily anneal to form a duplex, quantitatively liberating the bound protein (**Fig. 30**).

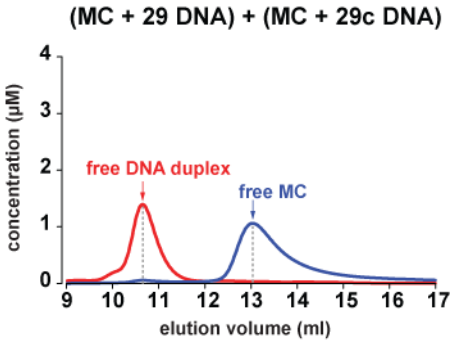

Fig. 30. The RRM-CTD construct releases DNA duplex. Size exclusion chromatography was done with 29 DNA and 29c DNA pre-mixed with the RRM-CTD construct separately and then mxed together. The resulting 29 DNA duplex is shown as red solid line and the released RRM-CTD construct as blue line. Elution volumes of the components are indicated by arrows and dashed gray lines.

In conclusion, hL1ORF1p-MC^{H6} preferably binds flexible, single-stranded nucleic acid, and the identical elution volumes of the 27U and 27L1 RNA complexes indicate that weakly base-paired structures like 27L1 RNA can be unwound by the protein. As a consequence, the RRM-CTD fragment could help resolve kinetically trapped nucleic acid structures, providing a path to the thermodynamically most favorable conformation.

3.1.6 Solution structures of RRM and CTD domains

Experiments showed that the RRM and CTD domains expressed in *cis* (i.e. on the same polypeptide chain) are required for the efficient nucleic acid binding. But the structural basis for this cooperativity, as well as for the selectivity for certain nucleic acids, was unknown. Therefore we were interested in obtaining the structure of the RRM and CTD domains in *cis*. Crystallization trials with various constructs containing both these domains were unsuccessful. We decided to use NMR for determining the structure of the RRM-CTD fragment as an alternative to crystallography. The molecular weight of the RRM-CTD construct is 21.5 kDa, which presents some difficulties for assigning the peaks in the NMR spectra. To overcome this problem we decided to first determine the NMR structures of the RRM and CTD domains separately, and then use the obtained spectra for facilitating the assignment of the

RRM-CTD fragment. (**Table 2**). The NMR structures were determined in collaboration with Dr. Murray Coles and Dr. Vincent Truffault.

The NMR structure of the human RRM domain (residues N157-D252) is very well defined and confirms the classical βαββαβ fold observed in the crystal structure (Khazina and Weichenrieder, 2009). In solution, the extended loop L(β2-β3) is flexible and disordered between residues P204 and T213 (yellow, **Fig. 31**), while the highly conserved salt bridges (E165-R215 and E169-R202) are still observed.

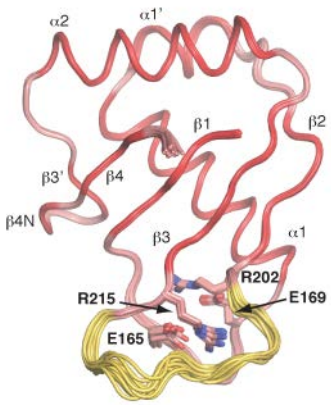

Fig. 31. NMR ensemble for the RRM domain. The disordered loop L(β2-β3) is highlighted in yellow (P204-T213) and the conserved salt-bridges are shown as sticks.

The NMR structure of the human CTD domain (residues S254-M323) is also well defined. It contains a highly conserved *cis*-proline (P283) as part of the α-hairpin (Chou, 2000) forming the loop L(β5-β6) that was not modeled as such in the murine homologue (Januszyk et al., 2007). The human CTD domain also contains a 40° kink after W264, separating helices α3N and α3 (green, **Fig. 32**). This kink might also exist in the murine homologue considering the deposited data and likely is highly relevant as a hinge (hinge 3, see below) in terms of L1ORF1p function. Furthermore, the C-terminal helix (helix α5) is longer in the human homologue and oriented to pack against helix α3 in an antiparallel fashion, leading to an overall ααβββαα fold for the human CTD domain. In strong contrast to the RRM domain, there are currently no further examples of this fold in the database.

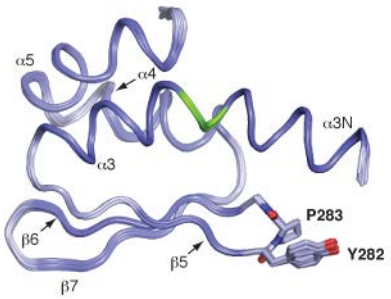

Fig. 32. NMR ensemble for the CTD domain. The hinge at W264 is highlighted in lime and important sidechains (Y282 and *cis*-proline P283) from the conserved α-hairpin are shown as sticks.

A construct containing both RRM and CTD domains on a single polypetide chain is monomeric in solution. The respective NMR spectra do not show any peak shifts with respect to the isolated RRM and CTD domains (**Fig. 33**).

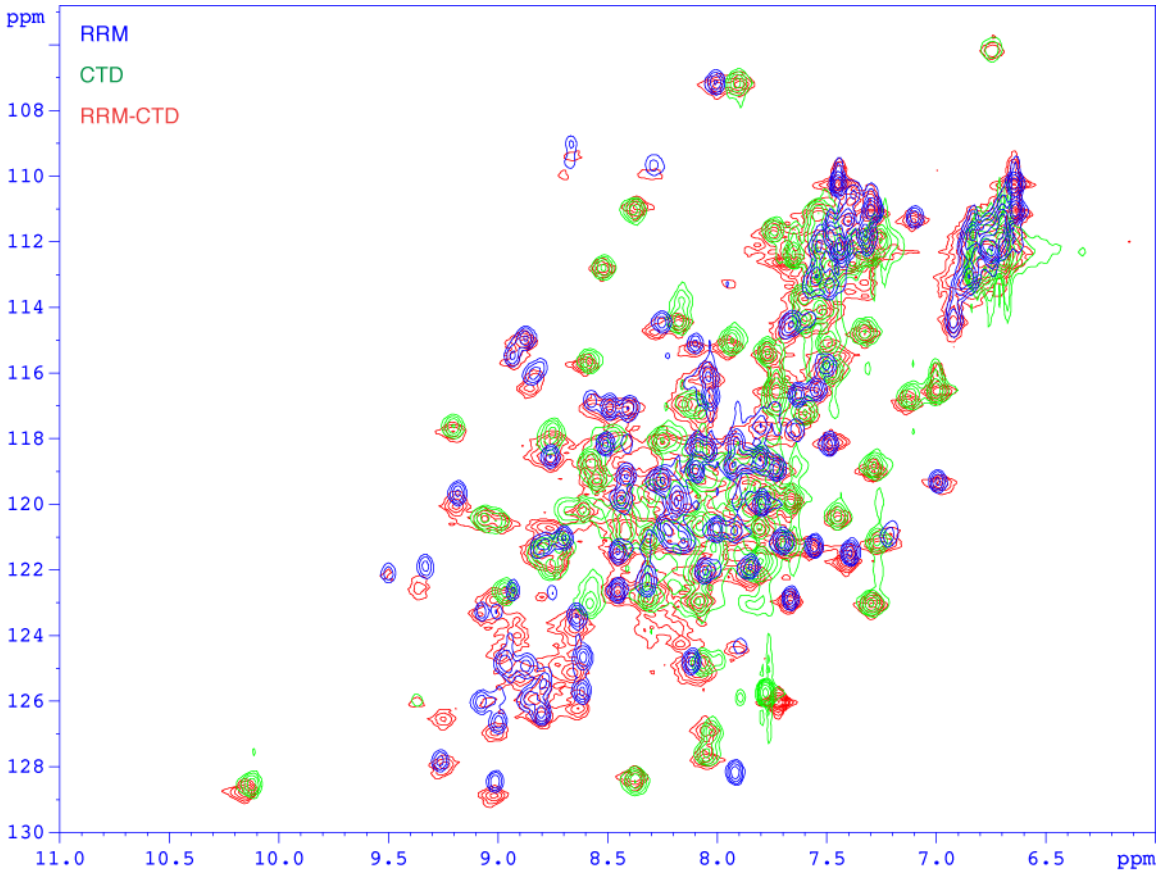

Fig. 33. NMR spectra (^{15}N-HSQC) of the isolated RRM and CTD domains superimposed on the spectrum of the RRM-CTD fusion construct. ^{15}N-HSQC spectra were recorded at 291K on a Bruker Avance III spectrometer at 600 MHz. Blue: Spectrum for the isolated RRM domain (hL1ORF1p-M). Green: Spectrum for the isolated CTD domain (hL1ORF1p-C). Red: Spectrum for the RRM-CTD fusion construct (hL1ORF1p-MC^{H6}). The spectrum for the RRM-CTD construct corresponds to the sum of the other two spectra and does not reveal additional peaks or significant peak shifts. This indicates that the RRM and CTD domains do not interact in a defined manner in solution and do not adopt a defined orientation with respect to each other.

This indicates that the RRM and CTD domains do not interact with each other in a defined way and that their relative orientations are not fixed in solution. These results raised the question if in the ORF1p trimer, which is the biologically relevant form of the protein, the orientation of the domains is not fixed either, or if it gets fixed in the presence of the coiled coil.

Table 2. Solution structure statistics for hL1ORF1p-MC^{H6}.

	SA	<SA>$_r$
A. Structural statistics[1]		
Distance restraints (Å)		
All (902)	0.010 ± 0.001	0.009
Intra-residue (133)	0.001 ± 0.001	0.001
Inter-residue sequential	0.011 ± 0.001	0.011
Medium range (153)	0.014 ± 0.001	0.013
Long range (248)	0.009 ± 0.001	0.009
H-bond (95)	0.003 ± 0.001	0.003
Persistent violation	0.05	
Dihedral restraints (°)		
All (456)	0.033 ± 0.002	0.031
Persistent violations[2]	0.2	
H-bond restraints[3] (Å/°) (89)		
Distance	2.16 ± 0.10	2.16 ± 0.10
Antecedent angle	13.7 ± 6.9	13.6 ± 7.0
Bonds (Å × 10^{-3})	11.4 ± 0.1	11.3
Angles (°)	0.57 ± 0.01	0.56
Impropers (°)	1.00 ± 0.01	0.99
Structure quality indicators[4]		
Ramachandran Map (%)	98.1 / 1.9 / 0.0	98.3 / 1.7 / 0.0

B. Atomic R.M.S. D (Å)[5]			
	SA vs <SA>	SA vs <SA>$_r$	<SA> vs <SA>$_r$
All residues	15		
Backbone heavy atoms	0.82 ± 0.17	1.43 ± 0.32	1.20
All heavy atoms	1.12 ± 0.20	1.77 ± 0.28	1.42
RMM domain			
Backbone heavy atoms	0.08 ± 0.02	0.11 ± 0.02	0.08
All heavy atoms	0.74 ± 0.09	0.98 ± 0.08	0.80
CTD domain			
Backbone heavy atoms	0.18 ± 0.04	0.20 ± 0.04	0.09
All heavy atoms	0.74 ± 0.05	0.98 ± 0.06	0.78

[1] Violations are expressed as RMSD ± SD unless otherwise stated. Numbers in brackets indicate the number of restraints of each type.

[2] Persistent violations are defined as those occurring in at least 75% of all structures. The thresholds at which no persistent violations occur are tabulated.

[3] Hydrogen bonds were treated as pseudo-covalent bonds. Deviations are expressed as the average distance/average deviation from linearity for restrained hydrogen bonds.

[4] Defined as the percentage of residues in the favoured/allowed/outlier regions of the Ramachandran map as determined by RAMPAGE (Lovell et al., 2003)

[5] Structures are labelled as follows: SA, the final set simulated of 15 annealing structures; <SA>, the structure calculated by averaging the coordinates of SA structures after fitting over secondary structure elements; <SA>$_r$, the structure obtained by regularising the mean structure under experimental restraints. RMSD values were obtained based on heavy atoms superimpositions over ordered residue (defined as N5-I9, G19-T51 and H64-V99 for RMM and S102-N170 for CTD).

3.2 Trimeric structure and flexibility of the L1ORF1p

3.2.1 Crystal structure of the trimer

The structures of the L1ORF1p domains separately or even the solution structure of the RRM-CTD fragment did not explain the molecular mechanism of the protein function in the cell. To understand this mechanism, we seeked determining the structure of the L1ORF1p in its trimeric state. After several trials we succeeded in crystallization of a construct, which is an N-terminal truncation of the human protein (hL1ORF1p-ΔN/1) retaining only the conserved, C-terminal half of the coiled coil (**Fig. 34**).

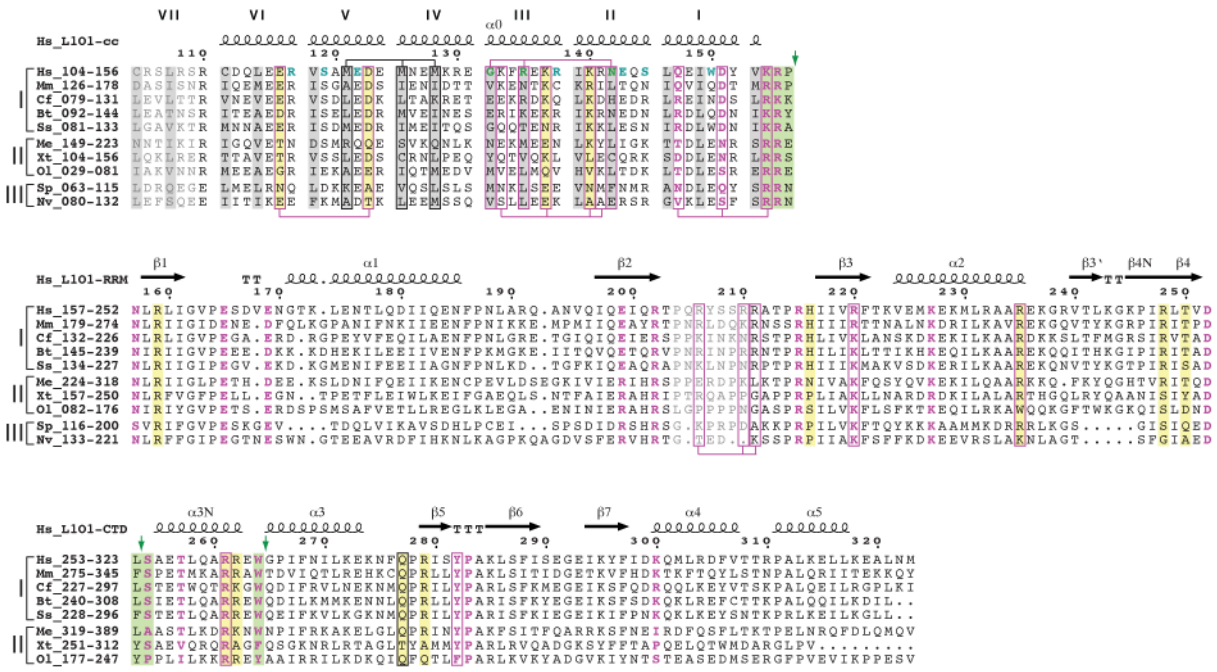

Fig. 34. Structure-based alignment of L1ORF1p homologues from selected non-LTR retrotransposons with placental mammals (group I) at the top, other vertebrates in the center (group II) and invertebrates (group III) at the bottom (compare to **Fig. 15**). Interdomain linkers are shaded in lime with hinges marked by green arrows. Residues discussed in the text are shaded yellow (conserved surface residues) or marked by magenta letters (structurally important sidechains). The *a* and *d* layers of the coiled coil are shaded in gray. Green and cyan letters indicate the ion-binding residues and RhxxhE trimerization motifs of the human sequence, respectively. Experimentally tested positions are boxed and colored magenta if important for retrotransposition. Secondary structure elements are according to monomer A of crystal form I (cfI) and letters are grey for residues that are unstructured in all of the monomers. For gene identifiers see **Fig.15**.

This construct still trimerizes in solution (**Fig. 35**) and crystallized in one of three crystal forms (crystal form I, cfI), containing one trimer per asymmetric unit.

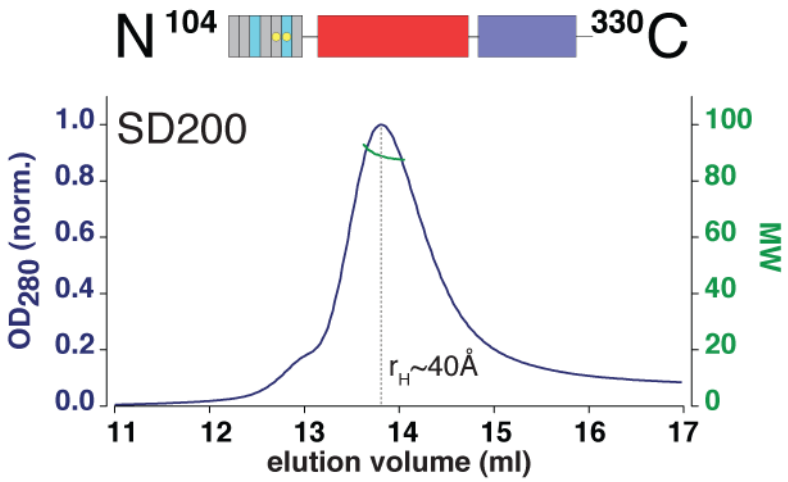

Fig. 35. Trimerization of human L1ORF1p-ΔN/1. Size exclusion chromatography was combined with multiangle static laser-light scattering (MALLS, green line, secondary axis). MALLS shows the trimerization of the protein, while the hydrodynamic radius (r_H) indicates a roughly spherical shape. The expected hydrodynamic radius (r_H) for a globular trimer of 3x28 kDa is 39 Å.

The structure was solved at 2.1 Å resolution by a combination of single wavelength anomalous dispersion and molecular replacement and was refined to an R_{free} of 26.1% (**Fig. 36, Table 3**).

Table 3. Data collection and refinement statistics for hL1ORF1p-ΔN/1 and hL1ORF1p-ΔN/2.

Data set	cf I peak	cf I remote	cf II	cf III
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell				
dimensions (a / b / c), Å	72.5/ 84.9/ 111.1	72.7/ 85.4/ 111.5	73.7/ 89.9/ 111.6	71.4/ 91.6/ 107.7
angles (α/ β/ γ), °	90/ 90/ 90	90/ 90/ 90	90/ 90/ 90	90/ 90/ 90
Data Collection				
Wavelength, Å	0.9790	0.9710	1.0070	0.9787
Resolution range, Å [a]	900-2.3 (2.30-2.36)	900-2.1 (2.10-2.15)	900-3.1 (3.10-3.18)	900-3.1 (3.10-3.18)
R_{sym}, % [a]	4.9 (45.5)	4.9 (49)	8.1 (60)	8.9 (59.1)
Completeness, % [a]	99.8 (99.9)	97.2 (97.6)	97.9 (99.6)	8.9 (59.1)
Completeness (anomalous), % [a]	98.8 (99.3)			
Mean I/σ(I) [a]	11.68 (2.1)	10.88 (1.95)	10.79 (1.71)	11.28 (2.23)
Number of unique reflections	31113 (2277)	40061 (2929)	13701 (1004)	13313 (966)
Multiplicity [a]	3.7 (3.8)	2.26 (2.29)	2.42 (2.42)	3.59 (3.69)
Multiplicity (anomalous) [a]	1.98 (1.97)			
Refinement				
Data range, Å		36.8-2.1	61.5-3.1	69.8-3.1
R_{cryst}, %		21.6	24	23.7
R_{free}, %		26.1	27.8	28.6
Number of molecules per asymmetric unit				
protein molecules		3	3	3
chloride ions		2	2	2
atoms (excluding water)		5206	4969	4948
water molecules		190	0	0
Average B-factor, $Å^2$				
all atoms		51.8	87.3	77.4
chloride ions		42.2	54.5	81.8
protein atoms		51.9	87.3	77.4
water molecules		48.8		
Ramachandran plot				
favored regions, %		98.3	96.6	97.6
disallowed regions, %		0	0	0
R.m.s.d. from ideal geometry				
bond lengths, Å		0.003	0.005	0.002
bond angles, °		0.60	0.48	0.51

[a] Values in parentheses correspond to those in the outer resolution shell

We find trimerization to be mediated indeed by a central parallel coiled coil of α-helices (helices α0, **Fig. 36a, b**) that C-terminally extends up to V153 as we predicted before and that is followed by a very short linker (K154-P156, for an alignment see **Fig. 34**). After P156, there is an almost 90° turn (hinge 1, **Fig. 36b**) leading directly into the RRM domains (N157-D252).

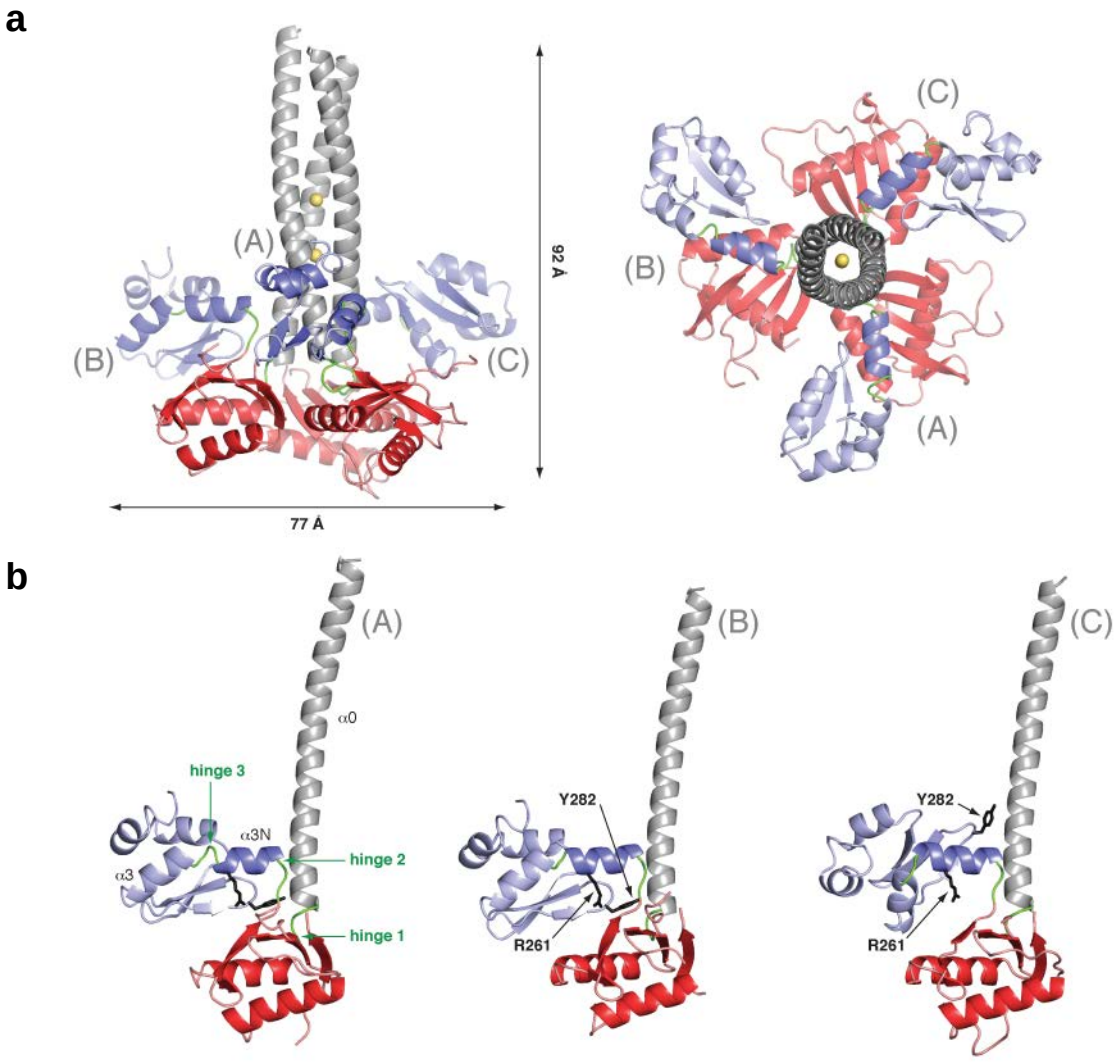

Fig. 37. Crystal structure of the human L1ORF1p trimer. (a) Overview of cf I with the coiled coil in grey, the RRM domains in red and the CTD domains in blue. Central chloride ions are shown as yellow spheres. Left: side view, right: top view. **(b)** Comparison of the three monomers with the interdomain linkers in lime and with the corresponding hinges marked by green arrows. Residues R261 and Y282 are shown as orange sticks to illustrate the different CTD orientations.

As a consequence, the RRM domains are tethered to the bottom end of the coiled coil with their β-sheets (i.e. their putative RNA-binding surfaces) facing up and describing a common plane that is perpendicular to the axis of the trimeric coiled coil.

In contrast to the solution structure, the RRM-CTD linker (L253-S254) now adopts a defined conformation and helps to anchor the CTD domains (S254-M323) to the outside of the coiled coil (hinge 2, **Fig. 36b**) from where helix α3N arcs out, suspending the CTDs over the RRM β-sheet plane. This mode of attachment allows for a considerable flexibility of the CTD domains that is further increased by the previously mentioned hinge 3 at W264 (**Fig. 36b**) that separates helix α3N from the CTD core.

3.2.2 The core of the coiled coil contains ions coordinated by polar residues

The sequence of the crystallized construct was designed to contain seven heptad repeats of the coiled coil (I-VII, counting backwards from the RRM domain, **Fig. 34, 37**). Heptad repeats are a sequence signature of coiled coils and positions within heptads are generally numbered *a-g*, where *a* and *d* form hydrophobic layers in the core of the coiled coil. The N-terminal heptad (VII) is disordered in the crystal, and hence only the second of two conserved RhxxhE trimerization motifs is visible (h, hydrophobic; x, any amino acid, Kammerer et al., 2005). It stabilizes the parallel, trimeric state of the coiled coil by interchain salt bridges and hydrogen bonds between R117, S119 and E122 (**Fig. 37a**).

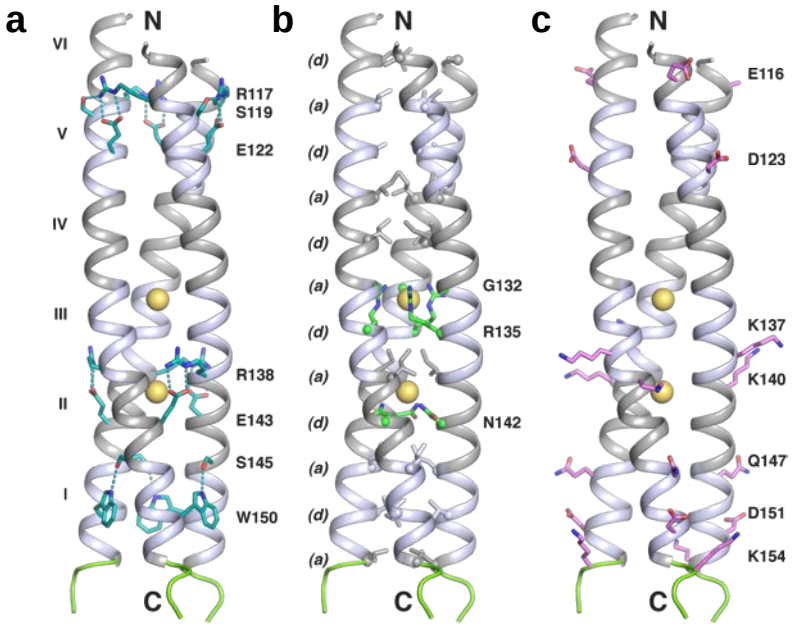

Fig. 37. Details of the coiled coil. Heptad repeats are numbered (I) to (VI) and are colored alternatingly grey and bluewhite, with central chloride ions shown as yellow spheres. **(a)** Externally stabilizing hydrogen bonds (including RhxxhE trimerization motifs in heptads II and V) are shown as dotted lines, with the corresponding residues as cyan sticks. **(b)** Sidechains in the *a* and *d* layers are shown as sticks with ion-coordinating residues in green. **(c)** Conserved and functionally important side chains on the surface of the coiled coil are shown as magenta sticks.

We also find interchain saltbridges between R138 and E143 (**Fig. 37a**), which belong to a similar RhxxhE trimerization motif where the second 'hydrophobic' position is occupied by a non-canonical asparagine (N142). In addition, the corresponding *d*-layer of the coiled coil (heptad II) contains a central chloride ion (**Fig. 37a, 37b**). Such N@*d* layers with a central anion are important motifs increasing the specificity for the trimeric state and they appear to be quite common in parallel trimeric coiled coils of bacterial adhesins, viral fibers and fusion proteins (Hartmann et al., 2009). However, they have not been described for an intracellular protein so far. Surprisingly, we identified a second chloride ion in the *a*-layer of heptad III (G132). In an unprecedented arrangement this chloride ion is coordinated by the guanidino groups of the three arginines (R135) that emanate from the following non-canonical *d*-layer (**Fig. 37b**).

3.2.3 The three RRM domains are in structurally distinct orientations resulting in asymmetric interfaces

The orientation of the RRM domains with respect to the coiled coil is stabilized by the highly conserved K227 that contacts the R155 carbonyl oxygen and by an intermolecular bond between N157 and Y152 from the neighboring protein monomer (**Fig. 38a**).

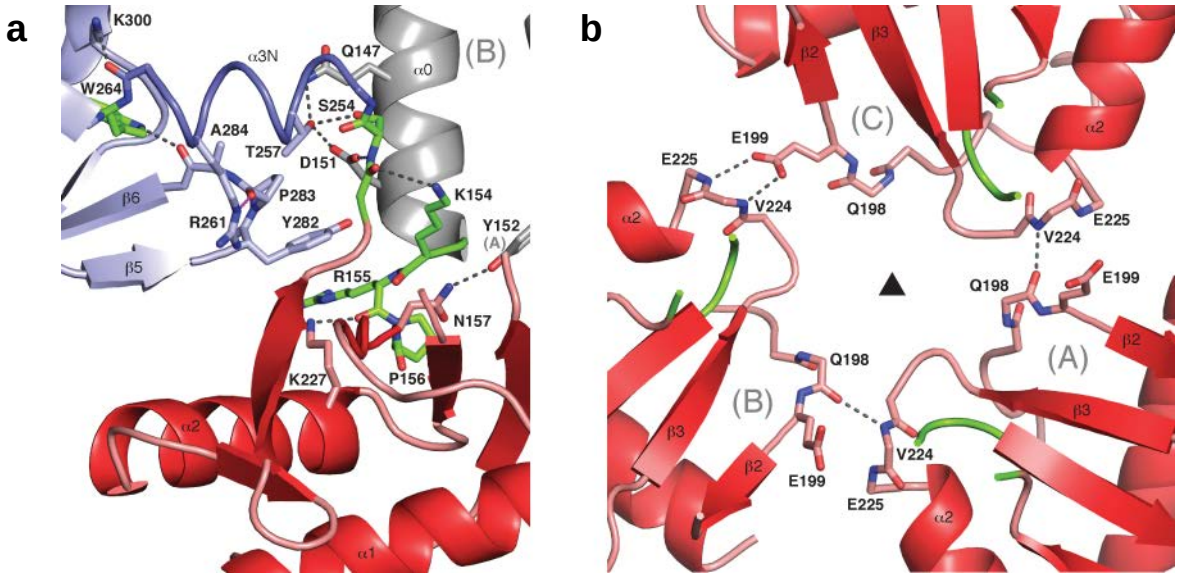

Fig. 38. Details of the hinges and asymmetric interfaces of the RRM domains. (**a**)Details of the hinges, side view. (**b**) Asymmetric interfaces of the RRM domains, top view with interdomain hydrogen bonds as dotted lines and with the corresponding interface residues as sticks. Helix α3N is drawn as a simple loop for clarity, and the fixation of the CTD core by R261 is highlighted in magenta. The black triangle indicates the threefold axis of the coiled coil.

Furthermore, the RRM domains contact their neighbors, but the interfaces are rather small (250 $Å^2$ or less of buried surface area). They are formed between the loop L(β3-α2) and the N-terminal part of helix α2 from one RRM domain (residues T222-E228) and the strand β2 (residues Q198-Q201) from the other RRM domain (**Fig. 38b**).

Most interestingly, the arrangement of the RRM domains deviates from the threefold symmetry of the coiled coil. This asymmetry is reflected by variable interface contacts. Comparing the three monomers A, B, and C, the primary interaction in the AB and CA interfaces is a main-chain hydrogen bond from V224 (A, C) to Q198 (B, A). In the BC interface this bond is broken. Instead, the main-chain nitrogens of V224 (B) and E225 (B) are contacted by the side-chain of E199 (C), which is invariant in placental mammals (**Fig. 38b**).

Finally, a series of rather conserved side-chains that are in positions to form additional contacts across domain interfaces were modeled in double conformations (R155 (A), Q201 (B), R220 (A, B), E256 (C)) or are even more strongly disordered (E175 (B), Q196, Q198 (B, C), K223 (A), E225 (B, C), Q277 (C), R279 (C), K285 and residues 204-213 from loop L(β2-β3)). This suggests alternative ways to optimize the interactions of neighboring domains in solution; interactions that are likely to be weak and dynamic.

3.2.4 The three CTD domains are flexibly attached to the coiled coil and lack defined contacts to their neighbors or to the RRM domains

The three CTD domains (**Fig. 36b, 38a**) are loosely suspended above the plane described by the RRM β-sheets, but without directly contacting the RRM domains. Furthermore, and in contrast to the RRM domains, there are also no contacts between neighboring CTD domains. As a consequence, the position of the CTD domains with respect to the RRM domains is only indirectly determined, allowing them to act even more dynamically than the RRM domains and to take individual orientations independently of the coiled coil symmetry (**Fig. 36a, b**).

Only the very tip of helix α3N gets anchored to the coiled coil (hinge 2) involving a network of hydrogen bonds that includes S254 and T257 in the CTD and Q147, D151 and K154 in helix α0 (**Fig. 37a**). This results in considerable leverage, such that subtle changes in the length and distance of individual hydrogen bonds translate into large reorientations of the CTD and vice versa.

Additionally, the rigid CTD core can move with respect to helix α3N, as observed in monomer C, where the CTD core twists around the internal hinge 3 by roughly 45° (**Fig. 36b**). In this case an important hydrogen bond is lost between the ε-nitrogen of the invariant R261 on helix α3N and the carbonyl oxygen of the equally invariant *cis*-proline P283 in loop L(β5-β6) (**Fig. 38a**). Notably, substitutions of R261 lead to a dramatic loss of retrotransposition of both human (Kulpa and Moran, 2005) and murine (Martin et al., 2005) L1 elements.

The only other detectable contact of the CTDs to the rest of the trimer is via the highly conserved Y282 that protrudes from the tip of the rigid α-hairpin (loop L(β5-β6)). It likely acts as a counter bearing, preventing the CTD core from collapsing onto the coiled coil and helping it to adopt a stable 'parking' position as in monomer B where Y282 neatly stacks on R155 (**Fig. 36b, 38a**).

3.2.5 The cleft between the RRM and CTD domains can open up considerably

Two additional crystal forms were solved at 3.1 Å resolution. They reveal additional, significantly different domain orientations, deepening the insight into the flexibility and dynamics of the trimer (**Fig. 39, Table 3**).

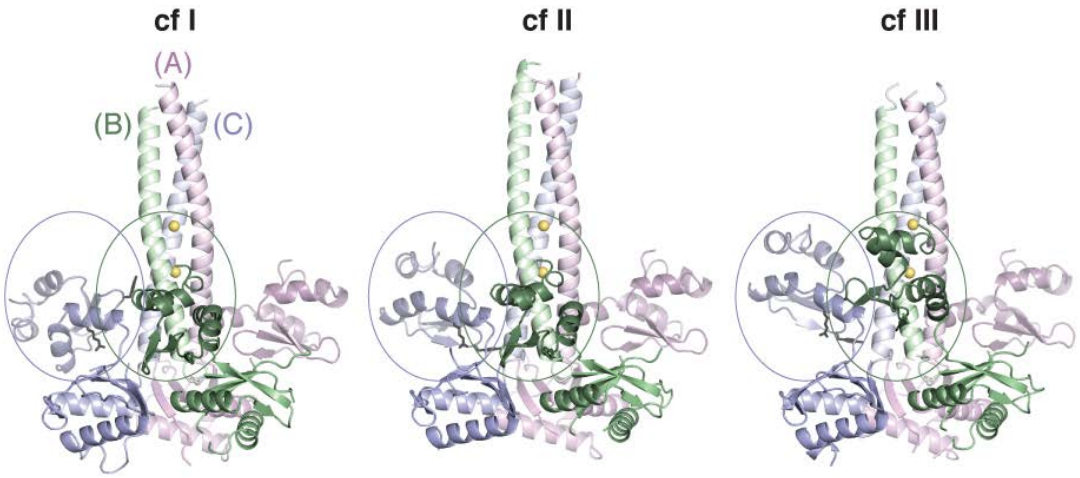

Fig. 39. Flexibility of the L1ORF1p trimer. Comparison of the three crystal forms, cfI (left), cfII (center) and cfIII (right), colored according to monomers. Circles highlight the different CTD domain orientations with residues R261(C) and Y282(C) as sticks.

Crystal form II (cfII, R_{free} = 27.8 %, **Fig. 39**) is the most symmetric of the three crystal forms and has all three CTDs in roughly similar orientations, in 'parking' positions like the ones described for monomers A and B of cfI. Thus, the CTD orientation observed in monomer C of cfI is not an inevitable consequence of trimerization, but can be adopted independently. The RRM domains of cfII remain in orientations comparable to cfI, but they move slightly with respect to each other, altering the distances and angles of the interface hydrogen bonds.

Crystal form III (cfIII, R_{free} = 28.6 %, **Fig. 39**) is the most asymmetric of the three crystal forms and is characterized by a pronounced upward rotation of the CTD domains in monomers B and C by roughly 30° around hinge 2, during which the guanidine group of R155 rotates by 90° and Y282 loses all contacts to the coiled coil. As a consequence, the cleft between the RRM and CTD domains opens considerably. Furthermore, the distance and orientations of the interface contacts between the RRM domains are altered yet again, revealing an intriguing adaptability.

3.2.6 Single-stranded nucleic acids are likely to bind in the deep basic clefts between the RRM and CTD domains

A molecular surface representation of the trimer, colored by the electrostatic surface potential shows deep, highly positively charged clefts; horizontally between the RRM and CTD domains and vertically between monomers (cfIII, **Fig. 40**).

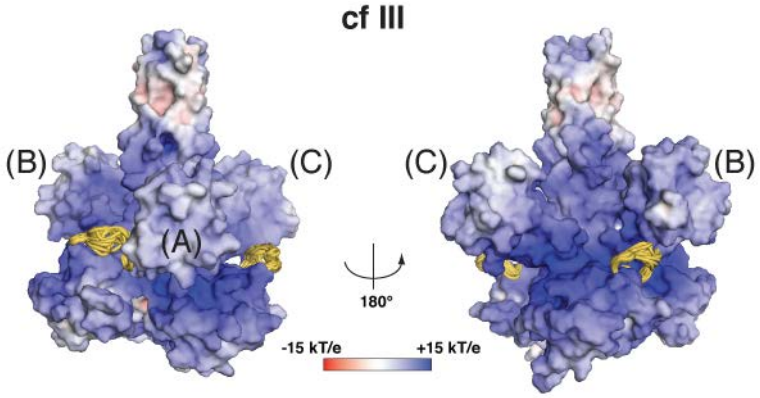

Fig. 40. Electrostatic potential mapped onto the molecular surface of the L1ORF1p trimer (cfIII). NMR ensembles are superimposed for the loops L(β2-β3) and are shown as yellow tubes. Potentials are contoured from -15kT/e (red) to +15kT/e (blue). Left: front view, Right: back view.

This suggests a direct interaction primarily with the flexible, negatively charged phosphoribose backbone of single-stranded nucleic acid substrates rather than with the bases. For the rigid backbones of structured and double-stranded nucleic acids, however, the depth and the curvature of these clefts would not allow a continuous interaction, effectively selecting against such substrates.

The horizontal cleft is lined with putative RNA binding residues that face each other from the RRM and CTD domains (**Fig. 41**).

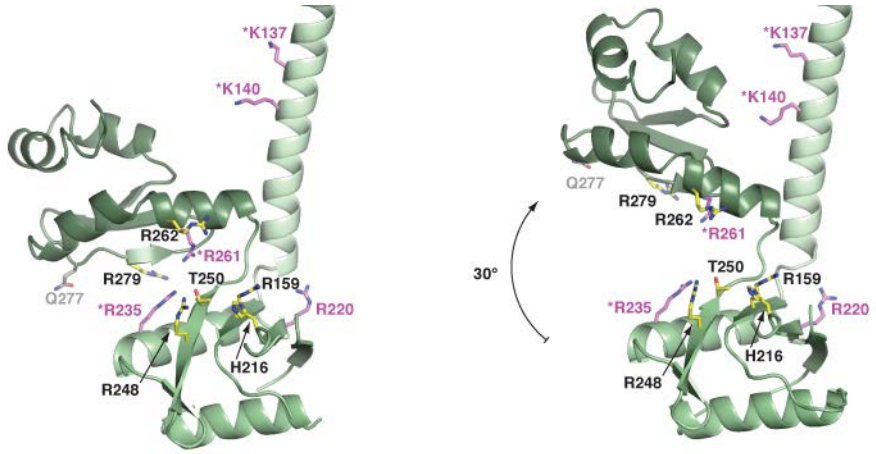

Fig. 41. Details and flexibility of the cleft between the RRM and CTD domains. Monomers (B) of cfI (left) and cfIII (right) are shown. Selected sidechains are shown as sticks and colored magenta if found important for retrotransposition, grey if found irrelevant, and yellow if not tested. Residues with an asterisk are found important for nucleic acid binding.

They are too far apart for directly contacting each other, but they could easily accommodate and reach a nucleic acid backbone from the two different sides already

when the CTD is in its 'parking' position, and even more so as the CTD is lifted up to its highest observed position in monomers B and C of cfIII.

The vertical clefts may be gated by the disordered loops L(β2-β3) of the RRM domain (yellow, **Fig. 40**) and potentially could accommodate nucleic acid strands that reach up to the two conserved lysines (K137 and K140) on the surface of heptads II and III of the coiled coil (**Fig. 37c, 41**).

3.2.7 Each L1ORF1p trimer binds 27-45 nucleotides of single-stranded nucleic acid

To test the trimer (hL1ORF1p-ΔN/1) and our mutants experimentally for nucleic acid binding we used size exclusion chromatography. This allows us to separate stable complexes from individual components and to estimate complex stoichiometry from the ratio of absorbances at different wavelengths.

Although each L1ORF1p monomer could bind to a separate 27-mer oligoU RNA on its own (Khazina and Weichenrieder, 2009), we never observe more than one nucleic acid per trimer. Furthermore, 27 or more nucleotides (oligoU RNA or oligoT DNA) are required for a stable interaction and shorter fragments start to dissociate in the assay (**Fig. 42**).

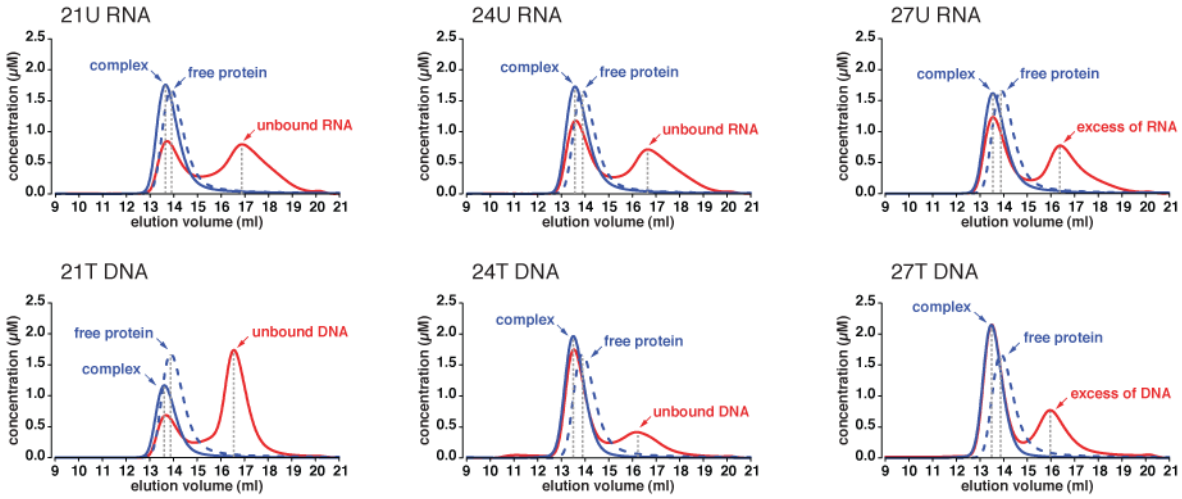

Fig. 42. Minimum length requirement for the nucleic acid binding of the L1ORF1p trimer. Size exclusion chromatography was done with various nucleic acid substrates (red lines) and L1ORF1p trimers (hL1ORF1p-ΔN/1, blue lines, dashed in the absence of nucleic acid substrate). Elution volumes of the complexes and of the free components are indicated by arrows and dashed grey lines, while apparent concentrations are calculated from the relative absorption properties of the components.

These observations indicate positive binding cooperativity between the individual monomers. Larger fragments of nucleic acid start to accommodate more than one trimer only if they are longer than 50 nucleotides (**Fig.43**).

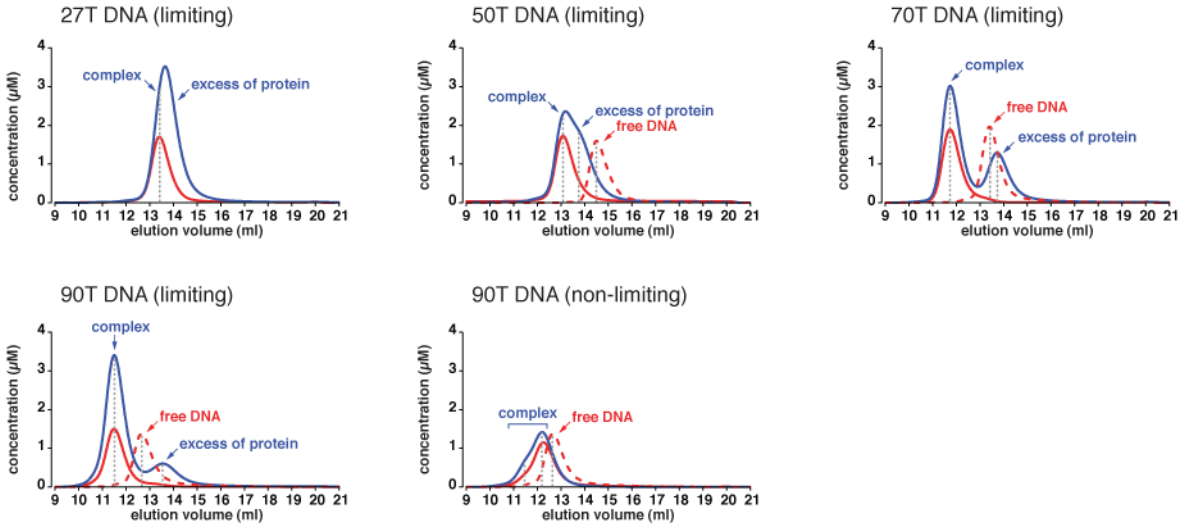

Fig. 43. Length requirements for binding more than one trimer, analyzed with poly T DNA. Size exclusion chromatography was done with various nucleic acid substrates (red lines) in the absence (dashed lines) or presence (solid lines) of L1ORF1p trimers (hL1ORF1p-ΔN/1, blue lines). Elution volumes of the complexes and of the free components are indicated by arrows and dashed grey lines, while apparent concentrations are calculated from the relative absorption properties of the components.

A fragment of 90 nucleotides stably binds two trimers if protein is in excess (i.e. ~45 nucleotides per monomer), but only one if protein is limiting. This indicates that the two trimers do not cooperate in nucleic acid binding.

3.2.8 L1ORF1p trimers distinguish nucleic acid substrates based on structure and sequence

L1ORF1p has been described to have little or no affinity for double-stranded substrates (Hohjoh and Singer, 1997; Kolosha and Martin, 1997; Martin et al., 2008). We observe the same behavior with the present, truncated trimer. A single-stranded 29-mer DNA sequence (5’ GCGAGTTGATGTTAGACTGTGTACTTTTT 3’, Martin and Bushman, 2001) binds the trimer with equimolar stoichiometry. The same is true for the reverse complement (rc29-mer DNA, data not shown). When both complexes are mixed together, the DNA strands readily anneal to form a duplex, quantitatively liberating the bound protein (**Fig. 44**).

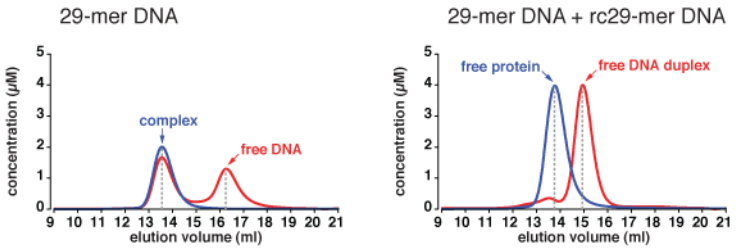

Fig. 44. Discrimination of L1ORF1p trimers against double-stranded substrates. Size exclusion chromatography was done with 29-mer DNA substrates (red lines) and of L1ORF1p trimers (hL1ORF1p-ΔN/1, blue lines). Elution volumes of the complexes and of the free components are indicated by arrows and dashed grey lines, while apparent concentrations are calculated from the relative absorption properties of the components.

Hence, there is an obvious balance between the stability of the double stranded DNA duplex and the stability of the individual protein-nucleic acid complexes. For the latter, both the sequence and the chemical structure of the backbone play an important role. Indeed we find that a 27-mer oligoA DNA does not bind the L1ORF1p trimer at all, while a 27-mer oligoA RNA forms a stable and separable complex (**Fig. 45**).

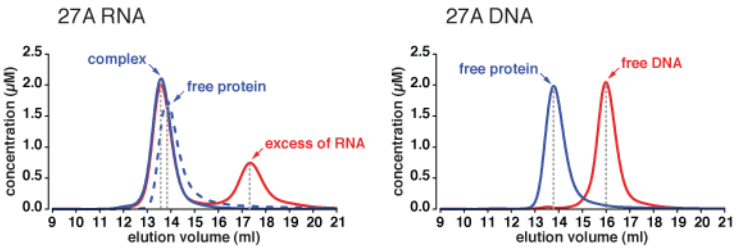

Fig. 45. Discrimination of L1ORF1p trimers against 27A DNA. Size exclusion chromatography was done with 27-mer RNA and DNA substrates (red lines) and of L1ORF1p trimers (hL1ORF1p-ΔN/1, blue lines, dashed in the absence of nucleic acid substrate). Elution volumes of the complexes and of the free components are indicated by arrows and dashed grey lines, while apparent concentrations are calculated from the relative absorption properties of the components.

In contrast, 27-mer oligoT DNA and 27-mer oligoU RNA are bound with no significant difference (**Fig. 42**).

Consequently, the sugar 2' hydoxyl group is an important but not a dominant binding determinant for L1ORF1p, and sequence composition is read out as well. The binding data show an exclusive binding to single-stranded substrates of 27 nucleotides or more. These might wrap around the trimer in the deep clefts between the RRM and CTD domains, where the flexibility of the protein permits an optimal

adjustment to structural requirements of the phosphoribose backbone, but apparently also a specific recognition of bases in certain positions of the sequence (**Fig. 46**).

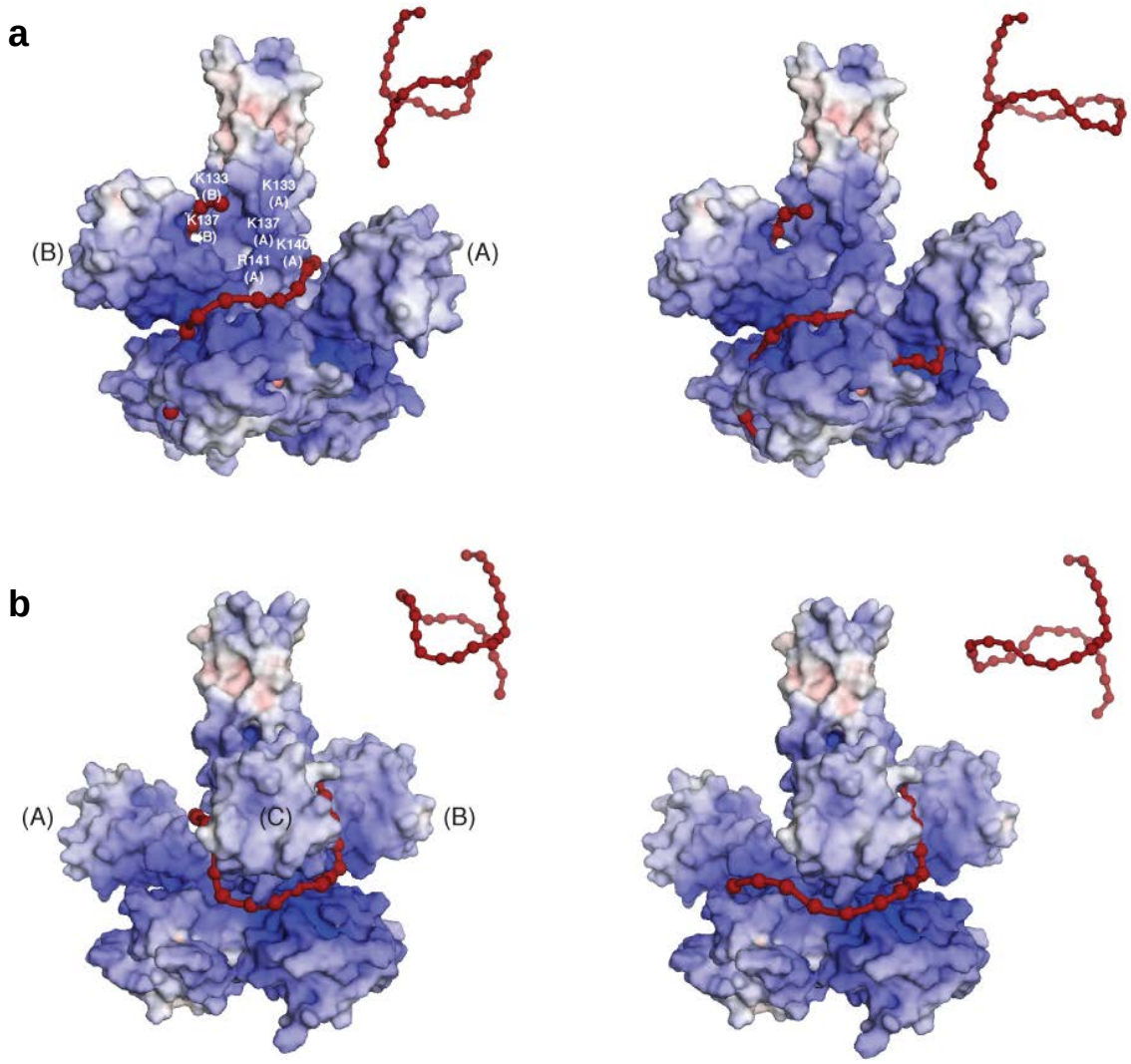

Fig. 46. Potential RNA-binding paths on the surface of the L1ORF1p trimer. (a) Alternative paths of a 27-mer RNA drawn as a tube with nucleotide spacing as spheres. The molecular surface of the L1ORF1p trimer (cfIII) is colored according to the electrostatic potential, contoured from -15kT/e (red) to +15kT/e (blue). White labels indicate basic residues at the bottom of the coiled coil that are important for nucleic acid binding. The inset shows the respective RNA path with the protein removed for clarity. Front view. **(b)** Back view.

3.2.9 Basic surfaces in all three structural domains mediate nucleic acid binding

The surface properties of the trimer and sequence conservation suggest that nucleic acids contact additional amino acids outside of the canonical RNA binding surface on

the β-sheet of the RRM domain. We therefore generated a series of mutants in all three structural domains and tested them for nucleic acid binding by size exclusion chromatography (**Fig. 34, 41, 47**).

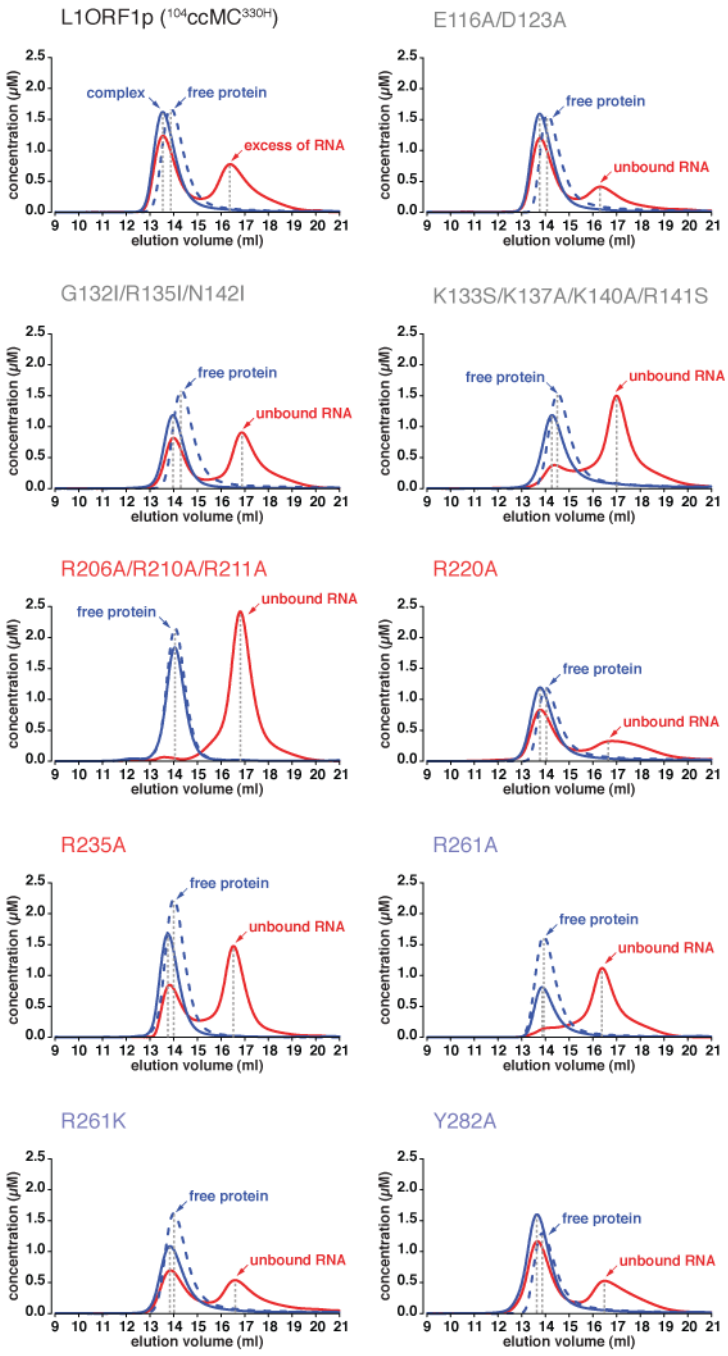

Fig. 47. Mutational analysis of L1ORF1p nucleic acid binding properties. Binding of single-stranded RNA (27U RNA, red lines) was tested for variants of L1ORF1p (L1ORF1p-ΔN/1, blue lines, see **Fig. 45** for details).

In the CTD domain, an R261A mutant (helix α3N) abolishes nucleic acid binding, while a R261K mutant does not. This is consistent with previous results (Kulpa and Moran, 2005; Martin et al., 2005) and with the present proposition of the arginine (or

lysine) to reach and fix the nucleic acid backbone from the opposite side of R235, explaining the requirement of both RRM and CTD domains in nucleic acid binding.

Within the RRM domain, an R235A mutant (helix α2) also shows a binding defect as expected from its position adjacent to the β-sheet and opposite R261. Surprisingly, an R206A/R210A/R211A triple mutant in the loop L(β2-β3) has an even stronger effect, although the loop is disordered in the crystal and NMR structures. This suggests the loop adapts to and binds the nucleic acid substrate, consistent with a proposed role in gating the vertical clefts of the trimer.

Finally, the neutralization of the lower, basic surface of the coiled coil (K133A/K137A/K140A/R141A) also abolishes nucleic acid binding. This is remarkable because, previously, nucleic acid binding was thought to be confined to the RRM and CTD domains. The respective surface may thus indeed assist the entry or exit of long single-stranded nucleic acids to wrap around the trimer in the clefts between the RRM and CTD domains (**Fig. 46**).

3.2.10 Retrotransposition critically depends on the structural integrity and flexibility of the trimer

To test the relevance of the identified nucleic acid binding surfaces and to test the significance of the observed domain motions for retrotransposition we used a well-established *in vivo* assay (Moran et al., 1996). In this assay successful genomic integration of a modified L1 element leads to G418-resistant HeLa cell colonies, allowing a quantitative comparison (**Fig. 34, 48**).

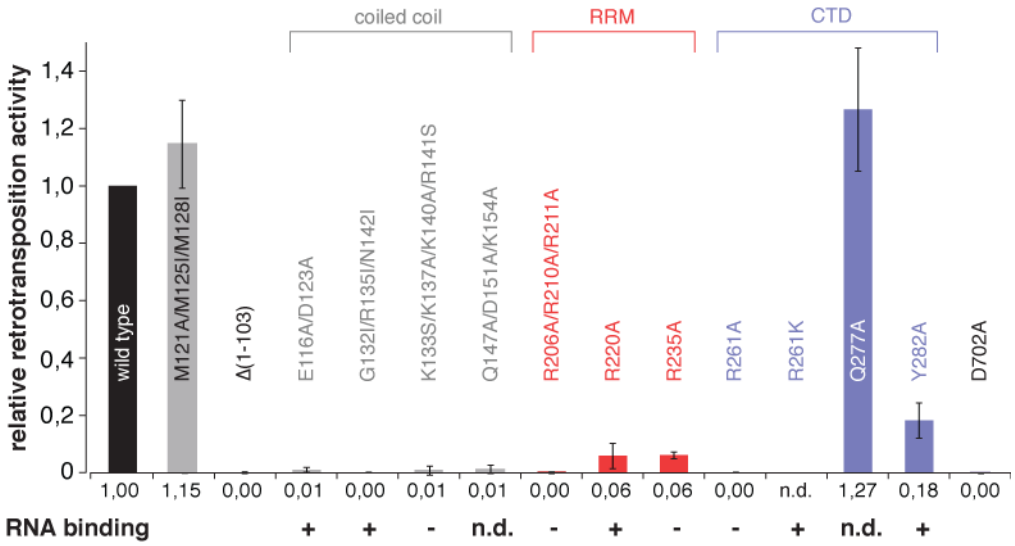

Fig. 48. Mutational analysis of L1ORF1p retrotransposition activity. Activity was scored in a cell-based assay relative to the wildtype protein, using an active site mutant of the L1ORF2p reverse transcriptase (D702A) as a negative control. Error bars are standard deviations from three independent experiments.

L1ORF1p mutants were tested in the context of the full-length protein, because an N-terminal deletion (Δ(1-103)) corresponding to the crystallized construct was not sufficient to promote retrotransposition.

Within the *a*- and *d*-layers of the coiled coil a M121A/M125I/M128I triple mutation corresponding to the murine sequence and used in crystallization was tolerated. In contrast, a G132I/R135I/N142I triple mutation, designed to eliminate the ions and to stabilize the core by canonical, non-polar residues does no longer allow retrotransposition. Although there is no apparent defect of this mutant in trimerization or nucleic acid binding *in vitro* (**Fig. 47, 49**), there could well be an assembly problem *in vivo* due to a lack of specificity or a functional defect because the coiled coil becomes too stable.

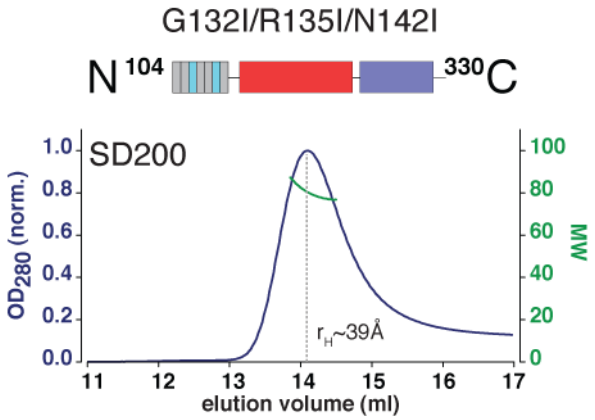

Fig. 49. Trimerization of the G132I/R135I/N142I mutant, designed to replace the chloride ions by canonical, non-polar side chains.

On the surface of the coiled coil the conserved acidic residues E116 and D123 are crucial as well. Their mutation does not affect RNA binding (**Fig. 47**), consistent with a D159H mutation in the context of the murine protein (D123 in the human sequence, Martin et al., 2008). The two residues do not serve any structural role within the crystallized trimer (**Fig. 37c**), but might do so in the context of the full-length protein. Alternatively, they could also mediate cooperation of the full-length trimers or important interactions with required partner proteins. Finally, a Q147A/D151A/K154A mutation (**Fig. 38a, 37c**), designed to liberate the CTD domain from the coiled coil

does not retrotranspose either, illustrating once more the need to structure the RRM and CTD domains with respect to the coiled coil scaffold.

In general, we find that any surface mutation that impairs nucleic acid binding of the trimer *in vitro* also impairs retrotransposition (**Fig. 47, 48**), demonstrating the importance of a proper association with nucleic acid and the involvement of the respective residues in that function. Most interesting, however, are three subtle mutations in the domain interfaces, which have no apparent defects in trimerization or nucleic acid binding, but which nevertheless severely reduce retrotransposition:

i) The R261K mutant, while still able to bind nucleic acid, probably can no longer hold on and stabilize the CTD core around the hinge 3 (**Fig. 38a**), explaining its previously described functional defect (Kulpa and Moran, 2005; Martin et al., 2005). ii) Similarly, retrotransposition efficiency of a Y282A mutant is lowered severely, probably because it affects the movements of the CTD core with respect to the coiled coil and around the hinge 2 (**Fig. 36b, 38a, 39**). iii) Finally, an R220A mutation also affects retrotransposition strongly. Although partially disordered in the crystals, this residue is expected to play a structural role at the tip of the RRM β-sheet from where it reaches towards the axis of the trimer, probably controlling the orientation of the RRM domains around the hinge 1 (**Fig. 40**).

Clearly, the precise control of domain orientations is very important. This strongly supports a crucial role for the flexibility of the L1ORF1p trimer in the retrotransposition cycle of the L1 element.

3.3 ORF1p proteins from many NLR clades contain RRM domains

To identify domain architecture of ORF1p proteins from other non-LTR retrotransposons, apart from L1, we again used HHpred (Söding et al., 2005) for remote protein domain homologues searches. This revealed potential RRM domains in nearly all of the major NLR clades that contain an ORF1p. The NLR clades and the query sequences used in this study are listed in **Tables 4 and 5**.

Table 4: Identification of RRM domains in selected NLR-ORF1p proteins.

name	species[a]	gi-number	query residues[b]	proba bility	E-value	P-value	PDB-ID residues	number of RRMs
Type I NLR ORF1p								
I clade								
I	Dr	20146016	081-233	96.1	0.13	3.3E-06	2adc_A 102-211	2
Jockey clade								
Jockey	Dm	157823	233-404	97.4	0.0045	1.1E-07	1cvj_A 235-401	2
TART-A	Dm	48596445	564-717	97.0	0.027	6.7E-07	2qfj_A 031-210	2
Het_A	Dm	14030851	471-623	97.1	0.022	5.4E-07	2qfj_A 031-208	2
R1 clade								
pilger	Dm	9369277	235-408	97.1	0.0023	5.8E-08	2dgx_A 326-396	1
Tad1 clade								
Tad1_1	Nc	409759	176-349	97.5	0.00094	2.4E-08	2dny_A 250-343	1
L1(plant) clade								
ATLINE1	At	12321249	082-258	97.0	0.021	5.4E-07	1fje_B 082-219	2
L1(Tx) clade								
Tx1L	Xl	214844	141-241	96.1	0.0088	2.2E-07	2o3d_A 149-240	1
Type II NLR ORF1p								
L1(vertebrate) clade								
L1.3	Hs	307098	157-252	77.5	8.3	0.00021	3bs9_A 164-237	1
L1(Tx) clade								

L1(Tx)	Nv	149795606	133-221	73.9	9.9	0.00025	2nlw_A 134-216	1
CR1 clade								
CR1	Nv	149844706	076-174	48.8	14	0.00035	2ghp_A 076-158(1)	1
CR1	Sp	111740418	116-200	23.1	88	0.0022	2dgp_A 117-212	1
Type III NLR ORF1p								
CR1 clade								
Q	Ag	432429	193-294	93.9	0.1	2.6E-06	2dng_A 193-287	1
T	Ag	159642	215-312	93.0	0.16	3.9E-06	2dng_A 220-305	1

a Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Nc, *Neurospora crassa*; At, *Arabidopsis thaliana*; Ag, *Anopheles gambiae*; Xl, *Xenopus laevis*; Sp, *Strongylocentrotus purpuratus*; Nv, *Nematostella vectensis*.

b amino acids are counted starting with the first methionine

Table 5: Query sequences[a] used for Table 4.

Type I NLR ORF1p

I_I-DR_Dr_gi_20146016_(081-233)

VFVRLVQEGATFEDWSPIQLTKALYKEIGEVRCAKKLRNGCLLVSCKDEAQQKKAIKVNKINGKKVKCSEVYD
RKLIRGVITGIPVSESLNNVIEGITNAKIKEAKRLKTRWNGAICDSLSIMLTFDETKLPDKVFIGYMSYEVKMYIP
PPVR

Jockey_Jockey_Dm_gi_157823_(233-404)

KPPAICVPSVSDPVTLERALNLSTGSSNYYIRISRFGVSRIYTANPDAFRTAVKELNKLNCQFWHHQLKEEKP
YRVVLKGIHANVPSSQIEQAFSDHGYEVLNIYCPRKSDWKNIQVNEDDNEATKNFKTRQNLFYINLKQGPNVK
ESLKITRLGRYRVTVERATRRKELLQ

Jockey_TART-A_Dm_gi_48596445_(564_717)

IFLSNIQQIIPLIEKLNYKAGVNSFTTKSELGNNIRIQAKTMDAYKAIQNVLLGANIPLHSHQPKSAKGFQIVIRHL
HQSTPTKWIESQLQDIGIATKFIRAMQFRDTRNPMRIHEVEVVPKADGSHLKVLLLKSLGGQTVKVERKRVSK
DPTQ

Jockey_HeT-A_Dm_gi_14030851_(471_623)

ILVNDVKEIVPLLEKLNYTAGVSSYTTRAIEGNGVRIQAKDMTAYNKIKEVLVANGLPLFTNQPKSERGFRVIIR
HLHHSTPCSWIVEELLKLGFQARFVRNMTNPATGGPMRMFEVEIVMAKDGSHDKILSLKQIGGQRVDIERKN
RTREP

R1_pilger_(waldo)_Dm_gi_9369277_(235-408)

AKVKPKRLRKKPEALILKKTGEVTYSDMLRKMKAEPSLTEFGKHVRKIRRTQQGELLLELEGKASEVIPSFKN
ELEATLKEIASVRTGAHRTALICSGLDETTTAQDLHNSLVSQFQGIRLEPEDVRGLRRRRDGTQIASVLMCAN
DAIAVINRGVVTVGWSRCRIAQDVRPIR

R1_TRAS1_Bm_gi_940388_(202-378)

RQPPKCTTLHSIMVSSKDENETGDGILTELRKTASEDEGWVRVERVRKIKDRKIIMSYRTEEERTKATQRLKK
SEGELVVEEIKNKDPLLILYNVLKMHSDEDLQKALRSKNKDLFRNLNKEDDRIEVKYKKSARNPHTHHVVLKV
SPTIWNRALSMGSLHIDIQPVRVADQTPLVQ

Tad1_Tad11_Nc_gi_409759_(176-349)

RQLTIKGATIAAEFVNRSNEDTKTTLATCLGKKKPGLIVRAATRMPTTGDYVIVFDEPTRTWCWRNQAWAKE
VFGPDAFITMSTVGVLVRGVPWDSVDNYTTAEAISNVAKERNPEASIIRVRPWKRRDGESRGLLLVEVATAS
AACFLQDNLFLWDGGAYPCEPFQASSNVQQ

L1_ATLINE1_1_At_gi_12321249_(082-258)

GLEVFEAMNSLWKNCMLVKVLGRSVPIAVLSKKLRELWKPIGAMHVVDLPRQYFMVRFESEEEYLTALTGGP
WRVFGSYLLVQAWSPDFDPMKDEIVTTPVWVRLSNIPLNLYHPSILMGITGGLGNLIKVDMTTLTCERARFAR
VCVEVNLRKPLKGTVMINEDRYFVAYEGLTNI

L1Tx_Tx1L_Xl_gi_214844_(141-241)

GGSYVPVEPLEGLGTRVVLSNVPPFLQDHLLYPHLQALGELKSNMSRIPLGCKESRLRHVLSFKRQVQLLLP
RGQDTIEGSFGVPFEGVLYKIFYSTEEVR

Type II NLR ORF1p

L1_L13_Hs_gi_307098_(157-252)

NLRLIGVPESDVENGTKLENTLQDIIQENFPNLARQANVQIQEIQRTPQRYSSRRATPRHIIVRFTKVEMKEKM
LRAAREKGRVTLKGKPIRLTVD

L1(Tx)_L1(Tx)_Nv_gi_149795606_(133-221)

NLRFFGIPEGTNESWNGTEEAVRDFIHKNLKAGPKQAGDVSFERVHRTGTEDKSSPRPIIAKFSFFKDKEEVR
SLAKNLAGTSFGIAED

Cr1_Cr1_Nv_gi_149844706_(076-174)

CLEFKGIPSLEDENTNDLVIQVAQLAGVELDEDDISISHRLPAANNREWSDYEGNVHPPSPPTIIAKFVRRDIK
DEIYKARFSLKDKTTQDLEHFNCTD

Cr1_Cr1_Sp_gi_111740418_(116-200)

SVRIFGVPESKGEVTDQLVIKAVSDHLPCEISPSDIDRSHRSGKPRPDAKKPRPILVKFTQYKKKAAMMKDRR
RLKGSGISIQED

Type III NLR ORF1p

CR1_Q_Ag_gi_432429_(193-294)

PFTDRIWIRLSAYQRPSLWNKWSLSVKRRLATDDVIAYCLLRRGVSVDSMNWLSFKVRVPAILRDAALTPST
WPVGIGVREFFQSRQHDHQTSSPIATRNRF

CR1_T_Ag_gi_159642_(215-312)

GIAEKVWLYFTNIKSHVSADDMRVWLKAVLPTDNIDVYRLTKKGANLDLMSFISFKVSIPKSLKDLALQSTIWP
VSLTVREFVDRGLPKQRIHERARF

[a] query sequences are labeled according to clade_name_organism_gi_number_(residue range)

Since no classic RNA binding domains could be identified in these proteins in the past, the discovery of RRM domains was unexpected. It provides an explanation for the RNA-binding properties of many NLR ORF1p proteins and clearly establishes that they are not related to the Gag proteins encoded by retroviruses and LTR-retrotransposons (Malik2002). According to the arrangement of the predicted structural domains we can roughly distinguish five types of NLR ORF1p proteins (**Fig. 50**).

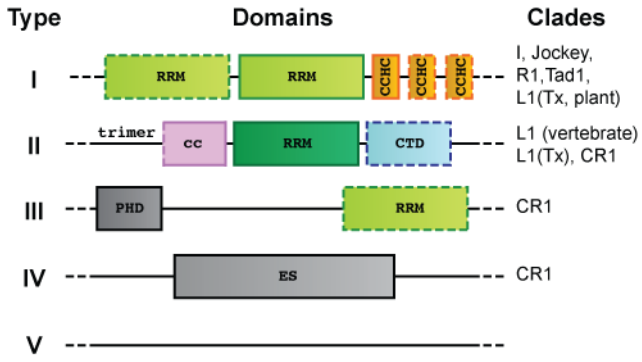

Fig. 50. Identification and organization of RRM domains in phylogenetically unrelated NLR-ORF1p proteins. Type I ORF1p is widespread and contains gag-like CCHC zinc-knuckles. Type II ORF1p is found in the human L1 element and trimerizes via a coiled coil (cc). Other types are described in the main text (see also **Table 4, 5**). CTD, C-terminal domain; PHD, plant homeodomain; ES, esterase domain.

Type I ORF1p is the most widespread type and contains at least one RRM domain immediately upstream of a gag-like CCHC zinc-knuckle. A second RRM domain and additional zinc-knuckles are frequent. The close association of the zinc-knuckle and RRM modules suggests a functional cooperation as observed frequently in other RRM proteins (Maris et al., 2005) (Lunde et al., 2007). Type I ORF1p is found from vertebrates to plants across at least five different clades, which indicates its ancient origin.

Type II ORF1p is found in the human L1 element. It contains a distinct RRM domain that is preceded by additional conserved amino acids leading to a trimerization of the molecule via a coiled coil. The CTD domain (Janusczyk et al., 2007) is conserved in vertebrate type II ORF1p proteins and characterizes the lineage of modern L1 elements (also referred to as mammalian-type L1 elements). This lineage is distinct from ancient members of the L1 clade that are found in amphibians (Garrett et al., 1989), fish (Kojima and Fujiwara, 2004), insects (Biedler and Tu, 2003) and plants (Wright et al., 1996) and that contain an ORF1p of type I.

For type III ORF1p we predict an occasional C-terminal RRM module in addition to a previously described N-terminal plant homeodomain (PHD, (Kapitonov and Jurka, 2003)). Type III ORF1p is found in the heterogeneous CR1 clade, which also harbors ORF1p proteins of type IV. These contain a functional esterase domain that enhances retrotransposition (Kapitonov and Jurka, 2003) (Sugano et al., 2006). Finally, there are numerous NLR ORF1p proteins (type V) that cannot be classified so far.

3.3.1 The ancient origin of the RRM domain in type II ORF1p supports a modular evolution of NLRs

So far, type II ORF1p has only been described in vertebrate members of the L1 clade. We were therefore surprised to identify homologues of the RRM domain in NLRs of the starlet see anemone *Nematostella vectensis* (a non-bilaterian animal) and of the purple sea urchin *Strongylocentrotus purpuratus* (a deuterostomian animal) (**Fig. 15, Tables 4 and 5**). This indicates a deeply rooted origin of this RRM-domain before the emergence of bilaterians approximately 750 million years ago and, possibly, a selective loss from the branch of protostomian animals. The respective NLRs do not seem to contain an equivalent for the CTD domain, and according to their reverse transcriptases they belong to the Tx group of the L1 clade and to the CR1 clade (Malik and Eickbush, 2002) (Jurka et al., 2005). The existence of such chimerical elements strongly supports the idea of a modular evolution of NLRs.

4 Discussion

4.1 Identification of RRM domains in NLRs and their significance for retrotransposition

For the last twenty years, NLR ORF1p proteins were studied in the absence of detailed structural information, and it was rather obscure if and how the various ORF1p proteins would bind RNA. We identified RRM modules in many NLR ORF1p proteins, which indicated that they could be generally important for retrotransposition. In this case, NLRs that do not encode their own RRM domains should depend on the proteins from other NLRs or recruit cell proteins.

The discovery of RRM domains in many NLR ORF1p proteins clearly demonstrates that non-LTR retrotransposons are not related to retroviruses and LTR-retrotransposons.

Furthermore, the modular nature of the ORF1p and ORF2p proteins and their respective combinations can be exploited technically to clarify ambiguous relations among NLRs and can ultimately help to regroup their phylogenetic tree with higher resolution.

4.2 The structure of the L1ORF1p trimer reveals the molecular basis for the cooperation between domains and for the mode of nucleic acid binding

The present crystal structures definitely establish the unusual trimerization of L1ORF1p and reveal a parallel arrangement of the monomers. They are the first structures of an RNA packaging protein from a non-LTR retrotransposon and illustrate the importance to analyze this multidomain protein in its trimerized state in order to understand its function. The structures rationalize a wealth of mutational data (Goodier et al., 2007; Kulpa and Moran, 2005; Martin et al., 2008; Martin et al., 2005; Moran et al., 1996) and provide a simple, topological explanation for how the RRM domain is assisted by the CTD domain in nucleic acid binding despite the fact that both domains do not directly interact with each other. Furthermore, only in the trimeric state do the surface properties indicate how a single, continuous strand of nucleic acid could wrap around the protein and occupy binding surfaces on each

monomer (**Fig. 46**). Assuming a minimum of ~45 nucleotides per trimer, a single L1RNA of ~6000 nucleotides could thus theoretically accommodate up to 130 copies of the trimer to protect the L1RNA from nuclease attack, separate it from the general mRNA metabolism and mark it for nuclear reimport and retrotransposition.

Nucleoproteins from single-stranded RNA viruses show conceptually similar architectures consisting of two lobes with a basic RNA binding cleft in between. They frequently multimerize continuously along the RNA substrate forming regular rings or spirals, but string-like RNPs without any helical symmetry have been observed as well (Albertini et al., 2008; Raymond et al., 2010). Whether L1ORF1p RNPs also can arrange into regular, higher order structures is not clear. However, atomic force microscopy on reconstituted murine L1RNPs rather suggests an asymmetric, irregular arrangement (Basame et al., 2006).

4.3 The flexibility of the structure is critical at possibly multiple steps of the L1 retrotransposition cycle

In contrast to most viral RNA packaging proteins (nucleoproteins), we did not only obtain a single snapshot of the structure but instead determined three distinct crystal structures plus NMR ensembles of the individual domains. This provides valuable and crucial additional insight into the molecular dynamics of the L1ORF1p trimer. We also observe conserved but malleable domain interfaces (e.g. between the RRM domains) as well as functionally important but unstructured loops (e.g. loop L(β2-β3)). Similar features have also been discussed for certain viral nucleoproteins (Tawar et al., 2009; Ye et al., 2006). They may be quite general for RNA packaging proteins that need to adapt to distinct sequence and structure contexts and for RNPs that need to undergo structural transitions. For the L1RNP such reorganizations might be crucial during the assembly and nuclear import processes. They might also be important in the TPRT reaction, where a gradual release of the L1RNA from the RNP would assure a smooth and continuous reverse transcription undisturbed by the formation of local RNA structures.

Furthermore, the observed flexibility likely contributes to the proposed function of L1ORF1p as a nucleic acid chaperone. First, it probably allows for a gradual, i.e. kinetically favorable unwinding of double-stranded nucleic acid substrates and second, it probably allows for an optimization of the interactions with the resulting

single-strands, keeping them in a hybridization-competent state for the formation of alternative structures (i.e. with the bases exposed). As a result, the L1ORF1p trimer can resolve kinetically trapped nucleic acid structures, providing a path to the thermodynamically most favorable conformation.

The dependence of L1 retrotransposition on L1ORF1p and the sensitivity to mutations is intriguing, especially since other non-LTR retrotransposons encode ORF1p proteins with completely different architectures, such as an esterase fold in the case of certain CR1/L2 non-LTR retrotransposons (Kapitonov and Jurka, 2003; Khazina and Weichenrieder, 2009; Sugano et al., 2006). Indeed, the analysis of protein variants affecting domain interfaces clearly demonstrates the functional importance of the complex architecture and of flexibility of the trimer. However, the requirement and conservation of surface residues like E116 and D123 also points to additional functions such as possible interactions with 'host' factors or with L1ORF2p that might be relevant at some stage of the retrotransposition cycle. Similarly, the N-terminal half of the coiled coil domain, which is missing from the structure has been implied in 'host' factor interaction (Boissinot and Furano, 2001).

Clearly, L1ORF1p has thus acquired specialized functions that go beyond simple nucleic acid binding.

4.4 Non-LTR retrotransposons and the viral world

LTR retrotransposons are clearly related to retroviruses. In contrast, for non-LTR retrotransposons no closely related viruses have ever been described, and none of their encoded protein sequences could be related to viruses so far (apart from the core of the reverse transcriptase). Also the present structure of the mammalian L1ORF1p has no clear sequence or structure homologs among viral proteins.

Nevertheless, there are three remarkable observations: i) A bilobal architecture with an RNA binding groove like the one in L1ORF1p (**Fig. 51**) has apparently evolved several times independently in the nucleoproteins of RNA viruses (Albertini et al., 2008).

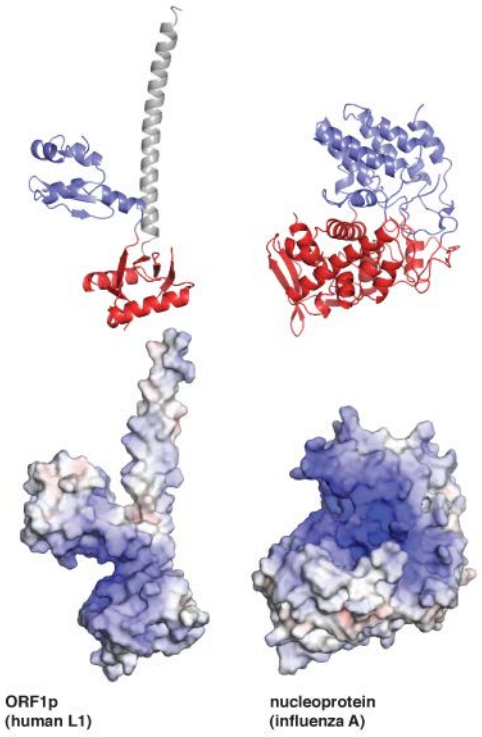

Fig. 51. Bilobal architecture and basic RNA binding groove of viral nucleoproteins. The ribbons representations (top) of L1ORF1p (cfIII, monomer C) and of the nucleoprotein (NP) from the influenza A virus (PDB-ID 2iqh) reveal two RNA binding lobes (red and blue) with a positively charged RNA binding cleft in between (bottom, electrostatic surface potential contoured between -15 kT/e, red and + 15 kT/e, blue).

ii) Trimeric coiled coils with ions in the core and stabilizing RhxxhE trimerization motifs are rarely found in cytosolic proteins, but frequently in viral fibers and membrane fusion proteins (**Fig. 52**) that undergo substantial rearrangements, such as influenza hemagglutinin, HIV gp41, or HTLV gp21 (see Hartmann et al., 2009, for a compilation).

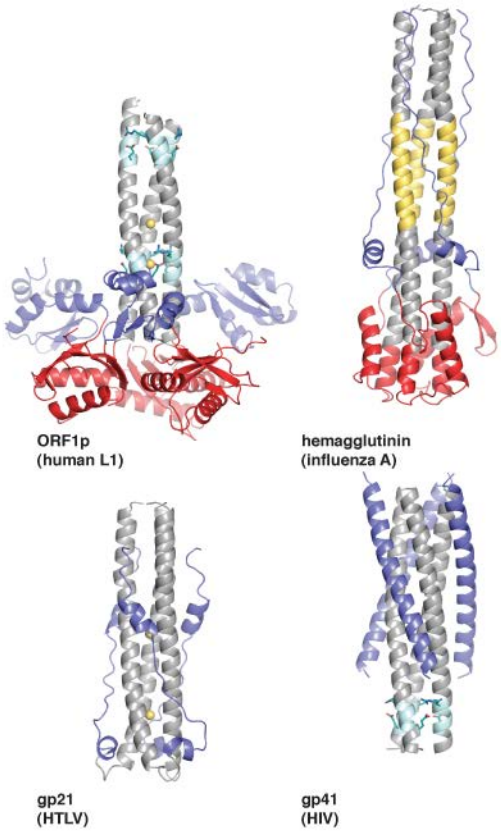

Fig. 52. Trimeric coiled coils of viral membrane fusion proteins. Post-fusion structures of HTLV gp21 (PDB-ID 1mg1), HIV gp41 (PDB-ID 1env) and influenza A haemagglutinin (PDB-ID 1qu1) reveal variable 'insertion domains' (red) and C-terminal sequences (blue) that are attached to the surface of the coiled coil, as well as central chloride ions (yellow spheres, gp21) and an RhxxhE trimerization motif (cyan sticks, gp41). Yellow helices: Regions in heamagglutinin undergoing a coil-to-helix transition upon membrane fusion.

iii) Certain viral membrane fusion proteins, the hemagglutinin-esterases of influenza C, of toroviruses, and of coronaviruses contain a receptor-destroying esterase (de Groot, 2006). Such an esterase (**Fig. 53**) is also found in the ORF1p protein from the ZfL2 non-LTR retrotransposon (Sugano et al., 2006) and in related elements, although these ORF1p proteins lack L1ORF1p-like features (Khazina and Weichenrieder, 2009).

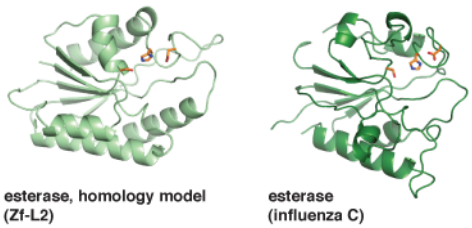

Fig. 53. Esterase domains of viral membrane fusion proteins. The fold and the active site residues (orange sticks) of the esterase domain from the influenza C virus (PDB-ID 1flc) are also found in the ORF1p protein from the ZfL2 non-LTR retrotransposon (PDB-ID 1es9 for a homology model) and in related elements that lack an L1ORF1p-like protein.

That the trimeric L1ORF1p is the evolutionary remainder of an ancient virus could thus be an intriguing hypothesis.

Abbreviations

Å	Ångström (1 Å = 10^{-10} m)
AFM	atomic force microscopy
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
cf	crystal form
CMV	cytomegalovirus
DCP1	decapping protein 1
DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6
DNA	deoxyribonucleic acid
DNase	deoxyridonuclease
Dnmt3L	DNA (cytosine-5)-methyltransferase 3-like gene
E.coli	*Escherichia coli*
EM	electron microscopy
GST	glutathione *S*-transferase
GW182	Gly-Trp repeat containing protein of 182 kDa
HA	hemagglutinin
HeLa	cell line derived from cervical cancer cells from the patient Henrietta Lacks
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HERV	human endogenous retrovirus
HIV	human immunodeficiency virus
HSQC	heteronuclear single quantum coherence
LINE-1 (L1)	long interspersed nuclear element 1

LTR	long terminal repeats
MALLS	multi-angle laser light scattering
MW	molecular weight
NLR	non-LTR retrotransposon
NMR	nuclear magnetic resonance
OD	optical density
oligoT	oligodeoxythymidine nucleotide
oligoU	oligouridine nucleotide
ORF	open reading frame
P-body	mRNA processing body
PDB	Protein Data Bank
pI	isoelectric point
pi-body	piRNA processing body
piRNA	Piwi-interacting RNA
PIWI	P-element induced wimpy testis
poly(A)	poly adenine
poly(T)	poly thymidine
poly(U)	poly uridine
r.m.s.d.	root mean square deviation
R_{cryst}/ R_{work}	crystallographic reliability factor
R_{free}	free crystallographic reliability factor
r_H	hydrodynamic radius
RNA	ribonucleic acid

RNP	ribonuclear particles
SDS	sodium dodecyl sulfate
siRNA	small interfering RNA
SRP	signal recognition particle
SV40	simian virus 40
SVA	SINE VNTR Alu
TEV-protease	tobacco etch virus protease
TPRT	target primed reverse transcription
Tris	tris(hydroxymethyl)aminomethane
UTR	untranslated region

References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." Acta Crystallogr D Biol Crystallogr **66**(Pt 2): 213-21.

Albertini, A.A., Schoehn, G., Weissenhorn, W. and Ruigrok, R.W. (2008). "Structural aspects of rabies virus replication." Cell Mol Life Sci **65**(2): 282-94.

Albertini, A.A., Wernimont, A.K., Muziol, T., Ravelli, R.B., Clapier, C.R., Schoehn, G., Weissenhorn, W. and Ruigrok, R.W. (2006). "Crystal structure of the rabies virus nucleoprotein-RNA complex." Science **313**(5785): 360-3.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T. and Hannon, G.J. (2008). "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice." Mol Cell **31**(6): 785-99.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. and Hannon, G.J. (2007). "Developmentally regulated piRNA clusters implicate MILI in transposon control." Science **316**(5825): 744-7.

Aravin, A.A., van der Heijden, G.W., Castañeda, J., Vagin, V.V., Hannon, G.J. and Bortvin, A. (2009). "Cytoplasmic compartmentalization of the fetal piRNA pathway in mice." PLoS Genet **5**(12): e1000764.

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

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

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proc Natl Acad Sci U S A **98**(18): 10037-41.

Basame, S., Wai-lun, Li P., Howard, G., Branciforte, D., Keller, D. and Martin, S.L. (2006). "Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition." J Mol Biol **357**(2): 351-7.

Batzer, M.A. and Deininger, P.L. (2002). "Alu repeats and human genomic diversity." Nat Rev Genet **3**(5): 370-9.

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

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

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O. and Devine, S.E. (2008). "Active Alu retrotransposons in the human genome." Genome Res **18**(12): 1875-83.

Biedler, J. and Tu, Z. (2003). "Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity." Mol Biol Evol **20**(11): 1811-25.

Boeke, J.D. (1997). "LINEs and Alus - the polyA connection." Nat Genet **16**(1): 6-7.

Boissinot, S., and Furano, A.V. (2001). "Adaptive evolution in LINE-1 retrotransposons." Mol Biol Evol **18**(12): 2186-94.

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

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

Chang, D.Y., Hsu, K. and Maraia, R.J. (1996). "Monomeric scAlu and nascent dimeric Alu RNAs induced by adenovirus are assembled into SRP9/14-containing RNPs in HeLa cells." Nucleic Acids Res **24**(21): 4165-70.

Chou, K.C. (2000). "Prediction of tight turns and their types in proteins." Anal Biochem **286**(1): 1-16.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E., Greally, J.M., Voinnet, O. and Heard, E. (2010). "LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation." Cell **141**(6): 956-69.

Cohen, S.X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T.K., Lamzin, V.S., Murshudov, G.N., and Perrakis, A. (2008). "ARP/wARP and molecular replacement: the next generation." Acta Crystallogr D Biol Crystallogr **64**(Pt 1): 49-60.

Cordaux, R. and Batzer, M.A. (2009). "The impact of retrotransposons on human genome evolution." Nat Rev Genet **10**(10): 691-703.

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

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V. and Gage, F.H. (2009). "L1 retrotransposition in human neural progenitor cells." Nature **460**(7259): 1127-31.

Cowtan, K. (2006). "The Buccaneer software for automated model building. 1. Tracing protein chains." Acta Crystallogr D Biol Crystallogr **62**(Pt 9): 1002-11.

Craig, N.L. (2002) "Mobile DNA: an Introduction." Mobile DNA II, eds Craig, N.L., Craigie, R., Gellert, M., Lambowitz, A.M. (ASM Press, Washington, D.C.), pp 3-11.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 3rd, Snoeyink, J., Richardson, J.S., et al. (2007). "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids." Nucleic Acids Res **35**(Web Server issue): W375-83.

de Groot, R.J. (2006). "Structure, function and evolution of the hemagglutinin-esterase proteins of corona- and toroviruses." Glycoconj J **23**(1-2): 59-72.

Deininger, P.L. and Batzer, M.A. (1999). "Alu repeats and human disease." Mol Genet Metab **67**(3): 183-93.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet **35**(1): 41-8.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

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

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., Grimmond, S.M. and Carninci, P. (2009). "The regulated retrotransposon transcriptome of mammalian cells." Nat Genet **41**(5): 563-71.

Furano, A.V., Duvernell, D.D. and Boissinot, S. (2004). "L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish." Trends Genet **20**(1): 9-14.

Garcia-Perez, J.L., Doucet, A.J., Bucheton, A., Moran, J.V. and Gilbert, N. (2007). "Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase." Genome Res **17**(5): 602-11.

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P. and Moran, J.V. (2010). "Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells." Nature **466**(7307): 769-73.

Garrett, J.E., Knutzon, D.S. and Carroll, D. (1989). "Composite transposable elements in the Xenopus laevis genome." Mol Cell Biol **9**(7): 3018-27.

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**(5): 1383-93.

Ge, P., Tsao, J., Schein, S., Green, T.J., Luo, M. and Zhou, Z.H. (2010). "Cryo-EM model of the bullet-shaped vesicular stomatitis virus." Science **327**(5966): 689-93.

Goodier, J.L., Zhang, L., Vetter, M.R. and Kazazian, H.H. Jr. (2007). "LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex." Mol Cell Biol **27**(18): 6469-83.

Green, T.J., Zhang, X., Wertz, G.W. and Luo, M. (2006). "Structure of the vesicular stomatitis virus nucleoprotein-RNA complex." Science **313**(5785): 357–360.

Han, J.S. and Boeke, J.D. (2005). "LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression?" Bioessays **27**(8): 775-84.

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

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

Hancks, D.C. and Kazazian, H.H. Jr. (2010). "SVA retrotransposons: Evolution and genetic instability." Semin Cancer Biol 2010 Apr 21.

Hartmann, M.D., Ridderbusch, O., Zeth, K., Albrecht, R., Testa, O., Woolfson, D.N., Sauer, G., Dunin-Horkawicz, S., Lupas, A.N., and Alvarez, B.H. (2009). "A coiled-coil motif that sequesters ions to the hydrophobic core." Proc Natl Acad Sci U S A **106**(40): 16950-5.

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

Hohjoh, H. and Singer, M.F. (1996). "Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA." EMBO J **15**(3): 630-9.

Hohjoh, H. and Singer, M.F. (1997). "Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon." EMBO J **16**(19): 6034-43.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. (1996). "Errors in protein structures." Nature **381**(6580): 272.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T., Wheelan, S.J., Ji, H., Boeke, J.D. and Burns, K.H. (2010). "Mobile interspersed repeats are major structural variants in the human genome." Cell **141**(7): 1171-82.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M. and Devine, S.E. (2010). "Natural mutagenesis of human genomes by endogenous retrotransposons." Cell **141**(7): 1253-61.

Januszyk, K., Li, P.W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J.A., Martin, S.L. and Clubb, R.T. (2007). "Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1." J Biol Chem **282**(34): 24893-904.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenet Genome Res **110**(1-4): 462-7.

Kabsch, W. (2010). "XDS." Acta Crystallogr D Biol Crystallogr **66**(Pt 2): 125-32.

Kammerer, R.A., Kostrewa, D., Progias, P., Honnappa, S., Avila, D., Lustig, A., Winkler, F.K., Pieters, J., and Steinmetz, M.O. (2005). "A conserved trimerization motif controls the topology of short coiled coils." Proc Natl Acad Sci U S A **102**(39): 13891-6.

Kapitonov, V.V. and Jurka, J. (2003). "The esterase and PHD domains in CR1-like non-LTR retrotransposons." Mol Biol Evol **20**(1): 38-46.

Khazina, E., and Weichenrieder, O. (2009). "Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame." Proc Natl Acad Sci U S A **106**(3): 731-6.

Khazina, E., Truffault, V., Büttner, R., Schmidt, S., Coles, M., Weichenrieder, O. (2011). "Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition." Nat Struct Mol Biol **18**(9): 1006-14.

Kirilyuk, A., Tolstonog, G.V., Damert, A., Held, U., Hahn, S., Löwer, R., Buschmann, C., Horn, A.V., Traub, P. and Schumann, G.G. (2008). "Functional endogenous

LINE-1 retrotransposons are expressed and mobilized in rat chloroleukemia cells." Nucleic Acids Res **36**(2): 648-65.

Kojima, K.K. and Fujiwara, H. (2004). "Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets." Mol Biol Evol **21**(2): 207-17.

Kolosha, V.O. and Martin, S.L. (1997). "In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition." Proc Natl Acad Sci U S A **94**(19): 10155-60.

Kriegs, J.O., Churakov, G., Jurka, J., Brosius, J. and Schmitz, J. (2007). "Evolutionary history of 7SL RNA-derived SINEs in Supraprimates." Trends Genet **23**(4): 158-61.

Kroutter, E.N., Belancio, V.P., Wagstaff, B.J. and Roy-Engel, A.M. (2009). "The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition." PLoS Genet **5**(4): e1000458.

Kulpa, D.A. and Moran, J.V. (2005). "Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition." Hum Mol Genet **14**(21): 3237-48.

Lander, E.S. et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

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**(4): 595-605.

Lunde, B.M., Moore, C. and Varani, G. (2007). "RNA-binding proteins: Modular design for efficient function." Nat Rev Mol Cell Biol **8**(6): 479-90.

Malik, H.S., and Eickbush, T.H. (2002). "Origins and Evolution of Retrotransposons." Mobile DNA II, eds Craig, N.L., Craigie, R., Gellert, M., Lambowitz, A.M. (ASM Press, Washington, D.C.), pp. 1111-1144.

Maris, C., Dominguez, C. and Allain, F.H. (2005). "The RNA recognition motif, a plastic RNAbinding platform to regulate post-transcriptional gene expression." FEBS J **272**(9): 2118-31.

Martin, S.L. (1991). "Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells." Mol Cell Biol **11**(9): 4804-7.

Martin, S.L. (2006). "The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition." J Biomed Biotechnol **2006**(1): 45621.

Martin, S.L. and Bushman, F.D. (2001). "Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon." Mol Cell Biol **21**(2): 467-75.

Martin, S.L., Branciforte, D., Keller, D. and Bain, D.L. (2003). "Trimeric structure for an essential protein in L1 retrotransposition." Proc Natl Acad Sci U S A **100**(24): 13815-20.

Martin, S.L., Bushman, D., Wang, F., Li, P.W., Walker, A., Cummiskey, J., Branciforte, D. and Williams, M.C. (2008). "A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity." Nucleic Acids Res **36**(18): 5845-54.

Martin, S.L., Cruceanu, M., Branciforte, D., Wai-Lun, Li. P., Kwok, S.C., Hodges, R.S. and Williams, M.C. (2005). "LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein." J Mol Biol **348**(3): 549-61.

Martin, S.L., Li, J. and Weisz, J.A. (2000) "Deletion analysis defines distinct functional domains for protein–protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1." J Mol Biol **304**(1): 11-20.

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

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

Mills, R.E., Bennett, E.A., Iskow, R.C. and Devine, S.E. (2007). "Which transposable elements are active in the human genome?" Trends Genet **23**(4): 183-91.

Miné, M., Chen, J.M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Férec, C., Abitbol, M., Ricquier, D. and 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**(2): 137-42.

Moran, J.V., DeBerardinis, R.J. and Kazazian, H.H. Jr. (1999). "Exon shuffling by L1 retrotransposition." Science **283**(5407): 1530-4.

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**(5): 917-27.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J. and Moran, J.V. (2007). "Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres." Nature **446**(7132): 208-12.

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**(2): 159-65.

Muller, M., Weigand, J.E., Weichenrieder, O. and Suess, B. (2006). "Thermodynamic characterization of an engineered tetracycline-binding riboswitch." Nucleic Acids Res **34**(9): 2607-17.

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

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). "Refinement of Macromolecular Structures by the Maximum-Likelihood Method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-55.

Ortín, J. (2003). "Unraveling the replication machine from negative-stranded RNA viruses." Structure **11**(10): 1194-6.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L. and Kazazian, H.H. Jr. (2002). "A mouse model of human L1 retrotransposition." Nat Genet **32**(4): 655-60.

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

Pace, J.K. 2nd and Feschotte, C. (2007). "The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage." Genome Res **17**(4): 422-32.

Raymond, D.D., Piper, M.E., Gerrard, S.R. and Smith, J.L. (2010). "Structure of the Rift Valley fever virus nucleocapsid protein reveals another architecture for RNA encapsidation." Proc Natl Acad Sci U S A **107**(26): 11769-74.

Repanas, K., Zingler, N., Layer, L.E., Schumann, G.G., Perrakis, A. and Weichenrieder, O. (2007). "Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease." Nucleic Acids Res **35**(14): 4914-26.

Reuter, M., Chuma, S., Tanaka, T., Franz, T., Stark, A. and Pillai, R.S. (2009). "Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile." Nat Struct Mol Biol **16**(6): 639-46.

Rubin, C.M., VandeVoort, C.A., Teplitz, R.L. and Schmid, C.W. (1994). "Alu repeated DNAs are differentially methylated in primate germ cells." Nucleic Acids Res **22**(23): 5121-7.

Rudolph, M.G., Kraus, I., Dickmanns, A., Eickmann, M., Garten, W. and Ficner, R. (2003). "Crystal structure of the borna disease virus nucleoprotein." Structure **11**(10): 1219-26.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). "Evaluation of comparative protein modeling by MODELLER." Proteins **23**(3): 318-26.

Schulz, W.A. (2006). "L1 retrotransposons in human cancers." J Biomed Biotechnol **2006**(1): 83672.

Schumann, G.G. (2007). "APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition." Biochem Soc Trans **35**(Pt 3): 637-42.

Shamoo, Y., Abdul-Manan, N. and Williams, K.R. (1995). "Multiple RNA binding domains (RBDs) just don't add up." Nucleic Acids Res **23**(5): 725-8.

Sheldrick, G.M. (2008). "A short history of SHELX." Acta Crystallogr A **64**(Pt 1): 112-22.

Shoji, M., Tanaka, T., Hosokawa, M., Reuter, M., Stark, A., Kato, Y., Kondoh, G., Okawa, K., Chujo, T., Suzuki, T., Hata, K., Martin, S.L., Noce, T., Kuramochi-Miyagawa, S., Nakano, T., Sasaki, H., Pillai, R.S., Nakatsuji, N. and Chuma, S. (2009). "The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline." Dev Cell **17**(6): 775-87.

Söding, J. (2005b). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-60.

Söding, J., Biegert, A. and Lupas, A.N. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic Acids Res **33**(Web Server issue): W244-8.

Sugano, T., Kajikawa, M. and Okada, N. (2006). "Isolation and characterization of retrotransposition-competent LINEs from zebrafish." Gene **365**: 74-82.

Tawar, R.G., Duquerroy, S., Vonrhein, C., Varela, P.F., Damier-Piolle, L., Castagné, N., MacLellan, K., Bedouelle, H., Bricogne, G., Bhella, D., Eléouët, J.F. and Rey, F.A. (2009). "Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus." Science **326**(5957): 1279-83.

Trelogan, S.A. and Martin, S.L. (1995). "Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis." Proc Natl Acad Sci U S A **92**(5): 1520-4.

Ullu, E. and Tschudi, C. (1984). "Alu sequences are processed 7SL RNA genes." Nature **312**(5990): 171-2.

Vonrhein, C., Blanc, E., Roversi, P. and Bricogne, G. (2007). "Automated Structure Solution With autoSHARP." Methods Mol Biol **364**: 215–230.

Wallace, N., Wagstaff, B.J., Deininger, P.L. and Roy-Engel, A.M. (2008). "LINE-1 ORF1 protein enhances Alu SINE retrotransposition." Gene **419**(1-2): 1-6.

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

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**(4): 1429-39.

Weichenrieder, O., Repanas, K. and Perrakis, A. (2004). "Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon." Structure **12**(6): 975-86.

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**(6809): 167-73.

Wright, D.A., Ke, N., Smalle, J., Hauge. B.M., Goodman, H.M. and Voytas, D.F. (1996). "Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana." Genetics **142**(2): 569-78.

Wu, T., Datta, S.A., Mitra, M., Gorelick, R.J., Rein, A. and Levin, J.G. (2010). "Fundamental differences between the nucleic acid chaperone activities of HIV-1 nucleocapsid protein and Gag or Gag-derived proteins: biological implications." Virology **405**(2): 556-67.

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

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**(9): 763-71.

Ye, Q., Krug, R.M. and Tao, Y.J. (2006). "The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA." Nature **444**(7122): 1078-82.

Academic teachers

Dr. Elisa Izaurralde, Prof. Dr. Andrei Lupas, Prof. Dr. Christiane Nüsslein-Volhard, Prof. Dr. Ralf J. Sommer, Prof. Dr. Thilo Stehle, Dr. Oliver Weichenrieder, Prof. Dr. Detlef Weigel.