GLOBULAR DOMAIN STRUCTURE AND FUNCTION OF RESTRICTION-LIKE-ENDONUCLEASE BEARING LONG INTERSPERSED NUCLEOTIDE ELEMENTS

by

MST MURSHIDA MAHBUB

DISSERTATION
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
August, 2017

Arlington, Texas

Supervising Committee:

Dr. Shawn Christensen, Supervising Professor
Dr. Clay Clark
Dr. Michael R Roner
Dr. Saiful Chowdhury
Dr. Todd Castoe

DEDICATION

I dedicate this dissertation to my parents and my four year old daughter as well as my husband.

Without their help I would not be able to pursue my Ph.D.

ACKNOWLEDGEMENTS

I express my heartfelt gratitute to my supervisor, Dr. Shawn Christensen for his continuous support throughout my Ph.D years. Without his motivation and inspiration I would not be able to believe in my capabilities. He not only guided me planning and implementing the experiments but also with his insightful comments helped me understand the complicated nature of the data. Thank you Dr. Christensen as you made me feel lucky by your mentoring.

Beside my supervisor, I would like to thank my committee member Dr. Saiful Chowdhury, whose guidance and suggestions helped me designing experiments and learning mass spectrometry. Without his support, it would be impossible for me to learn this whole new technique. Becaue of him I really could dare to use the mass spectrometry tool to investigate R2Bm protein structure.

My sincere thanks also goes to committee member: Dr. Clay Clark, Dr. Michael R Roner and Dr. Todd Castoe, for their valuable comments and encouragements in my Ph.D. research.

I am thankful to Abu Hena Mostafa Kamal, Post doc in Dr. Chowdhury lab, as well as graduate and undergraduate students, who not only helped me in my work but also provided time in nano-LC-ESI mass spectrometer.

Linda Taylor, Gloria Burlingham, Sherri Echols, Rachel Wostl and others from the Biology department's administration were always helpful since the beginning of my Ph.D program. My sincere gratitude to them who always work in background to make our life easier here at UTA. I would also like to extend my appreciation to the graduate school of UT Arlington for providing financial support for my doctoral studies.

Last but not least, I thank my lab mates who helped me in the laboratory work and also to involve myself in the scientific discussions and seminars. Especially, I would like to thank Dr. Kimberly Bowles, whose encourgement and thoughtful comments at the beginning of my Ph.D. helped me to understand and ask research questions. I am grateful to National Science foundation and Phi Sigma Biological Society, UTA chapter for funding our work.

ABSTRACT

GLOBULAR DOMAIN STRUCTURE AND FUNCTION OF RESTRICTION-LIKE-ENDONUCLEASE BEARING LONG INTERSPERSED NUCLEOTIDE ELEMENTS

MST MURSHIDA MAHBUB

The University of Texas at Arlington, 2017

Supervising Professor: Shawn Christensen

Long Interspersed Nucleotide Elements (LINEs) are a major group of eukaryotic transposable elements that have profoundly influenced and sculpted eukaryotic genome structure and function. LINEs replicate within the host genome, often to high copy number. Replication occurs through an RNA intermediate, which is integrated back into the host genome by target primed reverse transcription (TPRT). The element encoded protein is known to contain a restriction-like DNA endonuclease, a reverse transcriptase, and nucleic acid binding domains. However, the 2° and 3° structure of these domains as well as the overall protein is poorly understood. The protein encoded by the R2 element from *Bombyx mori* (R2Bm) is expressible and purifiable and has thus facilitated much biochemical studies of the integration reaction. Using limited proteolysis and mass spectrometry, I studied globular domain structure of the R2Bm protein. It was discovered that the protein had two major globular domains: the zinc finger/Myb domain and the reverse transcriptase/linker/endonuclease superdomain. An easily proteolytically cleaved region between these two globular domains mapped to an area previously implicated in RNA binding. The large domain structure is similar to eukaryotic splicing factor protein Prp8's reverse transcriptase/linker/restriction endonuclease superdomain. An updated model of the reverse transcriptase domain of R2Bm protein was also generated and presented. The model was generated by protein threading and homology modeling algorithms. The model was tested by mapping the proteolytic cleavages back onto the model. Protein sequence alignments and structural overlays of the R2 reverse transcriptase and DNA endonuclease onto the splicing factor Prp8 indicate that the R2 protein and Prp8 likely shared a common ancestor. The structural and functional similarities in the linker region of both Prp8 and R2Bm are similarly discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF TABLES

Chapter 1

Introduction

1.1 Background and significance of LINEs

Long Interspersed Nucleotide Elements (LINEs) are mobile genetic elements. Almost all eukaryotic genome possess LINEs[1]. The Human genome has, on average, ~ 100 active LINE-1s[2,3]. The replication machinery of LINEs, however, is also parasitized by non-autonomous elements that hijack LINE replication machinery for their own mobilization. LINE parasites are called SINEs and can be exemplified by human Alu elements[4,5]. LINE and SINE replication can represent a sizable portion of the genome. LINE and LINE mediated activity constitute at least one-third of human genomic DNA[4]. Activity of LINEs can be source of mutation, lead to altered gene expression, provide raw material for recombination, thereby continuously act to sculpt and diversify the host genome[6–8].

1.2 Secondary, tertiary, and quaternary structural information on LINE encoded protein is scarce

LINE encoded protein structure and protein-nucleic acid interactions are not well understood. Only a handful structural information is available (that includes atomic force microscopy of L1 ORF1[9] and NMR structures of L1 ORF1p's c-terminal domain[10] , L1 encoded endonuclease[11], R1Bm endonuclease[12]).

1.3 ORF structure of LINEs

LINEs either contain one or two open reading frames (ORF)s (see Figure 1-1 on page 3). Single ORF bearing LINEs encode a restriction like endonuclease (RLE) at their C-terminal end while the two ORF bearing LINEs encode an endonuclease (EN) that share homology to the apurinic/apyrimidinic endonucleases (APE). There are over 25 clades of LINEs including both RLE and APE-encoding LINEs[13,14]. Single ORF RLE-bearing-LINEs are phylogenetically earlier branching than the two ORF APE-bearing-LINEs[15]. The RLE LINEs are also called early branching LINEs and the APE encoding ones are known as late branching LINEs.

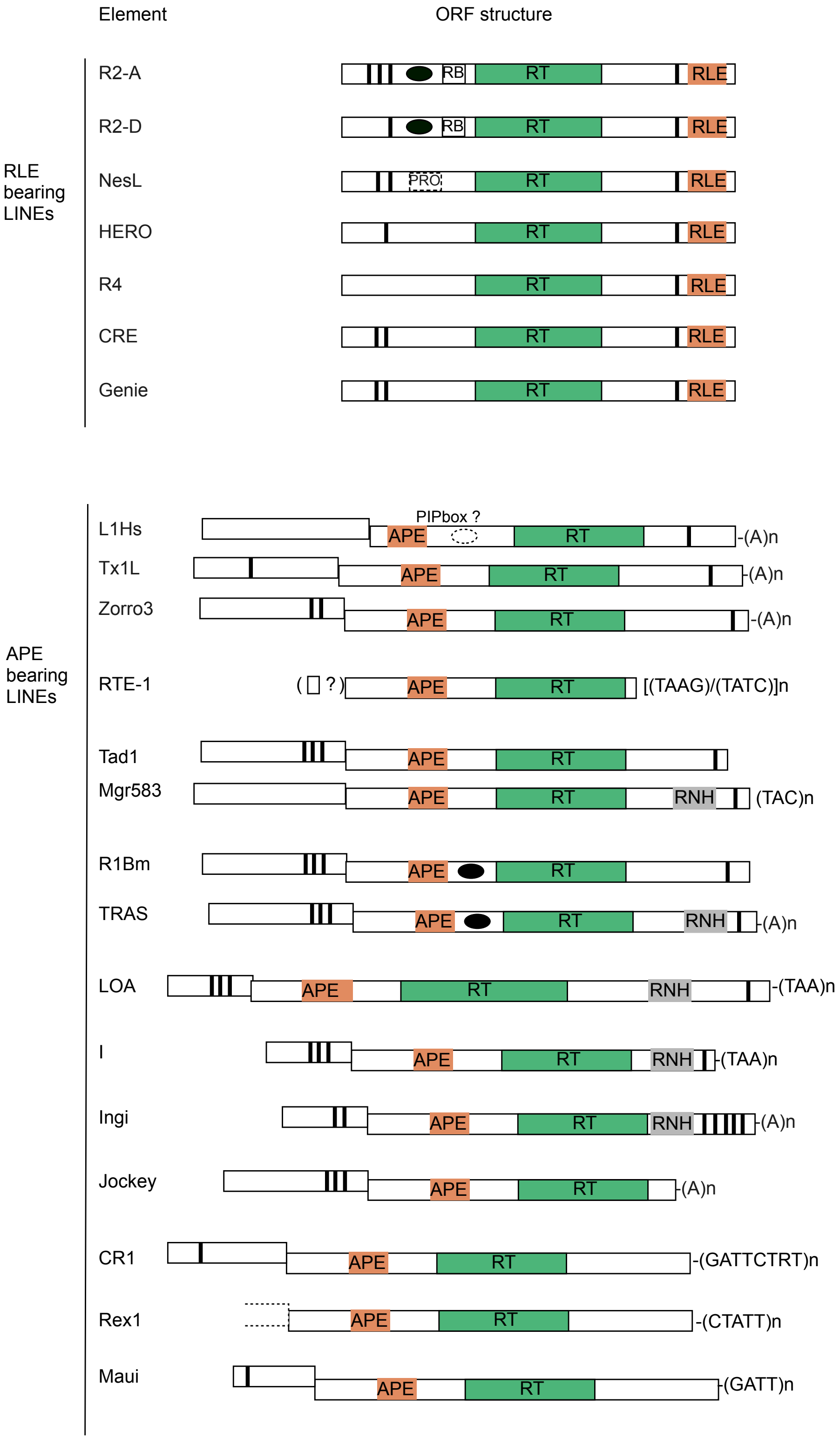


Figure 1-1. ORF structure of LINE encoded proteins. Zinc finger (ZF)/ gag knuckle like sequence motif is shown as vertical bold line; Myb is shown as ellipse; broken ellipse indicate

bioinformatically predicted myb three helix bundle; PIPbox = PCNA interacting protein domain; RB = RNA binding; RT = Reverse transcriptase; RLE = Restriction Like Endonuclease; APE = Apurinic-apyrimidinic endonuclease. The ORF1 for RTE element is putative and only 43 amino acids. The dotted open box in Rex1 indicates that the 5' end of this element has not been identified yet. The ORFs are not drawn to scale. Adapted from[13,15–23].

The single ORF of early branching LINEs typically encode one to three zinc finger(s) at the N-terminal end of the ORF, a centrally located reverse transcriptase (RT), a cysteine rich gag knuckle-like CCHC motif, and an RLE domain at the C-terminal end of the ORF. In addition, R2 clade elements contain a Myb motif and NeSL/Utopia clade contain a ULP-protease LINE elements[20]. Recently an RNA binding motif preceding the reverse transcriptase domain has been identified in R2 elements.[16,24–27]. The most extensively studied RLE bearing elements are the R2 clade elements, especially the R2 element from *Bombyx mori* (R2Bm).

The late branching LINEs contain two ORFs. The first open reading frame typically encodes 1-3 gag-knuckle zinc fingers. The second ORF encodes an N-terminal APE, a central reverse transcriptase, and a C-terminally located zinc knuckle-like CCHC motif. The L1 and I clades are the most well studied, particularly the human L1 element (L1Hs). L1Hs, is ~6kb long with two ORFs separated by a 63-nt intergenic region. 5'UTR with an internal promoter and 3' UTR with a poly(A) tail flank the L1Hs sequence[28,29]. The L-1Hs ORF1 generates ~40 kDa protein with n-terminal coiled coil domain, centrally located non-canonical RNA recognition motif (RRM) and a c-terminal basic domain. The amino terminal coiled coil domain facilitate trimer formation. The RNA recognition motif and the c-terminal domain are important for ORF1p's ability to bind RNA and RNP formation as well as nucleic acid chaperonin function [5,9,30,31].

The second ORF in L1Hs is ~150 kDa and contain amino terminal APE endonuclease, a central reverse transcriptase and the CCHC motif. The cysteine rich motif is required for retrotransposition[32] but its exact role is yet to be known. In the region between the APE and RT domain of L1, a PCNA interaction protein domain called PIPbox was found; PCNA is the sliding clamp protein essential for DNA replication. Exact role of PIPbox in L1 retrotransposition is unknown[5,19]. A c-myb like three helix motif has been predicted in area between APE and RT in TRAS, R1Bm, SART1, RT1Ag, TARTDm, and L1Hs element using secondary structure prediction program[33].

1.4 Shared domain (RT and H/RINALP and cysteine rich gag knuckle like motif)

All LINEs contain an RT domain, which is the key functional domain required for retrotransposition. A cysteine rich sequence motif (CCHC motif) near the C-terminal end of ORF (for APE LINEs) or preceding the C-terminal RLE domain is seen in most of the LINEs[18,34].The consensus sequence motif is $CX_{1-3}CX_{7-8}HX_4C$ or $CX_2CX_{12}HX_{3-5}H$[35,36]. Ingi elements carry five degenerate cysteine-rich regions[18]. Another conserved amino acid block that is common among all LINEs is H/RINALP or equivalent sequence containing alpha helical motif (see chapter 2). H/RINALP resides in the linker and is in juxtaposition to CCHC motif.

The CCHC ($CX_2CX_4HX_4C$) sequence has been shown to play key role in retroviral replication including RNA binding, genome packaging and chaperone activity[37]. Mutation in CCHC and H/RINALP equivalent in L1 interfere with L1 RNP (ribonucleoprotein) formation and retrotransposition in cultured cells[32,38]. Recombinantly expressed c-terminal segment of ORF2p containing the above motifs exhibits non-sequence specific RNA binding *in vitro* and cysteine to serine mutations in the C-domain do not adversely affect RNA binding[39].

1.5 Shared function (RNA binding, DNA binding, DNA cleavage and DNA polymerization)

All LINEs has to bind RNA, bind DNA and cleave DNA to release a 3'-OH group for priming DNA polymerization for retrotransposition. LINE protein shows cis-preference for the element RNA. The 3' UTR (untranslated region) of LINEs with specific sequence or structure is responsible for the recognition by the RT domain of LINEs and probably for cis-preference[40–44]. To be able to propagate in host DNA, every LINE has to find the suitable target DNA and bind without invoking host defense mechanism. Target site is specific for RLE containing LINEs and also for Tx1 and R1 clade elements among the APE containing LINEs[13]. For APE bearing LINEs, sequence or position specificity is lacking but weak target preference are found in A+T-rich regions[18].

All LINEs encode endonuclease either of RLE or APE type except for Dualen retrotransposons that bear both type of endonucleases flanking RT domain (not shown in figure 1-1). Dualen are located in the midpoint between early branched and late branched LINEs when the RT domains were analyzed phylogenetically[45]. The endonuclease domain of the LINE defines the distinctive feature of their retrotransposition, i.e., the target itself acts as primer after LINE mediated cleavage. The cleavage frees a 3'-OH group for the RT domain that can catalyze the transfer of nucleotidyl phosphate onto the primer for cDNA synthesis. This first strand cDNA synthesis begins the Target Primed Reverse Transcription (TPRT)[46].

1.6 Basic insertion mechanism

Target primed reverse transcription (TPRT) is the replication mechanism employed by LINEs, group II intron and telomerase[46–49]. R2Bm, an early branching LINE (and element studied in this dissertation), facilitated biochemical characterization of TPRT by reconstituting significant portions of the insertion reaction *in vitro*. A working model of the insertion reaction of R2Bm[50] includes four steps: 1) first strand DNA cleavage; 2) 1st strand DNA synthesis (TPRT); 3) 2nd strand DNA cleavage; and 4) 2nd strand DNA synthesis (Figure 1-2).

R2Bm protein appears to adopt different conformations upon binding to specific structured RNA components located in the 3' and 5' untranslated regions (UTRs) of the element RNA. R2 protein binds to these two segments of RNA and individual RNAs made to these two regions are called protein binding motif RNAs (5' and 3' PBM RNAs, respectively). DNase footprinting studies showed that when R2Bm protein is associated with the 3' PBM RNA has the RNP bound to target DNA upstream of insertion site. R2 protein bound to 5' PBM RNA was discovered to bind to target DNA downstream of the insertion site[50]. The upstream subunit's EN cleaves the antisense strand releasing the 3'-OH. The upstream subunit's RT uses the free -OH for 1st strand synthesis. First strand synthesis detaches the 5' PBM RNA from the protein in the downstream subunit. This triggers second strand cleavage by downstream subunit RLE. The second strand synthesis follows completing the integration reaction[50–52].

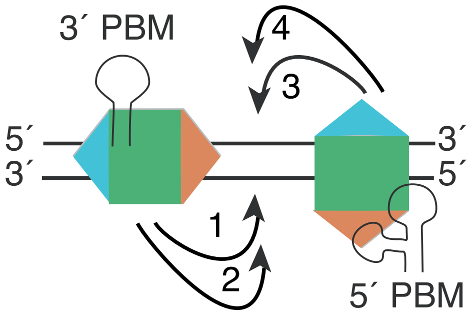


Figure 1-2. R2 insertion mechanism. 1) & 2) first strand DNA cleavage and first strand synthe-

sis. 3) & 4) second strand cleavage and second strand synthesis. PBM = Protein binding motif. Colored hexagons represent R2Bm protein. Black straight lines indicate double stranded target DNA. Turquoise and chocolate triangles represent N-terminal DNA binding and C-terminal endonuclease domains, respectively. Green rectangle is central RT domain. Adapted from[53].

DNA footprinting studies also indicate that the protein-DNA complex changes conformation after first strand DNA cleavage, shifting from the DNA cleavage to the DNA polymerization step of TPRT[51]. It is possible that the endonuclease and reverse transcriptase are on separate flexible arms to facilitate the ordered access to the insertion site. Knowledge of R2 protein globular domain structure would help elucidate how the different enzymatic active sites gain access to the insertion site and how they are coordinated.

1.7 Detailed TPRT mechanism

Life cycle of a LINE begins when a transcript is produced from an TPRT competent element. The RNA needs to travel to cytoplasm. Translation from the transcript yield element encoded protein that binds preferably the translation template. The RNP formed in this process finds the DNA target site to initiate first strand cleavage and first strand synthesis. With the second strand cleavage and second strand synthesis the replication process ends. The regulation of replication can be viewed both from host and retrotransposon perspective, which is discussed at the end of this chapter.

1.7.1 Transcription in LINEs

The site for R2 gene transcription and processing is the nucleolus[34]. R2 elements rely on the transcription of the rRNA gene unit for expression as rRNA/R2 cotranscript[54] (see Figure 1-3 on page 10). There is no internal promoter and RNA polymerase I is the responsible enzyme in the transcription process rather than the RNA polymerase III. There is a 5' ribozyme structure similar to hepatitis delta virus (HDV) that is autocatalytic and processes R2 transcript from the 28S/R2 cotranscript[55]. Structure based bioinformatics predicts that R4, R6 from the RLE-LINE group also possess self splicing ribozymes of HDV type for processing 5' end[56].

Location of RNA self-cleavage vary. *Drosophila simulans* R2 ribozyme cleaves at the precise 28S/R2 junction. Many other R2 ribozymes cleave in GC-rich area of the 28S rRNA 13-28 upstream of R2 insertion site[34,57].

The exact 3' end of the R2 transcript is not known. 3' UTR of ~250-nt length is necessary and sufficient for the TPRT reaction but short downstream 28S sequences help TPRT initiation at accurate position *in vivo.* Using only 3' UTR without any downstream 28S sequences *in vitro* appear to add nontemplated nucleotides before the cDNA [58].

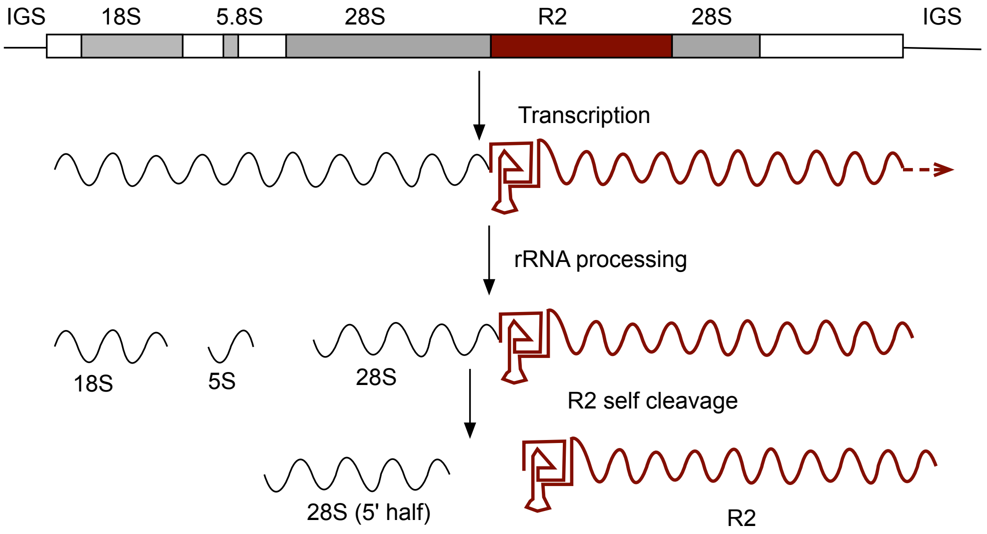


Figure 1-3. R2 ribozyme and self processing. Rectangle with gray shadings represent an rDNA unit with 18S, 5.8S and 28S genes. White boxes are transcribed spacers. A R2 insertion in the rDNA unit is shown with brick red box. R2 is co-transcribed with rRNA and following transcription, R2 splice out from the co-transcript using its 5' ribozyme. How R2 resolves its 3' end is not known[34]. IGS = Intergenic spacer.

Human L1 RNA transcript is bicistronic with two ORFs separated by 63-nt spacer containing two in-frame stop codons. The transcript starts with 5' UTR and ends in 3'UTR with poly(A) tail. There is also an out-of-frame AUG codon in a poor Kozak context in the spacer with a potential to encode a short ORF of 6 amino acid. Though the mRNA is bicistronic, the ORFs appear to be translated separately from each other, rather than being translated as a fusion protein[59,60].

L1 5' UTR is ~910 bp long with an internal RNA polymerase II promoter. A YY1-binding site at the 5' end of the 5' UTR is used for precise initiation of transcription[5,61].

Human L1 and silkworm SART1 elements have 3' UTR and poly (A) tail at their 3' end. SART1 needs both of these parts of the SART1 RNA for retrotransposition but for L1, the 3' UTR seems to be dispensable. R1Bm (a R1 clade element as is SART1), another APE containing LINE, does not contain an A-rich tract at its 3' end but requires a 3' UTR structure for retrotransposition. A spatial separation of 3'UTR and downstream 28S rRNA sequence leads to reduced retrotransposition suggesting some role of the downstream sequence in R1Bm retrotransposition[43] . SART element, Baggins and RTE possess ribozyme of the HDV type at the 5' terminus. *Trypanosoma cruzi* L1 elements also harvor HDV like ribozyme at the 5' end [56].

1.7.2 Translation in LINEs

Structure based searches identified HDV-like ribozyme structures through multiple types of retrotransposons in most of the animals phyla. The HDV-like ribozymes are active *in vitro* and *in vivo*. Liberation of the retrotransposons by ribozyme self scission activity from 28S-co-transcript as a 5' processing has been put forward. After releasing the transposon transcript from the precursor, HDV-ribozyme is set for its second function. HDV-ribozymes have a complex pseudoknot structure reminiscent of IRESes functional property. When tested *in vitro*, the ribozyme allowed translational machinery to carry out active heterologous protein expression. The aparent absence of 5'-mG cap or a conserved AUG codon in R2 elements thus explained well with the multifunctional HDV-ribozyme at the 5' end of many retrotransposons[56].

R2Bm elements has been reported to possess a 5' HDV-like ribozyme fold. However *in vitro* testing for self scission activity was not positive. But at 300 upstream, another non-canonical HDV-like ribozyme is also present which has the potential to form a stable catalytic core[56].

Human L-1 ORF2 translation occurs by an unconventional termination-reinitiation mechanism. *Bombyx mori* SART1 ORF2p translation uses translational coupling mechanism. In L1 translation, the conserved AUG code for ORF2p is not essential. ORF2p translation also does not need ORF1 or ORF1p. Rather, any translatable upstream ORF can help downstream ORF2p synthesis. Model for ORF2p translation by Alisch et al[59] postulates that the bicistronic mRNA is first used by the ribosome to translate protein from the upstream ORF; while the first ribosome continue to scan for ORF2 AUG and initiate translation of ORF2, more ORF1p form, which start coating the mRNA and thereby inhibiting any new ORF2p translation. ORF2p translation might vary among host and might exploit translation system inherent to the host system. Additional studies are needed[5]. A 7-mG (methyl Guanosine) cap is present in most LINE-1 element and is used for the element protein translation[5].

Silkmoth SART1 element however does use the AUG start codon. An overlapping UAAUG stop-start codons and downstream RNA secondary structure have been found as essential component for SART1 element translation[62] .

1.7.3 LINEs' RNPs

RNP formation is essential for TPRT. R2 protein finds its target site with ~150 fold increased affinity when provided with R2 RNA[63]. In presence of 3' PBM RNA, R2 protein can find its 28S rDNA target site in the crowd of genomic DNA and perform 1st strand synthesis. In absence of RNA, R2 protein can find and cleave the 28S rDNA target in total genomic DNA only at low level. The 2nd strand cleavage of TPRT is strictly dependent on presence of RNA[63]. The specific component of R2 RNA that plays role in 2nd strand cleavage is 5' PBM[50]. R2Bm RT can recognize and use R2Dm and even more distantly related arthropod 3'UTR for TPRT without any nucleotide sequence similarity[64,65]. The recognition appears to be dependent on the structure of 3' UTR RNA [64].

Human L-1 ORF1p and ORF2p both have *cis*-preference to L1 RNA. In L1 RNP, ORF1p is much more abundant than ORF2p. But the exact stoichiometry of ORF1p and ORF2p bound to LINE-1 mRNA is not confirmed[5]. The abundance of ORF1p probably lead to its easy detection in cytoplasmic RNPs[6]. In theory, as few as one ORF2p molecule could suffice for binding L1 RNA transcript and providing enzymatic machinery required for retrotransposition. The profound *cis*-preference likely allows the elements to retrotranspose the RNA transcripts from which they were translated[66]. In TART and HeT-A, ORF1p has been implicated in intracellular targeting (e.g., localization in Het dots)[67,68]. In human L1, ORF1p has RNA binding and chaperonin activity.

1.7.4 DNA recognition

DNA recognition appears to be largely separate from the catalytic domain of RLE in early branching LINEs. R2 recognizes 35 bp uptstream to 15 bp downstream of insertion site [34,69]. The n-terminal zinc finger (ZF) and Myb motifs contribute to downstream DNA binding in R2Bm, a R2-D clade element[70]. Protein region contributing to upstream DNA binding is not known for R2. But mutations in the α-helix-1 and loop region following the β-strand-4 of RLE domain affect overall DNA binding[53]. The element RNA might also affect protein conformation and contribute to target site recognition as the presence of RNA greatly enhances R2 proteins's ability to find 28S target region in the total genomic DNA[63].

Among the RLE LINEs, the single ORF structure varies primarily in the N-terminal domain. The N-terminal domain variation lead to different DNA target specificity (Table 1-1 on page 14 and Figure 1-4 on page 15). CRE, Genie and NeSL clade elements contain two zinc fingers (ZFs); R2-D clade elements have one ZF and one Myb motif while R2-A clade elements have three ZFs and one Myb motif (see figure 1-1 on page 3); HERO elements contain one ZF. R4 elements do not appear to contain any of the ZF or Myb motif[13,17].

Table 1-1: Target site of RLE bearing LINEs (Adapted from [13,17,71])

RLE bearing LINEs	Representative elements	Target site
R2-A	R2Lp, R8Hm, R9Av	rRNA gene (28S R2, 18S , 28S R9)
R2-D	R2Bm	rRNA gene (28S R2)
R4	R4Al, R4-2_Sra, Dong	rRNA gene (28S R4), tRNA-Asp gene, Microsatellite
NeSL	R5, NeSL-1Ce, R5-2_SM	rRNA gene, Spliced leader, Transposon
CRE	CRE2Cf, MoTeR, CRE-1_NV	Spliced leader, Telomeric repeat, Microsatellite
HERO	HERO-1_HR	Microsatellite
Genie	Genie-1Gl	771 bp repeat

Abbreviations: R9 element from *Adineta vaga* (R9Av), R8 from *Hydra magnipapillata* (R8Hm), R2 from *Limulus polyphemus* (R2Lp), R2 from *Bombyx mori* (R2Bm), R4 from *Ascaris lumbricoides* (R4Al), R4 from *Strongyloides ratti* (R4-2_Sra), NeSL from *Caenorhabditis elegans* (NeSL-1Ce), NeSL from *Schmidtea mediterranea* (R5-2_SM) CRE2 from *Crithidia fasciculata* (CRE2Cf), CRE from *Nematostella vectensis* (CRE-1_NV), HERO from *Helobdella robusta* (HERO-1_HR), Genie-1 from *Giardia lamblia* (Genie-1Gl).

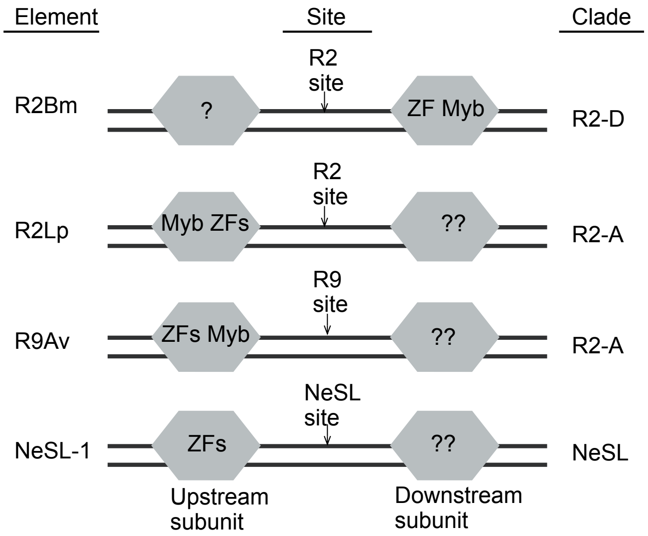


Figure 1-4. Differences in DNA targeting among RLE bearing LINEs. Abbreviations: R2 from *Bombyx mori* (R2Bm), R2 from *Limulus polyphemus* (R2Lp), R9 element from *Adineta vaga* (R9Av). Hexagons represent respective element protein. ZF and Myb = Zinc finger and Myb mediated target DNA binding. ? = protein subunit's position on DNA target is known, but the responsible protein motif for DNA binding unknown. ?? = hypothesized to bind template RNA, cleave the noncoding strand and begin TPRT as upstream R2Bm subunit does. Adapted from[17].

The R2-A group is more ancestral and binds DNA target differently than the R2-D group. R2Lp (R2 from *Limulus polyphemus*), a R2-A group element, uses the N-terminal domain (with three zinc finger and one Myb motifs) to bind upstream DNA sequences. Additionally, R9 from *Adineta vaga* (R9Av) and R8 from *Hydra magnipapillata* (R8Hm), two other R2-A clade elements, target non-canonical positions in the rRNA gene. R2 is the only clade amog RLE-containing LINEs that possess a Myb motif (see Figure 1-1 on page 3). Between the ZF and Myb motif, the later is used to attain major specificity in the target binding and DNA contact. NeSL, one of the two ZF containing RLE-LINE clade, uses the two ZF motifs in targeting DNA [17].

The L1 EN makes a single-strand nick at more or less random DNA sequence (5'-TTTT/AA-3'; while slash indicating the scissile phosphate). T_nA_n homopolymeric stretches create peculiar structural elements that can elicit response by L1 EN[72]. A wide minor groove tend to form in the TpA junction conducive to hairpin loop protusion of the protein surface[11]. L1 EN recognizes the extra helical flipped adenine residue 3' of the scissile bond to effect the cleavage[5]. Exchange of EN domain of SART1 and TRAS1 drives changes in the insertion site preference. This indicates that the primary determinant of APE-LINE mediated DNA recognition is the APE EN itself. For some groups of APE-LINEs, the target selection may additionally come from Myb like motif (R2Bm, Tx1L and Tx2L)[18]. Myb and myb-like proteins recognize specific DNA sequence using two helices of the three Myb motifs or by oligomerization of single Myb-motif containing protein. Probably the myb-like domain is used for primary recognition and then APE performs the final adjustment of the insertion position[18,73]. Homology of 3' end of the element RNA may dictate complementary base pairing with the target site. Chromatin formation may also play role. Transposon rich sites has shown positive correlation with heterochromatin regions. Specific phosphodiesters may get exposed upon nucleosome wrapping and become targeted for cleavage[18,74,75]. The R2 element has interestingly also been reported to recognize and cleave the target in the nucleosome wrapped DNA[69].

1.7.5 1st strand cleavage and 1st strand synthesis

Two symmetric half reactions constitute the RLE-LINE integration process. First half rection begins when R2 protein binds the 28S gene upstream of the cleavage site in presence of 3' PBM RNA. The R2 protein-3' PBM RNA complex cleaves the noncoding bottom strand with the RLE endonuclease. A free 3'-OH generated from this cleavage acts as primer for reverse transcription of cDNA from the RNA template using the RT activity of the upstream bound R2 protein[46]. 3'UTR is sufficient and required for this half reaction[40,63].

The initial stages of L1 retrotransposition was reconstituted *in vitro* using L1 ORF2p, L1 3' RNA, a target DNA and appropriate buffer components[76].

The 1st strand cleavage and 1st strand synthesis can be coupled or uncoupled. In an uncoupled reaction (endonuclease independent retrotransposition), cDNA synthesis is performed on a pre-nicked DNA as primer (for example, double strand break in the DNA can be used to initiate reverse transcription)[77–80].

1.7.6 2nd strand cleavage and 2nd strand synthesis

Second half reaction appears to begin when R2 protein is freed from 5' PBM RNA of the R2-5' PBM RNA complex[50], which triggers second strand cleavage.

Second strand synthesis so far has been inefficient *in vitro*. Variation in 5' junction of endogenous R2 insertion often is seen. In *Bombyx mori* and many animals, a particular length of 28S gene sequence is often duplicated at the 5' end, whereas in Drosophila and other animals, a common phenomenon is deletion of 28S gene and addition of non-templated nucleotides. Location of R2 ribozyme self-cleavage is correlated with these seen variations. Priming using a heteroduplex and microhomology were proposed for the observed duplications and deletion/non-templated addition, respectively[34].

1.8 Regulation of LINE replication

R2 copies are rapidly lost from rDNA loci by repeated recombination but new copies are also gained through new insertion events[81]. Higher number of insertions were found associated with higher rate of crossover in some *Drosophila simulans* stocks. Retrotransposition events mostly occured in female germ line and late in the development of egg. Transcriptional repression rather than post-transcriptional or post-translational degradaion regulates R2 expression. Only a subset of rDNA genes (30-40 out of several hundreds) need to be expressed in an individual[82]. All other rDNA genes are shut down by heterochromatin formation. Contiguous rDNA genes free of R2 insertions are activated to express ribosomal RNAs by euchromatin formation. If the R2 inserted rDNA genes are close to the expressed rDNA genes and if the expressed genes are short enough, R2 gets co-transcribed with the host rDNA genes. The mechanism to distinguish between inserted and uninserted rDNA units is not known but small RNA silencing pathways that induce heterochrmatin formation could be likely[34].

Cis-acting transcription factor binding sites are present in 5'UTR of LINE-1; mutation in these sites reduce LINE-1 transcription and retrotransposition. The transcription factors so far identified include Runx3, Sp1, and SRY-related (Sox) proteins. Other host factors are also likely to find these sites for controlling LINE-1 expression. Human LINE-1 5'UTR actually harbors two promoters: sense strand promoter as well as antisense promoter for RNA polymerase II. Mouse LINE-1 contain the antisense promoter in ORF1. The antisense promoter function is still speculative but in mammals such promoters have been implicated in noncoding RNA production; for LINE-1, the antisense promoter have been found to lead to chimeric transcript, which may lead to epigenetic silencing of L1 along with nearby cellular genes[5,60,83,84]. The bidirectionality of transcripts arising from sense and antisense promoter also have been found to suppress L1 by the RNA interference effect[85].

Human L1 3' UTR is ~206 bp in length with a conserved polypurine tract. A G-quadruplex structure is predicted to form by the 3' UTR of L1. The 3'UTR recently has been shown to possess promoter activity that can lead to generation of alternative L1 transcript in various tissues. Epigenetic modification of target DNA (nucleosome accessibility) might also affect L1EN cleavage activity[5].

1.9 LINE related RTs (group II intron protein, telomerase etc.) also use TPRT

Group II intron and telomerase use TPRT for replication[47,86,87] as do the LINEs. Group II intron employs TPRT mostly for retrohoming and occasionally for retrotransposition[88,89]. Retrohoming can be endonuclease dependent or independent[90,91]. In the retrohoming mechanism, the intron RNA reverse splices into target DNA after base pairing. The intron encoded protein cleaves the opposite strand releasing 3'OH group acting as primer. Synthesis of cDNA uses the reverse spliced RNA as template. Involvement of host enzymes, intron degradation and second-strand cDNA synthesis may complete the intron integration process. For group II introns that lack the endonuclease domain, DNA replication fork that is the source of nascent leading or lagging strand tend to act as source of primer for TPRT. A bias for dsDNA vs ssDNA as the target of endonuclease dependent and independent group II intron integration has been implicated[92], respectively. Telomerase RTs have the ability to extend the chromosome ends called telomeric repeats using short template RNA. The 3' OH of the DNA at the chromosome ends serve as primer and as LINEs RTs, telomerase RT synthesize the cDNA using the TPRT mechanism[87].

1.10 Dissertation goal

Despite of advancement of knowledge in LINE retrotransposition, several major aspects of detailed TPRT mechanism remain unanswered. Detailed mapping of interaction between LINE protein(s) and element RNA have yet to be worked out. The communication of protein, RNA, and DNA in the integration complexes at each stage of the integration reaction is still largely unknown. The dissection and elucidation of all of presumptive integration complex conformational changes would be aided by knowing more about the 2°, 3° and overall globular domain structure of the RLE LINE encoded protein. So far, the elucidation of ORF structures of RLE LINE encoded proteins are largely based on sequence alignment, mutational and biochemical data. Overall how the protein is folded into 3D globular domain structure is unknown. R2Bm, a model LINE, has faciliated much of the *in vitro* biochemical studies for detailed characterization of the RLE LINE replication integration. Recombinant R2Bm protein is heterologously expressed in *E. coli* cells; however, the current yield of purified R2Bm protein not amenable for crystallographic studies. Limited proteolysis offers useful, albeit low resolution, tool for investigation of R2Bm protein structure. We employed mass spectrometry to identify the limited proteolysis derived fragments of R2Bm protein to decipher its 3D globular domain structure. LINE related RTs' structures recently have been deposited in structure database. Phylogenetic connection[93,94] as well as structural (shared domain and amino acid motifs) and mechanistic similarities[47,95,96] among these different group of RTs indicate shared properties that might have interchangeable implications. Using homology driven protein threading algorithm[97], an updated model of RT domain has been built. Limited proteolysis derived cleavage data validates the model. We also show bioinformatically the relationship of LINEs with the central spliceosomal protein, Prp8. The detailed description of the technique and results sections are included in chapter 2. Finally in chapter 3, concluding remarks are stated along with limitations, implications and future directions.

References:

1. Arkhipova, I. & Meselson, M. Transposable elements in sexual and ancient asexual taxa. *Proc Natl Acad Sci U S A* **97**, 14473-14477 (2000).

2. Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**, 5280-5285 (2003).

3. Sassaman, D. M. et al. Many human L1 elements are capable of retrotransposition. *Nat Genet* **16**, 37-43 (1997).

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

5. Richardson, S. R. et al. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061 (2015).

6. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**, 187-215 (2011).

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

8. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268-274 (2004).

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

10. Januszyk, K. et al. 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**, 24893-24904 (2007).

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

12. Maita, N., Aoyagi, H., Osanai, M., Shirakawa, M. & Fujiwara, H. Characterization of the sequence specificity of the R1Bm endonuclease domain by structural and biochemical studies. *Nucleic Acids Res* **35**, 3918-3927 (2007).

13. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3**, MDNA3-0001 (2015).

14. Kapitonov, V. V., Tempel, S. & Jurka, J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207-213 (2009).

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

16. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res* **42**, 8405-8415 (2014).

17. Shivram, H., Cawley, D. & Christensen, S. M. Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements* **1**, 169-178 (2011).

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

19. Taylor, M. S. et al. Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155**, 1034-1048 (2013).

20. Malik, H. S. & Eickbush, T. H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. *Genetics* **154**, 193-203 (2000).

21. Burke, W. D., Malik, H. S., Rich, S. M. & Eickbush, T. H. Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. *Mol Biol Evol* **19**, 619-630 (2002).

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

23. Malik, H. S. & Eickbush, T. H. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol Biol Evol* **15**, 1123-1134 (1998).

24. Yang, J., Malik, H. S. & Eickbush, T. H. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A* **96**, 7847-7852 (1999).

25. Burke, W. D., Calalang, C. C. & Eickbush, T. H. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol* **7**, 2221-2230 (1987).

26. Burke, W. D., Muller, F. & Eickbush, T. H. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res* **23**, 4628-4634 (1995).

27. Burke, W. D., Singh, D. & Eickbush, T. H. R5 retrotransposons insert into a family of infrequently transcribed 28S rRNA genes of planaria. *Mol Biol Evol* **20**, 1260-1270 (2003).

28. Scott, A. F. et al. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**, 113-125 (1987).

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

30. Zingler, N. et al. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15**, 780-789 (2005).

31. Martin, S. L. et al. LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. *J Mol Biol* **348**, 549-561 (2005).

32. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).

33. Kubo, Y., Okazaki, S., Anzai, T. & Fujiwara, H. Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. *Mol Biol Evol* **18**, 848-857 (2001).

34. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).

35. Kajikawa, M., Ohshima, K. & Okada, N. Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. *Mol Biol Evol* **14**, 1206-1217 (1997).

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

37. Thomas, J. A. & Gorelick, R. J. Nucleocapsid protein function in early infection processes. *Virus Res* **134**, 39-63 (2008).

38. Doucet, A. J. et al. Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet* **6**, (2010).

39. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3**, 433-437 (2013).

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

41. Kajikawa, M. & Okada, N. LINEs mobilize SINEs in the eel through a shared 3’ sequence. *Cell* **111**, 433-444 (2002).

42. Osanai, M., Takahashi, H., Kojima, K. K., Hamada, M. & Fujiwara, H. Essential motifs in the 3’ untranslated region required for retrotransposition and the precise start of reverse transcription in non-long-terminal-repeat retrotransposon SART1. *Mol Cell Biol* **24**, 7902-7913 (2004).

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

44. Belancio, V. P., Whelton, M. & Deininger, P. Requirements for polyadenylation at the 3’ end of LINE-1 elements. *Gene* **390**, 98-107 (2007).

45. Kojima, K. K. & Fujiwara, H. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15**, 1106-1117 (2005).

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

47. Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545-554 (1995).

48. Nakamura, T. M. & Cech, T. R. Reversing time: origin of telomerase. *Cell* **92**, 587-590 (1998).

49. Eickbush, T. H. Telomerase and retrotransposons: which came first? *Science* **277**, 911-912 (1997).

50. Christensen, S. M., Ye, J. & Eickbush, T. H. RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* **103**, 17602-17607 (2006).

51. Christensen, S. & Eickbush, T. H. Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* **336**, 1035-1045 (2004).

52. Christensen, S. M. & Eickbush, T. H. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**, 6617-6628 (2005).

53. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* (2016).

54. Eickbush, D. G. & Eickbush, T. H. Transcription of endogenous and exogenous R2 elements in the rRNA gene locus of Drosophila melanogaster. *Mol Cell Biol* **23**, 3825-3836 (2003).

55. Eickbush, D. G. & Eickbush, T. H. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol* (2010).

56. Ruminski, D. J., Webb, C. H., Riccitelli, N. J. & Luptak, A. Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes. *J Biol Chem* (2011).

57. Eickbush, D. G., Burke, W. D. & Eickbush, T. H. Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441 (2013).

58. Luan, D. D. & Eickbush, T. H. Downstream 28S gene sequences on the RNA template affect the choice of primer and the accuracy of initiation by the R2 reverse transcriptase. *Mol Cell Biol* **16**, 4726-4734 (1996).

59. Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* **20**, 210-224 (2006).

60. Babushok, D. V. & Kazazian, H. H. J. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28**, 527-539 (2007).

61. Athanikar, J. N., Badge, R. M. & Moran, J. V. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* **32**, 3846-3855 (2004).

62. Kojima, K. K., Matsumoto, T. & Fujiwara, H. Eukaryotic translational coupling in UAAUG stop-start codons for the bicistronic RNA translation of the non-long terminal repeat retrotransposon SART1. *Mol Cell Biol* **25**, 7675-7686 (2005).

63. Yang, J. & Eickbush, T. H. RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol Cell Biol* **18**, 3455-3465 (1998).

64. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3**, 1-16 (1997).

65. Ruschak, A. M. et al. Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* **10**, 978-987 (2004).

66. Wei, W. et al. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**, 1429-1439 (2001).

67. Rashkova, S., Karam, S. E., Kellum, R. & Pardue, M. L. Gag proteins of the two Drosophila telomeric retrotransposons are targeted to chromosome ends. *J Cell Biol* **159**, 397-402 (2002).

68. Rashkova, S., Athanasiadis, A. & Pardue, M. L. Intracellular targeting of Gag proteins of the Drosophila telomeric retrotransposons. *J Virol* **77**, 6376-6384 (2003).

69. Ye, J., Yang, Z., Hayes, J. J. & Eickbush, T. H. R2 retrotransposition on assembled nucleosomes depends on the translational position of the target site. *EMBO J* **21**, 6853-6864 (2002).

70. Christensen, S. M., Bibillo, A. & Eickbush, T. H. Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468 (2005).

71. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).

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

73. Kyme, P. A. et al. Unusual trafficking pattern of Bartonella henselae -containing vacuoles in macrophages and endothelial cells. *Cell Microbiol* **7**, 1019-1034 (2005).

74. Lippman, Z. et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471-476 (2004).

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

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

77. Morrish, T. A. et al. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**, 208-212 (2007).

78. Srikanta, D. et al. An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93**, 205-212 (2009).

79. Sen, S. K., Huang, C. T., Han, K. & Batzer, M. A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35**, 3741-3751 (2007).

80. Teng, S. C., Kim, B. & Gabriel, A. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature* **383**, 641-644 (1996).

81. Jakubczak, J. L., Zenni, M. K., Woodruff, R. C. & Eickbush, T. H. Turnover of R1 (type I) and R2 (type II) retrotransposable elements in the ribosomal DNA of Drosophila melanogaster. *Genetics* **131**, 129-142 (1992).

82. Zhou, J., Eickbush, M. T. & Eickbush, T. H. A population genetic model for the maintenance of R2 retrotransposons in rRNA gene loci. *PLoS Genet* **9**, e1003179 (2013).

83. Matlik, K., Redik, K. & Speek, M. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006**, 71753 (2006).

84. Nigumann, P., Redik, K., Matlik, K. & Speek, M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628-634 (2002).

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

86. Belfort, M., Curcio, M. J. & Lue, N. F. Telomerase and retrotransposons: reverse transcriptases that shaped genomes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 20304-20310 (2011).

87. Blackburn, E. H. Telomerases. *Annu Rev Biochem* **61**, 113-129 (1992).

88. Ichiyanagi, K. et al. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* **46**, 1259-1272 (2002).

89. Coros, C. J. et al. Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol Microbiol* **56**, 509-524 (2005).

90. Cousineau, B. et al. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**, 451-462 (1998).

91. Martínez-Abarca, F., García-Rodríguez, F. M. & Toro, N. Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Mol Microbiol* **35**, 1405-1412 (2000).

92. Zhong, J. & Lambowitz, A. M. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J* **22**, 4555-4565 (2003).

93. Gladyshev, E. A. & Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* **108**, 20311-20316 (2011).

94. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**, 3353-3362 (1990).

95. Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**, MDNA3-0050 (2015).

96. Lingner, J. et al. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276**, 561-567 (1997).

97. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015).

Chapter 2

GLOBULAR DOMAIN STRUCTURE AND FUNCTION OF RESTRICTION-LIKE ENDONUCLEASE LINES: SIMILARITIES TO EUKARYOTIC SPLICING FACTOR PRP8

AUTHORS

Mst Murshida Mahbub[1], Saiful Chowdhury[2], and Shawn Christensen[1]

1. Department of Biology, University of Texas at Arlington, Arlington, TX 76010, U.S.A.

2. Department of Chemistry and Biochemistry, University of Texas at Arlington, Arlington, TX 76010, U.S.A.

Submitted in Journal of Biological Chemistry

2.1 Abstract

R2 elements are a clade of early branching long interspersed nucleotide elements (LINEs). LINEs are retrotransposable elements whose replication can have profound effects on the genomes in which they reside. No crystal or EM structures exist for the reverse transcriptase (RT) and linker regions of LINEs. Using limited proteolysis as a probe for globular domain structure, we show that the protein encoded by the *Bombyx mori* R2 element has two major globular domains: 1) a small globular domain consisting of the N-terminal zinc finger and Myb motifs, and 2) a large globular domain consisting of the RT, linker, and restriction-like endonuclease (RLE). Further digestion of the large domain occurs primarily within the fingers and palm of the RT. Mapping these RT cleavages onto an updated model of the RT indicated that a largely hydrophobic face of the RT and the thumb of the RT are largely protected from proteolytic cleavage. The thumb of the R2Bm RT is presumably protected by the linker. The RT model used in this study was generated using modern protein homology/threading algorithms. The crystal structure of Prp8, a eukaryotic splicing factor, was a major template used in building the R2Bm RT model, particularly the thumb region. The RT of Prp8 shares a number commonalities with the RT of LINEs (e.g., RT0, RT2a, RT3a, RT6a insertions and a palm traversing helix that connects -1 to RT0). Further, the large fragment of Prp8 consists not only of a RT but also an RLE and a linker connecting the two regions. The linker sequences adjacent to the RLE in LINEs and Prp8 share a set of two important α-helices and a (presumptive) ββα structural motif. Finally, the RLE found in LINEs is strikingly similar to the RLE found in Prp8. Both RLEs share a unique catalyt-

ic core residue spacing as well as several other key conserved residues. Prp8 and LINEs may have had a common ancestor.

2.2 Introduction

Long interspersed nucleotide elements (LINEs) are a major class of retrotransposable elements. LINEs package their transcribed RNA into ribonucleoprotein particles (RNP) using element encoded proteins translated from the mRNA being packaged. LINEs insert their genetic material back into the host genome at a new location by target primed reverse transcription (TPRT)[1–5]. TPRT is initiated by cleavage of one of the target chromosomal strands by an element encoded DNA endonuclease. The free 3'-OH DNA end generated by the DNA endonuclease is used to prime reverse transcription of the element RNA, thus inserting a new DNA copy of element into the host genome.

All LINEs are believed to require the same basic activities to integrate: RNA binding activity, DNA binding activity, DNA endonuclease activity, reverse transcriptase (RT) activity, and completion of integration by second strand synthesis. There are two major groups of LINEs. The two groups share a common RT and a gag knuckle-like CCHC motif. The two groups differ in their open reading frame (ORF) structures, RNA binding domains, DNA binding domains, and DNA endonuclease domains used to form the element RNP and to integrate into the host DNA.

The earlier branching group has a single ORF. The ORF encodes a multifunctional protein with N-terminal zinc finger and Myb motifs, a RT, a gag-knuckle like motif, and a restriction-like endonuclease (RLE) (reviewed in[6,7]). This group of LINEs is generally site-specific during integration. The insect R2 element is a well studied example of this early branching LINE group.

The later branching group has two open reading frames. The second open reading is similar to that of the earlier branching group. It encodes an apurinic-apyrimidinic family endonuclease (APE), a RT, and the gag knuckle-like motif (reviewed in[8–12]). The mammalian L1 element is a well studied example of this later branching LINE group.

While crystal structures exist for the APE endonuclease and for the protein product of the first ORF of APE LINEs, no crystal or cryo-EM structures exist for the RLE LINEs nor for the regions common between the two groups of LINEs[13–17]. Our previous paper reported a protein threading model for the restriction-like endonuclease of R2 elements[18]. This paper reports the globular domain structure of R2Bm as probed by limited proteolysis. An updated model of the R2 RT is also presented along with an analysis of the linker region between the RT and the endonuclease. The R2 proteolytic data in conjunction with sequence-structure alignments of the RT, linker, and RLE indicate that RLE LINEs share a number of commonalities with the large fragment of Prp8, a highly conserved eukaryotic splicing factor.

2.3 Materials and Methods

2.3.1 Protein expression and purification

R2Bm protein was expressed and purified as previously described[18,19]. Briefly, the R2Bm protein used in this study was ΔNR2Bm. The ΔNR2Bm construct removes the variable N-terminal found in R2 elements (the first 69 amino acids from R2Bm genbank entry M16558.1) and adds a six histidine tag on the C-terminal end of the protein[18,19]. The ΔNR2Bm expression construct was put into BL21 *Escherichia coli* cells. Five hundred mililiter of cultures were grown in LB broth, expressed with IPTG, lysed, and the soluble material purified over a Talon affinity column (Clontech #635501). The R2Bm protein was eluted off the column in 50 mM HEPES pH 7.5, 100 mM NaCl, 50% glycerol, 0.1% triton X-100, 150 mM imidazole. Proteins were stored in elution buffer supplemented with 1 mM DTT (final concentration) at -20° C. R2 protein was quantified by SYPRO Orange (Sigma #S5692) staining of samples run on SDS-PAGE relative to a BSA standard (Biorad #500-0202). All quantitations were done using Fiji software analysis of digital photographs[20].

2.3.2 Limited proteolysis of R2Bm protein and processing of the polypeptides

Limited digestion of purified R2 protein was carried out in the absence of nucleic acids using a trace amount of GluC (NEB, #P8100S) or LysC (Promega, #V1671) protease. Digestion was stopped using SDS loading buffer (to a final concentration of 50 mM Tris-Cl, pH 8.8 final; 4% SDS; 10 mM DTT) and heated. Proteolytic fragments were carbamidomethylated (55 mM final, Alfa aesar, #A14715) in the loading buffer in the dark at room temperature for 30 minutes prior to loading[21] onto a precast (Biorad criterion polyacrylamide gel, 18% and 4-15%) SDS-PAGE gel. The resolved protein fragments were stained with colloidal coomassie blue (Invitrogen, #LC6025). Prominent bands from across the proteolytic time course were excised from the gel, cut into 1 mm pieces, and destained using 25 mM NH_4HCO_3/50% Acetonitrile (ACN). Gel pieces were shrunk with 100% ACN (VWR, #BDH6002-4) and dried by Speed Vac[22–24] (Eppendorf).

The primary amines, including the amino-terminal end of the proteolytic fragments, were acetylated[25] in the gel slice using 15% acetic anhydride (Sigma, #320102) for 5 hours at room temperature within the individual excised gel fragments[26]. Acetylation was stopped by adding 1M NH_4HCO_3 (Sigma, #40867) solution. After 20 minutes, the gel pieces were shrunk by 100% ACN.

The dried gel pieces were swelled in 25 mM NH_4HCO_3 containing trypsin or GluC for 1.5 hours at 4^{o}C and any unabsorbed NH_4HCO_3 solution was then discarded. The gel pieces were covered with 25 mM NH_4HCO_3 and the in-gel digestion was carried out overnight at 37^{o}C. Peptides from in-gel digestion reaction were collected in the supernatant. Additional extractions with 0.1% Formic acid (FA)(Sigma, #399388) and 50% ACN/0.1% FA were also collected and added to the supernatant. The supernatant was dried in a Speed Vac and purified over C18 zip tip using standard procedures[22–24].

To catch any major cleavage sites that did not result in isolatable SDS-page bands, limited proteolysis reactions were run on SDS-PAGE gel for a very short time so as to not resolve bands, rather keeping them clustered near the well. The top portion of these lanes were excised and processed as above so as to remove triton and otherwise prepare the polypeptides for mass spectrometry thus avoiding the precipitation of the larger R2Bm protein fragments that occurs if the polypeptide processing (for mass spectrometry) was done in solution instead of in-gel. This abbreviated in-gel procedure is roughly equivalent to a direct “in-solution” detection of cleavage sites.

2.3.3 Mass spectrometry and Edman degradation

The eluted peptides were resuspended in 0.1% FA for sequencing by nanoLC-ESI-MS/MS using a Thermo Scientific LTQ Velos Pro ion trap mass spectrometer. R2 peptides were identified using Thermo Proteome Discoverer software (version 2.0); a database of R2Bm protein fragments was created, and a peptide was assigned as either N-term[25] end or internal peptide based on the position of acetyl groups in the peptide sequence. The internal peptides generated after trypsin (second) digestion will lack an acetyl group at the N-term end, as acetylation is performed prior to the second protease digestion step.

Amino-terminal sequencing of the separated proteolytic fragments was used to map the protease cleavage sites back onto the primary sequence of R2 and thus delimit globular domain boundaries. The internal peptides were also identified from the MS/MS spectrum. The internal peptide coverage and sequence was used to help verify the peptide location within the R2 ORF and act as a rough estimation of the C-terminal boundary of the fragment, along with SDS-PAGE estimation of the fragment's molecular weight.

For detection by Edman degradation, an SDS-PAGE gel was electrophoresed onto a PVDF membrane. Excised bands on the PVDF membrane were sent to UT Southwestern proteomics core for Edman sequencing.

2.3.4 3D modeling and multiple sequence alignments

Phyre 2.0 protein fold recognition server was used to model the RT domain of R2Bm protein[27]. Different length of R2Bm sequence upstream and downstream of RT domain were submitted for modeling. Model visualization was aided by UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization and Informatics at the University of California, San Francisco, CA, USA[28].

The PROMALS3D server was used for structure based alignment with minor manual adjustments. Seventy five LINE sequences were aligned first in PROMALS3D server that included 31 RLE LINE and 44 APE LINEs. Using this LINE alignment as constraint, an extended alignment was built with 3 nMat proteins, 5 group II introns, 7 RVT genes and 10 Prp8 proteins[29,30]. The secondary structure was plotted on the multiple sequence alignment (built by PROMALS3D) using Ali2D tool from MPI bioinformatic toolkit[31]. Aliview and clustalX softwares were used for viewing and editing of the PROMALS3D alignment[32,33]. Heat-map was generated using Gitools[34].

2.4 Results

2.4.1 Mapping and sequencing LysC protease resistant fragments of R2Bm protein

In order to probe the globular domain structure of R2Bm, R2Bm protein was subjected to limited proteolysis by one of several proteases. LysC, which cleaves on the C-terminal side of lysine residues, was one of these proteases. There are 42 lysine residues in the expressed and purified R2Bm protein. Aliquots from the digestion reaction were pulled at different time points and the reactions terminated. The digestion profile of R2 protein cleaved by LysC at the different time points were analyzed by SDS-PAGE (Figure 2-1A on page 35).

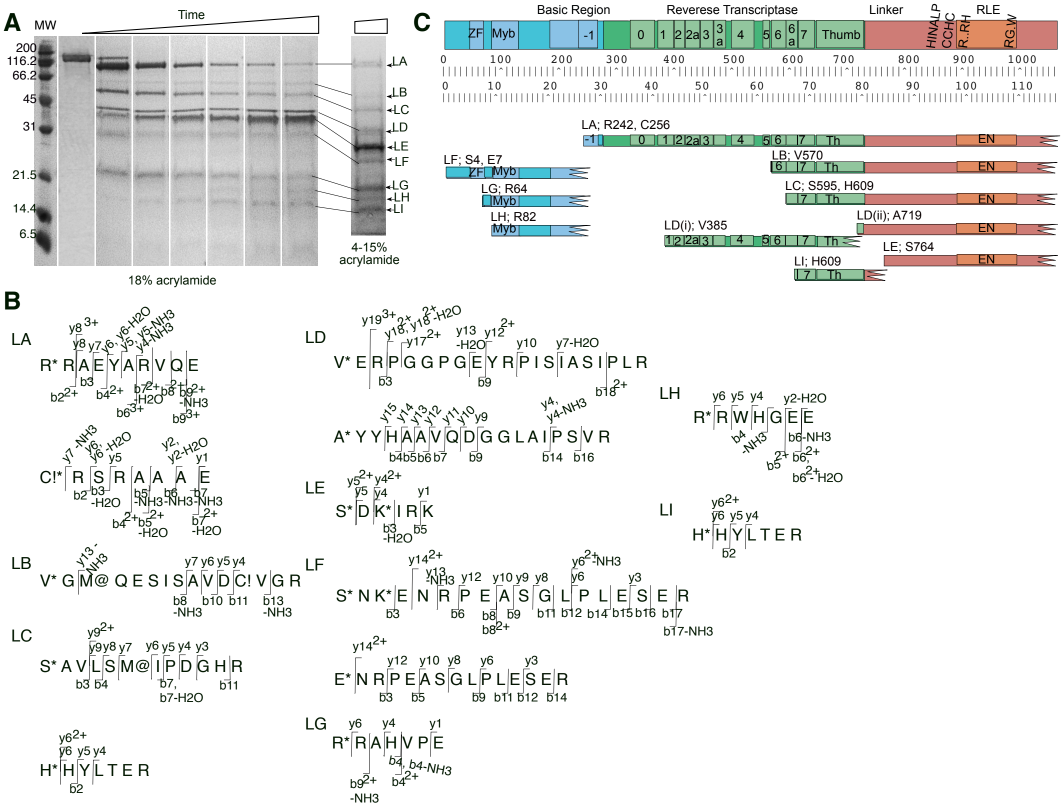


Figure 2-1. Mapping and sequencing LysC protease resistant fragments of R2Bm protein.
A) R2Bm protein was digested with LysC protease and analyzed by SDS-PAGE. Major observed bands were designated LA-LI. Triangle represents a time course of LysC digestion. The molecular weight (MW) marker values are given. B) Identification of proteolytic fragments of R2Bm protein. Bands from panel A were cut out, further processed, and analyzed by nano-LC-ESI-MS/MS sequencing. The N-terminal of the band producing fragment was identified by acetylation. Internal peptides were sequenced as well. The y and b ions that identified the N-terminal end are indicated. Symbols: * = acetylation; @ = oxidation; ! = carbamidomethylation. The spectrum is given in the supplemental data. C) Map of the band purified R2Bm fragments. A detailed diagram of the R2Bm open reading frame (ORF) is given along with amino acid and KD rulers. The boundaries of the R2Bm proteolytic resistant fragments LA-LI are mapped below along with the amino acid and primary sequence position of first amino acid of the fragment. The C-terminal ends were not exactly pinpointed but were roughly determined using the apparent MW from the SDS-PAGE gel and by the coverage of internal peptides sequenced by nano-LC-ESI-MS/MS. Abbreviations: zinc finger (ZF) and restriction-like endonuclease (RLE).

At least nine major bands (LA-LI) were observed. Some of these bands appeared early in the time course (e.g. LA, LC, LF and LG) while other bands appeared at later time points (e.g., LE, LH, and LI). Collectively, these bands represent protease resistant R2Bm fragments. The protease resistant fragments were excised from the gel, acetylated, and then digested to completion with trypsin. The peptides resulting from the trypsin digest were sequenced by nano-LC-ESI mass spectrometry. The original N-terminal end(s) of the protease resistant fragment (i.e., those ends resulting from LysC cleavage) were identifiable as they had been acetylated[25]. The N-terminal ends resulting from LysC cleavage in bands LA-LI are reported in Figure 2-1B on page 35. The y and b ion series that allowed the N-terminal peptide identification are given. The MS/MS spectrum in support of the peptide identification are provided in Supplemental Figure S1A (Appendix A). The internal peptides resulting from further trypsin cleavage of the LysC resistant bands were similarly sequenced by MS/MS (Figure S1B in Appendix A).

The approximate C-terminal end of LysC protease resistant fragments LA-LI were determined by sequencing of the internal peptides and by the apparent molecular weight of the original protease resistant bands on SDS PAGE gels given the experimentally determined N-terminal end. The peptide sequencing data derived from bands LA-LI have been mapped back onto the linear domain structure of R2Bm and are summarized in Figure 2-1C on page 35.

Full length R2Bm (118 kD) was quickly processed by LysC to form a large ~89 kD LA band and shorter ~29 kD LF and ~22 kD LG bands. The LF band was found to have fragment with alternative N-terminal ends that mapped near the beginning of the R2Bm ORF, at amino acid residues four and seven—a serine (S4) and a glutamic acid (E7), respectively. Internal peptides of the LF fragment included the ZF and Myb domains and ended within -1, a conserved basic region involved in RNA binding[35]. The fragment from band LG was similar to the LF fragment except that fragment LG was ~60 amino acids shorter. Fragment LG had an N-terminal end that mapped to amino acid R64 of the R2Bm ORF, removing most of the ZF from the fragment. The C-terminal end of the LG fragment appeared to be similar to LF.

The polypeptide that constituted the large ~89 kD LA band had two alternative N-terminal ends, R242 and C256. The LA fragment spanned from -1 to the end of the ORF. The LA fragment contained the entire RT, the endonuclease, and the linker region connecting the RT and endonuclease domains.

Another large prominent band appeared along with band LA at the earlier time points, band LC. The fragment from band LC consisted of part of the RT, starting within RT6, at amino acids S595 and H609, and ending at the end of the R2Bm ORF. Like the LA fragment, the LC fragment contained the endonuclease domain in addition to the RT.

Band LB was present at low amounts across the time series in Figure 2-1 A (page 35). At different protease ratios, however, band LB was only prominent at earlier time points (data not shown). Fragment LB had about 30 more amino acids of the RT than did fragment LC. Fragment LB is likely processed into fragment LC.

At later time points, fragments LA, LB, and LC were further processed. Band LD consisted of two non-overlapping fragments, LD(i) and LD(ii), of about the same size. In the 18% gel the LD fragments ran as a single band while on the gradient gel a doublet was observed (Figure 2-1A, page 35). The first LD fragment, LD(i), consisted of the bulk of the RT, from V385 which was located in RT1, through most of the thumb. The second LD fragment, LD(ii) started near the end of the thumb at amino acid A719 and continued through the end of the ORF. Fragment LD(ii) contained the endonuclease and the linker region that connects the endonuclease to the RT.

Fragment LC gets cleaved at K763 to generate bands LE and LI. Band LI consisted of the N-terminal portion of fragment LC with an N-terminal end of H609. The fragment in band LE had a N-terminal end of S764 and contained the linker and RLE. Band LE was a major late appearing band that accumulated over time. Fragments LC and LD(ii) are likely both processed into fragment LE.

Band LH consisted of a fragment with an N-terminal end located at the beginning of the Myb domain at amino acid R82. The polypeptide appeared to be derived from fragments LF and/ or LG but was further truncated at the N-terminal end.

As fragments from the RT and the ZF/Myb regions of the ORF were processed into smaller polypeptides, those polypeptides became difficult to resolve and visualise on SDS-PAGE, especially on preparative gels. Depending upon the gel percentage and band location, an excised gel slice can still contain signal from bands just above or below that area. In the later time points the background between bands increases due to non-banding polypeptides. We did not trust our ability to identify bands and N-terminal ends below about 18 kD.

2.4.2 Mapping and sequencing of GluC protease resistant fragments of R2Bm protein

The second protease used to probe globular domain structure of R2Bm was GluC. GluC cleaves on the C-terminal side of glutamic acid residues, and to a lesser extent (100-fold) aspartic acid residues. There are 69 glutamic acid residues and 47 aspartic acid residues in the R2Bm protein. Aliquots were pulled from the digestion reaction at different time points and terminated. The digestion profile of R2Bm protein cleaved by GluC at the different time points was analyzed by SDS-PAGE (Figure 2-2A).

The protease resistant bands visualised on the SDS-PAGE were labeled GA-GK. The A-K designators, however, do not necessarily equate to an equivalent LysC resistant R2 fragment as the designators are by order of apparent-molecular-weight and not by R2 ORF region. The protease resistant fragments were excised from the gel, processed and sequenced by nano-LC-ESI mass spectrometry. The y and b ion series that allowed the N-terminal peptide identification are given for each band (Figure 2-2B).

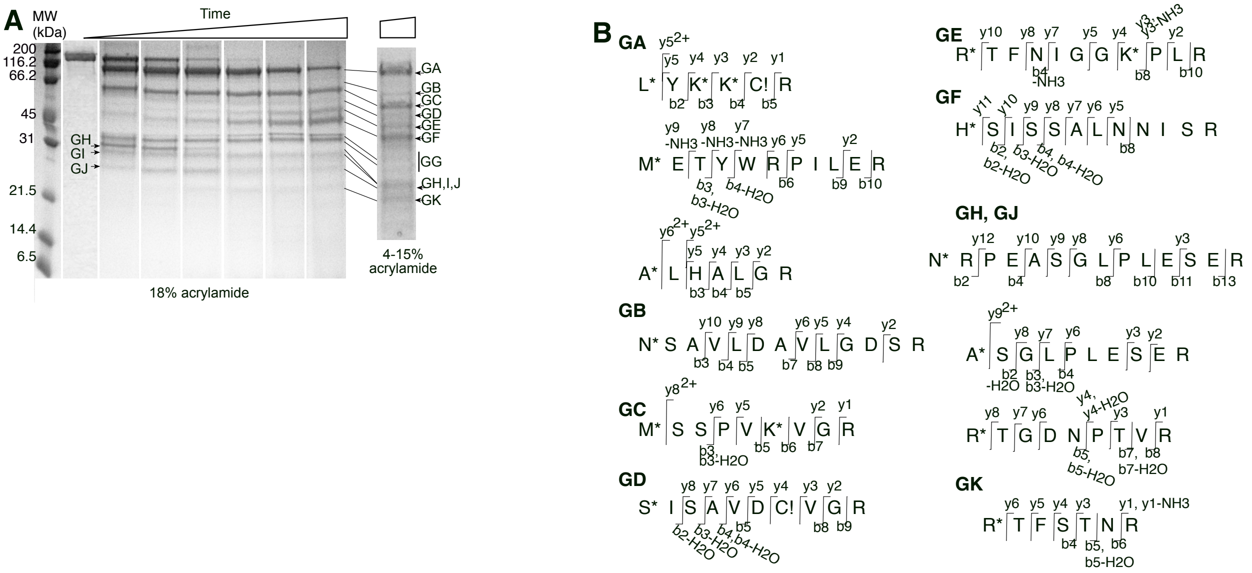


Figure 2-2. Limited proteolysis of R2Bm protein with GluC protease. A) R2Bm protein was digested with GluC protease and analyzed by SDS-PAGE. Major observed bands were designated GA-GK. B) Identification of proteolytic fragments of R2Bm protein. Bands from panel A were cut out, further processed, and analyzed by nano-LC-ESI-MS/MS sequencing. The N-terminal of the band producing fragment was identified by acetylation. The y and b ions that identified the N-terminal end are indicated. Symbols: * = acetylation; @ = oxidation; ! = carbamidomethylation. The spectrum is given in the supplemental data.

GluC, like LysC, quickly cleaved the R2Bm protein into a large fragment of about 87 kD (band GA) and a small fragment of about 30 kD (band GH). The large fragment, GA, consisted of the RT, linker, and RLE. The small fragment consisted of the N-terminal region of the R2Bm protein. The protein fragments isolated from bands GB-GF were, to a first approximation further truncations of the GA fragment where the truncating cleavages were located within the RT. The most prominent of these fragments and bands were GC, GE, and GF. Bands GE and GF appeared late in the time course. The fragments isolated from bands GJ and GK were, to a first approximation, further truncations of band GH. As band GH disappeared, band GJ became more prominent. As band GJ disappeared, band GK appeared. The two bands marked GG were prominent on the 18% acrylamide gel because of a band compression artifact. The GG area contained faint bands and a diffuse smear on the gradient gel. The lower of the two bands appeared to be GluC while the upper band could not be ascertained. Band GI also could not be ascertained.

There were several alternative N-terminal ends found for fragment GA: L252, M279, A300. Fragments GH and GJ also were found to have several alternative N-terminal ends: N8, A12, R21. In order to aid in interpreting the N-terminal ends of the protease resistant R2Bm fragments, especially early and late cleavage determinations, and to attain a more comprehensive accounting of cleavages that did not give rise to readily observable bands, an experiment was performed where GluC cleavages were detected at a given time point without separating individual proteolytic fragments. Instead of fractionating the fragments, the terminated protease reaction was run into the SDS-PAGE gel only a few millimeters. A fairly large section of gel near the wells was then excised and processed for cleavage detection (Figure 2-3B, page 41). This technique of running the reaction minimally into the gel is a near equivalent to direct detection in solution (i.e., no gel fractionation). For technical reasons (see materials and methods), however, it was necessary to have the proteolysis reaction processed through a gel slice. Each column of boxes below the ORF map is a potential GluC cleavage site (D/E), or rather the amino acids immediately following a GluC cleavage site that would become acetylated if the preceding D or E residue were cleaved. Each progressive row is a (longer) time point with identified cleavages reported as a heat map of peptide spectral match (psm) values for each site for each time point. In the heatmap data, there appeared to have been several pre-existing R2Bm N-terminal ends present in the R2Bm protein preparation as N-terminal signals at positions P36, P185, and S786 were detected in the zero time point on the heatmap. No major bands on the SDS-page gels, however, were attributable to these fragments.

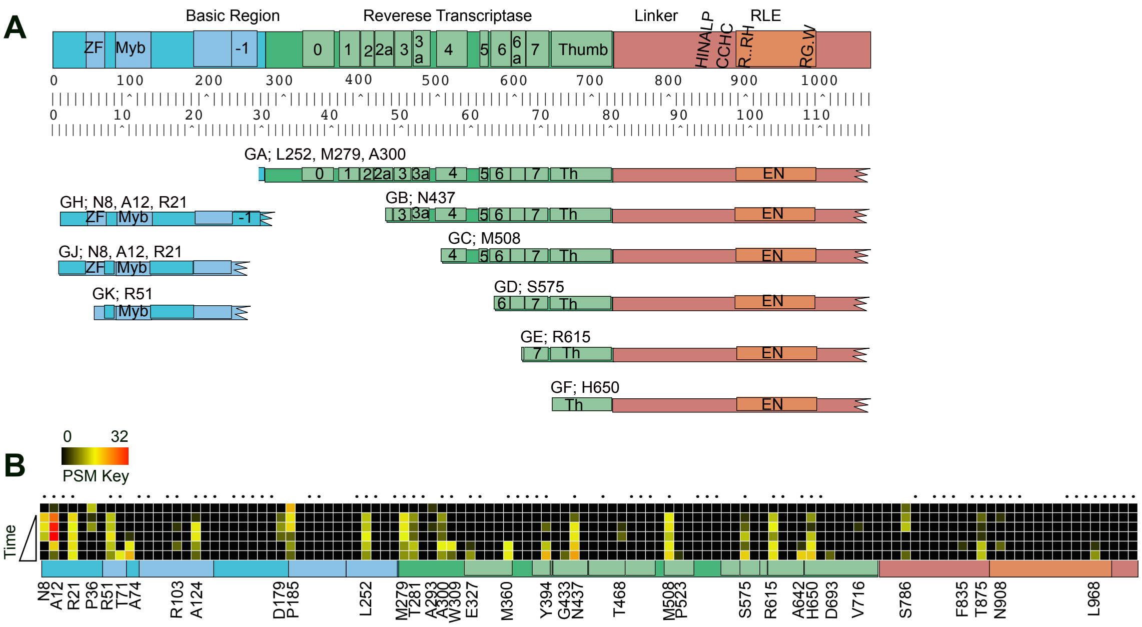


Figure 2-3. Map of GluC generated R2Bm fragments. A) Map of the band purified R2Bm fragments. The major GluC generated R2Bm fragments detected in Figure 2-2 are mapped below the ORF diagram and rulers. B) Heat map of gluC cleavages found in non-fractionated digestion re-

actions of R2Bm protein across time. Each column of boxes represents a GluC cleavage site. GluC cleaves after a E residue, indicated by a dot above the column, or a D residue (no dot). The positions of the amino terminal ends generated by observable GluC cleavages are given below the boxes. The number of peptide spectral matches (# PSM) are color coded as shown in key. The R2Bm ORF is diagramed below the heat map. Each row is a different time point with the top row being the zero time point (no GluC) and the bottom row being 8 hours. The triangle represents increasing time of GluC digestion.

Comparing the heat map results (Figure 2-3B) with the data derived from the SDS-PAGE bands (Figure 2-2, page 39) provided an extra window into the relative cleavability and timing of several important cleavage sites. It appeared that the major early cleavage events were near the start of the RT. Cleavage at E278 was the most robust cleavage event and gave rise to fragments GA and GH. The cleavage event in domain -1 (also called -1) at position E251 was also a major cleavage event. Cleavage at E251 peaked midway through the digestion reaction as band GJ become prominent. Cleavage at E251 occurred in the full protein as well as a C-terminal truncation of fragment GH. The A300 N-terminal end of band GA appeared to be the result of a later cleavage event that further truncated the original GA fragment. The A300 N-terminal end of a late time point GA fragment was confirmed by Edman degradation (data not shown).

Another major cleavage event in Figure 2-3B was an early event located at E507. Amino acid E507 is within RT4 and cleavage at this location resulted in fragment GC. Two other prominent cleavage locations in Figure 2-3B (page 41), E614 and E649, were later cleavage events and gave rise to fragments GE and GF, respectively. Amino acid E614 is located in RT6 and E649 is located at the beginning of the thumb of the RT. The N-terminal ends of fragments GC and GF were confirmed by Edman degradation (data not shown). Other cleavages were observed within the RT in Figure 2-3B, not all of which gave rise to major stable fragments visible on the SDS-PAGE gel. Interior RT fragments, i.e., those not associated with the linker and

RLE, were either heterogeneous in nature or unstable such that bands were not observed on an SDS-PAGE gel.

The N-terminal ends of the GH and GJ fragments, like GA, were ragged. The GH and GJ fragments had N-terminal ends of N8, A12, and R21. While all three positions were robust in Figure 2-3B, cleavage at E11 to generate the A12 end was the most prominent. It should be noted, however, that the original N-terminal end of the R2 protein was not tracked as the combination of proteases used in generating the peptides for MS/MS sequencing generated peptides too small to be readily detected. There is likely a time dependant shortening of the ends that we were unable to quantify accurately. Beyond the GH to GJ to GK progression, although Figure 2-3B would seem to indicate that there was progressive gnawing back of the N-terminal end in the E7-E50 region by the protease.

2.4.3 Protein threading model of the R2Bm RT and mapping of the protease cleavages onto the model:

It has been a nearly 20 years since a model of the R2 RT has been generated using homology modeling and protein threading[36]. The updated RT model shown in Figure 2-4 (page 45) was constructed using the Phyre 2.0 protein modeling server[27]. The model spanned amino acid residues Y246-P754 of the expressed ΔNR2Bm protein and spanned from the end of -1 through the thumb of the RT[18,19]. The initial residues, Y246-E263, and the final residues, R736-P754, were modeled by *ab initio* by the modeling program. Residues V264-V735 residues, however, were modeled with high homology confidence using four known protein structures as templates: 5hhl (chain A), 5g2X (chain C), 4i43 (chain B) and 1khv (chain A) (Figure 2-4A)[37–40]. The first two templates are group II intron RTs: the cryo-EM structure of *lactococcal* group II intron LtrA protein (5g2xC) and the crystal structure of the *Eubacterium rectale* group II intron RT (5hhlA). The third template is the RT found in the eukaryotic splicing factor Prp8. The fourth templates is the *caliciviral* RNA dependent RNA polymerase (d1khva). Only the high confidence regions of the R2Bm RT were kept in the final model, the *ab initio* regions were deleted from the 3D depictions presented in Figure 2-4B to 2-4E (page 45).

The region between -1 and RT0 (V264-P322) was modeled solely from the RNA dependent RNA polymerase (RdRP), but was of high confidence. The region from RT0 through RT2a (I323-R449) was built using the two group II intron RTs and the RNA dependent RNA polymerase. The RT3-RT6 area (K450-L602) was modeled using the group II intron structures, RNA dependent RNA polymerase, and Prp8. The area between RT6 and RT7 was modeled only from Prp8. RT7 was modeled by the group II intron structures as well as Prp8. The thumb was modeled using only the Prp8 crystal structure as a template.

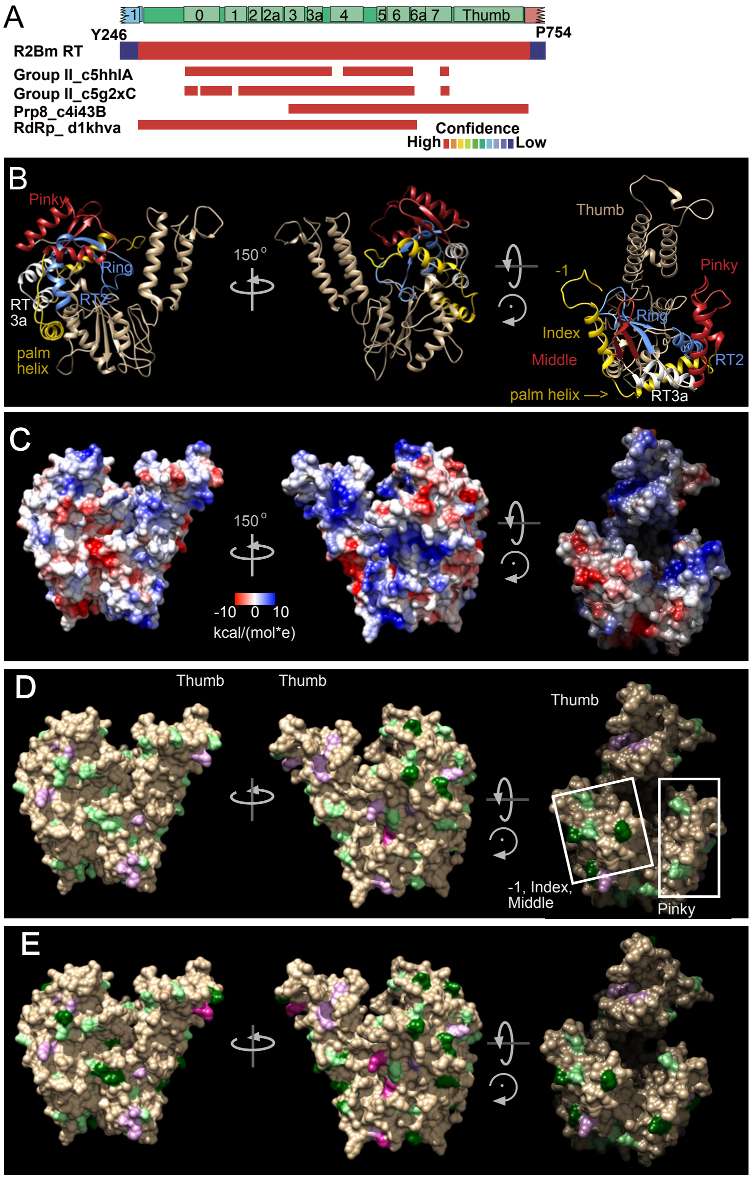


Figure 2-4. Modeling of the R2Bm RT domain and mapping of the proteolytic cleavages onto the RT model. A) R2Bm RT model construction and confidence report from Phyre2. 5hhl: crystal structure of the RT domain of the group II intron encoded protein from *Eubacterium rectale.* 5g2xC: the maturase protein in the cryo-EM structures of a spliced Lactococcus lactis group II A intron RNP. 4i43B: the splicing factor Prp8 protein large domain crystal structure. 1khv: the crystal structure of rabbit hemorrhagic disease virus RNA-dependent RNA polymerase. B) Ribbon model representation of R2Bm RT with several key regions highlighted. Pinky (RT0) is colored red as is the middle finger of RT4. The region spanning from a portion of the -1 to RT0 is in yellow. This region includes a remnant of -1 loop, the index finger α-helix, and the α-helix that

traverses palm. The ring finger (RT1) is in blue as is the RT2 α-helix. C) Coulombic surface rendering of the R2Bm model. D) Early proteolytic cleavage sites mapped onto the R2Bm RT model. Dark green coloring mark glutamic acid and aspartic acid residues that were cleaved. Pale green mark glutamic acid and aspartic acid residues that were not cleaved. Pink coloring mark lysine residues that were cleaved. Pale purple mark lysine residues that were not cleaved. E) Early plus later proteolytic cleavage sites mapped onto the the R2Bm RT model. Markings are as in panel D.

A ribbon diagram of the R2Bm RT model is presented in Figure 2-4B. The R2Bm RT assumed the canonical hand-like configuration, with fingers, palm, and thumb regions, and was very similar to RdRP[41–44]. The thumb region was very long and prominent in R2Bm. The -1, index finger, and middle finger formed one of two bulbous regions as in RdRP. The pinky finger (RT0) formed the second bulbous region. Just behind index and middle-finger was the ring finger (RT1 β-strands). The RT2 α-helix was positioned behind RT0. The region spanning from -1 to RT0 (yellow in the ribbon diagram) includes the index finger and the palm-traversing α-helix.

The index finger and RT0 are connected by the palm traversing α-helix, a feature shared between RdRP, Prp8 and, apparently, LINE polymerases (Figure 2-5A and 2-5B). Telomerases have the index finger α-helix, but lack RT0 (the pinky finger) as well as the palm-traversing α-helix (figure 2-5D). In group II intron RTs, the index finger and palm traversing helix are not present (figure 2-5C). Group II intron RTs do, however, have an RT0 and an extension to the RT0 termed NTD, both positioned on the pinky finger side (PDB ID: 5hhl and 5g2x)[37,38]. HIV-1 RTs do not contain IFD, RT0 and palm traversing alpha helix but the index finger is present (figure 2-5E).

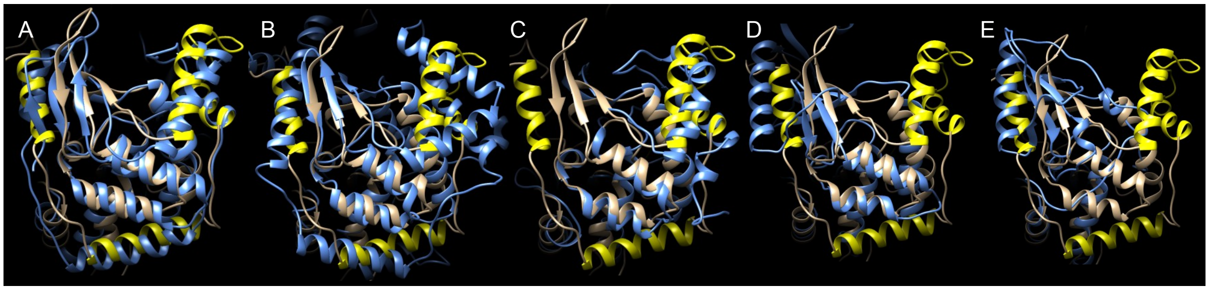


Figure 2-5. Structural overlays of R2Bm RT on A) hepatitis C virus NS5B RNA dependent RNA polymerase (PDB ID: 1c2p[45] Chain A), B) pre-mRNA splicing factor protein spp42 reverse transcriptase domain (PDB ID: 3jb9[46] Chain A), C) *Roseburia intestinalis* group II intron encoded reverse transcriptase (PDB ID: 5hhj[38] chain A), D) *Tribolium custaneum* telomerase catalytic subunit TERT (PDB ID: 3du5[47] chain A), and E) HIV-1 (PDB ID: 1rth[48] Chain A). R2Bm RT is shown in tan and all the rest of the RTs are shown in cornflower blue in the respective overlays. The index finger, palm traversing helix and RT0 are colored in yellow for R2Bm RT.

The index finger region is important for the polymerization functions. A monoclonal antibody directed against the vicinity of the index finger of the hepatitis C virus RdRP was found to inhibit both primer-dependent and *de novo* RNA synthesis[49].

The pinky finger region is also important for polymerization. The RT0 of R2Bm and group II intron RTs share a set of antiparallel α-helices connected by a loop[37,38,50,51]. In RdRP the RT0 homologue is the 'G-loop,' or "motif G'. The G-loop functions in template-RNA binding and translocation[41,52]. A monoclonal antibody directed against the G-motif was found to be inhibitory to primer-dependent RNA synthesis but not *de novo* RNA synthesis[49]. The RT0 domain of RLE LINEs contains a PGPD motif in the loop. The PGPD motif, when mutated in R2Bm, abolished template jumping activity of the RT and reduced, to some extent, overall polymerization activity[35]. Template jumping activity is also observed in RdRP, Mauriceville retroplasmid, and group II intron RTs[53–55]. Mutation of the PGPD motif in R2Bm also reduced the binding to the 5’ and 3’ PBM RNAs[35]. The group II inton protein's RT0 and its extension (the NTD) are involved in binding DIVa of the group II intron RNA[37,56,57]. The interaction between RT0 and DIVa is required for positioning the intronic-RNA-template for reverse transcription (TPRT), but is not strictly essential for splicing[56].

RLE LINEs, telomerase, and group II introns possess RNA binding domains upstream (N-terminal) of the reverse transcriptase. In the case of the group II intron protein, the N-terminal domain is an extension of RT0 and resides on that side of the RT (the pinky finger side). The extended RT0, along with IFD, bind to DIVa of the intron RNA[37,56]. In R2, the RNA binding domain -1 is on the opposite side of the fingers from RT0. The remnants of domain -1 is on the index finger side. Mutations in -1 abolished 5’ and 3’ PBM RNA binding[35]. Telomerases also contain an RNA binding region upstream of the RT that is involved in binding RNA. The RNA binding domains do not appear to be of similar origin[47,58,59].

A coulombic surface map is presented in Figure 2-4C (page 45). The R2Bm RT adopts an overall shape of a curved wedge with the backside of the thumb being the sharp edge. One of the two comparatively flat sides is the thumb-to-RT0 face. This face has a small central acidic patch surrounded by mostly hydrophobic residues in the model. The other fairly flat side is the thumb to index finger side and is predominantly basic. The third side is rounded. It spans from the index finger to RT0 and has a central vertical streak of acidic residues running through a central streak of (mostly) hydrophobic residues. The streaks are centered between below the ring finger. The hydrophobic regions and perhaps the acidic patches/streaks within them, are potential areas of further protein-protein interactions.

The R2Bm RT model was used for mapping the earlier cleavages (Figure 2-4D) as well as all of the cleavages (early plus later, Figure 2-4E on page 45) for both LysC and GluC proteases. LysC cleaves on the C-terminal side of K residues. There are 18 K residues in the R2Bm RT model, 6 of which are cleaved to some degree. Cleavage in the *ab initio* regions are included in the cleavage count, although the *ab initio* sequences have been deleted from the 3D models in the figure. GluC cleaves on the C-terminal of E residues, and much less often on the C-terminal side of D residues. There are 30 E residues in the R2Bm RT, 14 of which are cleaved to some degree. There are 26 D residues in the RT, six of which are weakly cleaved. Most of the early cleavages mapped to the -1 *ab initio* regions (not shown), the index finger and the tip of the middle finger. There was also a cleavage on the basic face between the thumb and -1. Some of the next cleavages were also on the basic face as well as on the RT0 protrusion and on the knife edge (the backside) of the thumb. Most of the prominent thumb was protected from cleavage. Later cleavages were found on the secondary structures just behind where the first cleavage were (e.g. the regions behind the index finger α-helix) and on the flat hydrophobic thumb-to-RT0 face inside the acidic patch.

2.4.4 The large fragment of the eukaryotic splicing factor Prp8 and restriction-like endonuclease bearing LINEs share a common set of sequence motifs and structure

RTs share a common set of sequence domains numbered 1-7 and a thumb region[50,51,60–62]. The thumb usually contains a three helix bundle. In addition to the thumb and RT1 through RT7, the RT of LINEs contain insertions: 0, 2a, 3a, and 6a. Several of these insertions are present in other eukaryotic RTs (Supplemental Figure S5 in Appendix C and[50,51,60–62]). The RT domain of

Prp8 is very similar to that of LINEs, having 0, 2a, 3a, 4a, and 6a insertions. The telomerase RT encodes 2a, and 3a. The RT of group II intron proteins encode 0, 2a, 3a, 4a, and 7a.

The area between the reverse transcriptase and the RLE in RLE LINEs is the linker region. The linker in RLE LINEs was predicted to be predominantly α-helical with six major helices with some groups having 2-3 additional helices (Supplemental Figure S5 in Appendix C). A weak scoring helix was also often observed in the highly-conserved (presumptive) gag-knuckle (see below). The linker of APE LINEs were more diverse with 5 -14 predicted helical regions. The crystal structure and EM-structures of Prp8 have about 13 helices. β-strands were less prevalent in the linker of RLE LINEs than APE LINEs (typically about 0-2 vs 4-6). Among the RLE LINEs, only Utopia may contain comparatively high number of linker β-strands.

A multiple alignment and Ali2D secondary structure prediction is presented for the most conserved portion of the linker for LINEs and Prp8 (Figure 2-6A). Near the end of the linker region of LINEs is the highly conserved cysteine-histidine rich motif (CCHC) with a typical spacing of $CX_{2-3}CX_{7-8}HX_4C$ (Figure 2-6A). In many of the RLE LINE clades (e.g., R2, Dong, NeSL, and Utopia) there was a conserved R residue between the first two C residues. APE LINEs, although lacking a downstream RLE, have a linker region that also ends with the CCHC motif. The spacing of the cysteines and histidine in the LINE CCHC motif is similar to that of IAP domains, although a bit smaller, or gag-knuckles, although a bit larger[63]. IAP domains form a ββα structure around zinc ion[63]. A gag-knuckle is β-strand followed by a knuckle (a sharp turn) with a less structured finish (e.g., coil with bends)[63]. The zinc ion is coordinated by the C and H residues of the motif[63]. The β-strands and α-helix are generally short. The canonical structure for an IAP domain is indicated above the R2Bm sequence listed in Figure 2-6A as is the predicted

(Ali2D) secondary structure for the linker region of RLE LINEs. For many RLE LINEs a short α-helix was predicted near the H residue (Supplementary Figure S5 in Appendix C). For APE LINEs, the area near the H was often predicted to be a β-strand. A β-strand was occasionally predicted near the first C residue using JPRED for RLE LINEs (data not shown).

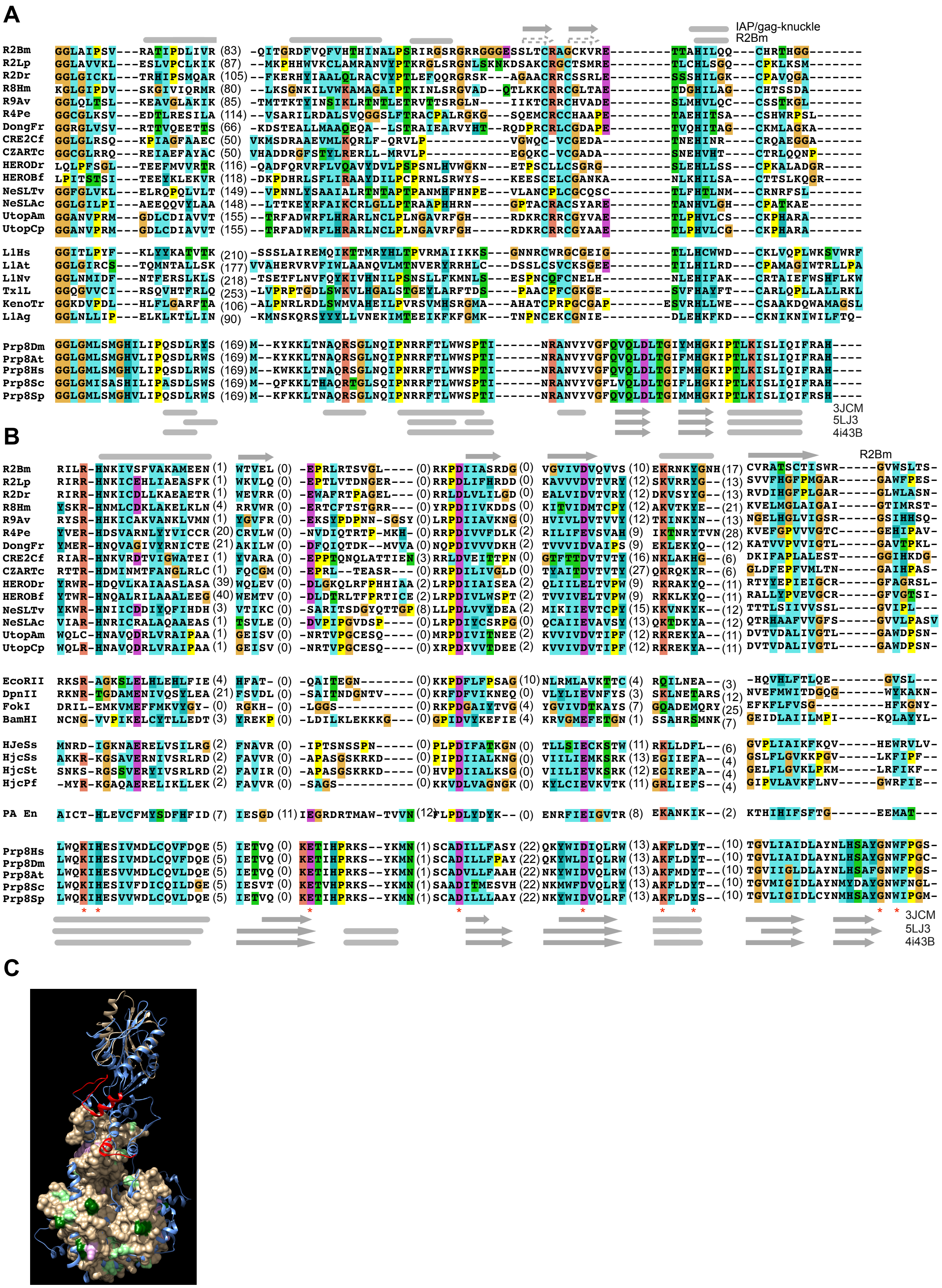

A
IAP/gag-knuckle
R2Bm
R2Bm GGLAIPSV---RATIPDLIVR (83) -QITGRDFVQFVHTHINALPSRIRGSRGRRGGGESSLTCRAGCKVRE--------TTAHILQQ---CHRTHGG-------
R2Lp GGLAVVKL---ESLVPCLKIK (87) --MKPHHWVKCLAMRANVYPTKRGLSRGNLSKNKDSAKCRGCTSMRE--------TLCHLSGQ---CPKLKSM-------
R2Dr GGLGICKL---TRHIPSMQAR (105) --FKERHYIAALQLRACVYPTLEFQQRGRSK---AGAACRRCSSRLE--------SSSHILGK---CPAVQGA-------
R8Hm KGLGIPDV---SKGIVIQRMR (80) --LKSGNKILVWKAMAGAIPTKINLSRGVAD--QTLKKCRRCRCGLTAE--------TDGHILAG---CHTSSDA-------
R9Av GGLQLTSL---KEAVGLAKIK (85) -TMTTKTYINSIKLRTNTLETRVTTSRGLN----IIKTCRRCHVADE--------SLMHVLQC---CSSTKGL-------
R4Pe GGCGLKSV---EDTLRESILA (114) --VSARILRDALSVQGGSLFTRACPALRGKG---SQEMCRCCHAAPE--------TAEHITSA---CSHWRPSL------
DongFr GGRGLVSV---RTTVQEETTS (66) -KDSTEALLMAAQEQA--LSTRAIEARVYHT--RQDPRCRLCGDAPE--------TVQHITAG---CKMLAGKA------
CRE2Cf GGLGLRSQ---KPIAGFAAEC (50) VKMSDRAAEVMLKQRLF-QRVLP-----------VGWQC-VCGEDA--------TNEHINR----CRQAEGGA------
CZARTc GGCGLRRQ---REIAEFAYAC (50) VHADDRGFSTYLRERLL-MRVLP-----------EGQKC-VCGADA--------SNEHVHT----CTRLQQNP------
HERODr LQLPFSGL---TEEFMVVRTR (116) -QADFQRVRFLVQAVYDVLPSPSNLHVWGKN---ETPSCLLCSGRG---------SLEHLLSS---CPKALADGR-----
HEROBf LPITSTSI---TEEYKLEKVR (118) -DKPDHRLSFLIRAAYDILPCPRNLSRWYS----KEESCELCGANKA--------NLKHILSA---CTTSLKQGR-----
NeSLTv GGFGLVKL---ELRQPQLVLT (149) --VPNNLYSAAIALRTNTAPTPANMHFHNPE---VLANCPLCGCQSC--------TLFHTLNM---CRNRFSL-------
NeSLAc GGLGILPI---AEEQQVYLAA (148) --LTTKEYRFAIKCRLGLLPTRAAPHHRN-----GPTACRACSYARE--------TANHVLGH---CPATKAE-------
UtopAm GGANVPRM---GDLCDIAVVT (155) --TRFADWRFLHRARLNCLPLNGAVRFGH-----RDKRCRRCGYVAE--------TLPHVLCS---CKPHARA-------
UtopCp GGANVPRM---GDLCDIAVVT (155) --TRFADWRFLHRARLNCLPLNGAVRFGH-----RDKRCRRCGYAAE--------TLPHVLCG---CKPHARA-------
L1Hs GGITLPYF---KLYYKATVTK (210) -SSSLAIREMQIKTTMRYHLTPVRMAIIKKS---GNNRCWRGCGEIG--------TLLHCWWD---CKLVQPLWKSVWRF
L1At GGLGIRCS---TQMNTALLSK (177) VVAHERVRVFIWLAANQVLMTNVERYRRHLC---DSSLCSVCKSGEE--------TILHILRD---CPAMAGIWTRLLPA
L1Nv GGLNMIDF---NTFERSLKLS (218) -TSETFLNVFQYKIVHNILPSNSLLFKMNLS---ESPNCQFCNELH---------NLEHIFAK---CRTAIEFWSHFLKW
Tx1L GGQGVVCI---RSQVHTFRLQ (253) -LVPRPTGDLSWKVLHGALSTGEYLARFTDS----PAACPFCGKGE---------SVFHAYFT---CARLQPLLALLRKL
KenoTr GGKDVPDL---HLFLGARFTA (106) -ALPNRLRDLSWMVAHEILPVRSVMHSRGMA---AHATCPRPGCGAP--------ESVRHLLWE---CSAAKDQWAMAGSL
L1Ag GGLNLLIP---ELKLKTLLIN (90) -KMNSKQRSYYYLLVNEKIMTEEIKFKFGMK---TNPNCEKCGNIE---------DLEHKFKD---CKNIKNIWILFTQ-
Prp8Dm GGLGMLSMGHILIPQSDLRYS (169) M--KYKKLTNAQRSGLNQIPNRRFTLWWSPTI------NRANVYVGFQVQLDLTGIYMHGKIPTLKISLIQIFRAH----
Prp8At GGLGMLSMGHVLIPQSDLRWS (169) M--KYKKLTNAQRSGLNQIPNRRFTLWWSPTI------NRANVYVGFQVQLDLTGIFMHGKIPTLKISLIQIFRAH----
Prp8Hs GGLGMLSMGHVLIPQSDLRWS (169) M--KWKKLTNAQRSGLNQIPNRRFTLWWSPTI------NRANVYVGFQVQLDLTGIFMHGKIPTLKISLIQIFRAH----
Prp8Sc GGLGMISASHILIPASDLSWS (169) M--QFKKLTHAQRTGLNQIPNRRFTLWWSPTI------NRANVYVGFLVQLDLTGIFLHGKIPTLKISLIQIFRAH----
Prp8Sp GGLGMLSMGHVLIPQSDLRWS (169) M--KFKKLTNAQRSGLNQIPNRRFTLWWSPTI------NRANVYVGFQVQLDLTGIMMHGKIPTLKISLIQIFRSH----
3JCM
5LJ3
4i43B
B
R2Bm
R2Bm RILR-HNKIVSFVAKAMEEN (1) WTVEL (0) -EPRLRTSVGL---- (0) RKPDIIASRDG (0) VGVIVIDVQVVS (10) EKRNKYGNH (17) CVRATSCTISWR-----GVWSLTS-
R2Lp RIRR-HNKICEHLIAEASFK (1) WKVLQ (0) -EPTLVTDNGER--- (0) RRPDILIFHRDD (0) KAVVVDVTVRY (12) SKVRRYG--(13) SVVFHGFPMGAR-----GAWFPES-
R2Dr RIRR-HNKICDLLKAEAETR (1) WEVRR (0) -EWAFRTPAGEL--- (0) RRLDLVLILGD (0) EALVIDVTVRY (12) DKVSYYG--(13) RVDIHGFPLGAR-----GLWLASN-
R8Hm YSKR-HNMLCDKLAKELKLN (4) RRVWR (0) -ERTCFTSTGRR--- (0) YRPDIIVKDDS (0) KITVIDMTCPY (12) AKVTKYE--(21) KVELMGLAIGAI-----GTIMRST-
R9Av RYSR-HHKICAKVANKLVMN (1) YGVFR (0) -EKSYPDPNN-SGSY (0) LRPDIIAVKNG (0) HVIVLDVTVVY (12) TKINKYN--(13) NGELHGLVIGSR-----GSIHHSQ-
R4Pe YVER-HDSVARNLYYVICCR (20) CRVLW (0) -NVDIQTRTP-MKH- (0) RRPDIVVFDLK (2) RILIFEVSVAH (9) IKTNRYTVN(28) KVEFGPVVVGTC-----GEHIPAV-
DongFr YMER-HNQVAGIVYRNICTE (21) AKILW (0) -DFQIQTDK--MVVA (0) NQPDIVVVDKH (2) TVVVIDVAIPS (9) EKLEKYQ--(12) KATVVPVVIGTL-----GAVTPKL-
CRE2Cf RIAR-HNKVRDTVIGWATEI (1) YVARA (0) -EPPTQNQLATTIEN (3) RRLDVEITTPN (0) GTFTTDVTVTY (16) NKLAKHG--(6) DKIFAPLALEST-----GGIHKDG-
CZARTc RTTR-HDMINMTFANGLRLC (1) FQCGM (0) -EPRL--TEASR--- (0) RRPDILIVGLD (0) TYAITDVTVTY (27) QKRQKYR--(6) GLDFEPFVMLTN-----GAIHPAS-
HERODr YRWR-HDQVLKAIAASLASA (39) WQLEV (0) -DLGKQLRFPHHIAA (2) LRPDIIAISEA (2) QLIILELTVPW (9) RKRAKYQ--(11) RTYYEPIEIGCR-----GFAGRSL-
HEROBf YTWR-HNQALRILAAALEEG (40) WEMTV (0) -DLDTRLTFPRTICE (2) LRPDIVLWSPT (2) TVVIVELTVPW (9) RKKLKYQ--(11) RALLYPVEVGCR-----GFVGTSI-
NeSLTv YKWR-HNIICDDIYQFIHDH (3) VTIKC (0) -SARITSDGYQTTGP (8) LLPDLVVYDEA (2) MIKIIEVTCPY (15) KKVNKYK--(12) TTTLSIIVVSSL-----GVIPL---
NeSLAc VIAR-HNRICRALAQAAEAS (1) TSVLE (0) -DVPIPGVDSP---- (0) LRPDIYCSRPG (0) QCAIIEVAVSY (13) QKTDKYA--(12) QTRHAAFVVGFS-----GVVLPASV
UtopAm WQLC-HNAVQDRLVRAIPAA (1) GEISV (0) -NRTVPGCESQ---- (0) MRPDIVITNEE (2) KVVIVDVTIPF (12) RKREKYA--(11) DVTVDALIVGTL-----GAWDPSN-
UtopCp WQLR-HNAVQDRLVRAIPAA (1) GEISV (0) -NRTVPGCESQ---- (0) RRPDXVVTDEE (2) KVVIVDVTIPF (12) RKREKYA--(11) DVTVDALIVGTL-----GAWDPSN-
EcoRII RKSR-AGKSLELHLEHLFIE (4) HFAT- (0) -QAITEGN------- (0) KKPDFLFPSAG (10) NLRMLAVKTTC (4) RQILNEA--(3) -HQVHLFTLQE-------GVSL---
DpnII RKNR-TGDAMENIVQSYLEA (21) FSVDL (0) -SAITNDGNTV---- (0) KRFDFVIKNEQ (0) VLYLIEVNFYS (3) SKLNETARS(12) NVEFMWITDGQG------WYKAKN-
FokI DRIL-EMKVMEFFMKVYGY- (0) RGKH- (0) -LGGS---------- (0) RKPDGAIYTVG (4) YGVIVDTKAYS (7) GQADEMQRY(25) EFKFLFVSG--------HFKGNY-
BamHI NCNG-VVPIKELCYTLLEDT (3) YREKP (0) -LDILKLEKKKG--- (0) GPIDVYKEFIE (4) KRVGMEFETGN (1) SSAHRSMNK(7) GEIDLAIILMPI-----KQLAYYL-
HJeSs MNRD-IGKNAERELVSILRG (2) FNAVR (0) -IPTSNSSPN----- (0) PLPDIFATKGN (0) TLLSIECKSTW (11) RKLLDFL--(6) GVPLIAIKFKQV-----HEWRVLV-
HjcSs AKKR-KGSAVERNIVSRLRD (2) FAVVR (0) -APASGSKRKD---- (0) PIPDIIALKNG (0) VIILIEMKSRK (13) EGIIEFA--(4) GSLFLGVKKPGV-----LKFIPF--
HjcSt SNKS-RGSSVERYIVSRLRD (2) FAVIR (0) -APASGSKRKD---- (0) HVPDIIALKSG (0) VIILIEVKSRK (12) EGIREFA--(4) GELFLGVKLPKM-----LRFIKF--
HjcPf -MYR-KGAQAERELIKLLEK (2) FAVVR (0) -SAGS---------- (0) KKVDLVAGNGK (0) KYLCIEVKVTK (11) GRLIEFS--(4) GIPVLAVKFLNV-----GWRFIE--
PA En AICT-HLEVCFMYSDFHFID (7) IESGD (11) IEGRDRTMAW-TVVN (12) FLPDLYDYK-- (0) ENRFIEIGVTR (8) EKANKIK--(2) KTHIHIFSFTG-------EEMAT---
Prp8Hs LWQKIHESIVMDLCQVFDQE (5) IETVQ (0) KETIHPRKS--YKMN (1) SCADILLFASY (22) QKYWIDIQLRW (13) AKFLDYT--(10) TGVLIAIDLAYNLHSAYGNWFPGS-
Prp8Dm LWQKIHESIVMDLCQVFDQE (5) IETVQ (0) KETIHPRKS--YKMN (1) SCADILLFPAY (22) QKYWLDIQLRW (13) AKFLDYT--(10) TGVLIAIDLAYNLHSAYGNWFPGC-
Prp8At LWQKIHESVVMDLCQVLDQE (5) IETVQ (0) KETIHPRKS--YKMN (1) SCADVLLFAAH (22) NKYWIDVQLRW (13) AKFMDYT--(10) TGVIIGLDLAYNLHSAFGNWFPGS-
Prp8Sc LWQKIHESIVFDICQILDGE (5) IESVT (0) KETVHPRKS--YKMN (1) SAADITMESVH (22) NKMWFDVQLRY (13) AKFLDYT--(10) TGVMIGIDLAYNMYDAYGNWFNGL-
Prp8Sp LWQKIHESVVWDLCQVLDQE (5) IETVQ (0) KETIHPRKS--YKMN (1) SCADILLLAAY (22) NKYWIDVQLRF (13) AKFLDYS--(10) TGVLIGIDLCYNMHSAYGNWIPGM-
* * * * * * * * *
3JCM
5LJ3
4i43B
C

Figure 2-6. Structure based alignments of linker and RLE regions of LINEs and Prp8. A) Multiple sequence alignment and Ali2D prediction of the most conserved region of the linker from LINEs and Prp8. The Ali2D predicted α helices and β-strands for R2Bm protein are marked above the R2Bm amino acid sequence. The canonical ββα structure of a gag-knuckle/IAP is also presented above the R2Bm structure. The secondary structure of Prp8 is given below the Prp8 sequences. The Prp8 secondary structures are from several reported crystal and cryo-EM structures: 3JCM (cryo-EM of U4/U6.U5 tri-snRNP), 5LJ3 (cryo-EM of core of spliceosome immediately after branching), 4i43B (crystal structure of Prp8 large domain). Rounded bars are α-helices and arrows are β-strands. B) Multiple sequence alignment of the core αβββαβ fold from RLE-LINEs, type II restriction enzymes, archeal Holliday junction resolvases, influenza PA endonuclease and Prp8. C) Overlay of the R2Bm RT and RLE onto the large fragment of Prp8. R2Bm RT contains the early and late cleavage information (from Figure 2-4E). The R2Bm RLE is the tan ribbon. The Prp8 reverse transcriptase and endonuclease is the blue ribbon. Prp8 ββα and dynamic helix and loop area are in red. Abbreviations: R2Bm = *Bombyx mori* R2 (M16558.1); R2Lp = *Limulus polyphemus* R2 (AF015814.1); R2Dr = *Danio rario* R2 (34392533); R8Hm = *Hydra magnipapillata* R8 (R8Hm-A from Repbase); R9Av = *Adenata vaga* R9 (ACV95454.1); R4Pe = *Parascaris equorum* R4 (AAB02297.1); DongFr = *Fugu rubripes* Dong; CRE2Cf = *Crithidia fasciculata* CRE2; CZARTc = *Trypanosoma cruzi* CZAR; HERODr = *Danio rario* HERO; HEROBf = *Branchiostoma floridae* HERO; NeSLTv = *Trichomonas vaginalis* NeSL; NeSLAc = *Acanthamoeba castellanii* NeSL; UtopAm = *Alligator mississippiensis* Utopia; UtopCp = *Crocodylus porosus* Utopia; Prp8Hs = *Homo sapiens* Prp8 (NP_006436.3); Prp8Dm = *Drosophila melangaster* Prp8 (NP_610735.1); Prp8At = *Arabidopsis thaliana* Prp8 (Q9T0I6); Prp8Sc = *Saccharomyces cerevisiae* Prp8 (P33334); Prp8Sp = *Schizosaccharomyces pombe* (O14187); PA En = Influenza virus PA endonclease (3HW4).

DongFr (Dong_FR), CRE2Cf (CRE2), CZARTc (CZAR), HERODr (HERO-2_DR), HEROBf (HERO-1_BF), NeSLTv (NeSL-1_TV), NeSLAc (NeSL-1_Aca), UtopAm (Utopia-1_Ami) and UtopCp (Utopia-1_Crp) sequences were collected from Repbase[64]. Holiday junction resolvases Ssol Hje (1ob8) and Ssol Hjc (1hh1) are from Sulfolobus solfataricus. Holiday junction resolvase Stok Hjc (2eo0) is from Sulfolobus tokodaii str. 7 Pfur Hjc (1gef) is from Pyrococcus furiosus.

It is not clear if Prp8 had a CCHC motif at one time that was subsequently mutated or not (Figure 2-6A). There are several possible ways to align the Prp8 sequence/structure with the LINE CCHC region. In Figure 2-6A, the $CX_{2\text{-}3}C$ of the LINE CCHC motif has been aligned with NRAN in Prp8. In this configuration, there is a conserved R between the N residues of Prp8 and a Y after the second N that may match the conserved R found in between the first two C's in a number of the RLE LINE CCHC motifs and the nearby Y in Utopia. In Prp8 there are two conserved H residues which one could potentially line up with the H of the LINE CCHC motif, one H residue near the NRAN and an H residue just prior to the endonuclease. In Figure 2-6A we aligned the H nearest to the NRAN to the H of the LINE CCHC motif. An alternative and perhaps better method of aligning Prp8 and LINEs in this area is by structure. The secondary structure of the region around the first Prp8 H residue mentioned above is a ββα structure, with the H being in the second β-strand.

The two predicted α-helices upstream of the CCHC in LINEs tended to be separated by LP. In R2 elements the sequence at the end of the first helix was highly conserved, being KXRINALP or similar. In R2Bm the sequence was HTHINALP (see also reference[36]). The two helices prior to the CCHC motif appear to be present in both RLE and APE LINEs. In the APE LINE L1Hs, the sequence upstream of the gag knuckle that PROMALS3D aligned to the R2 KXRINALP was TMRYHLTP of HMKKCSSSLIAREMQIKTTMRYHLTP. Conversion of HMKK to AAAA and SSS to AAA reduced retrotransposition activity[5]. A recombinant C-terminal 180 amino acid containing peptide (from SSS to the end of the ORF) bound RNA nonspecifically, but a mutation of the CCHC motif within the peptide did not affect RNA binding[65]. In the full-length protein, however, mutations of the conserved cysteines of the CCHC motif affected RNP formation and knocked out retrotransposition activity in cell culture assays[5,66]. In Prp8 the non-gag-knuckle ββα was at the RNA interface and was predicted to make contact with mRNA in the U4/U6.U5 tri-snRNP complex[67].

Prp8 also appeared to have similarly positioned helices important for binding RNA[67,68]. The KXRINALP helix equivalent in Prp8 was QFKKLTHAQRTGLS. The

QFKKLTHAQRTGLS region was found to be dynamic in Prp8. In the U4/U6.U5 tri snRNP (cryo-EM structure 3JCM) the area forms a loop (QFKK) plus an α-helix (HAQRTG). The loop residues contact RNA and the Dib1 protein and was involved in branch point selection[67,68]. After branching (cryo-EM structure 5LJ3), the area is not helical[69]. In the crystal structure (c4i43B), which lacks RNA, this area is unresolved and is thus likely unstructured[39,67,68,70].

RLE LINEs, like R2Bm, encode a restriction-like DNA endonuclease downstream of the CCHC motif (FIgure 2-6B). The DNA endonuclease found in RLE LINEs was found to have fairly canonical αβββαβ restriction endonuclease-like fold, although it had a unique variant of the PD-(D/E)XK catalytic core[18]. The catalytic K, which is usually near the D/E residue in the third β-sheet, was found to be located much farther away in LINE RLE. The catalytic K in the LINE RLE is the first K in the KX_2KY motif. The second K is less conserved across R2 elements and across RLE LINEs. The motif is located in the second α-helix[18]. The Y of the KX_2KY, when mutated, also reduces cleavage[18]. The catalytic K in Prp8 is located in an identical position as the RLE of LINEs. The Y residue is also present in Prp8 and is identically positioned relative to the catalytic K. The second K of the LINE KX_2KY is not present. The similarities between the Prp8 RLE and the LINE RLE go beyond the endonuclease fold and the positioning of the catalytic residues. At the far end of the endonuclease fold, just beyond the fourth β-strand, is a mutually conserved GXW motif. At the other end of the RLE fold–at the beginning of the first α-helix–is a conserved H residue and a conserved K residue. In R2Bm the equivalent is RH. Mutating the RH residues in R2Bm severely reduces DNA binding and DNA cleavage[18]. At the end of the first β-strand of both Prp8 and LINE RLEs is a conserved D/E that also appears to be unique to these two groups. Except for a 22 amino acid insertion between β-sheets 2 and 3 of Prp8, both Prp8 and LINE endonucleases are about the same size. Prp8 endonuclease appears to have the amino acid residues needed for the cleavage activity, but the residues do not appear to be involved in metal coordination in the crystal structure, rather these residues stabilize the polypeptide loop blocking the active site[39].

The model of the R2Bm RT, R2Bm RLE, and the large fragment of Prp8 are presented as a structural overlay in Figure 2-6C. The model of the R2Bm RLE is presented as a ribbon diagram as is the large fragment of Prp8. The ββα structure and the two α-helices upstream of the ββα structure are highlighted. The Prp8 equivalent of the R2Bm HINALP (QFKKLTHAQRTGLS) is positioned near the top of the thumb and is oriented toward the fingers of the RT. The R2Bm RT is presented as a surface model with the potential and actual protease cleavage sites highlighted as per Figure 2-4E. The protection from protease cleavage of the R2 thumb is, in part, explained by the positioning of the linker, assuming the linker surrounds much of the top half of the RT thumb as it does in Prp8 (Figure 2-6C).

2.5 Conclusions

The R2Bm protein was found to be comprised of two major globular domains: the ZF/Myb/-1 N-terminal domain and the RT/linker/RLE superdomain. The index finger of the RT and the -1 region were the most accessible areas for protease cleavages to occur, indicating that this region might represent a flexible conformational-switch area that may help coordinate the nucleic acid binding and cleavage activities of the two globular domains. The ZF/Myb/-1 region is present among all of the early branching LINEs with a variable number of ZFs and Myb motifs[71]. The primary variability in the RT/linker/RLE superdomain was in the linker, with Cre elements often having a deletion relative to R2 and Utopia having an insertion. The linker was predicted to be largely α-helical and across all RLE LINEs and, we hypothesize, closely associated with both the RT thumb and the RLE, similar to Prp8. The linker of R2 contained several highly conserved sequence motifs and secondary structures. Most notably the presumptive-ββα-forming gag knuckle-like CCHC motif. Just upstream of the CCHC motif in R2 elements are several predicted α-helices separated by a highly conserved KXRINALP motif. The two helices prior to the CCHC motif appear to be present in both RLE and APE LINEs. Mutations in the CCHC and preceding helical region affect retrotransposition[5,66].

The linker of Prp8 was also found to have an important ββα structure and an upstream dynamic helix region important for interacting with nucleic acids[67–70]. In Prp8 the dynamic helical region sits on top of the RT thumb and is oriented toward the fingers of the RT. It would appear that the region immediately upstream of the ββα/CCHC in RLE LINE, APE LINEs, and Prp8 might be structurally conserved and to a degree, functionally conserved. If the helices preceding the ββα/CCHC in LINEs were positioned as it is in Prp8, it is easy to envision the region participating in binding to element RNA or target DNA.

The gag knuckle-like motif and associated upstream helices might promote switching between polymerase active and endonuclease active conformations of the R2 protein in response to binding insertion reaction intermediates. Our proteolysis study was done in the absence of nucleic acids. In the absence of RNA, the R2 protein would be expected to adopt a conformation that would have characteristics of the conformation involved in second-strand cleavage. It is possible that in the presence of RNA or DNA our result would differ from those presented as the nucleic acid might block some sites from being cleaved while presenting other newly accessible sites due to protein conformational changes induced by nucleic acid binding.

The RT of Prp8 has been noted to share similarities with RdRP and to the RTs encoded by mobile group II introns and LINEs[38,50]. However, because Ppr8 and the group II intron protein both function as splicing maturases, the similarity to group II introns has been stressed. While it has been noted in the literature that the large fragment of Prp8 contains a RLE, the connection to LINEs had not been presented beyond noting that LINEs also contain a RT and a RLE. In this paper we shown through modeling, bioinformatics, and biochemistry that the LINE RT, linker, and RLE share more points of commonality to the large fragment of Prp8 than does the group II intron maturase.

References

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

2. Christensen, S. M., Ye, J. & Eickbush, T. H. RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* **103**, 17602-17607 (2006).

3. Christensen, S. M. & Eickbush, T. H. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**, 6617-6628 (2005).

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

5. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).

6. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).

7. Eickbush, T. H. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 813-835 (ASM Press, Washington, DC, 2002).

8. Moran, J. V. & Gilbert, N. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 836-869 (ASM Press, Washington, DC, 2002).

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

10. Richardson, S. R. et al. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061 (2015).

11. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3**, MDNA3-0001 (2015).

12. Babushok, D. V. & Kazazian, H. H. J. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28**, 527-539 (2007).

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

14. Repanas, K. et al. Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* **35**, 4914-4926 (2007).

15. Khazina, E. & Weichenrieder, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc Natl Acad Sci U S A* **106**, 731-736 (2009).

16. Schneider, A. M. et al. Structure and properties of the esterase from non-LTR retrotransposons suggest a role for lipids in retrotransposition. *Nucleic Acids Res* (2013).

17. Januszyk, K. et al. 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**, 24893-24904 (2007).

18. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* (2016).

19. Bowles, K. & Christensen, S. M. The variable amino-terminal end of R2 is dispensable for R2 protein activity. *submitted.*

20. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).

21. Lane, L. C. A simple method for stabilizing protein-sulfhydryl groups during SDS-gel electrophoresis. *Anal Biochem* **86**, 655-664 (1978).

22. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1**, 2856-2860 (2006).

23. Havlis, J., Thomas, H., Sebela, M. & Shevchenko, A. Fast-response proteomics by accelerated in-gel digestion of proteins. *Anal Chem* **75**, 1300-1306 (2003).

24. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **68**, 850-858 (1996).

25. Chowdhury, S. M. et al. Solid-phase N-terminal peptide enrichment study by optimizing trypsin proteolysis on homoarginine-modified proteins by mass spectrometry. *Rapid Commun Mass Spectrom* **28**, 635-644 (2014).

26. Celic, I. et al. The sirtuins hst3 and Hst4p preserve genome integrity by controlling histone h3 lysine 56 deacetylation. *Curr Biol* **16**, 1280-1289 (2006).

27. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015).

28. Pettersen, E. F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).

29. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295-2300 (2008).

30. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol* **1079**, 263-271 (2014).

31. Alva, V., Nam, S. Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res* **44**, W410-5 (2016).

32. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276-3278 (2014).

33. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

34. Perez-Llamas, C. & Lopez-Bigas, N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One* **6**, e19541 (2011).

35. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res* **42**, 8405-8415 (2014).

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

37. Qu, G. et al. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* **23**, 549-557 (2016).

38. Zhao, C. & Pyle, A. M. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565 (2016).

39. Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**, 638-643 (2013).

40. Ng, K. K. et al. Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. *J Biol Chem* **277**, 1381-1387 (2002).

41. Wu, J., Liu, W. & Gong, P. A Structural Overview of RNA-Dependent RNA Polymerases from the Flaviviridae Family. *Int J Mol Sci* **16**, 12943-12957 (2015).

42. Lu, G. & Gong, P. A structural view of the RNA-dependent RNA polymerases from the Flavivirus genus. *Virus Res* **234**, 34-43 (2017).

43. Lu, G. & Gong, P. Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. *PLoS Pathog* **9**, e1003549 (2013).

44. Thompson, A. A. & Peersen, O. B. Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *EMBO J* **23**, 3462-3471 (2004).

45. Lesburg, C. A. et al. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat Struct Biol* **6**, 937-943 (1999).

46. Yan, C. et al. Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**, 1182-1191 (2015).

47. Gillis, A. J., Schuller, A. P. & Skordalakes, E. Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature* **455**, 633-637 (2008).

48. Ren, J. et al. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat Struct Biol* **2**, 293-302 (1995).

49. Nikonov, A., Juronen, E. & Ustav, M. Functional characterization of fingers subdomain-specific monoclonal antibodies inhibiting the hepatitis C virus RNA-dependent RNA polymerase. *J Biol Chem* **283**, 24089-24102 (2008).

50. Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**, MDNA3-0050 (2015).

51. Zimmerly, S. & Wu, L. An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr* **3**, MDNA3-0058 (2015).

52. Shu, B. & Gong, P. Structural basis of viral RNA-dependent RNA polymerase catalysis and translocation. *Proc Natl Acad Sci U S A* **113**, E4005-14 (2016).

53. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase (3Dpol) is sufficient for template switching in vitro. *J Biol Chem* **274**, 2706-2716 (1999).

54. Chen, B. & Lambowitz, A. M. De novo and DNA primer-mediated initiation of cDNA synthesis by the mauriceville retroplasmid reverse transcriptase involve recognition of a 3’ CCA sequence. *J Mol Biol* **271**, 311-332 (1997).

55. Mohr, S. et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958-970 (2013).

56. Gu, S. Q. et al. Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16**, 732-747 (2010).

57. Watanabe, K. & Lambowitz, A. M. High-affinity binding site for a group II intron-encoded reverse transcriptase/maturase within a stem-loop structure in the intron RNA. *RNA* **10**, 1433-1443 (2004).

58. Huang, J. et al. Structural basis for protein-RNA recognition in telomerase. *Nat Struct Mol Biol* **21**, 507-512 (2014).

59. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol* **17**, 513-518 (2010).

60. Wyatt, H. D. M., West, S. C. & Beattie, T. L. InTERTpreting telomerase structure and function. *Nucleic Acids Research* **38**, 5609-5622 (2010).

61. Gladyshev, E. A. & Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* **108**, 20311-20316 (2011).

62. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**, 3353-3362 (1990).

63. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res* **31**, 532-550 (2003).

64. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).

65. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3**, 433-437 (2013).

66. Doucet, A. J. et al. Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet* **6**, (2010).

67. Wan, R. et al. The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science* **351**, 466-475 (2016).

68. Nguyen, T. H. et al. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530**, 298-302 (2016).

69. Galej, W. P. et al. Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537**, 197-201 (2016).

70. Bertram, K. et al. Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature* **542**, 318-323 (2017).

71. Shivram, H., Cawley, D. & Christensen, S. M. Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements* **1**, 169-178 (2011).

Chapter 3

Conclusions

3.1 Summary

Using limited proteolysis and mass spectrometry, we presented the 3D globular domain structure of R2Bm protein, a site specific LINE encoded protein. We showed that the protein has two major globular domains: the N-terminal ZF/Myb containing DNA binding domain and the large RT/linker/EN superdomain. The -1 RNA binding motif was the preliminary cleavage site indicating the motif is exposed and probably connects the two globular domains. An updated model of reverse transcriptase domain was constructed; cleavage data from limited proteolysis experiment was used to validate the model. Comparison among different groups of polymerases (including RNA dependent RNA polymerase, group II intron reverse transcriptase, telomerase, splicing factor protein Prp8's reverse transcriptase domain, HIV RT with LINE RT model) revealed both shared structural properties as well as unique features among these polymerases. Shared properties have implications of common functional role that can be explored using mutational studies (discussed more in future extension section later in this chapter).

The revelation that R2Bm protein possess a large RT/linker/EN superdomain led us to explore commonalities between RLE LINE and central splicing factor protein Prp8, which also possess a RT/linker/EN superdomain structure. Bioinformatic analysis showed striking similarity between these two superdomain structures particularly in the thumb domain, RLE domain and two conserved motifs in the linker region that bear functional importance for Prp8 protein in the spliceosomal complex. Similarity between R2Bm protein and Prp8 protein has been previously postulated in literature because of the presence of the RT and RLE. We showed biochemically and bioinformatically the close relationship between these two proteins. Early branching LINE and Prp8 may have had a common ancestor.

The dissertation also shows promise in exploration of LINE encoded protein-nucleic acid interaction by footprinting method and some other protein structural features using mass spectrometric technique (see the preliminary data discussed in the future extension section later in this chapter).

3.2 Limitations

Visualization of the protein structure is still sought through atomic force microscopy, cryo-EM or crystallography that can provide atomic details of protein model. Limited proteolysis can be considered the preliminary step before crystallography as the proteolytically resistant fragments are candidate for independently folded globular and functional domain expressible in higher amount than the full length protein. The limited proteolysis experiment was designed without nucleic acid; the proposed globular domain structure thus does not address protein conformational changes in response to RNA or DNA binding.

3.3 Implications for APE LINEs

Like the early branching LINE encoded proteins, we predict the RT/linker/CCHC also form a large superdomain in APE LINE ORF2 encoded proteins. The APE endonuclease is a recent addition in the phylogenetically late branching LINEs[1]. From multiple sequence alignment, the sequence conservation of RT, α-helix rich linker region, a conserved alpha helical sequence motif and CCHC motif may indicate a common structure-function unit among the RLE LINE and APE LINEs. Mutation in the CCHC and SSSS motifs of L1 linker have shown reduced retrotransposition activity *in vivo*[2], which indicate the above motifs' functional importance as well.

3.4 Future extension

3.4.1 Expression of independently folded globular domain in large amount for crystallography and cryo-EM and functional testing

Crystallization or NMR technique is difficult for larger protein with low yield like R2Bm. The limited proteolysis is a logical pre-step before these techniques. The proteolytically resistant fragments are candidate area of the protein for forming natively folded globular domain structure expressible in higher quantitiy. This can facilitate crystallization or NMR based structural studies. The candidate domains also can be expressed for testing of function. Cryo-EM, which does not require as much protein, can also be tried.

3.4.2 Footprinting of R2Bm protein for mapping nucleic acid interacting amino acid residues

Global analysis of amino acid residues in the R2Bm protein involved in nucleic acid interaction is possible with mass spectrometric footprinting analysis. Biotinylation of lysine residues in presence and absence of R2Bm interacting nucleic acid can show us a footprint of which lysine residues are protected from biotinylation in the presence of nucleic acid[3–5]. Adding target DNA, 3' PBM RNA, 5' PBM RNA, branched DNA in combinations and separately will led to the identification of lysine residues involved in interacting with corresponding nucleic acid. Arginine modification with p-hydroxyphenylglyoxal (HPG)/ 1,2-cyclo-hexanedione (CHD)[6] in similar manner will yield information on nucleic acid interacting arginine residues.

3.4.3 Preliminary data on lysine footprinting in presence of 3' PBM RNA

To carry out lysine footprinting analysis, optimization of biotin concentration for protein surface lysine residues biotinylation that does not protein native conformation was first attempted. R2Bm protein was first biotinylated at different concentration of biotin followed by binding reaction with RNA and DNA (Figure 3-1). The least concentration of biotin that hampered RNA and DNA binding was first chosen for protection assay (Figure 3-2 (A) and 3-2 (B)). Since single stranded RNA with primary amines groups in the bases can act as potential competitor for biotin with primary amines of lysine, a sink test was incorporated (figure 3-2(A)). Sink test indicated that RNA primary amines did not compete for biotin. Probably secondary structure hided the primary amine groups or otherwise bound and masked by the protein and thus rendering them inaccessible to the biotin reagent.

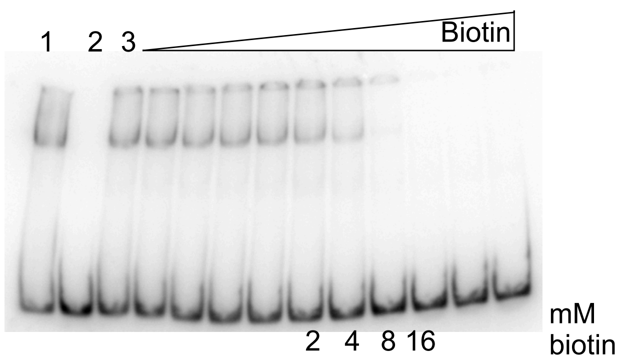


Figure 3-1. Effect of biotin concentration on RNA binding activity of R2Bm protein. R2Bm protein was first biotinylated or mock biotinylated followed by addition of quenching solution (10 mM Tris final) and then biotinylated protein was allowed to bind RNA + DNA. Protein bound to RNA was run on Electrophoretic Mobility Shift Assay (EMSA) gel. ^{32}P radiolabeled DNA was used as read-out of protein-RNA bound retarded DNA.
Lane 1 is regular EMSA control reaction. Lane 2 is DNA only control. Lane 3 is no biotin control. The rest of the lanes contained biotin at increasing concentration (64 to 0.125 mM).

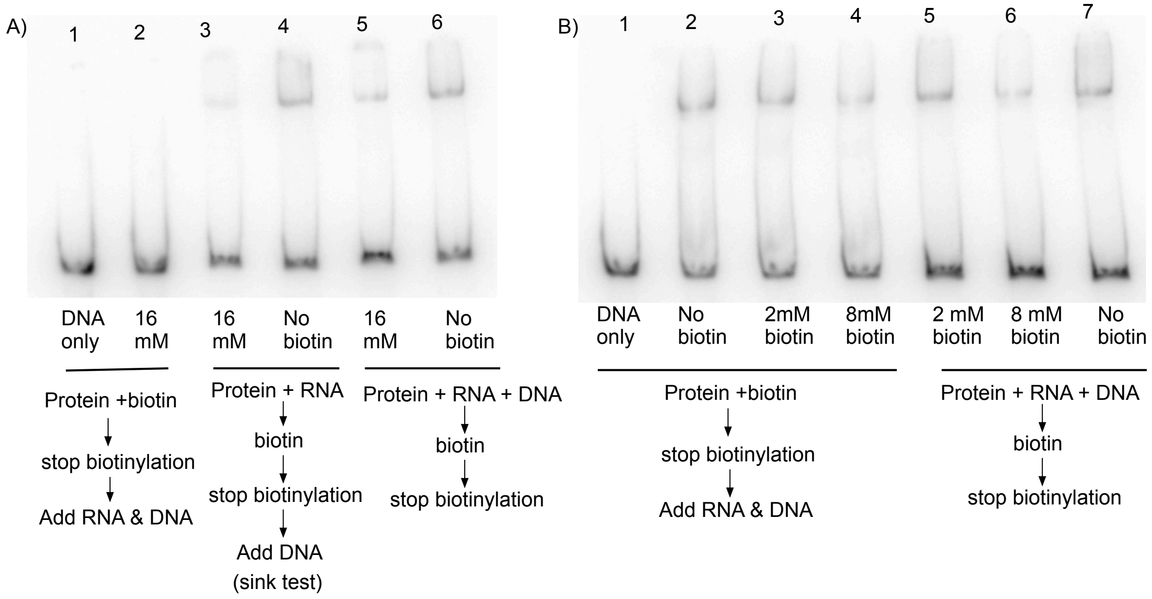


Figure 3-2. Test for preservation of nucleic acid binding activity. A) Sink test and biotin protection assay with 16 mM biotin. B) Biotin protection assay at 2 mM and 8 mM biotin.

Biotin protection assay and biotin titration experiment revealed some issues that need to be addressed in the future design of the experiment. We used sulfo-NHS-biotin for lysine modification. The reagent is readily soluble in water but after biotinylation the sulfo group leaves imparting an insoluble characteristic. This renders the biotinylated protein at a particular level precipitating out of solution, which we saw in our experiment as well. The precipitation correlated with biotin concentration. So, we did footprinting assay in three different biotin concentration (2 mM, 4mM and 8 mM) and with soluble or insoluble protein. At these three different concentration protein was biotinylated either in presence or absence of 3' PBM RNA. Biotinylated protein was loaded on pre-cast polyacrylamide gel. Only soluble fraction from 2 mM treatment, only insoluble fraction from 8 mM treatment and both soluble/insoluble protein from the 4 mM treatment was loaded. Protein bands were excised, reduced & alkylated and then completely digested with trypsin. Smaller trypsinized peptides were sequenced and identified by nano-LC-ESI-MS/MS. The biotinylation profile in the peptides is compared in table 3-1.

Table 3-1: Protection of R2Bm lysine residues by 3' PBM RNA from biotinylation

Peptide	position in R2Bm primary sequence	2 mM biotin (soluble)	4 mM biotin (soluble + insoluble)	8 mM biotin (insoluble only)
AHPVETNTDAAPMMVKR	N-terminal ZF	N/A	no?	no?
ACRAMRPKTAGR	-1 vicinity	N/A	yes	N/A
AMEENKWTVELEPR	starting of EN	no?	?	yes
AMRPKTAGR	-1 vicinity	N/A	yes	no/?
GVWSLTSYKELR	RGVW motif but in 3D model of EN possess similar position as the AMEENKWTVELEPR peptide	no?	yes	N/A
KSAVLSMIPDGHR	RT 6	N/A	yes	N/A

TASAHKTSR	-1 RNA binding motif	N/A	yes	yes?
VQELYKK	-1 RNA binding motif	yes	no?	no?
VQELYKKCR	-1 RNA binding motif	yes	yes?	yes?
WAWKQLR	linker	N/A	no	yes?

Table 3-2: Peptides that always or never biotinylated in the footprinting experiment

Peptides always biotinylated	peptides that never showed biotinylation
AAAKSDKIR AGCKVR APLKPQQR GLGVHKR HNKIVSFVAK KAVGQWLR KPDIIASR KSNKENRPEASGLPLESER LGLPKAECVR NKYGNHGELVELVAGR SNKENRPEASGLPLESER TFNIGGKPLR TLEAIKGQR TPTSTKWIR	DYTQLWKPISVEEIK ECHVAVLDFAK LPADVPK MLDVQIRK

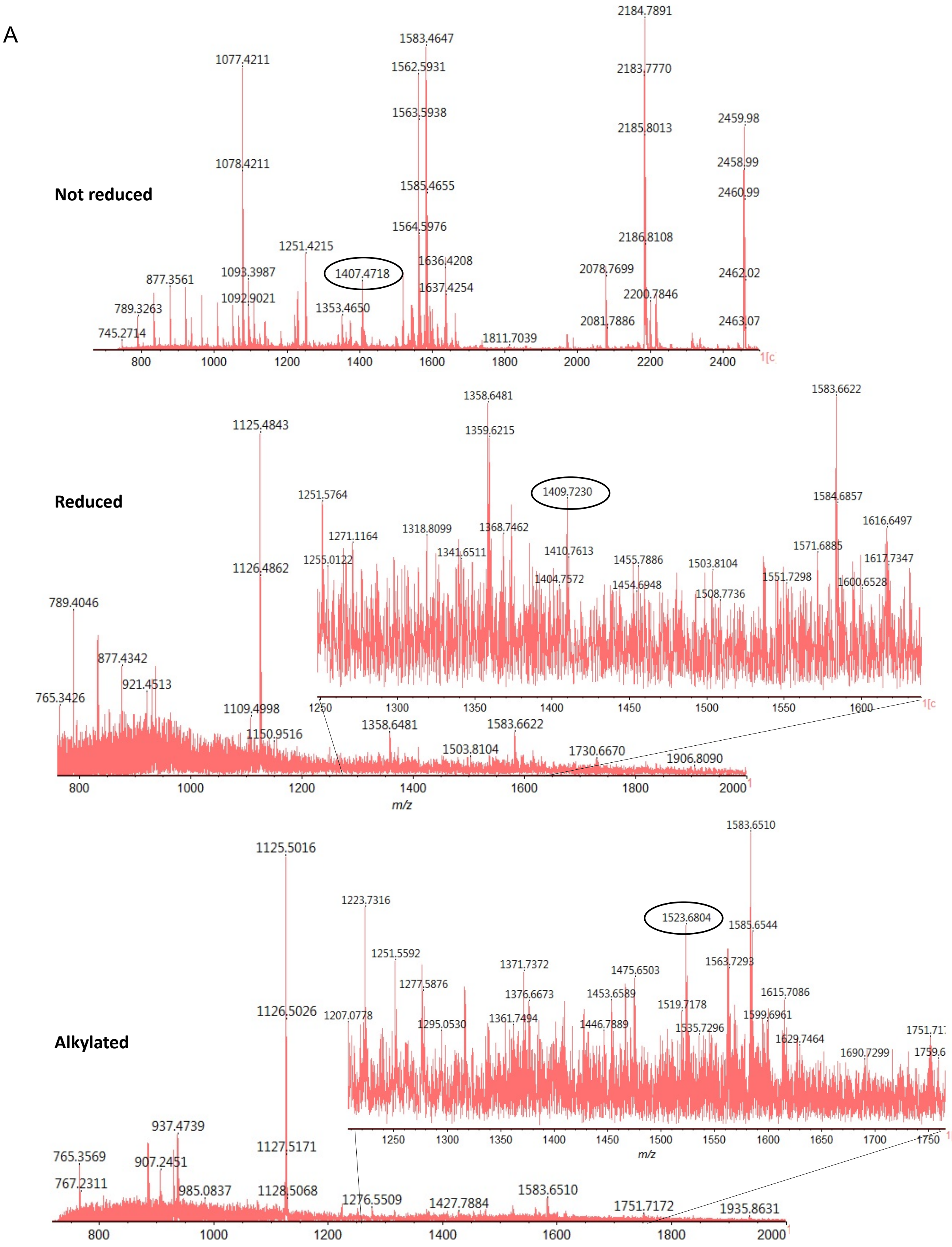
The data indicate more robust protection in -1 domain and RT 6 subdomain. A weaker protection is also likely in endonuclease domain (one of the endonuclease peptide being in the RGVW motif). Since the biotinylation leads to protein precipitation, explaining protection of peptide from biotinylation by RNA is difficult. A soluble PEG spacer containing biotin reagent might solve the issue.

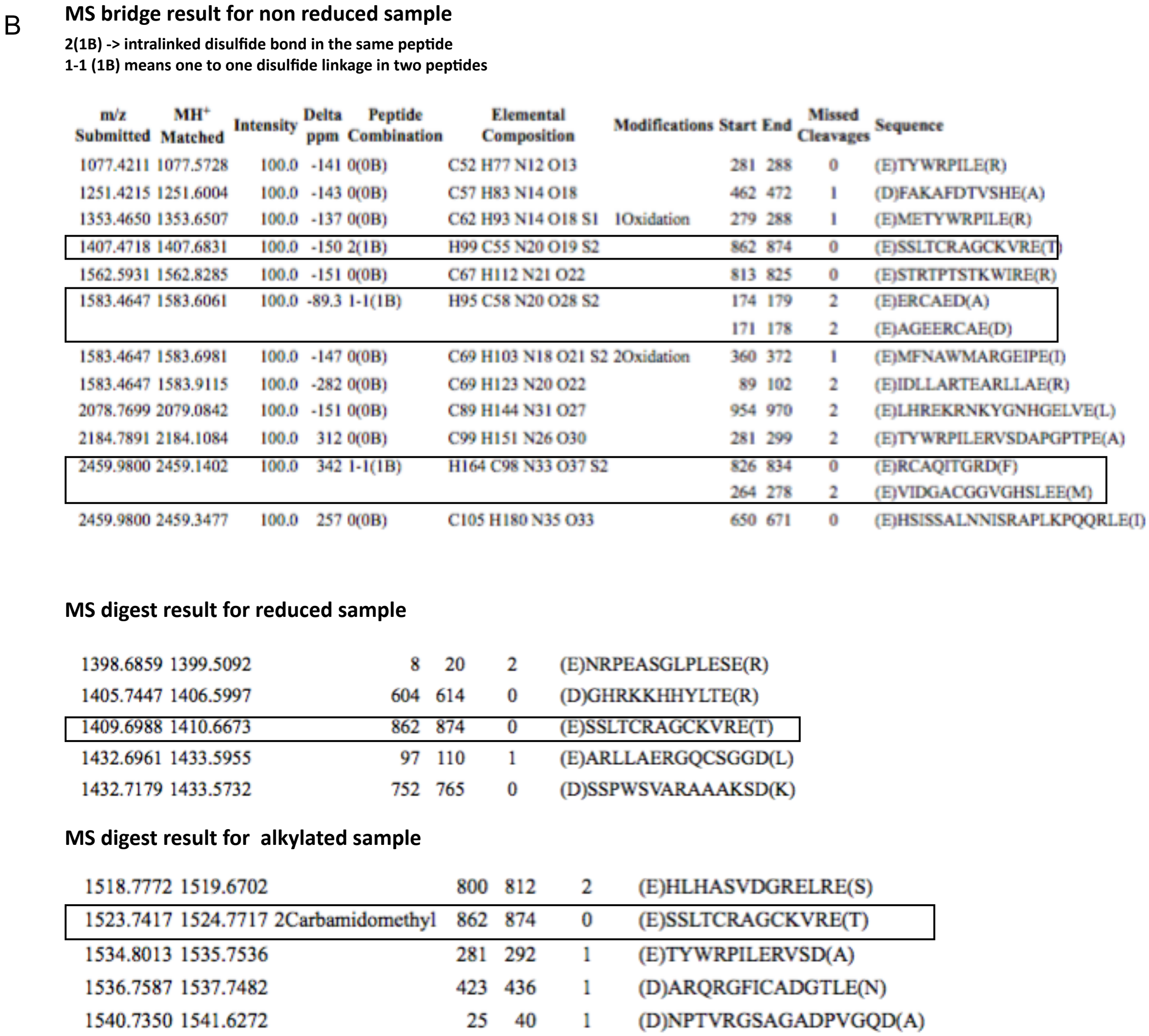
3.4.4 Identification of cysteine residues involved in disulfide bond formation in R2Bm protein

There are 28 cysteine residues in the R2Bm element encoded protein. It is possible that several of these cysteine residues are involved in disulfide linkage(s). Identification of disulfide linked cysteine residues would help strengthening the knowledge of the protein structure. Different approaches including the differential reduction and alkylation of the cysteine residues are noteworthy[7–11]. A preliminary experiment was carried out. R2Bm protein was digested in acidic pH by GluC protease. Detergent was removed using ethyl acetate after digestion with GluC protease and then dried in speed vac. Peptides were reconstituted in ammonium bicarbonate and aliquoted into 3. One of the aliquot was spotted directly on MALDI plate after zip-tiping. The second aliquot was only reduced and the third aliquot was both reduced and alkylated. After reducing, the peptides were immediately zip-tiped and spotted on MALDI plate. Alkylated peptides were zip-tiped and spotted on MALDI plate. Data acquisition was performed in Axima Resonance (MALDI-QIT-TOF) in Shimadzu institute at UTA. The molecular weight of the peptides were compared and analyzed using protein prospector's MS digest as well as MS bridge tools. The preliminary experiment (figure 3-3) suggested that the cysteine residues in the peptide SSLTCRAGCKVRE may, under the conditions tested, formed a disulfide bond, although MS/MS sequencing confirmation was not performed. It should be noted, however, that this disulfide bond may not be in the native structure of the protein. The two C's in question are part of the CCHC motif and are thought to coordinate a Zn^{++} ion. If the hypothetical zinc were lost, a disulfide bond might form. The protein is stored and reacted in buffers containing the reducing reagent DTT, but purified under partially oxidizing conditions. To elucidate if the SSLTCRAGCKVRE contain a true disulfide bond in the active protein solution, the future experiment can be designed such that the protein is always in either the presence or absence of the reducing reagent.

In an alternative experiment, the free cysteines containing peptides can be blocked by alkylation with an alkylating reagent 'A' in the absence of a reducing agent. The disulfide linked peptides then can be reduced and alkylated with a reducing agent and an alkylating reagent 'B'. The presence of different alkylation tag can thus help identify the disulfide linkage. If biotin-maleimide is used for alkylating disulfide linked cysteine residues after blocking the free cysteine residues with a different alkylating reagent, the biotin tagged peptide can be enriched with avidin from the crowd of non-disulfide linked peptides[8,10,12–14].

Also, instead of running the solution based experiment, a gel based approach can be performed to help remove the triton X-100 that complicates the mass spectra. The gel based approach would require in gel digestion of the protein and the differential reduction/alkylation treatment prior to gel loading to avoid disulfide scrambling. One possible issue is incomplete unfolding of the protein if alklyation is performed (for blocking the free cysteine residues) without reduction or any denaturant. Alkylation (+/-) reduction in presence of a gradation of denaturant can be performed.


A
Not reduced
2184.7891
1583.4647
1077.4211
1562.5931
2183.7770
1563.5938
2185.8013
2459.98
1078.4211
2458.99
1585.4655
2460.99
1564.5976
2186.8108
1251.4215
1636.4208
1093.3987
877.3561
1407.4718
2078.7699
1637.4254
2462.02
1092.9021
2200.7846
789.3263
1353.4650
2081.7886
2463.07
745.2714
1811.7039
800
1000
1200
1400
1600
1800
2000
2200
2400
Reduced
1583.6622
1358.6481
1125.4843
1359.6215
1409.7230
1251.5764
1584.6857
1368.7462
1616.6497
1318.8099
1271.1164
1571.6885
1410.7613
1341.6511
1455.7886
1503.8104
1617.7347
1255.0122
1126.4862
1404.7572
1454.6948
1551.7298
1600.6528
1508.7736
789.4046
877.4342
921.4513
765.3426
1250
1300
1350
1400
1450
1500
1550
1600
1109.4998
1150.9516
1358.6481
1583.6622
1503.8104
1730.6670
1906.8090
800
1000
1200
1400
1600
1800
2000
m/z
Alkylated
1583.6510
1125.5016
1223.7316
1523.6804
1585.6544
1251.5592
1371.7372
1475.6503
1563.7293
1277.5876
1376.6673
1453.6589
1615.7086
1126.5026
1207.0778
1519.7178
1599.6961
1295.0530
1361.7494
1446.7889
1535.7296
1751.717
1629.7464
1690.7299
1759.6
1250
1300
1350
1400
1450
1500
1550
1600
1650
1700
1750
m/z
937.4739
765.3569
907.2451
1127.5171
767.2311
985.0837
1128.5068
1583.6510
1276.5509
1427.7884
1751.7172
1935.8631
800
1000
1200
1400
1600
1800
2000



MS bridge result for non reduced sample

2(1B) -> intralinked disulfide bond in the same peptide
1-1 (1B) means one to one disulfide linkage in two peptides

m/z Submitted	MH+ Matched	Intensity	Delta ppm	Peptide Combination	Elemental Composition	Modifications	Start	End	Missed Cleavages	Sequence
1077.4211	1077.5728	100.0	-141	0(0B)	C52 H77 N12 O13		281	288	0	(E)TYWRPILE(R)
1251.4215	1251.6004	100.0	-143	0(0B)	C57 H83 N14 O18		462	472	1	(D)FAKAFDTVSHE(A)
1353.4650	1353.6507	100.0	-137	0(0B)	C62 H93 N14 O18 S1	1Oxidation	279	288	1	(E)METYWRPILE(R)
1407.4718	1407.6831	100.0	-150	2(1B)	H99 C55 N20 O19 S2		862	874	0	(E)SSLTCRAGCKVRE(T)
1562.5931	1562.8285	100.0	-151	0(0B)	C67 H112 N21 O22		813	825	0	(E)STRTPTSTKWIRE(R)
1583.4647	1583.6061	100.0	-89.3	1-1(1B)	H95 C58 N20 O28 S2		174	179	2	(E)ERCAED(A)
							171	178	2	(E)AGEERCAE(D)
1583.4647	1583.6981	100.0	-147	0(0B)	C69 H103 N18 O21 S2	2Oxidation	360	372	1	(E)MFNAWMARGEIPE(I)
1583.4647	1583.9115	100.0	-282	0(0B)	C69 H123 N20 O22		89	102	2	(E)IDLLARTEARLLAE(R)
2078.7699	2079.0842	100.0	-151	0(0B)	C89 H144 N31 O27		954	970	2	(E)LHREKRNKYGNHGELVE(L)
2184.7891	2184.1084	100.0	312	0(0B)	C99 H151 N26 O30		281	299	2	(E)TYWRPILERVSDAPGPTPE(A)
2459.9800	2459.1402	100.0	342	1-1(1B)	H164 C98 N33 O37 S2		826	834	0	(E)RCAQITGRD(F)
							264	278	2	(E)VIDGACGGVGHSLEE(M)
2459.9800	2459.3477	100.0	257	0(0B)	C105 H180 N35 O33		650	671	0	(E)HSISSALNNISRAPLKPQQRLE(I)

MS digest result for reduced sample

1398.6859	1399.5092	8	20	2	(E)NRPEASGLPLESE(R)
1405.7447	1406.5997	604	614	0	(D)GHRKKHHYLTE(R)
1409.6988	1410.6673	862	874	0	(E)SSLTCRAGCKVRE(T)
1432.6961	1433.5955	97	110	1	(E)ARLLAERGQCSGGD(L)
1432.7179	1433.5732	752	765	0	(D)SSPWSVARAAAKSD(K)

MS digest result for alkylated sample

1518.7772	1519.6702		800	812	2	(E)HLHASVDGRELRE(S)
1523.7417	1524.7717	2Carbamidomethyl	862	874	0	(E)SSLTCRAGCKVRE(T)
1534.8013	1535.7536		281	292	1	(E)TYWRPILERVSD(A)
1536.7587	1537.7482		423	436	1	(D)ARQRGFICADGTLE(N)
1540.7350	1541.6272		25	40	1	(D)NPTVRGSAGADPVGQD(A)

Figure 3-3. Preliminary results for disulfide bridged cysteine identification. A) Axima Resonance (MALDI-QIT-TOF) generated spectra from nonreduced, reduced and alkylated samples. B) MS bridge and MS digest analysis of spectra.

3.4.5 Mutational analysis of structurally conserved motifs in R2Bm RT and of nucleic acid interacting residues identified by protein footprinting experiments

Modeling of R2Bm RT provide us with the idea of overall fold the RT domain can adopt. Structural superposition of the RT model on homologous reverse transcriptases from PDB database further indicated some common components of reverse transcriptase structures. Mutant designing and testing for loss of function will be helpful for further validating the model and for investigating the RT and nucleic acid interactions and RT structure and function. In addition to the RT residues, mutants will be used to test possible nucleic acid binding residues that were identified by the protein footprinting studies proposed in thsi chapter.

Several mutations have been constructed already in our lab. One of the mutations, the arginine to alanine conversion in the YWR motif, of the RT index finger reduced RNA binding activity (Figure 3-4) in preliminary experiments.

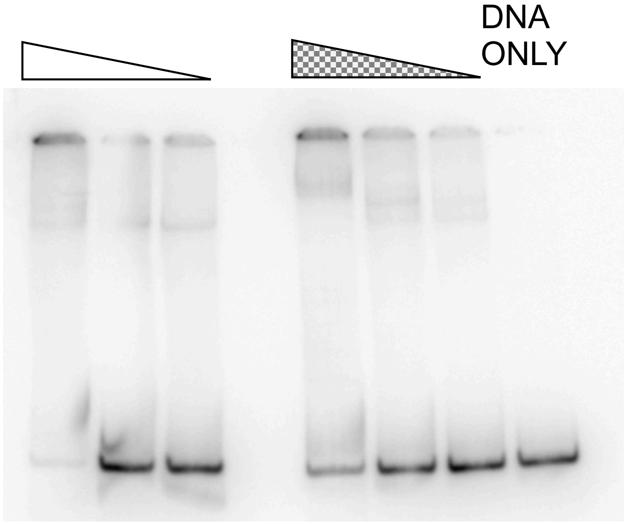


Figure 3-4. Effect of YWR/A mutation in the index finger of RT domain on the RNA binding

activity of R2Bm protein. Reaction formed between R2Bm protein (Wild type or mutant), 3' PBM RNA and DNA was loaded onto EMSA gel. The antisense strand of the target was ^{32}P radiolabeled. Gel was dried and phosphorimaged. Triangles represent protein titration series. Lanes indicated by white and pattern filled triangles represent wild type and mutant protein mediated band shift.

Several other already constructed mutants in the reverse transcriptase area are being currently tested in our lab. These previous mutants were designed based upon sequence conservation. Current R2 RT model and LINE protein's comparison to RdRP/Prp8/group II intron proteins give clue to additional area with potential function. A summerized information of these regions/amino acid residues is presented in Table 3-3.

Table 3-3: Designing mutations in R2Bm protein for possible loss of function

Amino acid residue in R2Bm protein	Function that can be tested	Rationale/ function in analogous structure
K in FVPK motif in the RT 1 finger	nucleotide binding? affect polymerization?	motif F or ring finger (corresponding finger is RT1 in R2 RT) has NTP binding role[15].
R in ERPGGP motif in the RT1 finger	nucleotide binding? affect polymerization?	motif F or ring finger (corresponding finger is RT1 in R2 RT) has NTP binding role.
KS in RT6	for RNA binding	footprinting indicates RNA binding function; mutation of K strongly affect polymerization in RdRPs[16].
QIT loop in the HINALP area	branched structure of nucleic acid binding?	Prp8 alpha finger aligns with QITG.........HINALP motif of R2Bm protein. In U4/U6.U5 tri-snRNP, the α-finger (helix-turn-helix) contacts U54-U55 of U4 snRNA near the three-way junction[17].

R in RGMP of IFD motif	Processivity? polymerization activity?	play role in repeat addition Processivity in telomerase [18]
R and R in RQRGF motif	polymerization activity?	form part of the exit channel for newly synthesized RNA in RdRP[16]
QQ in APLKPQQR motif	form template channel?	faces the interior side of the RT cavity.
R in RVSDAPGPTP motif and K in MSSPVKVGR motif	possibly bind RNA ?	as they are in similar position in 3D model and in the index finger vicinity that has been implicated in *de novo* and primer depenent polymerization[19]; our preliminary data also indicate index finger's YWR/A motif might have role in RNA binding

References

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

2. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).

3. Kvaratskhelia, M., Miller, J. T., Budihas, S. R., Pannell, L. K. & Le Grice, S. F. J. Identification of specific HIV-1 reverse transcriptase contacts to the viral RNA:tRNA complex by mass spectrometry and a primary amine selective reagent. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15988-15993 (2002).

4. Hilton, B. et al. A new structural insight into XPA-DNA interactions. *Bioscience reports* **34**, e00162 (2014).

5. Chen, T. H. et al. Use of chemical modification and mass spectrometry to identify substrate-contacting sites in proteinaceous RNase P, a tRNA processing enzyme. *Nucleic Acids Res* **44**, 5344-5355 (2016).

6. Wanigasekara, M. S. & Chowdhury, S. M. Evaluation of chemical labeling methods for identifying functional arginine residues of proteins by mass spectrometry. *Anal Chim Acta* **935**, 197-206 (2016).

7. Yen, T.-Y., Yan, H. & Macher, B. A. Characterizing closely spaced, complex disulfide bond patterns in peptides and proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. *Journal of mass spectrometry* **37**, 15-30 (2002).

8. Yen, T. Y. et al. Characterization of cysteine residues and disulfide bonds in proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. *J Mass Spectrom* **35**, 990-1002 (2000).

9. Smith, M. E. et al. Protein modification, bioconjugation, and disulfide bridging using bromomaleimides. *J Am Chem Soc* **132**, 1960-1965 (2010).

10. Mo, J., Tymiak, A. A. & Chen, G. Characterization of disulfide linkages in recombinant human granulocyte-colony stimulating factor. *Rapid Commun Mass Spectrom* **27**, 940-946 (2013).

11. Akter, S. et al. Cysteines under ROS attack in plants: a proteomics view. *J Exp Bot* **66**, 2935-2944 (2015).

12. Yen, T. Y., Yan, H. & Macher, B. A. Characterizing closely spaced, complex disulfide bond patterns in peptides and proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. *J Mass Spectrom* **37**, 15-30 (2002).

13. Jones, M. W. et al. Polymeric dibromomaleimides as extremely efficient disulfide bridging bioconjugation and pegylation agents. *J Am Chem Soc* **134**, 1847-1852 (2012).

14. Leonard, S. E. & Carroll, K. S. Chemical 'omics' approaches for understanding protein cysteine oxidation in biology. *Curr Opin Chem Biol* **15**, 88-102 (2011).

15. Lu, G. & Gong, P. Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. *PLoS Pathog* **9**, e1003549 (2013).

16. te Velthuis, A. J. Common and unique features of viral RNA-dependent polymerases. *Cell Mol Life Sci* **71**, 4403-4420 (2014).

17. Nguyen, T. H. et al. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530**, 298-302 (2016).

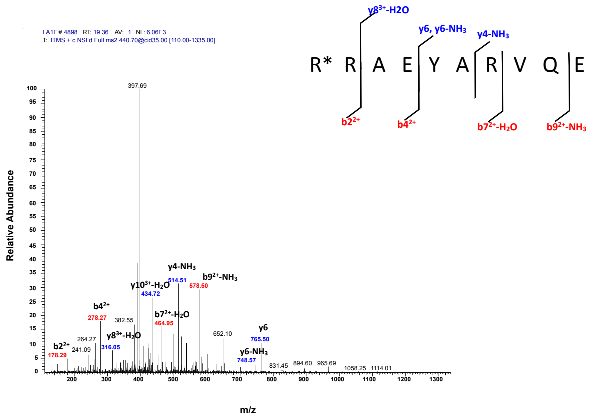
18. Lue, N. F., Lin, Y. C. & Mian, I. S. A conserved telomerase motif within the catalytic domain of telomerase reverse transcriptase is specifically required for repeat addition processivity. *Mol Cell Biol* **23**, 8440-8449 (2003).

19. Nikonov, A., Juronen, E. & Ustav, M. Functional characterization of fingers subdomain-specific monoclonal antibodies inhibiting the hepatitis C virus RNA-dependent RNA polymerase. *J Biol Chem* **283**, 24089-24102 (2008).

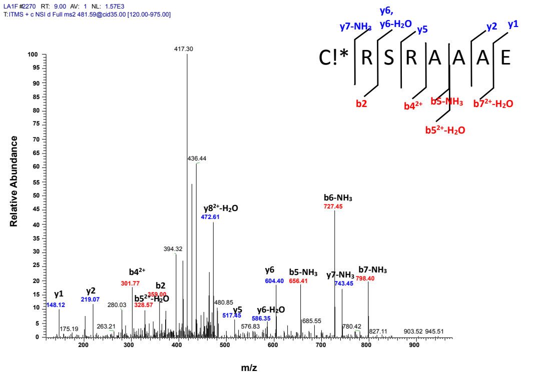
APPENDIX A

MS/MS SPECTRA OF N-TERMINAL END PEPTIDES AND MAP OF INTERNAL PEPTIDES ON R2Bm PROTEIN PRIMARY SEQUENCE FOR LYSC RESISTANT FRAGMENTS

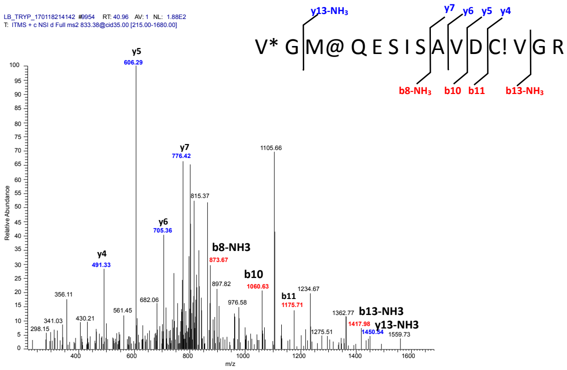
La(i) n-term end peptide = R*RAEYARVQE; z = 3+;
XCorr = 2.34; #PSMs= 3; [M+3H]3+ = 440.70; Theo. [M+H]1+ = 1319.6814; * = acetyl



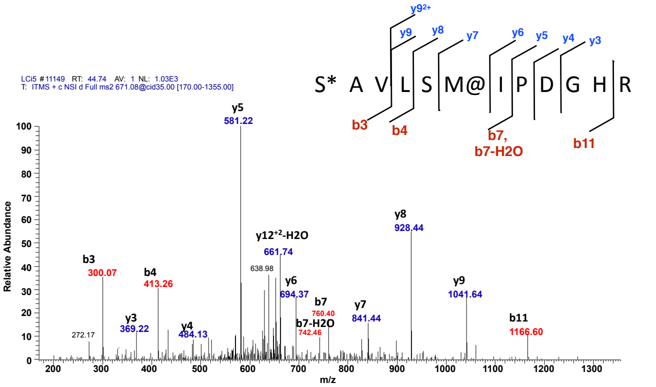
LA(ii) n-term end peptide = C!*RSRAAAE; z = 2+;
XCorr = 2.72; #PSMs= 1; [M+2H]2+ = 481.59; Theo. [M+H]1+ = 962.44 ; * = acetyl; ! = carbamidomethyl



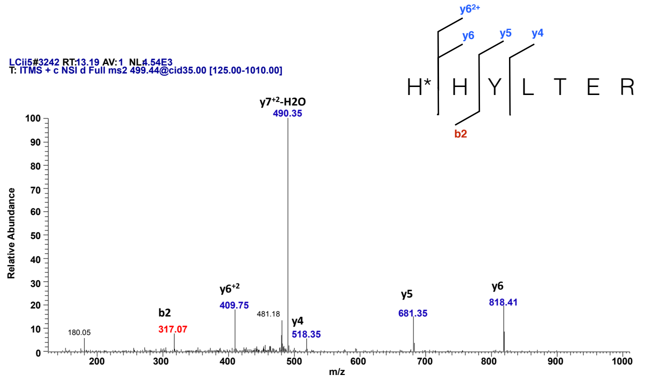
LB n-term end peptide = V*GM@QESISAVDC!VGR; z = 2+;
XCorr = 2.58; #PSMs= 1; [M+2H]2+ = 833.38; Theo. [M+H]1+ = 1665.75 ; * = acetyl; @ = oxidation; ! = carbamidomethyl



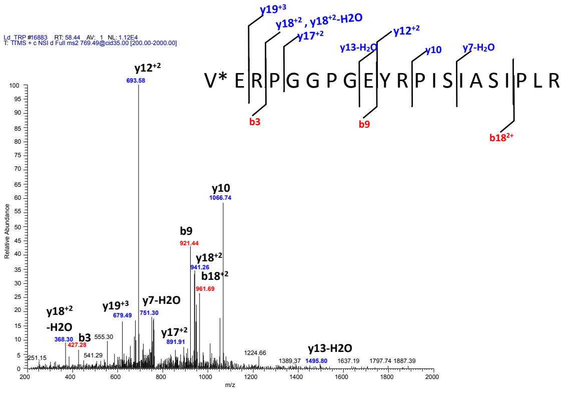
LC(i) N-term peptide: S*AVLSM@IPDGHR; z= +2; Xcorr = 3.22; Theo. [M+H]1+ = 1340.6521; [M+2H]2+ = 671.08; #PSMs = 3; * = acetyl; @ = oxidation



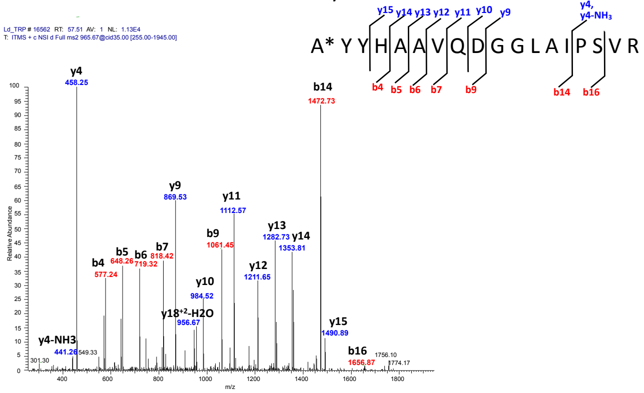
LC(ii) N-term peptide: H*HYLTER; z= +2; Xcorr = 2.067; Theo. [M+H]1+ = 997.4744; [M+2H]2+ = 499.44; #PSMs = 2; * = acetyl



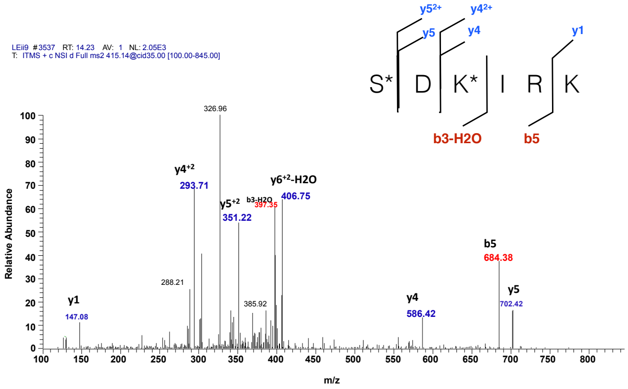
Band D(i) N-term peptide:V*ERPGGPGEYRPISIASIPLR; z= +3; Xcorr = 3.08; #PSM = 1; Theo. [M+H]1+ = 2305.24; [M+3H]3+ = 769.49; * = acetyl



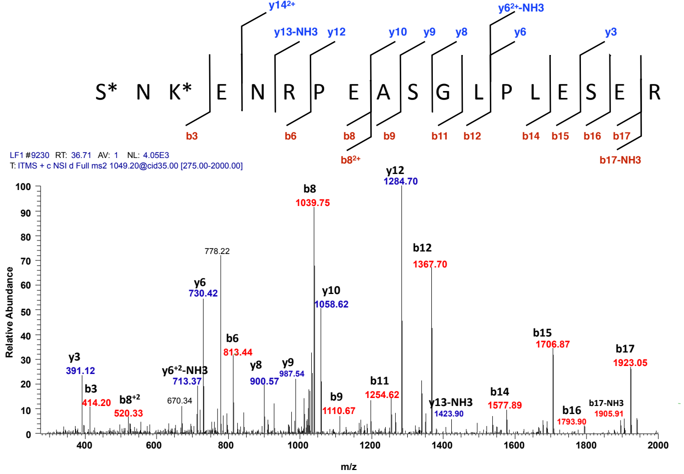
Band D(ii) N-term peptide: A*YYHAAVQDGGLAIPSVR; z= +2; Xcorr = 6.08; #PSM = 3 Theo. [M+H]1+ = 1929.97; [M+2H]2+ = 965.67; * = acetyl



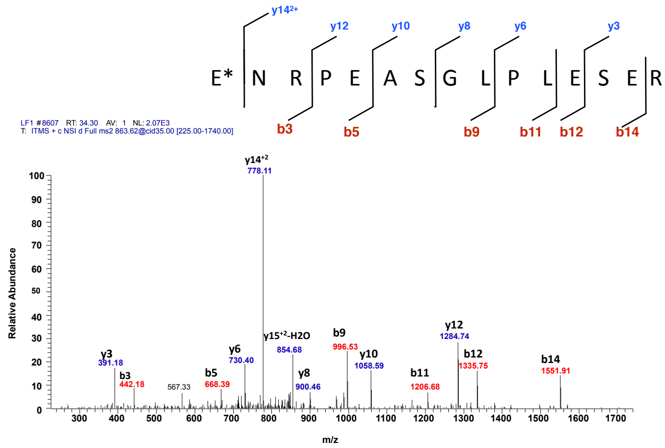
LE N-term peptide: S*DK*IRK; z= +2; Xcorr = 1.505; Theo. [M+H]1+ = 830.4518; [M+2H]2+ = 415.14; #PSMs = 3; * = acetyl



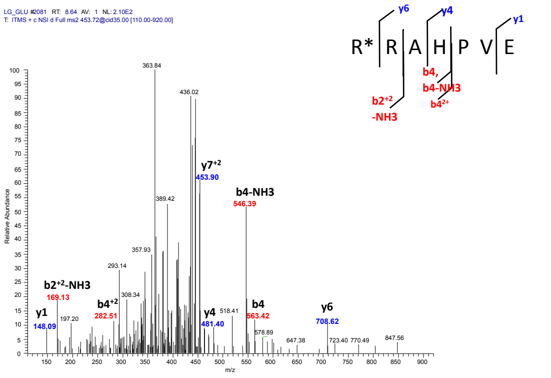
LF (i) N-term peptide: S*NK*ENRPEASGLPLESER; z= +2; Xcorr = 5.4277; Theo. [M+H]1+ = 2096.9995; [M+2H]2+ = 1049.20; #PSMs = 17; * = acetyl



LF(ii) N-term peptide: E*NRPEASGLPLESER; z= +2; Xcorr = 4.09; Theo. [M+H]1+ = 1725.8296; [M+2H]2+ = 863.62; #PSMs = 8; * = acetyl

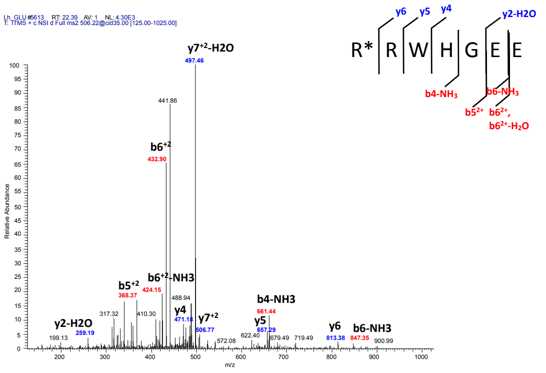


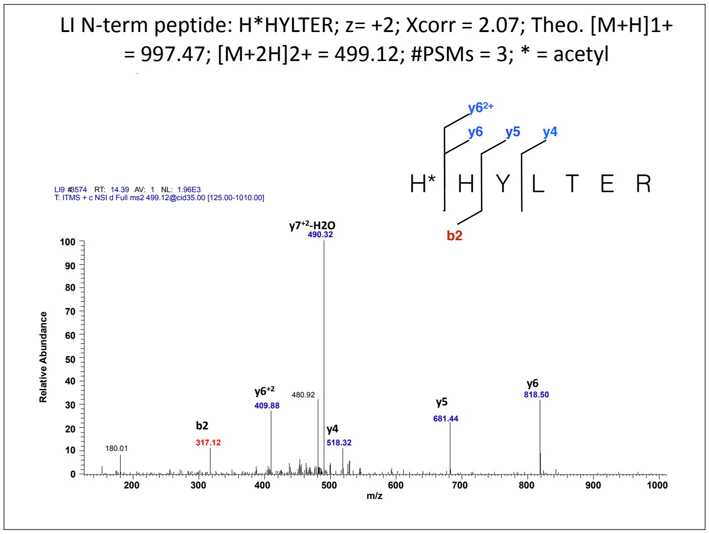
Band LG N-term peptide: R*RAHPVE; z= +2; Xcorr = 1.14; #PSMs = 1; Theo. [M+H]1+ = 906.48; [M+2H]2+ = 453.72; * = acetyl



The fragmentation show less confidence but there are two arginines and one histidine in the peptide sequence. Ammonia loss is seen as expected. The internal peptide sequence coverage also validates this peptides as N-terminal end of Band LG.

Band H N-term peptide: R*RWHGEE; z= +2; Xcorr = 1.65; Theo. [M+H]1+ = 1011.46; [M+2H]2+ = 506.22; #PSMs = 3; * = acetyl





LI N-term peptide: H*HYLTER; z= +2; Xcorr = 2.07; Theo. [M+H]1+ = 997.47; [M+2H]2+ = 499.12; #PSMs = 3; * = acetyl

Figure S1A: MS/MS spectra for N-term end peptides of LysC bands

Band LA

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LB

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
    HHHHHH
```

Band LC

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LD

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LE

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LF

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

red text + highlighted = N-terminal call for corresponding band
green only= gluc-trp; ESI run #1
green & italic = gluc_trp ; not in 1st run but in any of the following run of the band
green & underlined = glue as second digester
blue = common between two runs
orange = common among three runs

Band LG

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LH

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band LI

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

red text + highlighted = N-terminal call for corresponding band
green only= gluc-trp; ESI run #1
green & italic = gluc_trp ; not in 1st run but in any of the following run of the band
green & underlined = glue as second digester
blue = common between two runs
orange = common among three runs

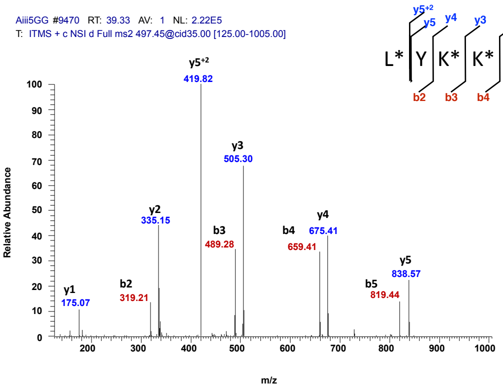
Number of ESI runs of the processed bands:
LA band: 2 ESI. 1 lysctrp. 1lyscgluc. blue is common between 1lysctrp & 1lyscgluc
LB band: 3 ESI. 2 lysctrp. 1lyscgluc. blue is common between 2lysctrp
LC band: 2 ESI. 2 lysctrp. blue is common between 2lysctrp
LD band: 3 ESI. 2 lysctrp. 1lyscgluc. blue is common between 2lysctrp
LE band: 2 ESI. 2 lysctrp. blue is common between 2lysctrp
LG band: 4 ESI. 2 lysctrp. 2lyscgluc but 1glucgluc run was bad. blue is common between 2lysctrp
LH band: 2 ESI. 1 lysctrp. 1lyscgluc. blue is common between 1 lysctrp & 1 lyscgluc

Figure S1B: Internal peptide for LysC bands

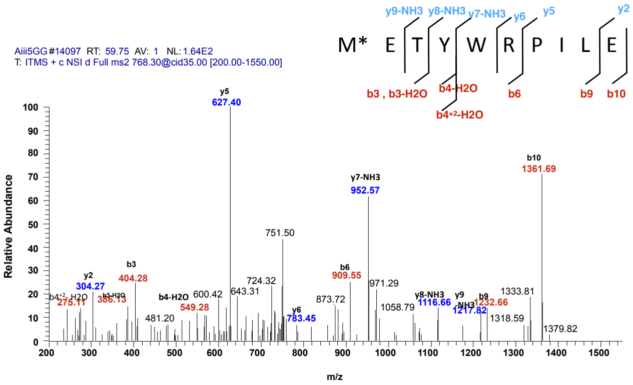
APPENDIX B

MS/MS SPECTRA OF N-TERMINAL END PEPTIDES AND MAP OF INTERNAL PEPTIDES ON R2Bm PROTEIN PRIMARY SEQUENCE FOR GLUC RESISTANT FRAGMENTS

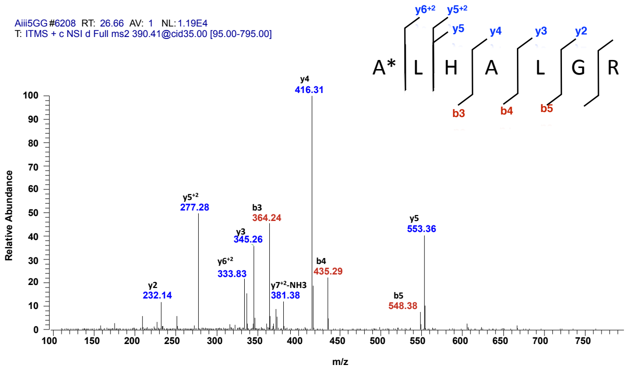
GA (i) N-term peptide: L*YK*K*C!R; z= +2; Xcorr =2.69; Theo. [M+H]1+ = 993.4868; [M+2H]2+ = 497.45 ; #PSMs = 37; * = acetyl; ! = carbamidomethyl



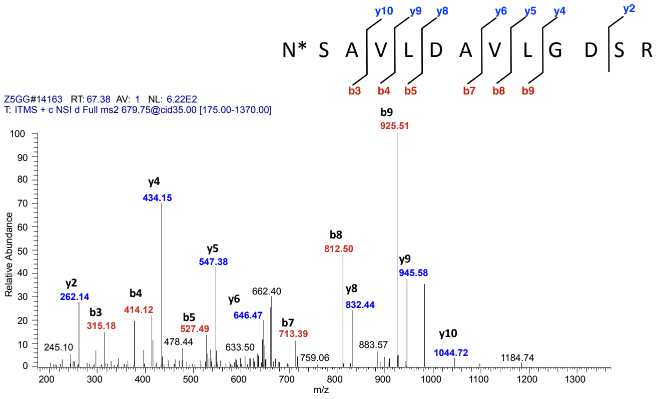
GA(ii) N-term peptide: M*ETYWRPILER; z= +2; Xcorr =2.59; Theo. [M+H]1+ = 1535.7569; [M+2H]2+ = 768.30; #PSMs = 3; * = acetyl



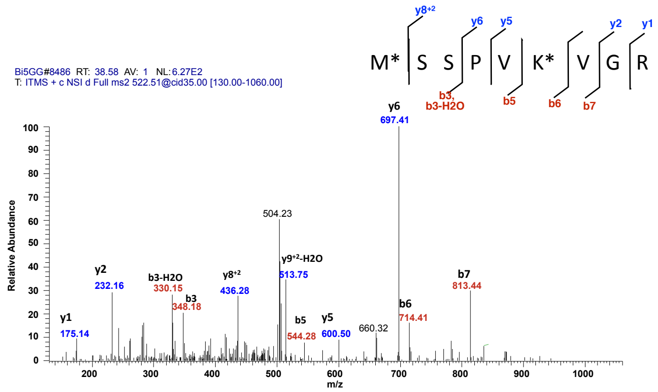
GA(iii) N-term peptide: A*LHALGR; z= +2; XCorr =2.41; Theo. [M+H]1+ = 779.4416; [M+2H]2+ = 390.41 ; #PSMs = 6; * = acetyl



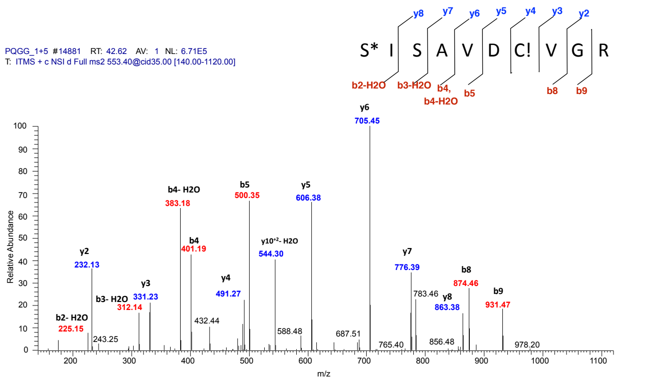
GB N-term peptide: N*SAVLDAVLGDSR; z= +2; XCorr =3.08; Theo. [M+H]1+ = 1358.6804; [M+2H]2+ = 679.75 ; #PSMs = 3; * = acetyl



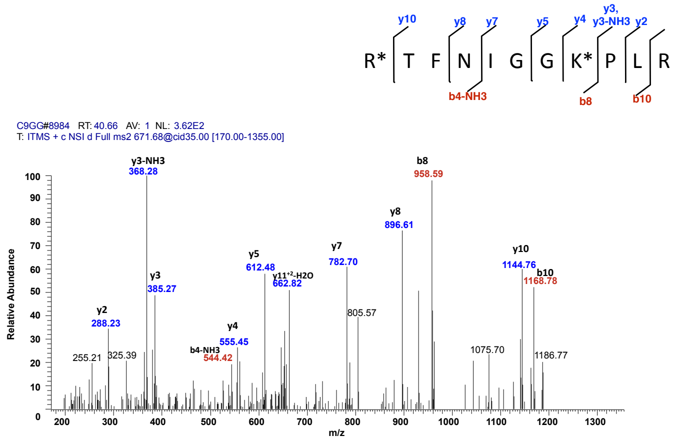
GC N-term peptide: M*SSPVK*VGR; z= +2; XCorr =2.90; Theo. [M+H]1+ = 1044.5294; [M+2H]2+ = 522.51 ; #PSMs = 2; * = acetyl



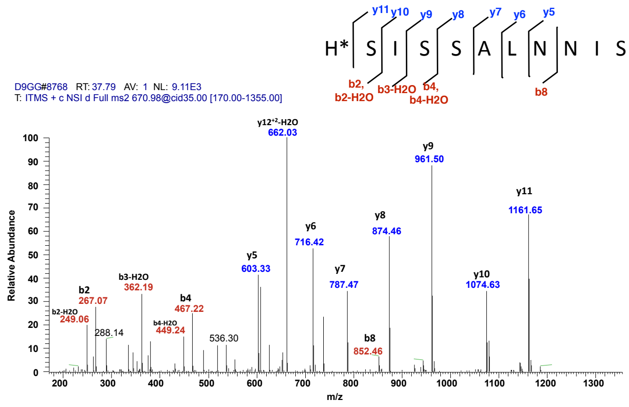
GD N-term peptide: S*ISAVDC!VGR; z= +2; XCorr = 3.67; Theo. [M+H]1+ = 1105.5200; [M+2H]2+ = 553.40; #PSMs = 31; * = acetyl; ! = carbamidomethyl



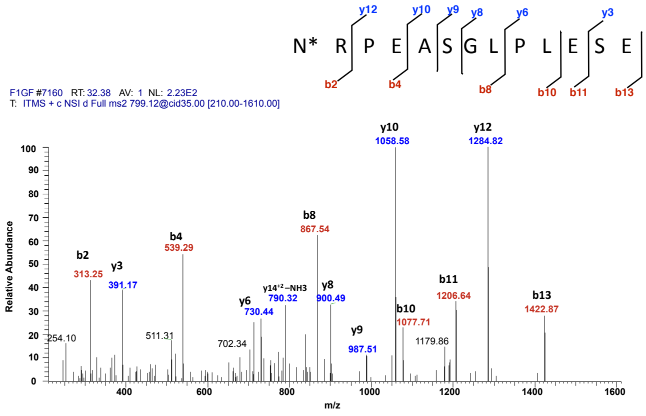
GE N-term peptide: R*TFNIGGK*PLR; z= +2; XCorr =2.65; Theo. [M+H]1+ = 1342.7378; [M+2H]2+ = 671.68 ; #PSMs = 4; * = acetyl



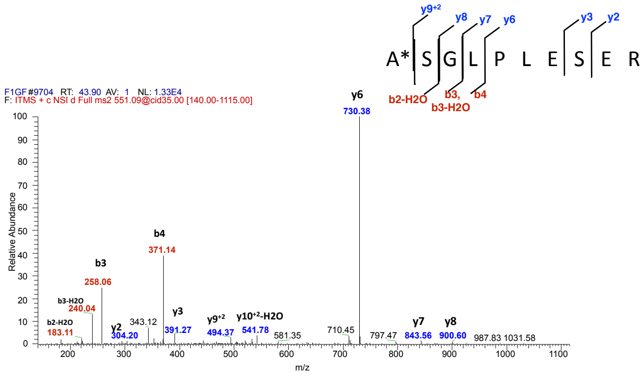
GF N-term peptide: H*SISSALNNISR; z= +2; XCorr =4.21; Theo. [M+H]1+ = 1340.6811; [M+2H]2+ = 670.98 ; #PSMs = 6; * = acetyl



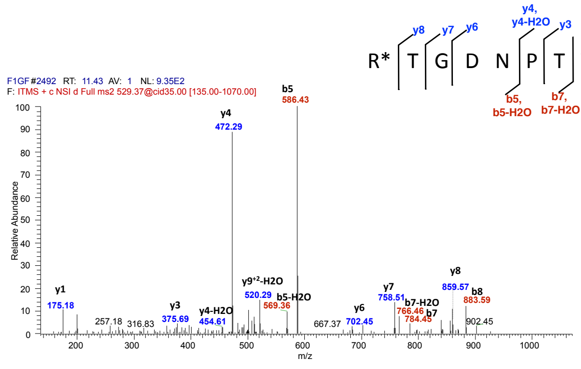
GH(i) N-term peptide: N*RPEASGLPLESER; z= +2; XCorr =3.68; Theo. [M+H]1+ = 1596.7870; [M+2H]2+ = 799.12 ; #PSMs = 2; * = acetyl



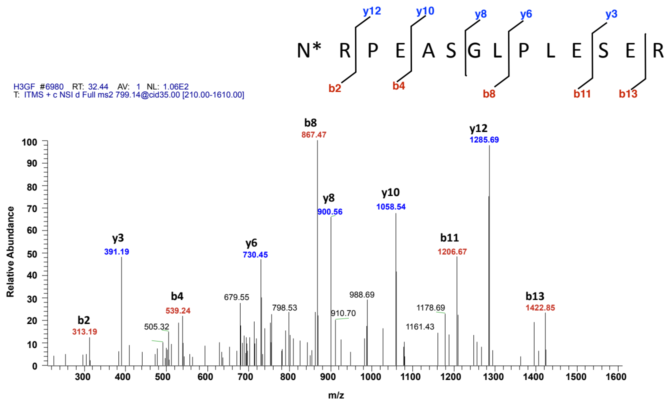
GH (ii) N-term peptide: A*SGLPLESER; z= +2; XCorr =2.47; Theo. [M+H]1+ = 1100.5476; [M+2H]2+ = 551.09 ; #PSMs = 13; * = acetyl



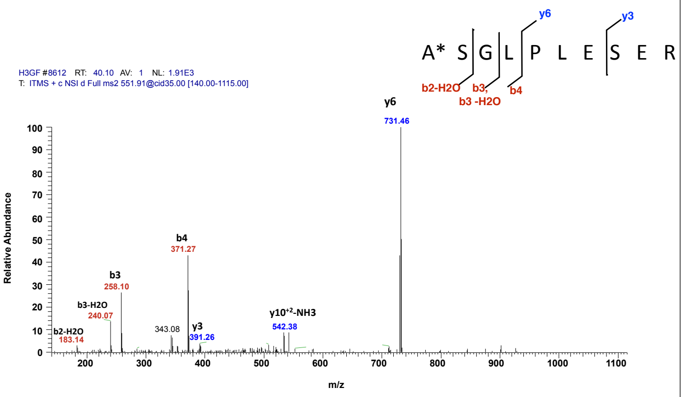
GH (iii) N-term peptide: R*TGDNPTVR; z= +2; XCorr =3.41; Theo. [M+H]1+ = 1057.5279; [M+2H]2+ = 529.37 ; #PSMs =5; * = acetyl



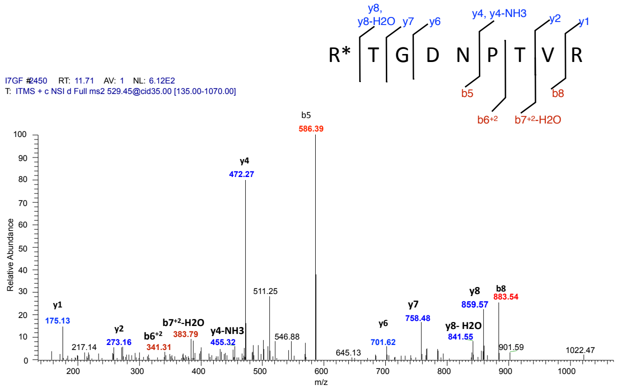
GJ (i) N-term peptide: N*RPEASGLPLESER; z= +2; XCorr =3.22; Theo. [M+H]1+ = 1596.7870; [M+2H]2+ = 799.14; # PSMs = 3; * = acetyl



GJ (ii) N-term peptide: A*SGLPLESER; z= +2; XCorr =2.53; Theo. [M+H]1+ = 1100.5476; [M+2H]2+ = 551.91 ; #PSMs = 10; * = acetyl



GJ (iii) N-term peptide: R*TGDNPTVR; z= +2; XCorr =2.83; Theo. [M+H]1+ = 1057.5279; [M+2H]2+ = 529.45 ; #PSMs =2; * = acetyl



GK N-term peptide: R*TFSTNR; z= +2; XCorr =3.02; Theo. [M+H]1+ = 923.4587; [M+2H]2+ = 462.42 ; #PSMs = 2; * = acetyl

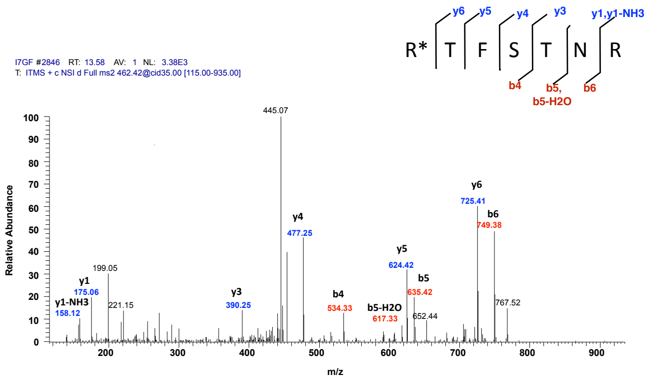


Figure S2A: MS/MS spectra for N-term end peptides of GluC bands

Band GA

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GB

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GC

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GD

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GE

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GF

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

red text + highlighted = N-terminal call for corresponding band
green only= gluc-trp; ESI run #1
green & italic = gluc_trp ; not in 1st run but in any of the following run of the band
green & underlined = glue as second digester
blue = common between two runs
orange = common among three runs

Band GH

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GI

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GJ

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

Band GK

```
MKKSNKENRP EASGLPLESE RTGDNPTVRG
SAGADPVGQD APGWTCQFCE RTFSTNRGLG
VHKRRAHPVE TNTDAAPMMV KRRWHGEEID
LLARTEARLL AERGQCSGGD LFGALPGFGR
TLEAIKGQRR REPYRALVQA HLARFGSQPG
PSSGGCSAEP DFRRASGAEE AGEERCAEDA
AAYDPSAVGQ MSPDAARVLS ELLEGAGRRR
ACRAMRPKTA GRRNDLHDDR TASAHKTSRQ
KRRAEYARVQ ELYKKCRSRA AAEVIDGACG
GVGHSLEEME TYWRPILERV SDAPGPTPEA
LHALGRAEWH GGNRDYTQLW KPISVEEIKA
SRFDWRTSPG PDGIRSGQWR AVPVHLKAEM
FNAWMARGEI PEILRQCRTV FVPKVERPGG
PGEYRPISIA SIPLRHFHSI LARRLLACCP
PDARQRGFIC ADGTLENSAV LDAVLGDSRK
KLRECHVAVL DFAKAFDTVS HEALVELLRL
RGMPEQFCGY IAHLYDTAST TLAVNNEMSS
PVKVGRGVRQ GDPLSPILFN VVMDLILASL
PERVGYRLEM ELVSALAYAD DLVLLAGSKV
GMQESISAVD CVGRQMGLRL NCRKSAVLSM
IPDGHRKKHH YLTERTFNIG GKPLRQVSCV
ERWRYLGVDF EASGCVTLEH SISSALNNIS
RAPLKPQQRL EILRAHLIPR FQHGFVLGNI
SDDRLRMLDV QIRKAVGQWL RLPADVPKAY
YHAAVQDGGL AIPSVRATIP DLIVRRFGGL
DSSPWSVARA AAKSDKIRKK LRWAWKQLRR
FSRVDSTTQR PSVRLFWREH LHASVDGREL
RESTRTPTST KWIRERCAQI TGRDFVQFVH
THINALPSRI RGSRGRRGGG ESSLTCRAGC
KVRETTAHIL QQCHRTHGGR ILRHNKIVSF
VAKAMEENKW TVELEPRLRT SVGLRKPDII
ASRDGVGVIV DVQVVSGQRS LDELHREKRN
KYGNHGELVE LVAGRLGLPK AECVRATSCT
ISWRGVWSLT SYKELRSIIG LREPTLQIVP
ILALRGSHMN WTRFNQMTSV MGGGVGIEGR
HHHHHH
```

red text + highlighted = N-terminal call for corresponding band
green only= gluc-trp; ESI run #1
green & italic = gluc_trp ; not in 1st run but in any of the following run of the band
green & underlined = glue as second digester
blue = common between two runs
orange = common among three runs

Number of ESI run of the bands:
Band GA: 5 ESIs. 3 gluctrp.2glucgluc. orange is common among 3 runs (2gluctrp (blue) &1glucgluc).
Band GC: 5 ESIs. 4 gluctrp.1glucgluc. orange is common among 3 runs (3gluctrp).
Band GE: 6 ESIs. 4 gluctrp.2glucgluc. orange is common among 3 runs (3gluctrp)
Band GF: 4 ESIs. 3 gluctrp.1glucgluc. blue is common among 2 runs (2gluctrp)
Band GH: 2 ESIs. 2 gluctrp 1 being enrichment experiment. blue is common among 2 runs (2gluctrp)
Band GI: 2 ESIs. 1 gluctrp 1 glucgluc. blue is common among 2 runs
Band GJ 2 ESIs. 1 gluctrp 1 glucgluc. blue is common among 2 runs

Figure S3A: Internal peptides for GluC bands

APPENDIX C

SUPPLEMENTARY INFORMATION ON MULTIPLE SEQUENCE ALIGNMENT

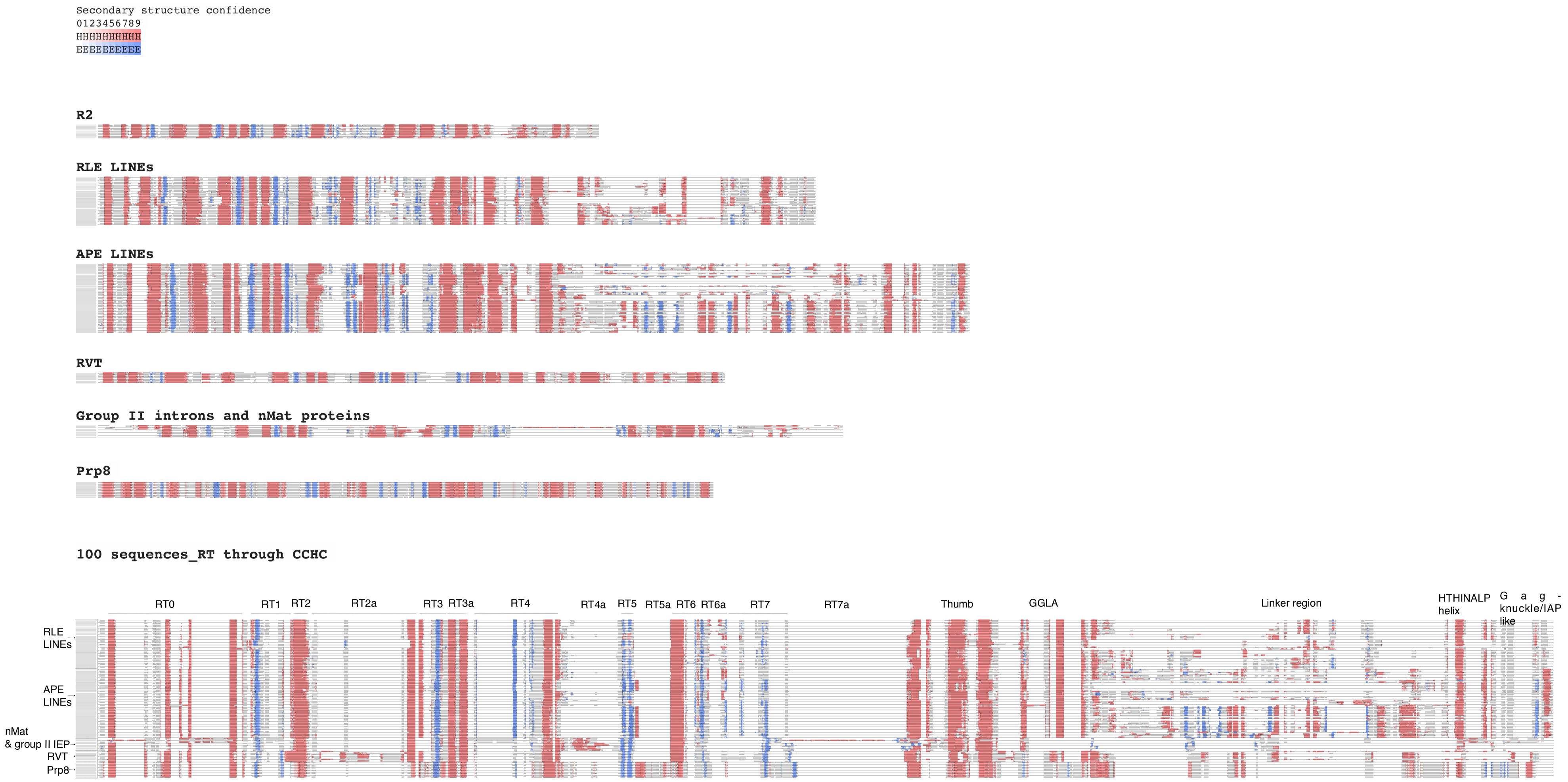


Figure S5: Multiple Sequence alignment (MSA) were generated in PROMALS3D server. Secondary structure information was scored on MSA using Ali2D tool of the MPI bioinformatic toolkits. The alignment begin at alpha helix of RT0 and do not include alpha helix of the index finger and alpha helix that traverses the palm.