

Ribonuclease H evolution in retrotransposable elements

H.S. Malik

Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (USA)

Manuscript received 12 January 2004; accepted in revised form for publication by J.-N. Voff 11 February 2004.

Abstract. Eukaryotic and prokaryotic genomes encode either Type I or Type II Ribonuclease H (RNH) which is important for processing RNA primers that prime DNA replication in almost all organisms. This review highlights the important role that Type I RNH plays in the life cycle of many retroelements, and its utility in tracing early events in retroelement evolution. Many retroelements utilize host genome-encoded RNH, but several lineages of retroelements, including some non-LTR retrotransposons and all LTR retrotransposons, encode their own RNH domains. Examination of these RNH domains suggests that all LTR retrotransposons acquired an enzymatically weak RNH domain that is missing an important catalytic residue found in all other RNH enzymes. We propose that this reduced activity is essential to ensure correct processing of the poly-

purine tract (PPT), which is an important step in the life cycle of these retrotransposons. Vertebrate retroviruses appear to have reacquired their RNH domains, which are catalytically more active, but their ancestral RNH domains (found in other LTR retrotransposons) have degenerated to give rise to the tether domains unique to vertebrate retroviruses. The tether domain may serve to control the more active RNH domain of vertebrate retroviruses. Phylogenetic analysis of the RNH domains is also useful to “date” the relative ages of LTR and non-LTR retroelements. It appears that all LTR retrotransposons are as old as, or younger than, the “youngest” lineages of non-LTR retroelements, suggesting that LTR retrotransposons arose late in eukaryotes.

Copyright © 2005 S. Karger AG, Basel

Retrotransposable elements are so defined because they possess a reverse transcriptase (RT) enzyme, responsible for copying genetic information from RNA to DNA. It is widely accepted that such an enzymatic activity must have been involved in the early transition from the RNA world to one in which genetic information was primarily inherited via DNA. The homology of reverse transcriptases to extant viral RNA-dependent RNA polymerases suggested their role as an evolutionary link between the RNA-dependent RNA polymerases and DNA-dependent DNA polymerases (Poch et al., 1989; Xiong and Eickbush, 1990).

This evolutionarily ancient transition from RNA to DNA may also have led to the origin of a very specialized nuclease activity. Priming DNA synthesis requires a 3'-hydroxyl group to initiate production of DNA. Several clever solutions to this problem exist in nature, but by far the most common one, settled upon by both RNA- and DNA-based genomes, was the use of short RNAs as primers. This required an enzymatic activity that would specifically “remove” the primer RNAs that ended up being covalently linked to the nascent DNA, in order to complete DNA replication. In most eukaryotes and prokaryotes, this is accomplished by two step-wise enzymatic activities. The first step could involve a nuclease activity, called the Ribonuclease H (RNH) activity, which removes most of the RNA primer (Sato et al., 2003) or a strand displacement activity that leads to a “single-stranded RNA flap” (Murante et al., 1998). The second step is carried out by a FEN-1 flap endonuclease activity that removes the terminal ribonucleotide (Murante et al., 1998). RNH activity thus emerged as a semi-strict requirement in RNA-primed DNA synthesis. In addition to its role in DNA replication and processing of Okazaki fragments,

Supported by the Helen Hay Whitney Foundation (postdoctoral fellowship in S. Henikoff's laboratory) and startup funds from the Fred Hutchinson Cancer Research Center.

Request reprints from Harmit S. Malik
Basic Sciences, Fred Hutchinson Cancer Research Center
1100 Fairview Avenue N. A1-162, Seattle, WA 98109-1024 (USA)
telephone: 1-206-667-5204; fax: 1-206-667-6497
e-mail: hsmalik@fhcrc.org

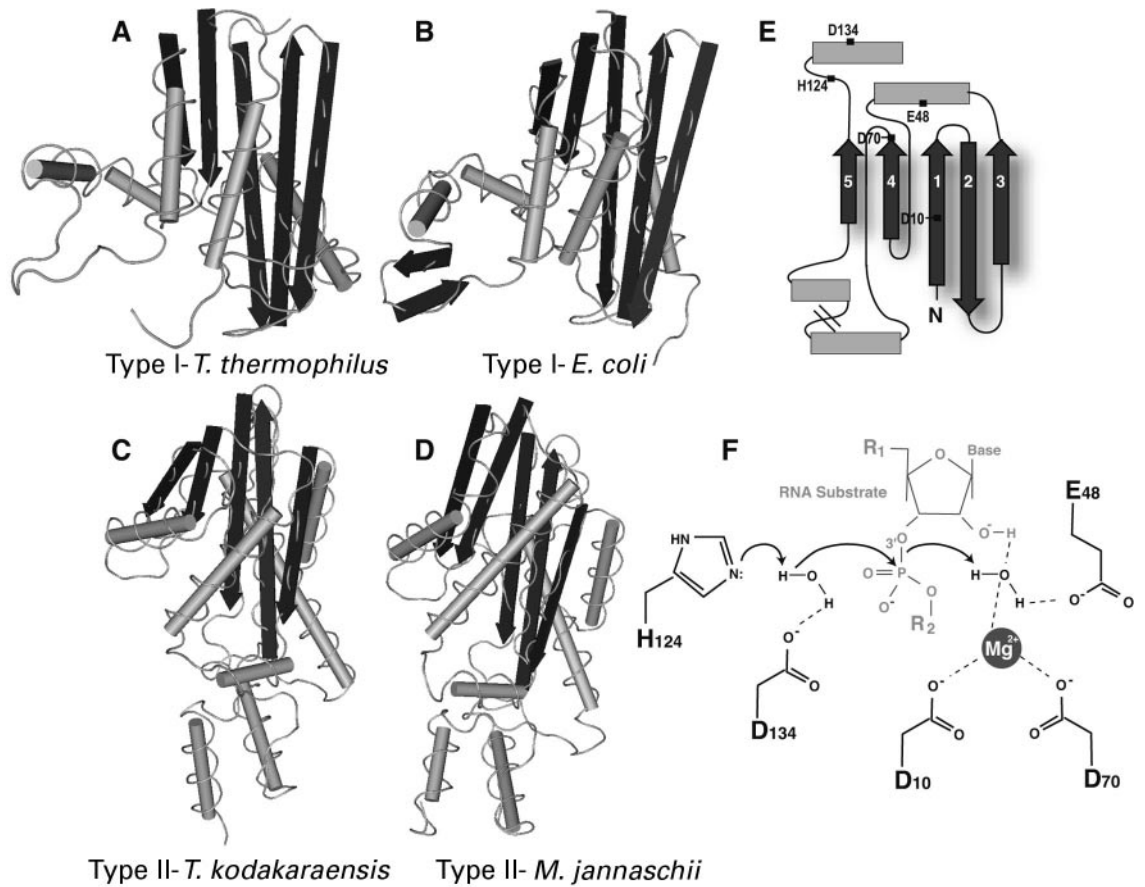


Fig. 1. Ribonuclease H structure and catalytic mechanism. Crystal structures of Type I RNH domains from (A) *Thermus thermophilus* (PDB: 1RIL) (Ishikawa et al., 1993) and (B) *E. coli* (PDB: 2RN2) (Katayanagi et al., 1990) compared to Type II RNH domains from (C) *Thermococcus kodakaraensis* (PDB: 1IO2) (Muroya et al., 2001) and (D) *Methanococcus jannaschii* (PDB: 1EKE) (Lai et al., 2000). Despite no obvious sequence similarities between Type I and Type II RNH domains, there is a great deal of structural similarity between them. Type II RNH domains have an extra C-terminal domain

characterized by three α -helices, that is believed to contribute to specificity of RNA:DNA hybrid recognition (Lai et al., 2000). (E) Schematized secondary structure of Type I RNH domains highlighting the typical 5 β strands, also indicating the location of the 5 catalytically important residues. (F) Proposed catalytic mechanism of Type I RNH domains (Kanaya et al., 1996). H124 initiates the nucleophilic attack on the water molecule that is responsible for cleaving the RNA backbone (gray).

RNH activity has also been implicated in choice of origins of replication (Hillenbrand and Staudenbauer, 1982; Dasgupta et al., 1987) and DNA repair (Arudchandran et al., 2000).

Since all RTs generate a template RNA:cDNA hybrid, all retroelements are influenced by RNH activity. However, because of variation in their method of priming DNA synthesis and other features of their life cycle, different retroelements vary in their dependence on RNH activity. Since most retroelements operate under constraints where their sizes are under strict selective constraints, their level of dependence on RNH activity is often reflected in whether they themselves encode RNH domains or instead depend on host genome-encoded enzymes. Nonetheless, many retroelements encode RNH domains that are an essential adjunct to their RT domains. By virtue of their abundance and evolutionary conservation, RNH domains provide us a second opportunity (after RT) to examine the details of the evolution of these retrotransposable elements (Malik and Eickbush, 2001). In several respects, the RNH analysis turns out to be more informative than that of the

RT, detailing an ancient chronology of events. Its history includes an ancient duplication in vertebrate retroviruses, and suggests a simple model for the evolutionary origin of long-terminal repeat (LTR) bearing retrotransposable elements.

RNH enzymes encoded by prokaryotic and eukaryotic genomes

Three evolutionarily distinct lineages of cellular RNH enzymes have been identified as a result of detailed studies (Ohtani et al., 1999). Type I or RNase HI (rnhA gene) enzymes constitute a lineage found in many Eubacteria, all Eukarya but not Archaea (the only exception being *Halobacterium* sp. NRC-1, Accession no. NP_279371.1). Type II enzymes consist of RNase HII (rnhB) and HIII (rnhC), which are homologous enzymes (Kanaya, 2001). RNase HII can be found in Archaea, Eubacteria and all Eukarya, while RNase HIII appears only in some Eubacteria. In eukaryotes and all Archaea, RNase HII

enzymes may constitute the bulk of all RNH activity, while the reverse is true in Eubacteria like *E. coli* where RNase HI is the major source of RNH activity. Type I enzymes are structurally similar to the N-terminal domain of Type II (Fig. 1A–D), suggesting common evolutionary ancestry although no primary sequence similarity can be discerned (Lai et al., 2000; Muroya et al., 2001). Biochemically, Type I enzymes are distinct from Type II, differing in their preferences of divalent cations for enzymatic activity (Katayanagi et al., 1993; Goedken et al., 2000).

All Ribonuclease H proteins adopt a similar fold that is typified by a mixed α helix β strand structure where the β strands adopt a characteristic 3-2-1-4-5 sheet structure (strands are numbered from N- to C-terminus) with the second β strand anti-parallel to the rest (Fig. 1E). In addition to this typical secondary structure, three carboxylates (D10, E48 and D70) that comprise the active site are arranged in identical fashion relative to the β -sheet (Fig. 1E) (Kanaya et al., 1996). Other proteins that adopt a similar fold (and belong to the structurally classified RNH superfamily) also appear to encode a similar suite of enzymatic activities as RNH. These proteins include the catalytic integrase/ transposase domains from most LTR-retrotransposons and DNA-mediated elements (Mu, Tn5), 3'–5' exonucleases (like DnaQ) and resolvases (RuvC, mitochondria) (Yang and Steitz, 1995).

The Type I RNH from *E. coli* (rnhA) remains the best studied. Previous studies have highlighted five catalytically important residues – D10, E48, D70 (that form the carboxylate triad typical of the RNH superfamily), H124 and D134. Of these, the four carboxylates are essential for RNH activity, and are highly conserved across all Type I RNH domains (Malik and Eickbush, 2001). In the proposed catalytic mechanism of RNH activity, H124 initiates the nucleophilic attack on the water molecule that will then attack the phosphate backbone of the RNA (Fig. 1F) (Kanaya et al., 1996). Consistent with this important role, an H124A substitution in the *E. coli* enzyme resulted in a large drop in kcat/Km (Kanaya et al., 1990), suggesting that loss of this histidine residue would result in severely impaired enzymatic activity.

From the perspective of their role in retroelement biology, Type I RNH enzymes have been found associated with the life cycle of a variety of retroelements (below) but Type II enzymes have not. This dichotomy is particularly noteworthy considering that retroelements, in general, have proven to be much less successful in Archaea (that typically lack Type I RNH) compared to Eubacteria. This has been attributed to the extreme environments that Archaea populate, that may inhibit reverse transcriptase activity. But several Archaea populate mesophilic environments and several Eubacteria are found in hyperthermophilic environments. Second, horizontal transfer can occur quite rapidly between different prokaryotic lineages in similar environments. Finally, Archaea can clearly possess retroelements, just less successfully than Eubacteria (Rest and Mindell, 2003). If RNH, more specifically Type I RNH, activity is crucial for retroelement biology, it is to be expected that the almost complete absence of Type I RNH enzymes in Archaea may have contributed to the detriment of retroelement propagation in Archaea.

The role of RNH activity in retroelement biology

Group II introns are mobile self-splicing introns found in eubacterial, archaeal and organellar genomes that often encode an RT activity and employ a reverse-transcription coupled mechanism to increase their copy number. **Non-LTR retroposons** are found in most eukaryotic genomes and mobilize by a similar mechanism of target-primed reverse transcription (TPRT, Fig. 2). Both Group II introns and non-LTR retroposons likely employ RNH activity at a similar stage of their life cycle (Belfort et al., 2002). In both cases, TPRT is initiated by a retroelement encoded nicking endonuclease which exposes a 3'-OH (Luan et al., 1993; Cousineau et al., 1998). This 3'-OH is then used to prime reverse transcription of the cDNA using either the reverse-spliced intron or the non-LTR retroposon RNA as a template. Plus strand DNA synthesis of Group II introns is accomplished either by a continued reverse transcription of the fully integrated intron into the upstream exon (recombination-independent) or by strand invasion of newly formed cDNA off the unspliced message RNA (recombination-dependent). The situation is less clear in the case of the non-LTR retroposons but may involve a template switch mechanism from RNA to DNA upstream of the insertion site. In either case, for second strand synthesis of DNA to occur, the template RNA must be displaced or digested. It is hard to imagine long stretches of RNA:cDNA hybrids not being attacked by RNH during first-strand synthesis. This suggests that while protein components may retain catalytic activity, the number of new insertions made may depend stoichiometrically on the number of template RNAs produced by transcription. Studies have not evaluated the integrity of the template RNA at the completion of first-strand DNA synthesis and it is unclear whether RNH activity is required for completion of TPRT, or whether a helicase activity that displaces the template RNA can suffice for this purpose.

One insight into RNH requirement at least for non-LTR retroposons comes from the observation that several elements encode RNH domains, just downstream of their RT (Malik et al., 1999). The RNH domains borne by these elements appear to be catalytically active (Olivares et al., 2002), and appear to participate in their life cycle. However, the appearance of RNH domains in non-LTR retroposons appears to have been a late event in the evolution of this lineage of retroelements, and it is clear that some elements may have lost their RNH domains (Eickbush and Malik, 2002; Malik et al., 1999). This suggests that the non-LTR retroposons rely extensively on host genome encoded RNH activity. Since their TPRT mode of mobilization involves reverse transcription at the future insertion site in genomic DNA, there is likely to be ready access to host encoded RNH activity (Malik and Eickbush, 2001). Thus, there may only be a small selective advantage conferred by harboring self-encoded RNH (offset by the slightly higher “cost” of replication). By analogy, Group II introns have similar access and may similarly depend on host encoded RNH activity.

LTR retrotransposons represent a monophyletic (on the basis of RT) group of eukaryotic retroelements whose prototypic members were characterized as having long terminal direct repeats. They can be classified on the basis of RT phylogeny

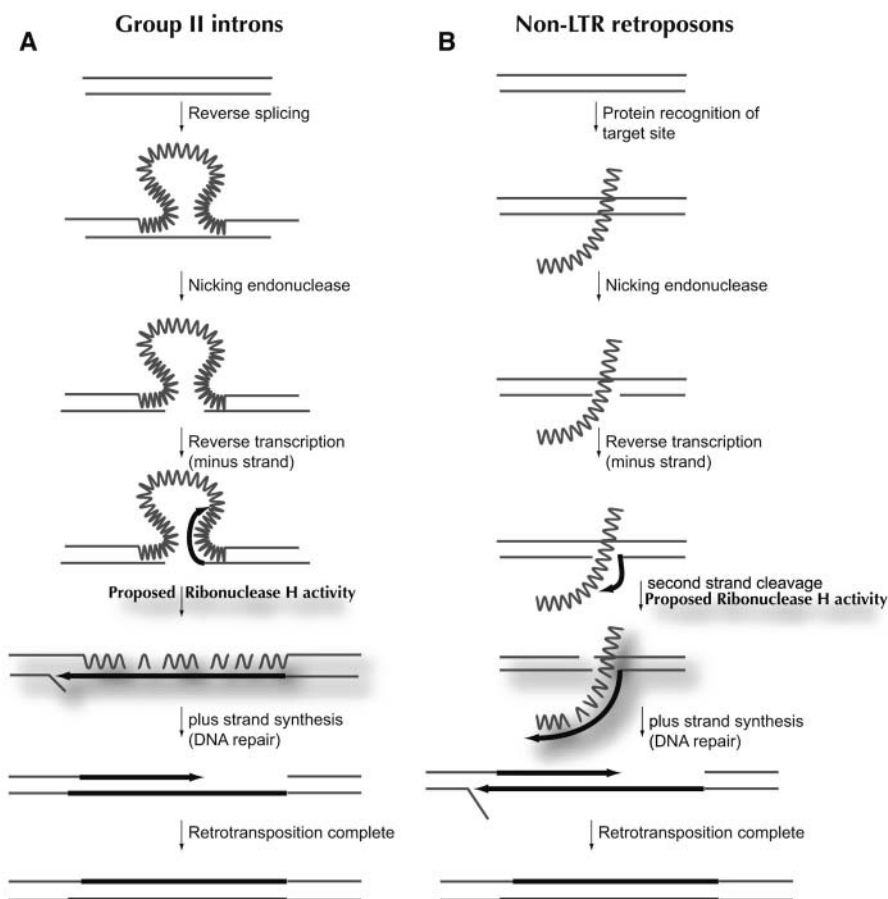


Fig. 2. Target-primed reverse transcription (TPRT) mechanism of Group II introns and non-LTR retroposons. **(A)** Group II intron retrotransposition in bacteria initiates by the reverse splicing of the intron into the plus strand of the target site followed by nicking the minus strand, exposing the 3'-OH that is used to prime minus strand cDNA synthesis into the upstream region of the target site (Cousineau et al., 1998). A proposed RNH activity then allows for priming plus strand synthesis by DNA repair enzymes. (An alternative would be plus strand nicking and displacement of the RNA by DNA synthesis of the plus strand.) A similar mechanism adopted by yeast mitochondrial group II introns also involves recombination with the donor element (not shown). **(B)** Like in **A**, R2 non-LTR retrotransposition initiates by nicking the minus strand, exposing a 3'-OH that primes minus strand synthesis (Luan et al., 1993). In the case of R2, a second strand cleavage does occur (Eickbush, 2002) so it is unclear whether RNH activity is necessary, as strand displacement would suffice to complete plus-strand synthesis. However, many non-LTR retroposons harbor their own RNH domains (Malik et al., 1999) indicating a dependence on RNH activity.

and ORF features into seven groups: the Ty1/copia group, the hepadnaviruses (Hepatitis B Virus), the DIRS1 group (that contain a tyrosine recombinase instead of integrase) (Goodwin and Poulter, 2001), the BEL group, the Ty3/ gypsy group, plant caulimoviruses and vertebrate retroviruses (Malik et al., 2000). Not all these lineages harbor LTRs, but most encode a core set of enzymatic activities. These include a protease (PR), RT, RNH and an integrase (IN), except DIRS1 and hepadnaviruses (which do not encode either PR or IN).

Prototypic LTR retrotransposons rely on an elaborate sequence of events involving their RT, RNH and LTRs to ensure synthesis of double stranded DNA starting from template RNA synthesis. This is illustrated in Fig. 3. The RNH activity is responsible for both the degradation of the template RNA and the release of the polypurine tract (PPT) that is particularly resistant to RNH activity and acts as the primer for synthesis of the plus-strand of the LTR retrotransposon. By virtue of its activity and its ability to generate the PPT, RNH defines the edges of the LTRs in the daughter elements, and strongly influences the ability to generate the double-stranded DNA intermediate, crucial to the life cycle of LTR retrotransposons. Most of this life cycle of LTR retrotransposons is carried out in the cytoplasm of eukaryotic cells or in virus-like particles, while the host RNH activity is restricted to the eukaryotic nucleus and organelles. Thus, among retroelements, LTR retrotransposons rely most heavily on RNH activity, but have least

access to host genome encoded RNH. It is thus imperative for LTR retrotransposons to possess their own RNH domains, and this is reflected in the high conservation of RNH domains in these elements.

Retrons are retroelements that are responsible for production of multicopy single stranded DNA (msDNA) in a number of eubacterial lineages, including myxobacteria and enteric bacteria, and at least one Archaeal genome (Yamanaka, 2002; Rest and Mindell, 2003). While the biological significance of msDNA is not clear, it is presumed that a selective advantage is conferred by msDNA to its carrier genome since the reverse transcriptase activity does not contribute to an increase in copy number of the retron itself. Instead, the retron-encoded reverse transcriptase utilizes an associated but exogenous RNA as template to produce the msDNA, sometimes at high copy number. This is analogous to the case of the eukaryotic telomerase, which also utilizes an exogenous RNA template to heal the ends of linear chromosomes, but does not increase its own copy number.

msDNA production proceeds by transcription of three elements, msr that encodes the mature RNA component, msd that will contribute to mature DNA component of msDNA and the open reading frame (ORF) encoding RT (Fig. 4) (Yamanaka, 2002). The cDNA synthesis terminates at a specific position for each retron, leaving a cDNA:template RNA hybrid, composed of the msd:msr regions respectively, that represents mature

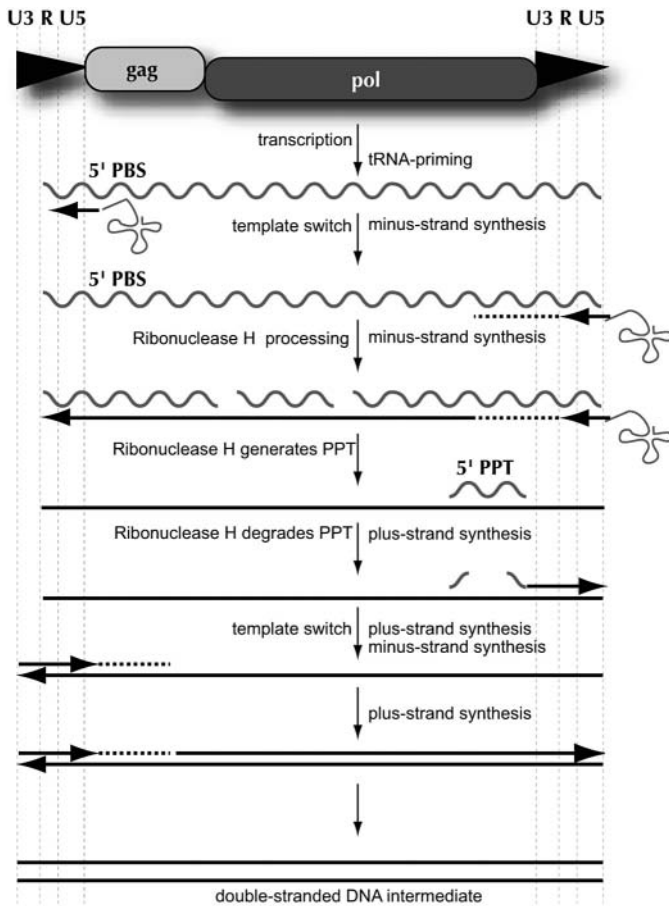


Fig. 3. LTR retrotransposition mechanism of Ty3. Long terminal repeats (LTRs) are composed of three distinct segments – U3, R, and U5. Transcription initiates in the 5' R segment, and proceeds through to the 3' U5. A tRNA molecule then hybridizes at its 3' end with the primer binding site (PBS) downstream of the 5' LTR. The 3' end of the tRNA molecule provides the 3'-OH required for initiating reverse transcription (note that different variations for initiating this priming exist in LTR retrotransposons) that proceeds through to the 5' end of the transcript, ending in U5-R. The newly synthesized minus strand then switches templates by virtue of the direct homology between the R and U5 regions to the 3' end of the transcript, and primes the reverse transcription of the minus strand. The ribonuclease H processes the RNA template, exposing a Polypurine Tract (PPT) that is resistant to RNH activity. This PPT then primes replication of the plus strand, which after template switching then leads to the double-stranded DNA intermediate, which is the hallmark of both DNA-mediated transposons and most LTR retrotransposons. An integrase/transposase then integrates this intermediate into genomic DNA (not shown).

msDNA. Thus, RNH activity is believed to play a vital role in the completion of the msDNA synthesis. Secondly, because of RNH activity, the number of mature msDNA molecules becomes stoichiometrically dependent on the number of msr-msd transcripts made (one transcript cannot be used to make several msDNA).

Genetic screens for insertion mutations in the *E. coli* genome that led to defects in msDNA synthesis only recovered three mutations, all in the *rnhA* gene that encodes Type I RNH activity (Lima and Lim, 1995). Under wildtype conditions, msDNA produces large amounts of homogeneous reverse tran-

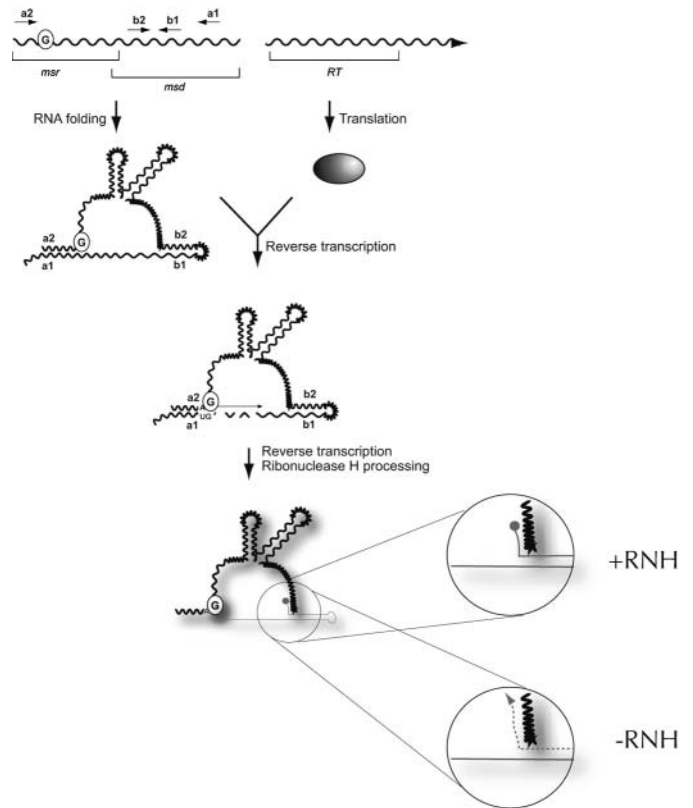


Fig. 4. RNH activity in msDNA production by retrons (Yamanaka, 2002). Retrcons consist of three components – the msr, msd (on the same transcript) and the ORF encoding RT (often on the same transcript). The msr-msd transcript consists of two pairs of highly conserved nested short, inverted repeats – a2, b2 and b1, a1. On folding of this transcript, the a1-a2 RNA stem abuts a highly conserved AG* dinucleotide at the end of a2 versus GU in a1. The 2'-OH of the G* branching residue (shown circled) is then used to prime reverse transcription (a biological novelty) that proceeds to make a DNA copy of the msd region, while the template RNA is processed by RNH activity. Loss of RNH activity results in premature stop of reverse transcription as well as reverse transcription beyond the “stop site” (gray dot) (Lima and Lim, 1995; Shimamoto et al., 1995).

scription products, but defects in Type I RNH activity (but not Type II RNH) resulted in both smaller and larger products (Lima and Lim, 1995; Shimamoto et al., 1995). This implies that lack of RNH activity contributed both to premature termination of reverse transcription (perhaps because of torsion introduced by a larger RNA:cDNA hybrid) and failure to arrest reverse transcription at the “stop site” (Fig. 2). In some instances, reverse transcription proceeds through the “stop site” all the way up to the branched G residue that primed the RT (Shimamoto et al., 1995). Since the biological function of msDNA is not elucidated, it is unclear what effect this read-through would have on “function” of the msDNA.

Its strong reliance on RNH activity might suggest that retrons should harbor their own RNH domains (with RT).

However, there is only one report of an RNH-containing retron, Ec67, where a C-terminal extension to the ORF encoding RT has been suggested to be an RNH (Lampson et al., 1989). However, careful analyses of this domain do not support the conclusion that Ec67 encodes an RNH domain of either Type I or Type II. Retrons are often found in close proximity to prophages in bacterial genomes, and might mobilize along with them. This would place severe constraints on the size of retrons, perhaps precluding them from encoding their own RNH. Unlike eukaryotes, there is relatively little partitioning of cellular processes in prokaryotes, which may allow ready access of retrons to host genome-encoded RNH activity, obviating the need for retrons to encode their own RNH. Thus, host genomes that do not encode Type I RNH are more likely to be deficient in the biosynthesis of msDNA, and less likely to reap any “benefits” from it.

Retroplasmids are enigmatic retroelements found in fungal mitochondria. These include the Mauriceville and Varkud plasmids in *Neurospora* species (Akins et al., 1986). These are found in both circular and linear forms, and their RT is phylogenetically closely related to group II introns, although their method of priming reverse transcription and other features of life cycle make them distinct from all other lineages of retroelements (Wang et al., 1992). To a first approximation, however, the reliance of retroplasmids on RNH activity could be considered analogous to the TPRT-dependent retroelements, except that this RNH activity needs to be present in mitochondria (Wang and Lambowitz, 1993).

Telomerase is a eukaryotic specific reverse transcriptase that is essential for replenishing of linear chromosomes, which have no other means to recover DNA lost due to “lagging-strand” RNA-primed DNA replication at their ends. Like retrons, RT activity of telomerase does not increase its copy number. Instead, it uses an exogenous RNA (Singer and Gottschling, 1994) to add short oligonucleotides to the ends of chromosomes. Ironically, RNH activity would be considered counterproductive to telomerase function, since it would destroy template RNA that is paired with newly formed telomeric DNA. Since these hybrids are typically short (6–8 nt) they may be resistant to RNH action, or RNH may be sequestered away from telomeres in order to “protect” the RNA template.

Penelope retroelements were first discovered in *Drosophila virilis* (Pyatkov et al., 2002), but are now known to be widespread in a number of eukaryotic lineages. While clearly possessing an RT activity, they are unique in encoding an endonuclease related to the UvrC proteins involved in DNA repair in bacteria (Lyozin et al., 2001). Amazingly, Penelope elements clearly harbor introns (Arkhipova et al., 2003), sometimes in the middle of their ORFs (ability to lose introns is one of the hallmarks of retrotransposition). Thus, it is not even clear whether the Penelope elements mobilize primarily using an RNA intermediate, so the potential effect of RNH activity is unclear. Based on the other retroelements, if RNH activity is required for retrotransposition, and since Penelope elements do not encode RNH, this might imply that Penelope elements transpose by a mechanism more closely related to the TPRT elements, i.e. at the genomic DNA rather than away from genomic DNA, like LTR retrotransposons.

A note about methodology

Bioinformatic techniques have advanced quickly in the last ten years. The use of position-specific scoring matrices (PSSMs) (Henikoff and Henikoff, 1996) in exhaustive iterative searches (PSI-BLAST) (Altschul et al., 1997; Altschul and Koonin, 1998) have greatly reduced the possibility of missing primary sequence-based homology. Despite these advances, structural homologies sometimes cannot be recapitulated using sequence-based techniques. For instance, despite the obvious structural similarities between Type I and Type II RNH domains (Fig. 1) (Lai et al., 2000), we cannot assign sequence homology between them. Nonetheless, the use of hand-made alignments and hidden Markov-based matrices (HMMs) with very low gap penalties in the early stages of retroelement biology may have led to unacceptably high levels of false positives, which do not stand the rigor of current methodology. For instance, earlier assignments of RNH domains to the Ec67 retron (Lampson et al., 1989) and to non-LTR retrotransposons R2, Line1 and Cin4 (McClure, 1991) are not justifiable based on sequence alone. In some cases, the assignment is clearly incorrect based on detailed biochemistry (Yang et al., 1999) and is sometimes not even consistent in different iterations of the same method (McClure, 1991; McClure et al., 2002). We have employed a relatively conservative approach in assigning RNH function that is consistent with known methodology. It warrants mentioning that iterative searches are sufficient to elucidate homology of all Type I RNH domains in the database. These approaches may lower the possibility of finding extremely remote homology, but help eliminate the possibility of false matches.

RNH evolution in retroelements

Type I RNH domains have been found in all Eukarya, one Archaea, many Eubacteria, a few non-LTR retrotransposons and all LTR retrotransposons. In spite of RNH domains being limited to only two groups of retroelements, their evolution within these two groups is nonetheless insightful in reconstructing early events in the evolution of eukaryotic retroelements.

Non-LTR retroposon RNH domains possess all five of the important catalytic residues (Malik et al., 1999) and at least one of the element-borne RNH domains has been shown to be catalytically active (Olivares et al., 2002). RNH domains were acquired relatively late in non-LTR retroposon evolution, most likely from an early eukaryotic RNH. This is schematized in the reconstructed chronology of non-LTR ORF evolution (Fig. 5) based on phylogenetic analysis of all the enzymatic domains (Malik et al., 1999). The oldest lineages of non-LTR retrotransposons contain a site-specific endonuclease, while later lineages have acquired an apurinic endonuclease that is largely non site-specific. Only the latter third of non-LTR lineages possess an RNH domain, and even among these, there is variable retention, possibly reflective of the dependence of non-LTR elements on host encoded RNH activity (above). Non-LTR elements rarely undergo horizontal transfers and their phylogenies are largely congruent to those of their host genomes. Their

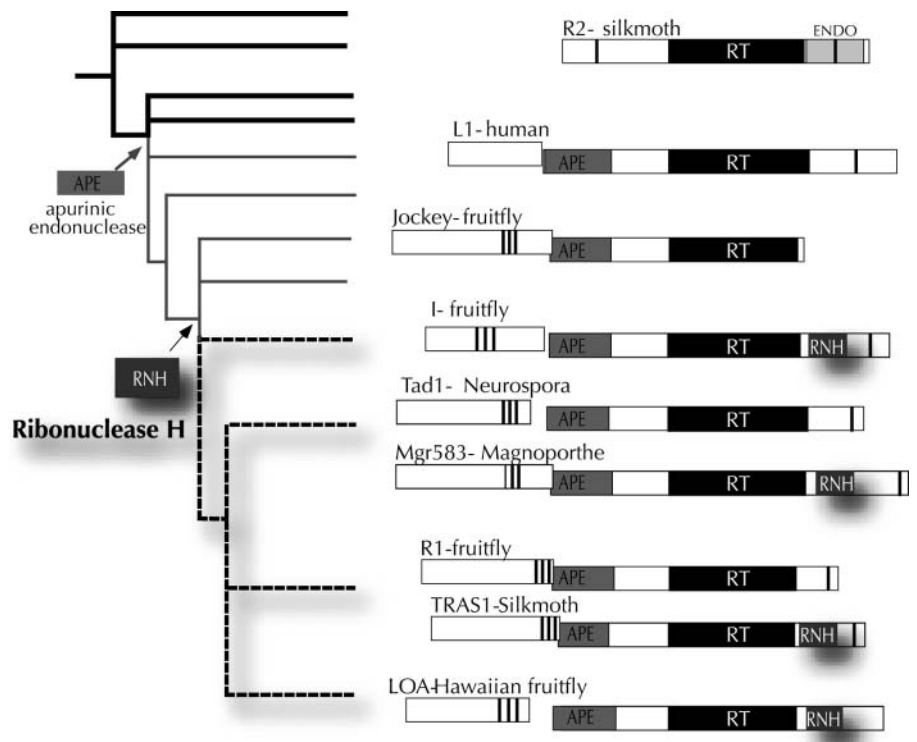


Fig. 5. Schematic representation of non-LTR retroposon evolution (Malik et al., 1999; Eickbush and Malik, 2002). Based on a composite phylogenetic analysis of the RT domain (common to all non-LTR retroposons) and additional enzymatic domains found only in some lineages, we can reconstruct the chronology of domain acquisition and loss in non-LTR retroposons. The earliest branching non-LTR elements (bold black lines) had a single ORF with a central RT and a C-terminal site-specific endonuclease (ENDO) domain. Subsequently, non-LTR elements acquired a largely non site specific endonuclease (APE) related to the apurinic endonucleases involved in DNA repair, upstream of the central RT, along with a first ORF (gray lines). Later lineages (dashed black lines) of non-LTR elements also acquired an RNH domain, but there has been variable retention of RNH in these lineages (Tad1 vs Mgr583, R1 vs TRAS etc.).

abundance in some of the earliest branching eukaryotic genomes (Burke et al., 2002) suggests that they have been vertically inherited since the origin of Eukarya (Eickbush and Malik, 2002; Malik et al., 1999).

Alignment of RNH domains from LTR retrotransposons revealed the surprising lack of conservation of the catalytically important histidine residue corresponding to the H124 residue in *E. coli* (Fig. 1F) (Malik and Eickbush, 2001). All LTR retrotransposon lineages, except vertebrate retroviruses appear to have lost this residue, which might suggest an impaired catalytic activity for them. Phylogenetic reconstructions of the LTR retrotransposons based on either the RT or RNH domains find them to be in remarkable agreement with each other (Fig. 6) except for the vertebrate retroviruses that branch much earlier than all other LTR retrotransposons in the RNH phylogeny (Malik and Eickbush, 2001; Eickbush and Malik, 2002). Phylogenetic reconstructions of other catalytic domains and general life cycle features strongly suggest the close relationship of vertebrate retroviruses to the Ty3/ gypsy group (Eickbush, 1999; Malik and Eickbush, 1999). Thus, overall it appears that the vertebrate retroviruses have acquired an older, more catalytically active RNH domain (by virtue of retaining the H124 residue) most likely from a non-LTR retroposon.

Comparison of the ORF features of the LTR retrotransposons suggested another piece of evidence pointing towards a secondary RNH acquisition in vertebrate retroviruses. In non-LTR retroposons and most LTR retrotransposons, the RT and RNH domains (boundaries defined by similarity to other elements) directly abut each other, but this is not the case for any of the vertebrate retroviruses. Instead, they each have a variable length linker domain between the RT and RNH domains

that has been referred to as the **tether or connection domain**. The structure of the tether domain from HIV-1 is strongly reminiscent of the core backbone of the RNH superfamily (Fig. 7A), including the characteristic 3-2-1-4-5 β sheet with strand 2 anti-parallel to the rest (Kohlstaedt et al., 1992). There is no primary sequence similarity to suggest that the tether is homologous to RNH domains, and none of the catalytically important residues are conserved, but the structures are homologous and core backbones of the five β strands and the single α helix of the tether can be superimposed onto the HIV RNH structure with an RMS deviation of only 1.77 Å over 48 Ca atoms (Artymiuk et al., 1993).

This allows us to propose the following parsimonious chronology for the evolution of RNH domains in retroelements (Fig. 7B) (Malik and Eickbush, 2001). The first retroelements to acquire RNH domains were late branching lineages of non-LTR retroposons. The common ancestor to all LTR retrotransposons then acquired RNH domains as part of an RT-RNH pair but lost one of the catalytically important histidine (H124) residues. This RNH domain was then inherited by all subsequent lineages of LTR retrotransposons, including vertebrate retroviruses. However, in a secondary acquisition (or several secondary acquisitions), vertebrate retroviruses acquired an H124-containing and presumably more active RNH domain, most likely from non-LTR retroposons. The ancestrally inherited (H124 lacking) RNH domain was then under relaxed constraints to maintain catalytic activity, and degenerated in primary sequence but maintained structural information to retain the characteristic structure of an RNH.

Why is the tether domain still retained in vertebrate retroviruses? It is clear from studies in HIV-1 that the tether domain is

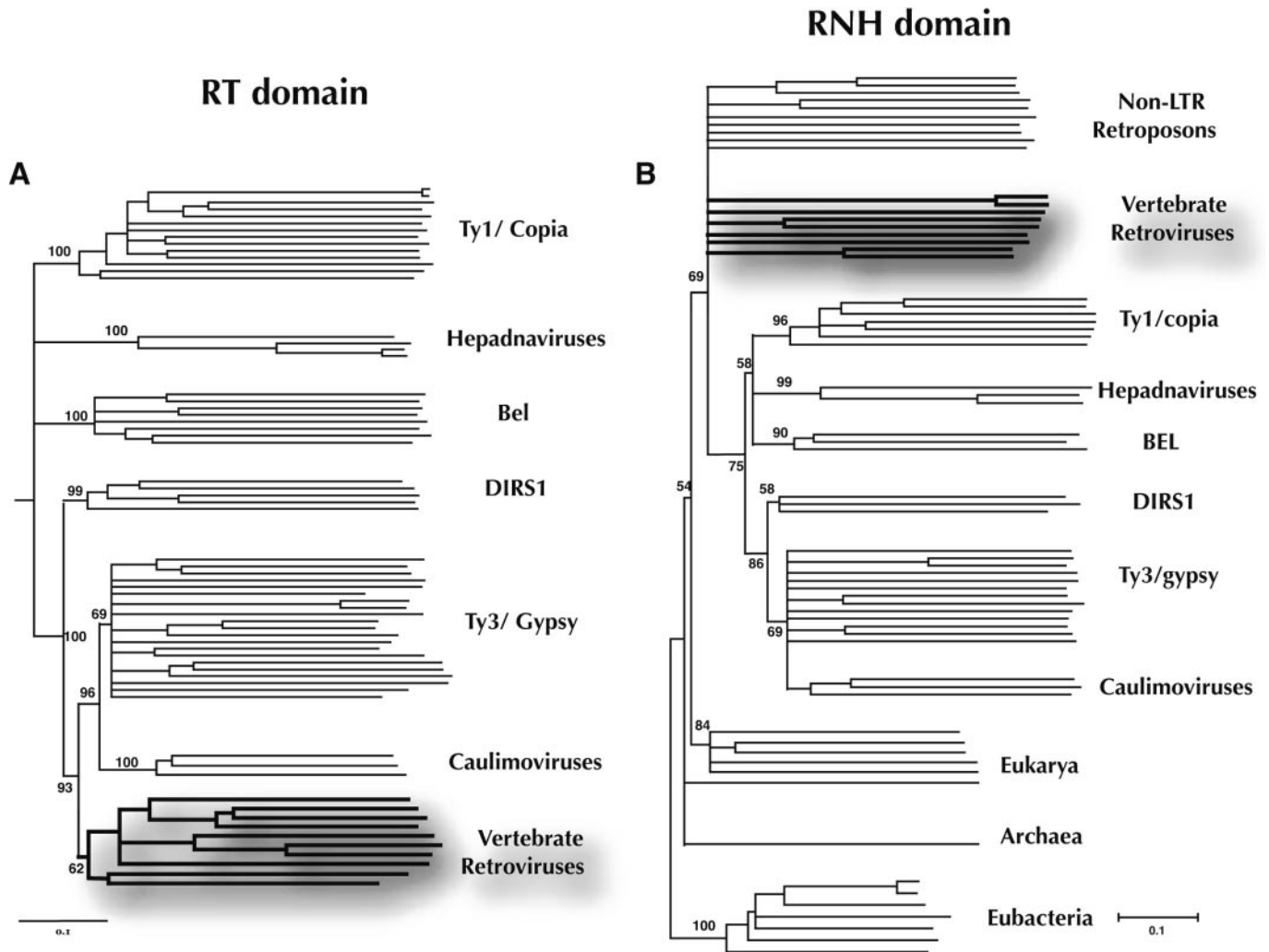


Fig. 6. Comparison of RT and RNH phylogenies of the LTR retrotransposons (Malik and Eickbush, 2001). **(A)** Neighbor-Joining analysis of RT domain provides good resolution within the LTR element lineage, with the Ty1/Copia, Hepadnaviruses and BEL lineages branching ancestrally, and the Ty3/gypsy group, plant Caulimoviruses and vertebrate retroviruses forming a well supported clade. Bootstrap support is indicated adjacent to the main nodes. **(B)** The RNH phylogeny rooted on eubacterial RNH domains,

shows that archaeal (one representative) and eukaryotic lineages branched ancestrally to all retroelements, with the non-LTR retroposons and vertebrate retroviruses being the earliest to branch. The rest of the LTR retrotransposons phylogeny is in good agreement with **A**, suggesting that the only incongruence can be explained by vertebrate retroviruses having acquired a different lineage of RNH domains.

involved in the heterodimerization of the p66 and p51 subunit in the active heterodimer (note that the p51 subunit lacks the RNH domain, but retains the tether) and in modulating RNH activity (Kohlstaedt et al., 1992). The retention of the tether domain in vertebrate retroviruses may be consistent with the model of subfunctionalization (Force et al., 1999) after gene duplication (or in this case acquisition) where it is likely that the tether domains are now performing a structural role while the “new” RNH domains are only involved with enzymatic function (Fig. 7C). One of these “structural roles” may involve organizing the RT-bearing heterodimer.

However, the finding that LTR retrotransposons chose to acquire a less active RNH domain early in their evolution raises another intriguing possibility. As evident from the life

cycle of LTR retrotransposons (Fig. 4), the precise generation of a PPT is crucial for the priming of the plus-strand and thereby to the survival of the LTR retrotransposon. A hyperactive RNH might result in the premature or incorrect processing of the PPT, which would be lethal to the element. Consistent with this, the introduction of the *E. coli* RNH domain is inhibitory to the retrotransposition of LTR retrotransposons including vertebrate retroviruses (Ma and Crouch, 1996). On the other hand, there appears to be only a minimal cost associated with a drop in RNH activity since it does not appear to be rate-limiting (Sevilya et al., 2003). When vertebrate retroviruses acquired an active RNH domain, perhaps the tether domain was retained to modulate the activity of the newly acquired RNH perhaps by affecting its affinity for RNA:DNA hybrids, so that

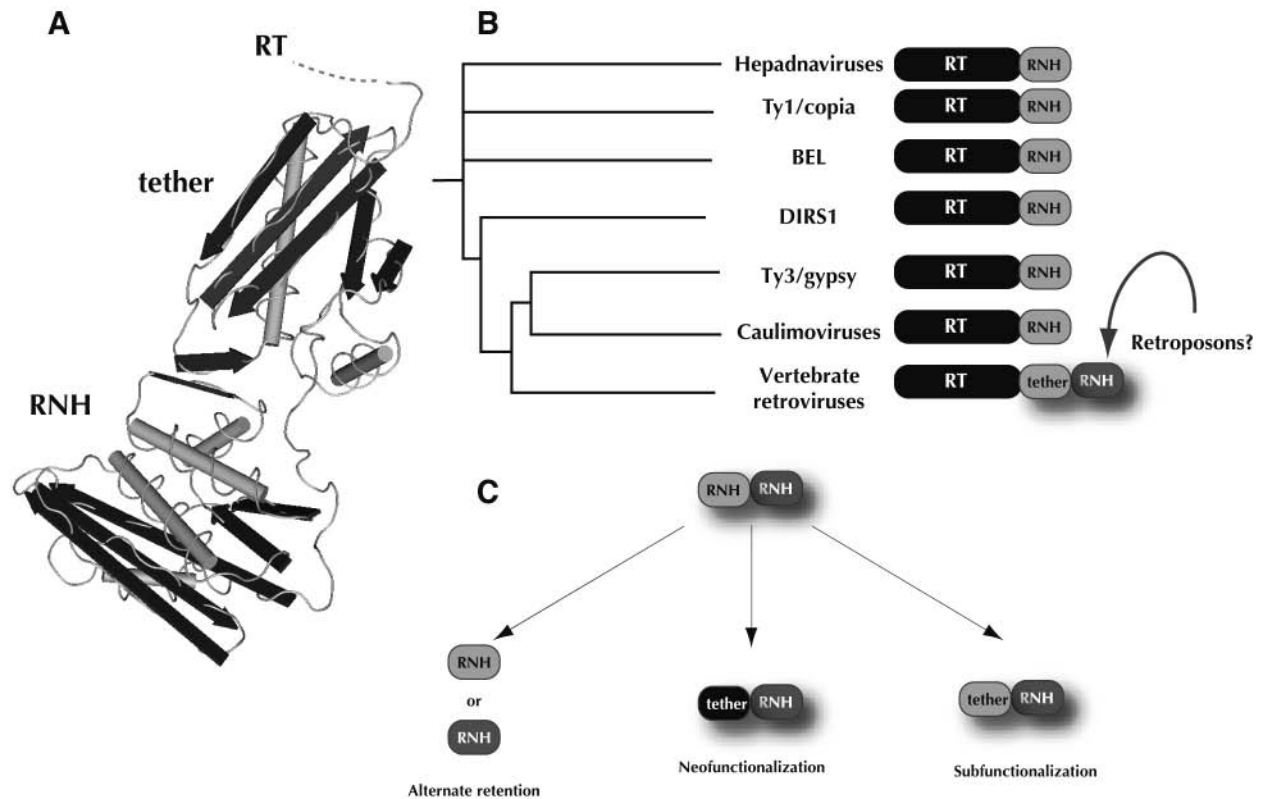


Fig. 7. Evolutionary origin of the tether domain in vertebrate retroviruses. **(A)** Crystal structure of the tether and RNH domains from HIV-1 (PDB: 1IKX) indicates that the tether domain adopts the characteristic fold of the RNH superfamily (compare to Fig. 1E). **(B)** A schematic of RNH evolution in LTR retrotransposons shows that the tether domain in vertebrate retroviruses most likely represents the evolutionary remnant of the ancestrally inherited RNH domain found in other LTR retrotransposons while the newly acquired RNH domain may have originated from a non-LTR retroposon source. **(C)** Following the acquisition of a newer RNH domain, vertebrate retroviruses could have chosen to keep one and delete the other (alternate retention). Instead, the tether domain could have acquired a new role (neofunctionalization) perhaps to negatively regulate RNH activity and ensure proper PPT generation. Finally, the tether and RNH domains could have partitioned ancestral functions (subfunctionalization) such that the tether domain carries out many of the structural functions (dimerization interface) while the RNH domain is responsible for enzymatic activity. The last two scenarios could both explain the maintenance of tether domains in vertebrate retroviruses and are not mutually exclusive.

it did not negatively impact the life cycle of the retrovirus (Seviya et al., 2001). Thus, the tether domain may have counteracted the negative effects of the more active RNH domain acquired by vertebrate retroviruses.

Concluding comments

Detailed phylogenetic analyses of RNH domains in retroelements have strong implications for the relative age of non-LTR and LTR retrotransposons (Malik and Eickbush, 2001; Eickbush and Malik, 2002). It appears evident from the RNH phylogeny (Fig. 6B) that the origin of the entire lineage of LTR domains is as old as or younger than the non-LTR retrotransposons. Recalling that only the youngest non-LTR elements harbor their own RNH domains (Fig. 5), this strongly implies that the LTR retrotransposons were a relatively late invention in the eukaryotic lineage that likely came about by the fusion of a non-LTR retroposon and a DNA-mediated transposon carrying an integrase domain. The relative age of LTR retrotransposons has generated a lot of controversy due to different interpreta-

tions of the RT phylogeny (choice of outgroup), and even different RT phylogenies! Since all LTR retrotransposons carry RNH domains, the RNH phylogeny proves to be informative and unambiguous in its conclusions. This also leads to the rather remarkable possibility that the invention of the eukaryotic nucleus may have selected for the origin of this chimeric retrotransposon in eukaryotes, owing to the obvious disadvantages incurred by both DNA-mediated transposons (sponsoring “dead” elements) and non-LTR retrotransposons (stability of RNA template) due to the physical and temporal separation of transcription and translation by the nuclear envelope (Malik and Eickbush, 2001; Eickbush and Malik, 2002).

Acknowledgements

The ideas presented in this review were developed in close collaboration with Tom Eickbush, and supported by funds from NSF to Tom Eickbush. I am especially grateful to my colleagues at the 2002 RNase H meeting (Tsuruoka, Japan) for educating me about various aspects of RNH biology. I thank Sara Sawyer and Danielle Vermaak for their comments on this manuscript.

References

- Akins RA, Kelley RL, Lambowitz AM: Mitochondrial plasmids of *Neurospora*: integration into mitochondrial DNA and evidence for reverse transcription in mitochondria. *Cell* 47:505–516 (1986).
- Altschul SF, Koonin EV: Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem Sci* 23:444–447 (1998).
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402 (1997).
- Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB: Retroelements containing introns in diverse invertebrate taxa. *Nat Genet* 33:123–124 (2003).
- Artymiuk PJ, Grindley HM, Kumar K, Rice DW, Willett P: Three-dimensional structural resemblance between the ribonuclease H and connection domains of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. *FEBS Lett* 324:15–21 (1993).
- Arudchandran A, Cerritelli S, Narimatsu S, Itaya M, Shin DY, Shimada Y, Crouch RJ: The absence of ribonuclease H1 or H2 alters the sensitivity of *Saccharomyces cerevisiae* to hydroxyurea caffeine and ethyl methanesulphonate: implications for roles of RNases H in DNA replication and repair. *Genes Cells* 5:789–802 (2000).
- Belfort M, Derbyshire V, Parker MM, Cousineau B, Lambowitz AM: Mobile introns: pathways and proteins, in Craig NL, Craigie R, Gellert M, Lambowitz AM (eds): *Mobile DNA II*, pp 761–783 (ASM Press, Washington DC 2002).
- Burke WD, Malik HS, Rich SM, Eickbush TH: Ancient lineages of non-LTR retrotransposons in the primitive eukaryote *Giardia lamblia*. *Mol Biol Evol* 19:619–630 (2002).
- Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M: Retrohoming of a bacterial group II intron: mobility via complete reverse splicing independent of homologous DNA recombination. *Cell* 94:451–462 (1998).
- Dasgupta S, Masukata H, Tomizawa J: Multiple mechanisms for initiation of CoIE1 DNA replication: DNA synthesis in the presence and absence of ribonuclease H. *Cell* 51:1113–1122 (1987).
- Eickbush TH: Mobile introns: retrohoming by complete reverse splicing. *Curr Biol* 9:R11–14 (1999).
- Eickbush TH: R2 and related site-specific Non-Long Terminal Repeat retrotransposons, in Craig NL, Craigie R, Gellert M, Lambowitz AM (eds): *Mobile DNA II*, pp 813–835 (ASM Press, Washington DC 2002).
- Eickbush TH, Malik HS: Origins and evolution of retrotransposons, in Craig NL, Craigie R, Gellert M, Lambowitz AM (eds): *Mobile DNA II*, pp 1111–1146 (ASM Press, Washington DC 2002).
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary degenerative mutations. *Genetics* 151:1531–1545 (1999).
- Goedken ER, Keck JL, Berger JM, Marqusee S: Divalent metal cofactor binding in the kinetic folding trajectory of *Escherichia coli* ribonuclease HI protein. *Science* 9:1914–1921 (2000).
- Goodwin TJ, Poulter RT: The DIRS1 group of retrotransposons. *Mol Biol Evol* 18:2067–2082 (2001).
- Henikoff JG, Henikoff S: Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* 12:135–143 (1996).
- Hillenbrand G, Staudenbauer WL: Discriminatory function of ribonuclease H in the selective initiation of plasmid DNA replication. *Nucleic Acids Res* 10:833–852 (1982).
- Ishikawa K, Okumura M, Katayanagi K, Kimura S, Kanaya S, Nakamura H, Morikawa K: Crystal structure of ribonuclease H from *Thermus thermophilus* HB8 refined at 2.8 Å resolution. *J Mol Biol* 230:529–542 (1993).
- Kanaya S: Prokaryotic type 2 RNases H. *Meth Enzymol* 341:377–394 (2001).
- Kanaya S, Kohara A, Miura Y, Sekiguchi A, Iwai S, Inoue H, Ohtsuka E, Ikehara M: Identification of the amino acid residues involved in an active site of *Escherichia coli* ribonuclease H by site-directed mutagenesis. *J Biol Chem* 265:4615–4621 (1990).
- Kanaya S, Oobatake M, Liu Y: Thermal stability of *Escherichia coli* ribonuclease HI and its active site mutants in the presence and absence of the Mg²⁺ ion. Proposal of a novel catalytic role for Glu48. *J Biol Chem* 271:32729–32736 (1996).
- Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Ikehara M, Matsuzaki T, Morikawa K: Three-dimensional structure of ribonuclease H from *E. coli*. *Nature* 347:306–309 (1990).
- Katayanagi K, Ishikawa M, Okumura M, Ariyoshi M, Kanaya S, Kawano Y, Suzuki M, Tanaka I, Morikawa K: Crystal structures of ribonuclease HI active site mutants from *Escherichia coli*. *J Biol Chem* 268:22092–22099 (1993).
- Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA: Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* 256:1783–1790 (1992).
- Lai L, Yokota H, Hung LW, Kim R, Kim SH: Crystal structure of archaeal RNase HII: a homologue of human major RNase H. *Structure Fold Des* 8:897–904 (2000).
- Lampson BC, Sun J, Hsu MY, Vallejo-Ramirez J, Inouye S, Inouye M: Reverse transcriptase in a clinical strain of *Escherichia coli*: production of branched RNA-linked msDNA. *Science* 243:1033–1038 (1989).
- Lima TM, Lim D: Isolation and characterization of host mutants defective in msDNA synthesis: role of ribonuclease H in msDNA synthesis. *Plasmid* 33:235–238 (1995).
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH: 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).
- Lyozin GT, Makarova KS, Velikodvorskaja VV, Zelentsova HS, Khechumian RR, Kidwell MG, Koonin EV, Evgen'ev MB: The structure and evolution of Penelope in the *virilis* species group of *Drosophila*: an ancient lineage of retroelements. *J Mol Evol* 52:445–456 (2001).
- Ma WP, Crouch RJ: *Escherichia coli* RNase HI inhibits murine leukaemia virus reverse transcription in vitro and yeast retrotransposon Ty1 transcription in vivo. *Genes Cells* 1:581–593 (1996).
- Malik HS, Eickbush TH: Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR-retrotransposons. *J Virol* 73:5186–5190 (1999).
- Malik HS, Eickbush TH: Phylogenetic analysis of ribonuclease H domains suggests a late chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* 11:1187–1197 (2001).
- Malik HS, Burke WD, Eickbush TH: The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* 16:793–805 (1999).
- Malik HS, Henikoff S, Eickbush TH: Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 10:1307–1318 (2000).
- McClure MA: Evolution of retrotransposons by acquisition or deletion of retrovirus-like genes. *Mol Biol Evol* 8:835–856 (1991).
- McClure MA, Donaldson E, Corro S: Potential multiple endonuclease functions and a ribonuclease H encoded in retrotransposon genomes. *Virology* 296:147–158 (2002).
- Murante RS, Henricksen LA, Bambara RA: Junction ribonuclease: an activity in Okazaki fragment processing. *Proc Natl Acad Sci USA* 95:2244–2249 (1998).
- Muroya A, Tsuchiya D, Ishikawa M, Haruki M, Morikawa M, Kanaya S, Morikawa K: Catalytic center of an archaeal type 2 ribonuclease H as revealed by X-ray crystallographic and mutational analyses. *Protein Sci* 10:707–714 (2001).
- Ohtani N, Haruki M, Morikawa M, Crouch RJ, Itaya M, Kanaya S: Identification of the genes encoding Mn²⁺-dependent RNase HII and Mg²⁺-dependent RNase HIII from *Bacillus subtilis*: classification of RNases H into three families. *Biochemistry* 38:605–618 (1999).
- Olivares M, Garcia-Perez JL, Thomas MC, Heras SR, Lopez MC: The non-LTR (long terminal repeat) retrotransposon L1Tc from *Trypanosoma cruzi* codes for a protein with RNase H activity. *J Biol Chem* 277:28025–28030 (2002).
- Poch O, Sauvaget I, Delarue M, Tordo N: Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J* 8:3867–3874 (1989).
- Pyatkov KI, Shostak NG, Zelentsova ES, Lyozin GT, Melekhin MI, Finnegan DJ, Kidwell MG, Evgen'ev MB: Penelope retroelements from *Drosophila virilis* are active after transformation of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 99:16150–16155 (2002).
- Rest JS, Mindell DP: Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol* 20:1134–1142 (2003).
- Sato A, Kanai A, Itaya M, Tomita M: Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon *Pyrococcus furiosus*. *Biochem Biophys Res Commun* 309:247–252 (2003).
- Sevilya Z, Loya S, Hughes SH, Hizi A: The ribonuclease H activity of the reverse transcriptases of human immunodeficiency viruses type 1 and type 2 is affected by the thumb subdomain of the small protein subunits. *J Mol Biol* 311:957–971 (2001).
- Sevilya Z, Loya S, Adir N, Hizi A: The ribonuclease H activity of the reverse transcriptases of human immunodeficiency viruses type 1 and type 2 is modulated by residue 294 of the small subunit. *Nucleic Acids Res* 31:1481–1487 (2003).
- Shimamoto T, Shimada M, Inouye M, Inouye S: The role of ribonuclease H in multicopy single-stranded DNA synthesis in *retron-Ec73* and *retron-Ec107* of *Escherichia coli*. *J Bacteriol* 177:264–267 (1995).
- Singer MS, Gottschling DE: TLC1: template RNA component of *Saccharomyces cerevisiae* telomerase. *Science* 266:404–409 (1994).
- Wang H, Lambowitz AM: Reverse transcription of the Mauriceville plasmid of *Neurospora*. Lack of ribonuclease H activity associated with the reverse transcriptase and possible use of mitochondrial ribonuclease H. *J Biol Chem* 268:18951–18959 (1993).
- Wang H, Kennell JC, Kuiper MT, Sabourin JR, Saldanha R, Lambowitz AM: The Mauriceville plasmid of *Neurospora crassa*: characterization of a novel reverse transcriptase that begins cDNA synthesis at the 3' end of template. *RNA Mol Cell Biol* 12:5131–5144 (1992).
- Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 9:3353–3362 (1990).
- Yamanaka K, Shimamoto T, Inouye S, Inouye M: Retrons, in Craig NL, Craigie R, Gellert M, Lambowitz AM (eds): *Mobile DNA II*, pp 784–795 (ASM Press, Washington DC 2002).
- Yang J, Malik HS, Eickbush TH: Identification of the endonuclease domain encoded by R2 and other site-specific non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci USA* 96:7847–7852 (1999).
- Yang W, Steitz TA: Recombining the structures of HIV integrase RuvC and RNase H. *Structure* 3:131–134 (1995).