Acquisition of Endonuclease Specificity during Evolution of L1 Retrotransposon

Kenji Ichiyanagi,*1 Hidenori Nishihara,* David D. Duvernell,† and Norihiro Okada*†

*Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Japan; and †Department of Biological Sciences, Southern Illinois University, Edwardsville

L1 is the most proliferative autonomous retroelement that comprises about 20% of mammalian genomes. Why L1s have proliferated so extensively in mammalian genomes is an important yet unsolved question. L1 copies are amplified via retrotransposition, in which the DNA cleavage specificity by the L1-encoded endonuclease (EN) primarily dictates sites of insertion. Whereas mammalian L1s show target preference for 5′-TTTTAA-3′, other L1-like elements exhibit various degrees of target specificity. To gain insights on diversification of the EN specificity during L1 evolution, ENs of zebrafish L1 elements were analyzed here. We revealed that they form 3 discrete clades, M, F, and Tx1, which is in stark contrast to a single L1 clade in mammalian species. Interestingly, zebrafish clade M elements cluster as a sister group of mammalian L1s and show target-site preference for 5′-TTTTAA-3′. In contrast, elements of the clade F, the immediate outgroup of the clade M, show little specificity. We identified certain clade-specific amino acid residues in EN, many of which are located in the cleft that recognizes the substrate, suggesting that these amino acid alterations have generated 2 types of ENs with different substrate specificities. The distribution pattern of the 3 clades suggests a possibility that the acquisition of target specificity by the L1 ENs improved the L1 fitness under the circumstances in mammalian hosts.

Introduction

Mammalian genomes contain over half a million copies of long interspersed nuclear elements (LINE)-1 (L1), which account for 17–23% of the respective genomes (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Mikkelsen et al. 2005, 2007). An intact L1 element encodes 2 proteins. The product of the first open reading frame (ORF), ORF1p, has RNA-binding and nucleic chaperon activities and has been proposed to form a nucleoprotein complex with the L1 RNA, an intermediate of L1 retrotransposition (Hohjoh and Singer 1996; Kolosha and Martin 1997; Martin and Bushman 2001; Martin et al. 2005). The ORF2-encoded protein, ORF2p, contains reverse transcriptase (RT) and endonuclease (EN) domains, the latter of which belongs to the apyrimidine/apurine (AP) endonuclease superfamily. L1s and other autonomous non–long-terminal repeat retrotransposons (hereinafter referred to as LINEs) proliferate via retrotransposition. During the process, the DNA sequence of the original element is first copied into RNA, then the RNA is reverse transcribed into cDNA at a genomic site where a new LINE copy becomes inserted. In retrotransposition, ORF2p nicks the target DNA duplex and initiates reverse transcription using the 3′-OH end of the cleaved DNA strand as a primer—a mechanism called target-primed reverse transcription (TPRT; Luan et al. 1993; Cost et al. 2002).

Because of this TPRT mechanism, the DNA cleavage specificity of the EN domain primarily determines sites of LINE insertion (Luan et al. 1993; Feng et al. 1996, 1998; Takahashi and Fujivara 2002). Human L1 preferentially inserts at 5′-TTTTAAA-3′, where "T" indicates the site of insertion (Moran et al. 1996; Gilbert et al. 2002, 2005; Morrish et al. 2002; Symer et al. 2002), and its EN cleaves the TpA bond in 5′-TTTTAA-3′ on the complementary strand (Feng et al. 1996; Cost and Boeke 1998). The mammalian L1s belong to the L1 clade, which includes numerous LINEs of a variety of organisms (Malik et al. 1999). Phylogenetic analysis has suggested that ENs of L1-clade LINEs are the oldest of LINE-encoded AP endonuclease–like ENs (Malik et al. 1999), which share a common ancestral origin with cellular AP endonuclease and DNase I. Because L1-clade elements show varying degrees of target-site specificity (Zingler et al. 2005), important unresolved issues include how the L1-encoded ENs evolved to acquire the target specificity and whether such specificity is implicated in the explosive L1 proliferation in mammals.

To better understand L1 dynamics and evolution, it is necessary to study L1s in the nonmammalian vertebrate classes as well as mammalian L1s. The zebrafish genome, for example, has a variety of L1 elements. These L1s form multiple clades, each of which contains retrotranspositionally active elements (Duvernell et al. 2004; Furano et al. 2004; Ichiyanagi and Okada 2006), which is in contrast to that mammalian L1s generally form a single clade despite their enormous copy numbers (Smit et al. 1995; Furano 2000; Furano et al. 2004; Khan et al. 2006). Phylogenetic relationships among these fish L1 clades and the mammalian clade are uncertain based on the analyses of the RT sequences. In the present study, we analyzed the ENs of these zebrafish L1 elements. We revealed the presence of 3 clades, phylogenetic relationships among these and mammalian L1s, their target-site specificity, and some clade-specific amino acid residues in the EN domain that might be involved in defining L1 target specificities. Based on these results, we discuss an impact of the acquisition of target specificity on the L1 evolution.

Methods

Phylogenetic Analysis

The amino acid sequences of the EN domains of all LINEs analyzed were deduced from the representative nucleotide sequences available in RepBase (Jurka et al. 2005) or those identified previously (Duvernell et al. 2004; Khan et al. 2006). Swimmer (SW)-like elements previously reported as BX088527, AL845351, AL928908, and BX005012 in the

1 Present address: Division of Human Genetics, Department of Integrated Genetics, National Institute of Genetics, Mishima, Shizuoka, Japan.

Key words: cleavage specificity, endonuclease, L1 evolution, LINE, retrotransposon.

E-mail: nokada@bio.titech.ac.jp.

doi:10.1093/molbev/msm130

Advance Access publication June 30, 2007

© The Author 2007. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.

For permissions, please e-mail: journals.permissions@oxfordjournals.org
teleost clade I (Duvernell et al. 2004) were designated here as L1-SW1_DR, L1-SW2_DR, L1-SW3_DR, and L1-SW4_DR, respectively. L1-1_DR, L1-6_DR, L1-8_DR, and L1-10_DR are the same sequences as BX005008, BX088694, AL772298, and AL929331 in the teleost clade II in the previous report, respectively, and L1-7_DR is most closely related to AL929396, BX005317, and AL929206 in the teleost clade II (Duvernell et al. 2004). To obtain the full-length EN sequence of L1-7_DR, we used the consensus nucleotide sequence of 3 genomic copies (in chr1:15522525-15527721, chr4:6767066-6772462, and chrUn:126631142-126635178 of denRer3). Because of the high divergence of L1-Tx1-1_DR elements, we constructed consensus nucleotide sequences using elements in chr7: 49336755-49340964, chr10:37893109-37897365, chr22:26853260-26857469, and chrNA:198797577-198801836 (for L1-Tx1-1a_DR) and chr7:55417053-55421090, chr24:10108301-10112062, and chr25:19590484-19594521 (for L1-Tx1-1b_DR) to deduce the amino acid sequences. The sequence alignment was made on ClustalX with default parameters (Thompson et al. 1997). The Neighbor-Joining tree was constructed using MEGA 3.1 (Kumar et al. 2004) with the p-distance model and the pairwise deletion option. Support for each node was estimated by a bootstrap analysis with 10,000 replicates.

Analysis of Target Sites of Genomic Zebrafish L1s

The locations and nucleotide sequences (with 60-bp flanking regions) of L1-1_DR to L1-10_DR and L1-Tx1-1_DR elements in the zebrafish genome (danRer3, May 2005) were downloaded from the University of California, Santa Cruz, genome browser at http://genome.ucsc.edu/ (Hinrichs et al. 2006). Along with these L1 copies, we selected L1_DRs that carry the complete 3' terminus. The genomic copies of L1-SW1_DR to L1-SW4_DR were previously reported (Duvernell et al. 2004). The target-site duplication (TSD) was identified using GENETYC-MAC 10.1.1 (software development) with the criteria: 1) the unit of direct repeat is ≥12 bp, 2) the divergence of the 2 units is ≤10% of their lengths, and 3) the repeated sequences are located within the 60-bp region flanking the LINE (in some cases, the TSD region overlapped the L1 sequence). Then, we inferred the target-site sequences for these L1 copies with a TSD. For instance, if an L1 sequence is flanked by 5'-ATACGTATTGTTGTTTAAATAGAGTAAATG-3' at the 5' end and 5'-AAAAATAGATTAGTTAATGAAATGG-3' at the 3' end (the underlines indicate a TSD), the inferred target sequence for this L1 insertion is 5'-ATACGTTTAAAATAGAGTAAATGTAATGG-3' (↓ indicates the site of insertion).

Results and Discussion

The Endonuclease Phylogeny

It has been reported that the zebrafish genome harbors various L1 elements (Duvernell et al. 2004; Furano et al. 2004); however, either their insertion sites or their coding endonuclease domains have not been characterized. To initiate the study on the EN domains of the mammalian and fish L1s, we constructed a Neighbor-Joining tree of these ENs as well as the ENs of zebrafish L2- and RTE-clade elements (ca. 230 amino acids corresponding to amino acid positions 8–238 of the human L1 ORF2 protein); DNase I and AP endonuclease were also included (fig. 1). The ENs of zebrafish L1s form 3 discrete clades, namely M, F, and Tx1. The clade M includes L1-1_DR, L1-6_DR to L1-8_DR, L1-10_DR (in the teleost clade II; Duvernell et al. 2004), and SW-like elements, L1-SW1_DR to L1-SW4_DR (in the teleost clade I; Duvernell et al. 2004). The clade Tx1 consists of elements of L1-Tx1-1_DR, whereas the clade F includes L1-2_DR to L1-5_DR, which were not included in the previous analysis (Duvernell et al. 2004), thereby newly identified by this study. The averages of amino acid identity in pairwise comparisons within clades are 40 (M; zebrafish elements only) and 50 (F), and 48% (Tx1), whereas those between clades are 24%–26%.

Interestingly, the clade M also includes mammalian L1s. The sibling relationship of the mammalian and fish members of clade M has been seen in Neighbor-Joining trees of the RT sequences, but bootstrap supports were low (Furano et al. 2004; Ichiyanagi and Okada 2006). A Bayesian analysis has suggested this relationship as well (Duvernell et al. 2004). Here, our analysis of the EN sequences strongly supports the relationship with a high bootstrap value of 94 (fig. 1, highlighted in pink). The relationship is also supported by the maximum likelihood analysis of the EN sequences and by the Neighbor-Joining analysis of the full-length ORF2p sequences (data not shown). Thus, the zebrafish members of the clade M are related more closely to mammalian L1 elements than to the zebrafish members of the clades F and Tx1. The topology of the tree suggests that the L1 endonuclease diverged in a common ancestor of fish and mammals and that one of the siblings (clade M) has been inherited and amplified in both fish and mammalian lineages, whereas the other siblings (clades F and Tx1) has been amplified in fish but not in mammals.

Compilation of Target Sites for Zebrafish L1 Insertions

Next, we tried to investigate cleavage specificities of the endonucleases encoded by the 3 clades of L1 elements. Our attempts to obtain recombinant endonucleases of the zebrafish L1s were not successful. Thus, we determined the sequences of their insertion sites because L1 insertion sites reflect the cleavage specificity of the element-encoded ENs (Feng et al. 1996; Cost et al. 2002). Our previous analysis (Ichiyanagi and Okada 2006) revealed that about half of the collected zebrafish L1 copies (clades M and F) have a TSD of 11–20 bp, although this analysis could identify only 20 target-site sequences. Thus, we further collected ~200 genomic L1 copies and searched for a direct repeat of 12 bp or longer (allowing <10% divergence between the repeat units) in 60-bp regions flanking each end of an L1 copy. Target sequences were then inferred from the sequence information of genomic copies containing an obvious TSD. We could identify TSDs in 67 and 14 copies of the clades M and F, respectively, whereas no members of the clade Tx1 have an obvious TSD. Our subsequent analysis, therefore, focused on the clades M and F.
The compilation of the target-site sequences revealed striking differences between the clades M and F. All L1s in the clade M have a target preference for sites resembling \(5\text{'-TTYAAAA-3'}\) (fig. 2A), which is the consensus sequence for human L1 targets as well. On the other hand, L1s in the clade F exhibit only very weak specificity, if any, and thus do not show the preference for \(5\text{'-TTAAAA-3'}\) (fig. 2B). The difference in the conservation among these target sequences was validated by a Mann–Whitney test for the numbers of nucleotides identical to \(5\text{'-TTAAAA-3'}\) (P, 0.00001; fig. 2C). Therefore, L1s of the clades M and F have significantly different target preference for their insertion, implying different degrees of cleavage specificity of their ENs.

Clade-Specific Amino Acids on the DNA Recognition Cleft of the EN Domains

Human L1 EN specifically cleaves \(5\text{'-TTTTYAAAA-3'}\) on the complementary strand of the consensus target sequence, \(5\text{'-TTAAAA-3'}\). Rather than the simple base sequence, this enzyme has been proposed to recognize the special geometry of the \(5\text{'-TnAn-3'}\) duplex, which has a narrow minor groove in the A tract and structural flexibility at the T-A step (Cost and Boeke 1998). Based on the crystal structure of the free EN domain of human L1, a structural model for substrate recognition has been proposed (Weichenrieder et al. 2004). In the model, the adenosine downstream of the scissile bond is flipped out from the helix and the EN domain recognizes this extrahelical adenosine with Phe-193 and Ile-204 interacting with the sugar moiety, and Arg-155 and Ser-202 making hydrogen bonds to the base moiety.

To infer which amino acid alterations are ascribable to the difference in the degree of target-site specificity between the clades M and F, we aligned the amino acid sequences of their ENs (fig. 3). Whereas many residues are conserved among all or most ENs (fig. 3; gray-shaded residues), some residues are clade specific (fig. 3; residues
shaded by light pink, pink, light blue, or blue; see caption for color codes). Interestingly, some of these clade-specific residues are clustered in the regions around residues Arg-155 and Phe-193 of human L1 EN. When mapped onto the crystal structure of human L1 EN, the conserved residues are clustered in the protein interior, catalytic center, and bottom of the DNA-binding cleft (fig. 4B and C). In contrast, the clade-specific residues constitute the wall of the DNA-binding cleft, which contains Arg-155 and Phe-193 (fig. 4C). In addition, the finger-like β hairpin (residues from Phe-194 to Tyr-201), which comprises a part of the wall (fig. 4C), carries 4 clade-specific residues and 2 amino acid additions in clade F (fig. 3). Therefore, its configuration is likely altered severely in the clade F ENs. In addition, the unique target specificity at the clade-specific positions is responsible for generating the clade M ENs with the unique target specificity.

Target Specificity and L1 Proliferation

Each zebrafish L1 subfamily seems to include currently active elements, as judged by the presence of genomic copies that are identical, or almost identical, to the consensus sequences. As discussed above, L1s diverged into the 3 clades in the common ancestor of fish and mammals. In the fish lineage, these L1s have retained their activity of proliferation regardless of the degree of target specificity, although the copy numbers are lower than in mammals (100–1,000 copies, fig. 1). Why fish have tolerated the proliferative activities of both sequence-specific and nonsequence-specific L1s is currently unknown. However, the tolerance for nonsequence-specific transposable elements seems a character of zebrafish because it also harbors total ~70,000 copies of L2-, CR1-, and RTE-clades of LINEs, all of which exhibit nonsequence-specific insertions (Ichyanagi et al. 2007; Ichyanagi K, Okada N, in preparation).

On the other hand, only the clade M elements have maintained their proliferative activity in the mammalian lineage to occupy a substantial fraction of the genomes (~20%, >500,000 copies in total). Therefore, it is conceivable that the clade M members have gained much better fitness in mammalian hosts. Such better fitness should involve several cumulative factors, such as the regulation of L1 transcription, the efficiency of retrotransposition, and the neutralization of the harmful potential of insertions. We propose that the acquisition of target specificity, rather than a diminution thereof, by L1 endonucleases is one of the factors for the successful amplification of mammalian L1s because the moderate restriction of insertion targets provides a better chance for L1 to be tolerated by host genomes. For example, the preference for 5′-TTAAAA-3′ substrates directs the L1 insertion toward noncoding regions, which partly neutralizes the mutagenic toxicity of the L1 insertion (Cost and Boeke 1998), thereby better than random insertion. It may be also possible that the cleavage

![FIG. 2. —L1_DR target sites. (A and B) Alignment of target-site sequences for L1s in the clades M (panel A) and F (panel B). The sites of insertion are indicated by hyphens. Nucleotides that are identical to 5′-TTAAAA-3′ are highlighted. (C) The target sequences for the clades M and F of zebrafish L1s were categorized by the number of nucleotides that matched the 5′-TTAAAA-3′. The P value (by Mann–Whitney test) represents the significance of the differences in target-site distribution in the clades M and F.]
specificity for 5′-TTTT↑AA-3′ on the primer strand generates a more efficient L1 retrotransposition machinery by providing a better probability of annealing of the target DNA and the polyA tail of the L1 RNA, which assists the initiation of reverse transcription (Ostertag and Kazazian 2001; Kulpa and Moran 2006). Consistently, zebrafish clade M members have higher copies numbers than clade F elements (fig. 1; P = 0.045 and 0.034 by U and t tests, respectively), although these numbers show relatively large variance.

Mammalian L1 elements have undergone multiple waves of amplification in the last 100 Myr (Furano 2000). After each wave, a new subfamily emerged from a preexisting active subfamily, then it predominated over the predecessor for the replication process possibly by acquiring a different regulatory sequence in its 5′ untranslated region and/or the rapid evolution of the ORF1 protein (Martin et al. 1985; Adey et al. 1994; Furano 2000; Boissinot and Furano 2001; Khan et al. 2006). Because these L1s are all clade M members (fig. 1 and the supplementary figure, Supplementary Material online), the acquisition of target specificity made little impact on this phenomenon. Rather, as discussed above, our results suggest that the acquired target specificity played an important role in the evolution of the DNA substrate. Numbers in parentheses on the right indicate amino acid positions in ORF2p.
role in the predomination of clade M members over other clades at earlier stages. In closing, our study underscores that better understanding of the dynamics of nonmammalian vertebrate L1s will provide important information on why L1s have proliferated so extensively in the course of mammalian evolution.

Supplementary Material

In addition to elements shown in figure 1, all other human L1 subfamilies (except for L1PA17 and L1MA4 and 5) were included in the phylogenetic analysis. Multiple alignment, Neighbor-Joining tree construction, and bootstrap calculations were carried out as described in Methods. Supplementary figure is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Mr Mitsuhiro Nakamura for helpful comments on the manuscript. This work was supported by a Grant-in-Aid to N.O. from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the 21st Century Center of Excellence program of the ministry.

Literature Cited


FIG. 4.—Amino acid conservations and variations mapped on the human L1 EN structure. (A) A ribbon representation of the crystal structure of the human L1 EN (Protein Data Bank code: 1VYB; Weichenrieder et al. 2004). The catalytic side chains, Glu-43, Asp-145, Asp-229, and His-230, and side chains possibly interacting with the flipped-out adenosine of the DNA substrate, Arg-155, Phe-193, Ser-202, and Ile-204, are shown in a stick model. (B and C) Amino acid positions that are conserved (gray), clade-specific (magenta), or not conserved (green) are mapped onto the protein backbone (B) and surface (C) of the human L1 EN crystal structure. The structure representations were drawn on PyMOL (DeLano Scientific, Palo Alto, CA).


William Martin, Associate Editor

Accepted June 20, 2007