Minireview

DNA "Fossils" and Phylogenetic Analysis

USING L1 (LINE-1, LONG INTERSPERSED REPEATED) DNA TO DETERMINE THE EVOLUTIONARY HISTORY OF MAMMALS*

Anthony V. Furano and Karen Usdin

From the Section on Genomic Structure and Function, Laboratory of Biochemical Pharmacology, NIDDK, National Institutes of Health, Bethesda, Maryland 20893-0830

The L1 element (LINE-1, long interspersed repeated DNA) is the mammalian version of the non-long-terminal repeat class of transposable elements that replicate via an RNA intermediate (retrotransposons) (1). Every modern mammalian species studied to date contains a distinctive L1 family consisting of tens of thousands of members, which are interspersed throughout the genome. Despite their distinctiveness, all full-length mammalian L1 elements share the same organization: a 5′-UTR, 1 which includes a regulatory sequence; ORF I, which encodes a protein of unknown function; ORF II, which encodes an RT (2); and a 3′-UTR that contains a G-rich polypurine-polypyrimidine tract and terminates in an A-rich sequence (Fig. 1).

Each of the modern L1 families evolved independently in the various mammalian lineages from a common ancestral L1 element that dates back to sometime before the mammalian radiation — 100 million years ago (3–5). Being capable of prodigious amplification, the modern L1 elements and their evolutionary antecedents (see below) now account for at least 30% of the mass of mammalian DNA. In addition, L1 elements are active in present day species and are a frequent cause of genetic polymorphisms including a number of non-inherited genetic defects in humans (6–8). It is also possible that the L1 RT catalyzed the retrotransposition of elements that do not encode their own RT such as the mammalian SINE families (e.g. Alu in primates, B1, B2, I D, etc., in rodents) (5, 9–11). Since these families can reach copy numbers as high as 1 × 10^9 and alone contribute up to 5% of mammalian DNA (e.g. Alu (9)), L1 elements quite likely had, and continue to have, a profound effect on the structure, function, and evolution of mammalian genomes.

In spite of their prominence, most of the biochemical and molecular details of L1 regulation, replication, and transposition remain unknown. To a large extent, what is known has been derived from evolutionary studies, and these have yielded two kinds of information. The first is derived from comparisons between different mammalian L1 families or between L1 elements and their counterparts in other organisms. This comparative biochemical approach identified and assigned possible functional significance to different features of non-long-terminal repeat retrotransposons.

The second type of information, generated by the analytical techniques of evolutionary biology, revealed the evolutionary dynamics of L1 families. These studies suggest that L1 evolution is a paradigm for a novel, but as yet incompletely understood, evolutionary process that is taking place within the “ecosystem” of the mammalian genome and that L1 evolution is quite dynamic, with novel L1 variants continually emerging over relatively short periods of time. As a consequence, L1 evolution has generated a rather complex family structure, and it has become apparent that this feature of L1 evolution can be exploited to examine the evolutionary (phylogenetic) history of the mammalian hosts that harbor these elements (12–16). It is this last aspect of L1 biology that will be the focus of this review. By way of introduction, we will briefly summarize some results derived from the comparative biochemical analysis and the evolutionary studies of L1 families.

Comparative Biochemistry of L1 Elements

Evolutionary comparisons have shown that the L1 RT is seemingly of very ancient lineage since transposable elements encoding an homologous protein have been found in bacteria, Group II introns, plants, fungi, and invertebrates (1). Elegant biochemical studies on the L1-like RTs from invertebrates including insects, fungi, some Group II introns, and bacteria revealed several intriguing mechanistic properties of this class of RT, which may bear directly on the biochemical properties of the L1 RT. Although this is the subject of a recent review (17), two properties of the RT are worth mentioning here. First, efficient cDNA synthesis by the RT depends on recognition of a structural feature near the 3′-end of the transposon transcript (10, 18–21). Second, the RT of the L1-like R 2b element of Bombyx mori tends to incorporate non-templated bases (mainly, but not only, As) at the 3′-end of the transposed cDNA (21). These properties could explain two evolutionarily conserved features of the mammalian L1 3′-UTR. The first is a G-rich polypurine stretch, which can form various unusual folded structures whether present as DNA (22–24) or as RNA. In the latter case such structures could possibly act as a recognition site for the L1 RT. The second is the A-rich terminus of L1 elements. While originally thought to have originated as the poly(A) tail of the retrotranscribed L1 transcript (25, 26), the A-rich terminus could have been generated during the retrotransposition process, as has been found for the R2b element (21). Such a mechanism could account for the fact that even recently transposed L1 elements do not always terminate in a pure poly(A) sequence (e.g. see Refs. 27 and 28).

One of the more striking findings revealed by the comparisons of different mammalian L1 families is that, in contrast to the rest of the element, the 5′-UTRs of even very closely related L1 families are not homologous (29–33). This indicates that the evolutionary origin of the 5′-UTR region is independent of the rest of the L1 element and that novel 5′-UTRs have been repeatedly acquired by the various mammalian L1 families. Since the 5′-UTR includes a region that has regulatory properties (34–38), the repeated acquisition of a novel regulatory sequence could be a means whereby the element bypasses either inactivating mutations in the L1 element (38) or a host-encoded repressive mechanism. Either explanation is consistent with the fact that sense strand-specific L1 transcripts are produced mainly from the most recently evolved L1 elements (39, 40). Although the evolutionary source for the novel L1 regulatory sequences is not known, they share certain sequence features with viral and housekeeping promoters in that they are CpG islands (41, 42) and lack many of the traditional transcription factor binding motifs found in RNA polymerase II promoters (e.g. TATA and CAAAT boxes).

The Evolutionary Dynamics of L1 Families

L1 replication generates two types of progeny: replication-competent copies and, in far greater numbers, defective copies, e.g. 5′-truncated, rearranged, etc. (25, 26, 29). For the most part, these defective copies were neither excised (4, 5, 11, 12) nor homogenized by postreplicative events such as gene conversion (11, 12, 43–46) but have diverged from each other due to the accumulation of random mutations over time. Therefore, the extent of divergence

---

* This minireview will be reprinted in the 1995 Minireview Compendium, which will be available in December, 1995.

† To whom correspondence should be addressed. Tel.: 301-496-6180; Fax: 301-402-0240; E-mail: avf@helix.nih.gov; ku@helix.nih.gov.

1 The abbreviations used are: UTR, untranslated region; ORF, open reading frame; RT, reverse transcriptase; SINE, short interspersed repeated DNA; Ma, million years ago.

2 R. Howell and K. Usdin, unpublished observations.
Examples of Using L1 as a Phylogenetic Character

The use of L1 DNA as a phylogenetic character is relatively simple in both principle and practice and depends on obtaining enough DNA sequence information to prepare clad-specific hybridization probes. Although probes cognate to any region of L1 DNA can be used (e.g. see below and the legend to Fig. 2), those specific to the 3'UTR are most generally useful, especially for recently evolved clades. This is because the 3'UTR evolves for the most part more rapidly than most of ORF I and all of ORF II (e.g. Refs. 5, 12, 13) and is not replaced wholesale during evolution as can be the case for the 5'UTR (see “Comparative Biochemistry of L1 Elements”). In spite of the relatively rapid evolutionary change in the 3'UTR, clades that are as old as 12–15 million years can be readily distinguished (see below).

For older L1 clades, we have found probes of ~200 base pairs to be both specific yet long enough to hybridize efficiently to the divergent members of a given clade. For the younger families oligonucleotide probes are essential. Oligonucleotide probes of ~20 bases cognate to regions of clades that differ by 2 or more diagnostic nucleotides are ideal. In cases where the multiple diagnostic base differences between clades are further apart than can be accommodated on a single oligonucleotide more than one oligonucleotide should be used to eliminate the possibility that the shared hybridization signal is due to chance mutation in precisely the same position in two otherwise different clades (but see below). We have obtained excellent discrimination using oligonucleotides to probe for the base change difference as long as the difference lies in the middle of the oligonucleotide and the hybridization is carried out in the presence of a large excess of the appropriate competitor oligonucleotide, i.e. one that has the same sequence as the probe except for the distinguishing base change.

Hybridizations are most conveniently carried out using dot blots of genomic DNA. However, hybridization to blots of electrophoretically separated fragments of genomic DNA that had been digested with restriction endonucleases, which recognize conserved sites within the 3'UTR, greatly increases both the sensitivity and specificity of the method. The appearance of novel restriction fragments is indicative of subdivisions within a given clade due to the loss or gain of a particular restriction enzyme site. Therefore, a shared novel restriction fragment detected even by a probe specific for just a single base difference would be highly specific for a given clade. This is because the presence of the novel restriction fragment would have required at least two base changes: the one detected by the oligonucleotide and the one that either added or destroyed a given restriction enzyme site. The sensitivity of the method is increased because the presence of subdivisions within a given clade could be evidence of recently evolved (or evolving) L1 clades. In the two sections below we demonstrate the use of L1 as a phylogenetic character to examine an evolutionary event that occurred about 12 Ma and one that began 1–3 Ma.

Phylogenetic Analysis Using an Ancient Murine L1 Clade

Murinae, a rodent subfamily, which includes Old World rats (Rattus) and mice (Mus) and many other genera, first appeared 12–15 Ma. The classification of Murinae is traditionally based on several cranial and dental characters (57) and in a number of cases has been problematic (58). A few years ago we discovered the relics of an ancient L1 clade (referred to as LX) in the genomes of mice and rats (11, 12, 15). Based on the extent of nucleotide divergence between Lx members and the murine neutral nucleotide substitution rate, we estimated that the Lx amplification coincided with the mus racialization (15). Therefore, we expected that the relic copies of LX would be present in all modern day murines but absent from the non-murine taxa.

We found LX to be present in 24 unambiguously classified mu-

---

Footnotes:

1 A. V. Furano and K. Usdin, unpublished data.

2 The complex L1 DNA composition of mammals presents a horrendous nomenclature problem, especially if one attempts to use a naming convention that accounts for all possible ancestor/descendant relationships between distinct "families" of L1 elements. For example, family "A" from which descendant subfamilies "B" and "C" arose, could itself have been a subfamily of an older family "Y." Subfamily "B" could have gone on to split into sister subfamilies "B-1" and "B-2." Furthermore, these four "generations" ("Y," "A," "B," and "B-1/"B-2") of L1 families could exist in one species of animal or be shared by several species. To avoid confusion, from here on we will use the word "clade" to refer to any distinct group of L1 elements, implying nothing about ancestry. Descendant relationships between the clades. Therefore, in our usage a clade could be synonymous with "family" or "subfamily" or even "superfamily." In those cases where ancestor/descendant relationships are important we will explicitly state them. In standard evolutionary usage a clade designates all of the descendants of a given ancestor.

3 No common ancestry.

4 The complex L1 DNA composition of mammals presents a horrendous nomenclature problem, especially if one attempts to use a naming convention that accounts for all ancestor/descendant relationships between distinct "families" of L1 elements. For example, family "A" from which descendant subfamilies "B" and "C" arose, could itself have been a subfamily of an older family "Y." Subfamily "B" could have gone on to split into sister subfamilies "B-1" and "B-2." Furthermore, these four "generations" ("Y," "A," "B," and "B-1/"B-2") of L1 families could exist in one species of animal or be shared by several species. To avoid confusion, from here on we will use the word "clade" to refer to any distinct group of L1 elements, implying nothing about ancestry. Descendant relationships between the clades. Therefore, in our usage a clade could be synonymous with "family" or "subfamily" or even "superfamily." In those cases where ancestor/descendant relationships are important we will explicitly state them. In standard evolutionary usage a clade designates all of the descendants of a given ancestor.

5 No common ancestry.

6 The complex L1 DNA composition of mammals presents a horrendous nomenclature problem, especially if one attempts to use a naming convention that accounts for all ancestor/descendant relationships between distinct "families" of L1 elements. For example, family "A" from which descendant subfamilies "B" and "C" arose, could itself have been a subfamily of an older family "Y." Subfamily "B" could have gone on to split into sister subfamilies "B-1" and "B-2." Furthermore, these four "generations" ("Y," "A," "B," and "B-1/"B-2") of L1 families could exist in one species of animal or be shared by several species. To avoid confusion, from here on we will use the word "clade" to refer to any distinct group of L1 elements, implying nothing about ancestry. Descendant relationships between the clades. Therefore, in our usage a clade could be synonymous with "family" or "subfamily" or even "superfamily." In those cases where ancestor/descendant relationships are important we will explicitly state them. In standard evolutionary usage a clade designates all of the descendants of a given ancestor.

7 No common ancestry.

8 The complex L1 DNA composition of mammals presents a horrendous nomenclature problem, especially if one attempts to use a naming convention that accounts for all ancestor/descendant relationships between distinct "families" of L1 elements. For example, family "A" from which descendant subfamilies "B" and "C" arose, could itself have been a subfamily of an older family "Y." Subfamily "B" could have gone on to split into sister subfamilies "B-1" and "B-2." Furthermore, these four "generations" ("Y," "A," "B," and "B-1/"B-2") of L1 families could exist in one species of animal or be shared by several species. To avoid confusion, from here on we will use the word "clade" to refer to any distinct group of L1 elements, implying nothing about ancestry. Descendant relationships between the clades. Therefore, in our usage a clade could be synonymous with "family" or "subfamily" or even "superfamily." In those cases where ancestor/descendant relationships are important we will explicitly state them. In standard evolutionary usage a clade designates all of the descendants of a given ancestor.

9 No common ancestry.
DNA "Fossils" and Phylogenetic Analysis

Fig. 2. Distribution of L1 clades in various rodents. A, diagrammatic representation of the presence or absence of an ancient murine L1 clade, Lx, and several more recent clades, L13, L1, and L1mlvi2. The original data presented, as a phylogenetic character helped resolve two problems in rodent classification. The distribution of Lx in murine and non-murine species (11, 15) and the presence or absence of an amplified L1 clade can be used as a novel and robust phylogenetic character. We should also mention that individual transposition events can be used for phylogenetic analysis. Batzer et al. (64) showed that the frequency of a SINE insertion at four different loci in the human genome distinguishes Mus domesticus and Mus spretus (13) and have been used to detect M. spretus genomic sequences present in an inbred strain of Mus musculus (63). Additionally, recent work on modern M. musculus L1 DNA has revealed emerging and apparently competing L1 clades in this clade that may be useful in defining subpopulations of this species as well (46, 49). Humans also contain a very complex L1 DNA composition (5) including a number of distinct replication-relevant L1 clades (6, 51).

Concluding Remarks

As a consequence of their long replicative history in mammalian genomes, L1 elements have generated a rich collection of DNA "fossils" that can be used to determine the phylogenetic history of mammals. Here we have shown how the presence (or absence) of an amplified L1 clade can be used as a novel and robust phylogenetic character. We should also mention that individual transposition events can be used for phylogenetic analysis. Batzer et al. (64) showed that the frequency of a SINE insertion at four different loci in the human genome distinguishes Mus domesticus and Mus spretus (13) and have been used to detect M. spretus genomic sequences present in an inbred strain of Mus musculus (63). Additionally, recent work on modern M. musculus L1 DNA has revealed emerging and apparently competing L1 clades in this clade that may be useful in defining subpopulations of this species as well (46, 49). Humans also contain a very complex L1 DNA composition (5) including a number of distinct replication-relevant L1 clades (6, 51).

Phylogenetic Analysis with Modern L1 Clades

The distribution of recently amplified L1 clades can be used to resolve the taxonomy of more recently diverged animals. The genus Rattus contains about 50 species, considered to be Rattus sensu strictu. Single copy DNA hybridization is unable to establish a branching pattern for many of these species, and the systematics of this group remains largely unresolved (16, 58). We can distinguish at least five relatively modern L1 clades in Rattus norvegicus. One of the species containing Rattus norvegicus at least 3.5 million years ago when the species comprising Rattus sensu strictu began emerging. As Fig. 2 illustrates, the L14 clade is present only in animals classified as Rattus sensu strictu (16). Therefore, the L14 clade probably arose in the common ancestor of Rattus sensu strictu some time after the divergence of these animals from the ancestor they shared with Rattus sensu lato.

By contrast, two younger rat L1 clades, L13 and L1mlvi2, are present only in R. norvegicus and in animals identified as Rattus rattus moluccarius, a presumed subspecies of Rattus rattus (16). Although R. rattus moluccarius specimens contained both the L13 and L1mlvi2 clades, these L1 clades were absent from a number of other R. rattus specimens (Fig. 2). This result was quite surprising and suggested that the R. rattus moluccarius specimens were misclassified and represent a sister taxon of R. norvegicus rather than a subspecies of R. rattus (16). Further analysis using mitochondrial DNA sequences and our finding that R. norvegicus and R. rattus moluccarius share a satellite DNA sequence supported this conclusion (15). Therefore, the L13 and L1mlvi2 clades are markers for a new taxon within Rattus sensu strictu; this taxon contains R. norvegicus and R. rattus moluccarius.

The L1mlvi2 clade has evolved rapidly, and two descendents of the L1mlvi2 clade can be distinguished: L1mlvi2molv1 and L1mlvi2molv2. While R. rattus moluccarius contains only the L1mlvi2molv1 clade, R. norvegicus contains some members of this clade but far greater numbers of the L1mlvi2molv2 clades (Fig. 2B). This indicates that the L1mlvi2molv2 clade either arose in or began amplifying in R. norvegicus soon after it and R. rattus moluccarius diverged from their common ancestor. Furthermore, it is possible that the L1mlvi2molv1 clade may have expanded at the expense of the L1mlvi2molv2 clade in the R. norvegicus genome since this clade has not amplified to the same extent as the L1mlvi2molv1 clade in R. norvegicus or as the L1mlvi2molv2 clade in R. rattus moluccarius. These results suggest that very closely related L1 clades can exclude each other perhaps by competing for limiting host factors.

Some early L1 DNA of Mus have revealed a similar picture of L1 evolution and have demonstrated the usefulness of L1 DNA as a phylogenetic character in this species. Species-specific L1 clades distinguish Mus domesticus and Mus spretus (13) and have been used to detect M. spretus genomic sequences present in an inbred strain of Mus musculus (63). Additionally, recent work on modern M. musculus L1 DNA has revealed emerging and apparently competing L1 clades that may be useful in defining subpopulations of this species as well (46, 49). Humans also contain a very complex L1 DNA composition (5) including a number of distinct replication-relevant L1 clades (6, 51).

Minireview: DNA "Fossils" and Phylogenetic Analysis

The distribution of recently amplified L1 clades can be used to resolve the taxonomy of more recently diverged animals. The genus Rattus contains about 50 species, considered to be Rattus sensu strictu. Single copy DNA hybridization is unable to establish a branching pattern for many of these species, and the systematics of this group remains largely unresolved (16, 58). We can distinguish at least five relatively modern L1 clades in Rattus norvegicus. One of the older ones, L1x, amplified about 3.5 million years ago when the species comprising Rattus sensu strictu began emerging. As Fig. 2 illustrates, the L1x clade is present only in animals classified as Rattus sensu strictu (16). Therefore, the L1x clade probably arose in the common ancestor of Rattus sensu strictu some time after the divergence of these animals from the ancestor they shared with Rattus sensu lato.

By contrast, two younger rat L1 clades, L13 and L1mlvi2, are present only in R. norvegicus and in animals identified as Rattus rattus moluccarius, a presumed subspecies of Rattus rattus (16). Although R. rattus moluccarius specimens contained both the L13 and L1mlvi2 clades, these L1 clades were absent from a number of other R. rattus specimens (Fig. 2). This result was quite surprising and suggested that the R. rattus moluccarius specimens were misclassified and represent a sister taxon of R. norvegicus rather than a subspecies of R. rattus (16). Further analysis using mitochondrial DNA sequences and our finding that R. norvegicus and R. rattus moluccarius share a satellite DNA sequence supported this conclusion (15). Therefore, the L13 and L1mlvi2 clades are markers for a new taxon within Rattus sensu strictu; this taxon contains R. norvegicus and R. rattus moluccarius.

The L1mlvi2 clade has evolved rapidly, and two descendents of the L1mlvi2 clade can be distinguished: L1mlvi2molv1 and L1mlvi2molv2. While R. rattus moluccarius contains only the L1mlvi2molv1 clade, R. norvegicus contains some members of this clade but far greater numbers of the L1mlvi2molv2 clades (Fig. 2B). This indicates that the L1mlvi2molv2 clade either arose in or began amplifying in R. norvegicus soon after it and R. rattus moluccarius diverged from their common ancestor. Furthermore, it is possible that the L1mlvi2molv1 clade may have expanded at the expense of the L1mlvi2molv2 clade in the R. norvegicus genome since this clade has not amplified to the same extent as the L1mlvi2molv1 clade in R. norvegicus or as the L1mlvi2molv2 clade in R. rattus moluccarius. These results suggest that very closely related L1 clades can exclude each other perhaps by competing for limiting host factors.

Some early L1 DNA of Mus have revealed a similar picture of L1 evolution and have demonstrated the usefulness of L1 DNA as a phylogenetic character in this species. Species-specific L1 clades distinguish Mus domesticus and Mus spretus (13) and have been used to detect M. spretus genomic sequences present in an inbred strain of Mus musculus (63). Additionally, recent work on modern M. musculus L1 DNA has revealed emerging and apparently competing L1 clades that may be useful in defining subpopulations of this species as well (46, 49). Humans also contain a very complex L1 DNA composition (5) including a number of distinct replication-relevant L1 clades (6, 51).

Concluding Remarks

As a consequence of their long replicative history in mammalian genomes, L1 elements have generated a rich collection of DNA "fossils" that can be used to determine the phylogenetic history of mammals. Here we have shown how the presence (or absence) of an amplified L1 clade can be used as a novel and robust phylogenetic character. We should also mention that individual transposition events can be used for phylogenetic analysis. Batzer et al. (64) showed that the frequency of a SINE insertion at four different loci in the human genome distinguishes human populations and used their results to further support the African origin of modern humans. Comparisons between mammalian β-globin loci have shown that different species can be distinguished by the pattern of L1 insertions at this site (4, 65, 66). For example, an ancient L1 insertion between the e and γ genes distinguishes eutherians (mammals) from metatherians (marsupials) (4, 66), and two independent L1 insertions flank the γ-globin gene in simians but not in prosimians (66). However, independent insertional events could be problematic for phylogenetic analysis. First, they are much harder to identify or characterize initially (though once detected, relatively easy to screen for) than the presence or absence of an amplified L1
clade. Second, any individual insertion or site that is being scored for the presence of the insertion could be subject to re-arrangement, e.g. deletion of the inserted element. Therefore, both the problems of homoplasy and of determining whether the character is an ancestral or acquired one could theoretically afflict the use of individual insertion events.

Finally, we would like to close with a comment about the possible effect of L1 transposition on mammalian evolution. Because L1 insertions are random and potentially either beneficial or deleterious, it is easy to visualize how an L1 amplification event introduces genetic diversity into an extant animal population. Depending on a number of extrinsic (e.g. geographical isolation, population size) and intrinsic (e.g. changes in fitness caused by an L1-induced genetic effect) factors, a given animal population could become differentiated into subpopulations as a consequence of the difference between their pattern of L1 insertions. Moreover, depending on the rate at which novel L1 clades emerge and amplify, it would be quite possible that subpopulations could also differ by their content of distinct L1 clades, which, depending on the relative transposition rate of the distinct L1 clades, further enhances the content of distinct L1 clades, which, depending on the relative genetic effect factors, a given animal population could become differentiated into subpopulations as a consequence of the difference between their pattern of L1 insertions. In this regard, we note the apparent correlation, at least during rodent evolution, between the genera-