The Evolution of Modern Lineages of Mouse L1 Elements

Michael L. Mears, Clyde A. Hutchison, III

1 Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290, USA
2 Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290, USA

Received: 21 December 1999 / Accepted: 23 August 2000

Abstract. There are three known families of L1 elements in the Mus genome, V, F, and A. An individual L1 element is classified as a member of one of these families based on which of three different types of transcription promoters is at its 5’ end. Initial evidence suggested that only actively transposing L1 elements in the modern mouse genome were a young subfamily of A-type elements. That belief was overturned when a transposing F subfamily, TF, was discovered. We used molecular phylogenetic methods to investigate the emergence of the two currently transposing L1 lineages, young A’s and TF’s. Both of these subfamilies appear to be direct descendents from a specific clade of F-type L1’s. Our results imply that recombination between L1 sequences occurred in the lineage from which the TF subfamily evolved. We also found that phylogenetic analysis of a L1 3’ untranslated region (UTR) is diagnostic for the promoter type at the 5’ end of the sequence and, therefore, for the family to which it belongs. As part of this investigation, we developed a set of full-length L1 elements, which may serve as a general reference set for phylogenetic analyses in Mus. Our analyses included 21 full-length L1 elements from the GenBank nonredundant database that had not been phylogenetically analyzed previously.

Key words: L1 — LINES-1 — Mus — Mouse — Retrotransposon — TF — Phylogenetic analysis — Recombination — Horizontal transfer

Introduction

All mammalian genomes contain approximately 10^5 copies of highly conserved retrotransposable sequences referred to as L1 or LINES-1 elements. This implies that the L1 progenitor predates the mammalian radiation (Burton et al. 1986). The L1 population comprises a larger portion (>10%) of the mammalian genome than all of the other gene exons combined (~3%) (e.g., Dunham et al. 1999).

Several human genetic diseases have been associated with the insertion of L1 elements in, or near, the malfunctioning gene. Diseases associated with L1 insertion include specific cases of hemophilia A (Kazazian et al. 1988), Duchenne muscular dystrophy (Narita et al. 1993), breast cancer (Morse et al. 1988), colon cancer (Miki et al. 1992), and β thalassemia (Gilman 1987). Highly conserved L1 sequences, widely distributed throughout the genome, likely contribute to chromosomal rearrangements. Moran et al. (1999) suggested that L1’s might represent a general mechanism for the evolution of new genes.

Since L1 elements are in the DNA of all mammals, what is learned about L1 in one mammal may help us understand how it functions in other mammals. We have studied L1’s in the genome of Mus musculus domesticus. Major features of a typical full-length mouse L1 element are illustrated in Fig. 1. Full-length L1’s are 6–7 kb in length, although most are variably 5’ truncated (Voliva et al. 1983). There are three L1 transcription promoter types: A, F, and V. A typical A or F promoter is comprised of several tandemly repeated sequences called monomers, each containing approximately 200 nucleotides. The structure of V promoters is less certain due to
Fig. 1. This schematic (not to scale) illustrates a typical mouse full-length L1 element. L1's contain two open reading frames (ORFs). ORF1 encodes a single-stranded mRNA binding protein (Martin 1991; Kolosha and Martin 1997) and includes a variable length polymorphic region (LPR) of unknown function (Schichman et al. 1992). ORF2 encodes a protein that has both endonuclease (E) (Feng et al. 1996) and reverse transcriptase (RT) (Mathias et al. 1991) activities. Both proteins are required for retrotransposition (Moran et al. 1996). Each element is bracketed by short direct repeats (SDRs) resulting from target site duplication, typically 5–15 bp in length. Mouse L1's have promoters comprised of multiple direct repeats of individual promoter elements called monomers (see text). The length of the 5' untranslated region (5' UTR) between the promoter and ORF1 is about 200 bp and the 3' UTR is about 650 bp. The 3' UTR ends with a variable-length poly(A) tail, evidence of transposition via a mRNA intermediate. The lines below the L1 schematic indicate subsequences for which molecular phylogenetic analyses were performed. The subsequences of the analysis alignment are referenced in the text by the Greek letter above each. The analogous delimiting nucleotides in L1Md-A2 (accession number M13002) to which the subsequences correspond are as follows: α, 1539–1740; β, 1741–2058; γ, 2059–2814; δ, 2798–4115; ε, 4116–5455; η, 5456–6700; θ, 6701–7333; and τ, 2798–7333.

Full-length L1 elements contain two open reading frames (ORFs) (Loeb et al. 1986). Both ORFs are required for transposition (Moran et al. 1996). The ORF1 and ORF2 regions of L1's of the same family are often greater than 90% identical, and these regions in L1s of two families are usually >70% identical. Likewise, the 5' and 3' untranslated regions (UTRs) are more similar within a family than between different families but are less conserved than the ORFs. Most L1's of all three families have significant 5' truncations and therefore are inactive (i.e., they cannot transpose).

Phylogenetic analyses of mouse L1 sequences indicated that the three families of L1's represent three phases in the evolution of one dominant lineage (Aden et al. 1994b). This analysis also showed that V elements arose early in the lineage. F elements arose later and A elements arose most recently. These results were interpreted as evidence that actively transposing L1 elements had one of three distinctly different promoter sequences and that the period during which each promoter was active was limited to a specific phase of L1 evolution. The implication was that each promoter functioned during a specific time interval before ultimately becoming inactive. Before all elements of a family became inactive, a new active promoter was acquired by a conserved L1 that became the progenitor of the next family. All known F elements and most A elements had inactivating mutations (Aden et al. 1991b). F elements were relatively more divergent than A elements (Aden et al. 1994b). Isolated F monomers did not have promoter activity but a consensus of these sequences did (Aden et al. 1994a). Therefore, Aden et al. (1994b) suggested that only a few L1's with A promoters were currently actively transposing in the mouse genome. They also concluded that most A elements and all F and V elements were inactive.

Therefore, it was a surprise when recent mouse mutations were shown to have been caused by insertion of L1 sequences with monomers that are clearly homologous to F-type promoters. Insertion of two different full-length L1's (hereafter referred to as L1spa and L1Orl in the text and as spa and orl in the tables and figures) was the cause of specific cases of the spastic and Orleans reeler mutant phenotypes (Kingsmore et al. 1994; Takahara et al. 1996). Comparison of these L1 sequences showed that they were 99.4% identical over their entire lengths and that their promoters were related to the an-
Fig. 2. L1 length polymorphic region groups (based on Adey 1991c). Every L1 has an LPR, of unknown function, about 150 nucleotides from the 5′ end of ORF1. A L1 is classified into one of four groups based on which of four known LPR sequence patterns it contains. Although we previously designated these groups with Arabic numerals, here we use Roman numerals to avoid confusion of element names and family group designations. The family and group of an L1 element are designated with capital letters, and the designators are separated by a hyphen. Currently, the nine family groups for which sequences have been identified are A-I (i.e., A family with LPR group I (or 1)), A-II, A-III, F-I, F-II, T-F-II, F-III, F-IV, and V. All four groups begin with a very highly homologous 66-base pair sequences. In addition, the Group I LPR contains a repeat of the last 42 bases of the initial 66-base sequence. Group II LPRs contain two copies of the 42-base sequence in Group I. Group III LPRs are comprised of two nearly identical copies of a 66-base pair sequences. Group IV contains three copies of the 66-base sequence. The gaps in the pattern illustrations, for Group I and Group II, are inserted to show the sequences align; there are no actual gaps in the sequences. All four LPR patterns maintain the translation reading frame for ORF1.

Material and Methods

Sequence Alignment and Analysis. All sequence alignments were done using GCG programs (Deverux et al. 1984). Multiple sequence alignments for phylogenetic analysis were done using PILEUP, LINEUP, and PRETTY. Pairwise sequence alignments to confirm consistent identity over entire full-length L1 elements were done with GAP. Graphs were produced with COMPARE and DOTPLOT and visually checked for sequence continuity.

Phylogenetic Analysis. All phylogenetic analyses were performed with PAUP 4.0 b2a (Swofford 2000). Primary phylogenetic analyses were done using the maximum-parsimony heuristic search option. CpG positions that showed hypermutability to CpA or to TpG were excluded. Transversions were weighted twice as much as transitions. A strict consensus of the most parsimonious trees from each analysis was generated. Each consensus tree was then subjected to 100 rounds of bootstrap analysis. Neighbor-joining and maximum-likelihood analyses were also performed for all subsequences in Fig. 1. The results of both of these last analyses were consistent with those from maximum-parsimony analysis. Sequence divergence was also analyzed using data generated using PAUP’s pairwise difference option.

Phylogenetic Analysis Reference (PAR) Testset. As a standard for analyzing L1 sequence ancestry, we developed a reference testset. To
find candidates for the testset, we searched the GenBank nonredundant DNA sequence database for previously unanalyzed full-length L1 sequences. Several hundred candidate files were identified by querying with the ORF1 coding region from an A-type L1 element. The same list of files, with slightly different scores, was obtained when the query was made with the same region of an F-type element. From this list, 51 unique full-length L1 sequences were identified. The 35 mouse and 1 rat sequence used in this study are listed in Table 1. Previous studies of some of these sequences have been published (those for which “L1” appears in the last column in Table 1). Many L1’s that were in the database as part of sequences obtained for other purposes had not been phylogenetically analyzed.
The family of each sequence was determined by analyzing the monomers in the promoter at the 5' end of the sequence. The letter (A, F, or T) in the second column in Table 1 indicates the monomer type(s), A, F, or T_F, respectively, for the sequences chosen for this study. All 36 sequences in Table 1, excluding their monomers and poly(A) tail, were aligned throughout their entire lengths. The group of each sequence was determined by visual inspection of the ORF1 LPR alignment (see Fig. 1) (Schichman et al. 1992; Adey et al. 1994b). The Roman numeral in column 2 in Table 1 indicates the group for each element.

Evolutionary trees produced by phylogenetic analysis were generated to determine whether or not the sequences segregated as expected based on the family and group assigned to each. Separate trees were generated for A and F family sequences. In each tree we included previously studied sequences of that family and a few previously studied sequences of the other family as known references. A few T_F sequences were also included in both trees to help us identify portions of the trees most critical to our analysis. These analyses were performed for different regions of L1 defined by subsequences α through τ in Fig. 1. For many L1 elements, the results of the analyses of all subsequences were consistent with the family and LPR group assignments in Table 1. However, some L1’s segregated to different regions of the evolutionary tree, depending on which of their subsequences the tree was based on. This implied that they were recombinant.

The results from the family and group analyses were used to identify a set of test sequences, containing as many of the known L1 family groups as possible (see Fig. 2). The testset would be used for generating informative trees for subsequences over their entire length (except for the monomer region). The testset sequences were also pairwise compared to other testset sequences with GAP, and with COMPARE and DOTPLOT to ensure that they were full-length elements with no obvious points of recombination. Some recombinant sequences of special importance to our investigation were included in the testset as discussed below. Trees were then generated for only the testset sequences for all of the subsequences indicated in Fig. 1 to confirm that they segregated as expected for all of the subsequences analyzed.

**Results**

**Testset Development**

We identified a testset of 20 full-length L1’s representing as many of the known family groups as possible. L1 family elements are further classified into one of four groups based on the structure of the length polymorphic region (LPR) in ORF1 of the sequence (Schichman et al. 1992; Adey et al. 1994b). The location of the LPR is indicated in Fig. 1 and the known LPR structures are illustrated in Fig. 2. Previously analyzed full-length mouse L1 elements represented only three of the nine known family and subfamily groups. We searched the GenBank nonredundant DNA sequence database for additional full-length L1 sequences as candidates for the testset. Information about the testset and other L1 sequences used in this investigation is given in Table 1. The testset includes a newly identified T_F-1 sequence, Tna1, which is slightly divergent from the T_F-2 L1_Qal and L1spa sequences. This discovery expands the number of known L1 family groups to 10.

The sequences in Table 1 include 3 full-length sequences and 4 partial sequences containing only ORF1 that were studied previously by Adey et al. (1994b) and 28 additional full-length L1’s. The additional full-length elements included 11 A’s, 7 F’s, and 10 T_F’s with 7 of the 10 known family groups represented. The only F-I sequence found, F15, is the only full-length F-I sequence published to date (Adey et al. 1991a). We found no A-III, F-IV, or V full-length L1 sequences. Less than full-length A-III’s segregate with F-III’s (Adey et al. 1994b). These are some of the oldest mouse L1 sequences. It was to be expected that older sequences would have incurred a relatively large number of mutations including insertions and deletions. We found two F-III sequences that consistently segregated in the oldest portion of the trees, with the F-I and the oldest F-II sequences. For most subsequences, there was little resolution among those older sequences, and the segmentation observed was not identical in all trees. In time, we hope to add more old elements to the testset to improve resolution in the parts of our trees representing early mouse L1 evolution. We found two or more sequences each for the youngest F-II’s and for all T_F and A family groups.

**Phylogenetic Tree Generation and Sequence Divergence Calculation**

We then used the testset sequences to generate maximum-parsimony phylogenetic trees for each subsequence identified in Fig. 1 and computed divergence values for each pair of sequences. For each subsequence, the consensus tree was subjected to 100 rounds of bootstrap analysis. The results of the parsimony analyses were confirmed by neighbor-joining and maximum-likelihood analyses. Analysis of the trees and divergence values showed that there were two distinct clades in the F-II family group and two clades in the T_F subfamily. The nomenclature we use for these clades and examples of each are presented in Table 2. We computed average sequence divergence both within, and between, each family group and clade represented by a full-length element in our testset. Table 3 contains average divergence information showing that divergence within these clades is significantly less than it is between clades. Maximum-
parsimony trees for a set of subsequences spanning the entire L1 element, except for the promoter monomers, are shown in Fig. 3.

**T F and F-II 2 L1’s Are Recombinant**

During testset development, it became apparent that some sequences of special interest were recombinant. All TF sequences were in a very closely related cluster in the trees for all subsequences but segregated in two distinct patterns relative to the rest of the tree. They segregated with older F elements in trees for the 5’ UTR and all of ORF1 (subsequences α through δ). In trees for all of ORF2 and the 3’ UTR (subsequences ε through η), they segregated with younger F elements. This confirmed that TF L1’s are recombinant as Saxton and Martin (1998) concluded. The F-II 2 clade segregated similarly and is recombinant also. The TF and F-II 2 cladades always had a common ancestor but were strongly separated from each other (bootstrap values of 92–100%), with neither ance-

### Table 3. Average family-group/clade divergences (%)a

<table>
<thead>
<tr>
<th>Family groups or clades compared</th>
<th>5’ UTR</th>
<th>ORF1</th>
<th>ORF2</th>
<th>3’ UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>β</td>
<td>γ</td>
<td>η</td>
</tr>
<tr>
<td>A-I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-I</td>
<td>0.25</td>
<td>0.21</td>
<td>0.26</td>
<td>0.73</td>
</tr>
<tr>
<td>A-II</td>
<td>1.92</td>
<td>1.11</td>
<td>1.10</td>
<td>1.21</td>
</tr>
<tr>
<td>F-I</td>
<td>16.28</td>
<td>11.88</td>
<td>7.58</td>
<td>7.10</td>
</tr>
<tr>
<td>F-II</td>
<td>18.19</td>
<td>8.93</td>
<td>6.91</td>
<td>6.41</td>
</tr>
<tr>
<td>F-II 2</td>
<td>19.20</td>
<td>10.03</td>
<td>7.64</td>
<td>4.13</td>
</tr>
<tr>
<td>F-III</td>
<td>19.95</td>
<td>10.55</td>
<td>7.90</td>
<td>7.94</td>
</tr>
<tr>
<td>Tp1</td>
<td>21.58</td>
<td>9.09</td>
<td>6.56</td>
<td>6.92</td>
</tr>
<tr>
<td>Tp2</td>
<td>20.84</td>
<td>9.78</td>
<td>7.80</td>
<td>5.35</td>
</tr>
<tr>
<td>A-II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-II</td>
<td>1.33</td>
<td>0.59</td>
<td>0.86</td>
<td>1.36</td>
</tr>
<tr>
<td>F-I</td>
<td>16.14</td>
<td>11.68</td>
<td>7.69</td>
<td>7.21</td>
</tr>
<tr>
<td>F-II</td>
<td>18.31</td>
<td>8.41</td>
<td>7.24</td>
<td>6.61</td>
</tr>
<tr>
<td>F-II 2</td>
<td>18.97</td>
<td>9.28</td>
<td>7.37</td>
<td>3.54</td>
</tr>
<tr>
<td>F-III</td>
<td>19.88</td>
<td>10.36</td>
<td>8.18</td>
<td>7.97</td>
</tr>
<tr>
<td>Tp1</td>
<td>21.16</td>
<td>8.19</td>
<td>6.85</td>
<td>5.48</td>
</tr>
<tr>
<td>Tp2</td>
<td>20.49</td>
<td>9.77</td>
<td>7.87</td>
<td>5.32</td>
</tr>
<tr>
<td>F-II 2a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-II 2</td>
<td>8.12</td>
<td>9.34</td>
<td>7.87</td>
<td>7.84</td>
</tr>
<tr>
<td>F-II 2</td>
<td>10.29</td>
<td>9.64</td>
<td>7.52</td>
<td>7.50</td>
</tr>
<tr>
<td>F-III</td>
<td>10.49</td>
<td>11.70</td>
<td>7.72</td>
<td>8.56</td>
</tr>
<tr>
<td>Tp1</td>
<td>13.63</td>
<td>8.49</td>
<td>7.37</td>
<td>8.45</td>
</tr>
<tr>
<td>Tp2</td>
<td>12.60</td>
<td>9.39</td>
<td>8.38</td>
<td>8.20</td>
</tr>
<tr>
<td>F-II 2b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-II 2</td>
<td>10.69</td>
<td>7.73</td>
<td>7.88</td>
<td>7.49</td>
</tr>
<tr>
<td>F-II 2</td>
<td>12.76</td>
<td>7.60</td>
<td>7.92</td>
<td>6.70</td>
</tr>
<tr>
<td>F-III</td>
<td>11.38</td>
<td>9.76</td>
<td>8.42</td>
<td>8.52</td>
</tr>
<tr>
<td>Tp1</td>
<td>15.53</td>
<td>6.80</td>
<td>8.13</td>
<td>8.00</td>
</tr>
<tr>
<td>Tp2</td>
<td>14.34</td>
<td>8.13</td>
<td>8.63</td>
<td>7.80</td>
</tr>
<tr>
<td>F-III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-III</td>
<td>2.48</td>
<td>2.95</td>
<td>2.74</td>
<td>2.54</td>
</tr>
<tr>
<td>F-III</td>
<td>10.22</td>
<td>9.06</td>
<td>6.70</td>
<td>7.54</td>
</tr>
<tr>
<td>Tp1</td>
<td>13.82</td>
<td>4.82</td>
<td>3.68</td>
<td>4.96</td>
</tr>
<tr>
<td>Tp2</td>
<td>12.89</td>
<td>6.52</td>
<td>4.27</td>
<td>4.69</td>
</tr>
<tr>
<td>Tp1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tp1</td>
<td>0.00</td>
<td>0.87</td>
<td>0.52</td>
<td>0.63</td>
</tr>
<tr>
<td>Tp2</td>
<td>2.35</td>
<td>2.30</td>
<td>3.17</td>
<td>2.12</td>
</tr>
<tr>
<td>Tp2</td>
<td>0.75</td>
<td>0.30</td>
<td>0.43</td>
<td>0.31</td>
</tr>
</tbody>
</table>

*a Average percentage pairwise divergence values between the family groups and/or clades indicated in the first column. The values were computed for the same subsequences of the alignments used to generate the trees in Fig. 3.

b There are no intra-family-group values for F-I sequences because we had only one F-I sequence.
Fig. 3. Shown are the bootstrap-supported strict consensus, maximum-parsimony trees generated for all subsequences in Fig. 1 (except d and t of the testset L1's). The numbers beside some branches are the percentages of the bootstrap replicates, which support the cosegregation of the elements under the node below it. The results of both neighbor-joining and maximum-likelihood analyses of the subsequences were consistent with those from maximum-parsimony analysis. One rat sequence was included as a reference in these unrooted trees.
tral to the other. We consider these two segregation patterns when we use the testset to analyze other sequences. Fortunately, the F-II, clade and both TF clades segregate very predictably into distinct clusters strongly supported by bootstrap analysis in both patterns.

**F-IId’s Are Ancestral to A’s, the 3’ End of TF’s, and the 3’ End of F-II2’s**

In trees generated from the 5’ UTR and ORF1 (subsequences α through δ), there were two primary branches, one containing all the A’s and the other containing the TF’s and the F-II2’s. All A elements were in a strongly supported distinct cluster in the trees for all subsequences. For subsequences β, γ, and δ, Fmv1 (F-II1) descended from a node on the branch to the A elements, indicating that it might be related to a near ancestor of the A’s. For ORF2 and the 3’ UTR (subsequences e through τ), all A, TF, and F-II2 L1’s are on branches of the tree below the node for Fmv1. Sequence divergence computed for these subsequences shows that the A, TF, and F-II2 L1’s are all very similar to F-II1’s, and even more similar to each other. This implies that the F-II1 clade is ancestral to the entire sequence of A elements, except for the promoter, and to the ORF2 and 3’ UTR portions of TF and F-II2 elements.

**Older F Is Ancestral to the 5’ End of TF’s and the 5’ End of F-II2’s**

In trees generated from the 5’ UTR and ORF1 (subsequences α through δ), the TF’s and the F-II2’s segregated on a separate branch away from the A family. For those same subsequences, the F-III element, Fmn4, segregated to the same branch as the TF’s and the F-II2’s. When portions of known older F sequences were added to the trees, they also segregated to the branch with the TF’s and F-II2’s (not shown). This implies that an older F element is ancestral to the 5’ UTR and ORF1 portions of TF and F-II2 elements. This conclusion is supported by sequence divergence data. The 5’ UTR and at least the first half of ORF1 of TF’s and the F-II2’s are clearly less diverged from F-III’s than they are from A-II’s.

**Recombination Occurred Between an Older F and a Descendant of F-II1**

Taken together, these segregation patterns indicate that TF’s and F-II2’s are descended from a common ancestor. The founder of that lineage was formed by the recombination of the 5’ portion of an older F element and the 3’ end of one of the last F-II1 elements or one of the first A-II’s.

**Discussion**

We have conducted molecular phylogenetic and sequence divergence analyses of a larger and more representative set of full-length L1 sequences than had been used in previous investigations of *M. m. domesticus* L1 element evolution. A testset of intact full-length L1 sequences was carefully selected to include as many of the sublineages that are, or have been, transpositionally active. These L1 elements were analyzed for a series of subsequences spanning the entire body of the L1 element. We identified novel clades of L1 elements, some of which may be currently proliferating. We obtained evidence concerning ancestral relationships between these clades and have shown that recombination generated an ancestor to both the modern TF elements and the relatively young F-II1 clade. We propose the scheme shown in Fig. 4 to summarize the major events of *Mus* L1 evolution. This proposal integrates our new information, with that previously known, concerning ancestral relationships among L1 lineages. We also include estimates of the ages of most of the clades based upon the amount of sequence divergence within and between clades (see the legend to Fig. 4). We discuss this proposal starting with the currently active modern lineages, then working back in evolutionary time.

**Modern L1’s**

Two clades have been shown experimentally to have active members, A-I and TF2 (DeBerardinis et al. 1998; Saxton and Martin 1998). A third clade, TF1, also appears to be active for the following reasons. Hardies et al. (2000) identified a new “Z family,” which we believe is the same as our TF1 clade. They noted that the recent de novo insertion of a Z element, L1_beige, responsible for the beige mutation, indicates that this clade is also currently active. They observed that Z elements are close relatives of TF’s, but classified those three sequences separately because they did not fall within the 0.2% sequence divergence range reported by Naas et al. (1998) for TF’s. All three Z family sequences were 5’ truncated. Therefore, Hardies et al. (2000) were not able to determine the monomer based family of the elements. The Z fragments have higher sequence identity with the 3’ end of elements of our TF1 clade than with any other family group or clade we used. Our TF1 sequences are no more divergent from Z sequences than Z sequences are from each other. Our TF1 sequences also have the four, single-nucleotide substitutions that Hardies et al. (2000) determined were diagnostic for Z elements. We therefore conclude that the Z elements of Hardies et al. (2000) are the same as our TF1 clade. We classify these sequences as TF1’s based on the close similarity of their monomers to known TF2 monomers and their strongly bootstrap supported cosegregation with known TF1’s in maximum-parsimony trees.

Hardies et al. (2000) further proposed that TF L1’s (equivalent to our TF2 clade) were horizontally transferred to *M. m. domesticus* in the last 400,000 years from...
Origins of the Modern L1 Clades

A-I and A-II sequences show very high sequence identity throughout except for the 42-base difference in their LPRs. Sequence divergence data indicates that A-I’s are younger than A-II’s and evolved from A-II’s just over 200 KYA. Our analysis indicates that the A-II’s are direct descendants of an F-II element. They segregate as descendants of F-II’s in phylogenetic trees for all sub-sequences except the 5’ UTR. Sequence divergence by subsequence between F-II’s and A-II’s varies less than it does between them and any of the recombined clades. We expected that the estimated age of our A-II sequences, based on sequence divergence, would indicate how long ago the A promoters were acquired. However, A-II sequence divergence implies a date (600 KYA) that seems improbably recent. This implies that F-II’s were still active 600 KYA also. The youngest L1’s are the ones most likely to be intact. However, we found no very young F elements that were not recombined. The way we selected the testset may be of importance in evaluating our results. By choosing full-length, preferably unrecombined, L1’s for our testset, it is very likely that we introduced a bias for relatively young elements. This would result in lower divergences and underestimation of how long ago the major evolutionary events depicted in Fig. 4 occurred. Nonetheless, our conclusions about the times of other events in Fig. 4 are consistent with those of Saxton and Martin (1998) and Hardies et al. (2000).

Using probes specific for the A monomers, Saxton and Martin (1998) found evidence of abundant A monomers in Mus cervicolor. This implies that the A expansion began over 2 MYA unless there was a more recent horizontal transfer of A L1’s. The dotted portion of the arrow from F-II to A-II is an indefinite interval in the transition during which the A monomer was acquired. Annotation beside an arrow indicates significant insertion, deletion, and/or recombination events, in addition to base substitutions. The zigzagged arrow indicates an apparent recombination event between a late F-II or an early A-II element and an older F element. The approximate time of expansion, in millions of years ago (MYA), was calculated as 1 MYA/1% divergence from a common ancestor (Li et al. 1987). The percentage divergence from a common ancestor was assumed to be one-half of the weighted average of the intradrome or intraspecies group subsequence divergence values from Table 3. Because we had only one each of the TpI, Tna1, and Tpa2 elements, the divergence between the Tpa2-HII’s and Tna2 was used to calculate how long before the Tp2’s that Tna2 arose (~1.2 million years earlier). This leads to an estimate of 1.3 MYA (1.2 MYA + 140 KYA) for the age of the single TpI-II sequence. We then compared the divergence between Tna1 and Tna2 to calculate how long after the TpI-II’s that the Tpa2 sequence gives 850 KYA as the approximate age of the TpI-II sequence.

Fig. 4. Mus L1 evolution. This figure depicts a simple model to explain the relationships among known full-length L1 sequences. Family-group names (with clade designations where appropriate) are given for the major subclasses of currently known Mus L1 elements. Asterisks indicate currently active L1 clades. The question mark indicates a clade that may be currently active (see text and Hardies et al. 2000). Solid, dashed, and dotted arrows indicate evolutionary change by base substitutions. In addition, dashed arrows indicate alternate transition possibilities for which there is insufficient evidence to choose between. The dotted portion of the arrow F-II-I to A-II is an indefinite interval in the transition during which the A monomer was acquired. Annotation beside an arrow indicates significant insertion, deletion, and/or recombination events, in addition to base substitutions. The zigzagged arrow indicates an apparent recombination event between a late F-II or an early A-II element and an older F element. The approximate time of expansion, in millions of years ago (MYA), was calculated as 1 MYA/1% divergence from a common ancestor (Li et al. 1987). The percentage divergence from a common ancestor was assumed to be one-half of the weighted average of the intradrome or intra-family group subsequence divergence values from Table 3. Because we had only one each of the TpI-I (Tna1) and Tpa2-II (Tna2) elements, the divergence between the Tpa2-II’s and Tna2 was used to calculate how long before the Tp2’s that Tna2 arose (~1.2 million years earlier). This leads to an estimate of 1.3 MYA (1.2 MYA + 140 KYA) for the age of the single TpI-II sequence. We then compared the divergence between Tna1 and Tna2 to calculate how long after the TpI-II’s that the Tpa2 sequence gives 850 KYA as the approximate age of the TpI-II sequence.

another mouse species. It seems certain from the very low divergence between the two M. m. domesticus Tp clades that the Tp2 is a direct descendant of Tp1. Since we have no domesticus sequences intermediate between these clades, the implied evolution could have occurred in another mouse species and then have been horizontally transferred back to domesticus. Our estimate of 1.3 MYA for the appearance of Tp1’s suggests that that clade arose prior to the spretus/domesticus split. Therefore, it seems plausible that Tp2 evolved from Tp1 in spretus or some more recently diverged species, following the split, and then was transferred back to domesticus.
the earliest that the recombination event might have occurred. Therefore, the recombination that produced L1Rec likely occurred after the A monomer was acquired by L1. The young element involved in the recombination could have been either an A-II or an F-II1, if this lineage continued to proliferate after the A monomer was acquired. No A-type elements appear ancestral to the recombinant clades (F-II2, Tp1, and Tp2) in our trees. However, it is quite possible that the young element involved in the formation of L1Rec was an A-type element older than any of those represented in our testset.

Some observations suggest that the older element that donated the 5′ portion of L1Rec might have been an F-III L1. The Tp and F-II2 clades descended from, or with, Fnn4 (F-III) in maximum-parsimony trees for the 5′ UTR and all of ORF1 (subsequences α through δ). Divergence values for the same subsequences are consistent with the suggestions of the parsimony trees. However, the conclusion is problematic because the recombinant portion of the sequence includes the LPR region of ORF1. Therefore, we would expect the recombination product to be an F-III L1 rather than an F-II.

The Origin of the F-II1 Clade

It is not clear which L1 clade is immediately ancestral to F-II1. We depict the two most obvious possibilities, F-IV and F-III. Transition from either an F-IV or an F-III element requires two precise sequence changes in the LPR in addition to base substitutions. Transition from an F-IV to an F-II should produce a, so far undetected, new intermediate family group with two 66-base pair LPR modules and one 42-base pair module. Transition from an F-III to an F-II would produce an F-I intermediate. The only full-length F-I sequence we have, F15, is >7% divergent from each of the other family groups and clades. Though highly divergent, some portions of F15 are more like A elements than F’s. Therefore, F15 does not provide a basis for deciding to what family group the immediate ancestor of the F-II1′s belonged. Although we show F-I′s evolving from F-II1′s, there is no clear evidence that this is correct. Divergence values suggest that the F-III L1′s we used may have been active at the same time as the F-II1 sequences.

The Origin of the F Family and the Older F-IV and F-III

We did not find any full-length F-IV or V L1′s, therefore we could not generate additional information about early L1 evolution in domesticus. Adey et al. (1994b) showed that L1′s in the rat and all subsequent L1′s in mice evolved from V-family elements, active at the time of the Rattus/Mus split. They also demonstrated that the F family descended from the V family, the A′s descended from the F′s, and each of those families was composed of closely related clades each having the same LPR group in common. The most divergent, and presumably the oldest, F sequences analyzed by Adey et al. (1994b) were F-IV′s.

Open Questions and Future Directions

Soon, enough of the M. m. domesticus genome will be sequenced to allow identification of several more full-length elements of each of the most recently arising clades. Then, consensus sequences for these clades should provide a more reliable basis for further resolving the details of recent L1 evolution in the domesticus genome. We anticipate finding older A-II sequences that support the acquisition of the A monomer promoter between 2 and 3 MYA. We also expect to find F-II, sequences that are much closer ancestors of the F-II2, Tp, and A L1′s than those in this study are. We also expect to find intermediate F-II1 to older F recombinant L1′s. If so, these intermediates might have group III LPRs or provide evidence that recombination took place by multiple recombination events involving fractions of the total sequence, which appears to be recombinant. Failure to find sequences intermediate to the two Tp clades would argue in favor of the horizontal transfer of the Tp2 clade into domesticus as Hardies et al. (2000) concluded.

We did not identify any full-length A-III, F-IV, or V elements. If suitable full-length representatives of these older L1′s are found, it may be possible to get better overall resolution in the trees and more clearly deduce the order of earlier events leading to the birth of the modern Tp and A lineages. Among these sequences, we expect to find sequences clarifying whether F-IV-to-F-II1 evolution occurred directly or through an F-III intermediate.

Our conclusions provide clarification about the portion of the F family from which the A family descended but still leave the questions of how the A monomers were acquired and from where. Our conclusions also leave unanswered the question of the precise event that led to the rapid expansion of Tp L1′s. The 5′ end of these elements may represent an F lineage that remained transposition competent at a very low frequency after the end of the F family expansion and underwent an accelerating event less than 3 MYA. The 5′ end of the Tp lineage also may be descended from an F element that was inactive and underwent a reactivating event less than 3 MYA. The second scenario seems less likely because there would have been no selective pressure to preserve the required open reading frames. Another, possibly related, question is whether the recombination event provided a selective advantage for the Tp′s. If so, why did they flourish, while the F-II1′s apparently did not? They also may have evolved as an active element in another mouse species and have been horizontally transferred into domesticus less than 400 KYA as concluded by Hardies et
al. (2000). This is an interesting suggestion in light of the theory that L1’s may be down-regulated by cytosine methylation in their promoters (Yoder et al. 1997). According to this theory, transposing repeat elements are more likely to be methylated as their numbers increase. Horizontal transfer of T<sub>F</sub> into Mus musculus could be the activating event which led to the rapid expansion of the T<sub>F</sub> subfamily. The first several generations after introduction would dilute the number of T<sub>F</sub>’s to a fraction of their number in the species from which they were transferred. This would allow them to expand, unchecked, until their numbers evoked down-regulating methylation. This mode of activation could equally apply to A- and F-family activation.

The results of our analyses indicate that all L1 sub-sequences are potentially diagnostic for the L1 family, defined by promoter type, to which the sequence belongs. That the 3’ UTR portion of L1 elements is diagnostic for the L1 family to which the sequence belongs may be of special relevance. This is especially true for the very youngest groups of A and T<sub>F</sub> elements. For full-length A-I, A-II, and T<sub>F</sub> elements, at least, the type of LPR of the very youngest groups of A and T<sub>F</sub> elements. For full-length A-I, A-II, and T<sub>F</sub> elements, at least, the type of LPR that is at the 5’ end and the type of LPR that is in ORF1 can be predicted by the activation of the 3’ UTR. This should be true for truncated elements as well. Because the majority of mouse L1’s have 5’ truncations, this relationship should make possible a more accurate estimate of the proportions of the total number of L1’s comprised by each of the young clades.

We are currently identifying 3’ fragments of young L1’s inserted throughout genomic sequences, which are expected to be very polymorphic (DeBerardinis et al. 1998). By designing PCR primers based on sequences flanking truncated L1’s, we hope to generate a large set of polymorphic markers. Such a set of markers would have a number of valuable applications. These applications include use in gene mapping, in fine-tuning our knowledge of recent mouse evolution, and possibly in determining the ancestry of lab strains. Because of the high degree of similarity among L1’s of all mammals, it is likely that similar relationships exist in other mammals, in which distinct subsets of L1’s exist, possibly including humans.

Acknowledgments. We thank Dr. Nils B. Adey for his advice and sharing data from his previous work with us. We also appreciate helpful criticism of the manuscript from Dr. Michael J. Conrad, Michael Monteagudo, Tony Cesare, and Gwynedd Smith. This work is supported in part by NIH Grant GM21313.

References

Adey NB (1991) Phylogenetic analysis of mouse L1 elements, PhD dissertation. Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill
Adey NB, Comer MB, Edgell MH, Hutchison CA III (1991a) Nucleotide sequences of a mouse full-length F-type L1 element. NAR 19:2497
Morse B, Rothberg PG, South VJ, Spandorfer JM, Astrin SM (1988)
Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. Nature 333(6168):87–90
Saxton JA, Martin SL (1998) Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. JMB 280:611–622