Positive Selection and Multiple Losses of the LINE-1-Derived \textit{L1TD1} Gene in Mammals Suggest a Dual Role in Genome Defense and Pluripotency

Richard N. McLaughlin Jr.\textsuperscript{1,3}, Janet M. Young\textsuperscript{1,3}, Lei Yang\textsuperscript{2,6}, Rafik Neme\textsuperscript{3}, Holly A. Wichman\textsuperscript{2}, Harmit S. Malik\textsuperscript{1,4*}

\textsuperscript{1}Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, \textsuperscript{2}Department of Biological Sciences & Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, United States of America, \textsuperscript{3}Max-Planck Institute for Evolutionary Biology, Plön, Germany, \textsuperscript{4}Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

Abstract

Mammalian genomes comprise many active and fossilized retroelements. The obligate requirement for retroelement integration affords host genomes an opportunity to ‘domesticate’ retroelement genes for their own purpose, leading to important innovations in genome defense and placentation. While many such exaptations involve retroviruses, the \textit{L1TD1} gene is the only known domesticated gene whose protein-coding sequence is almost entirely derived from a LINE-1 (L1) retroelement. Human \textit{L1TD1} has been shown to play an important role in pluripotency maintenance. To investigate how this role was acquired, we traced the origin and evolution of \textit{L1TD1}. We find that \textit{L1TD1} originated in the common ancestor of eutherian mammals, but was lost or pseudogenized multiple times during mammalian evolution. We also find that \textit{L1TD1} has evolved under positive selection during primate and mouse evolution, and that one pseudomammalian \textit{L1TD1} has ‘replenished’ itself with a more recent L1 ORF1 from the prosimian genome. These data suggest that \textit{L1TD1} has been recurrently selected for functional novelty, perhaps for a role in genome defense. \textit{L1TD1} loss is associated with L1 extinction in several megabat lineages, but not in sigmodontine rodents. We hypothesize that \textit{L1TD1} could have originally evolved for genome defense against L1 elements. Later, \textit{L1TD1} may have become incorporated into pluripotency maintenance in some lineages. Our study highlights the role of retroelement gene domestication in fundamental aspects of mammalian biology, and that such domesticated genes can adopt different functions in different lineages.

Introduction

Retroelements have profoundly shaped mammalian genomes over millions of years. Insertion of these selfish elements can lead to gene inactivation or changes in transcriptional profiles of neighboring genes. Moreover, the presence of large stretches of almost identical sequence distributed across the genome poses recombinational hazards, leading to chromosomal rearrangements, often with pathological consequences. With a few celebrated exceptions [1], the retrotransposition activity of retroelements is rarely beneficial [2]. However, individual transposition events can occasionally drive dramatic episodes of adaptation through the generation of genetic novelty [3]. Indeed, repetitive elements have often been co-opted as transcriptional regulatory elements like promoters, enhancers, and insulators [4].

One particularly striking class of such genetic novelty results from the ‘domestication’ or ‘capture’ of retroelement- and retrovirus-derived protein-coding genes by host genomes, a process by which these domesticated coding regions are exapted for a new function that is beneficial to the host [5,6]. Such instances are often recognized by the preservation of individual protein-coding frames despite the mutational decay of the rest of the parental retroelement. This extinction of the replication-competent retroelement but not its protein-coding gene implies that selection for the benefit of the host genome must have prevented the mutational attrition of that particular gene.

Retroelements and retroviruses must integrate into the host genome as part of their replication cycle, presenting the host with a source of potentially advantageous protein-coding sequences. The \textit{syncytin} genes of eutherian mammals represent some of the best characterized examples of such domestication events in which
Transposable elements comprise major portions of most animal genomes and are selfish genetic elements that may encode proteins needed for their own spread to new genomic locations. Though often considered genomic parasites, these elements also occasionally create novel genes that prove beneficial to the host, a process called ‘domestication’. Here, we describe the evolution of a gene, L1TD1, which is derived from the protein-coding regions of the L1 mobile element family. We show that L1TD1 was born in the common ancestor of placental mammals. L1TD1 expression in stem cells and its requirement to maintain the pluripotent state of human embryonic stem cells suggested it might have been originally domesticated for such a pluripotency role. We find that L1TD1’s evolution does not fit with the predictions of this model; in fact, L1TD1 has rapidly evolved in primates and mice and has been lost several times in mammals. We suggest an alternate model that L1TD1 was born as a means to defend genomes against transposable elements, perhaps L1 itself. We propose that following this initial domestication, L1TD1 later became incorporated into pluripotency programs in some mammalian lineages.

envelope genes from ancient retroviruses have been preserved for their membrane-fusing and/or immunosuppressive activities in the syncytiotrophoblast, the layer of the placenta which mediates maternal-fetal nutritional transfer [7–9]. In fact, loss of syncytin-A in mice is embryonic lethal, consistent with its indispensable role in placental function [10]. Syncytin genes thus represent a dramatic example of the maintenance and possibly invention of an essential function, placentation, via retroviral gene domestication. Importantly, syncytin is not unique in this regard; similar domestications of sushi-ichi LTR retrotransposon protein-coding regions have also created multiple host genes (e.g., Peg10) involved in genomic imprinting and placentation [11,12].

Though syncytin represents a dramatic and beneficial genetic innovation, not all domesticated retroviral genes serve conserved, essential functions. The Fv1 gene in mice represents a domesticated gag (viral capsid and nucleocapsid encoding) gene, which can actively interfere with the uncoating of incoming retroviral capsids [13,14]. As expected for this genome defense role, Fv1 is under strong diversifying selection, presumably as a result of constant innovation required to recognize and block different retroviruses [15]. As a result of this diversifying selection to chase the sequence of incoming capsids, orthologous genes can have dramatically different antiviral specificities [16–19]. Fv1 has also been lost or pseudogenized at least twice in the Mus genus [15], perhaps because retention of Fv1 may depend on persistence of selection from incoming viruses.

Mammalian genomes acquired both syncytin and Fv1 genes as a result of the insertion of a retrovirus into the germline. Though such endogenous retroviruses have markedly impacted the mammalian genome, these elements have spent relatively little time coevolving with their host compared to the non-LTR retroposons, which date back at least to the origin of Metazoa [20]. The LINE-1 (Long INTerspersed Element-1, L1) non-LTR retroposons make up a significant fraction of the human genome. Given their ancient history of coevolution with mammalian genomes, it is not surprising that there are numerous examples of exaptation of non-LTR retroelements into non-coding RNAs, as promoter or other regulatory elements, or as small portions of a coding region [4,6,21–24].

Results

L1TD1 originated in the common ancestor of placental mammals

To gain insight into the cellular function of L1TD1, we decided to date its evolutionary origin and examine how its gene structure has changed since its birth. Since the only published analysis examined five mammalian genomes for its presence [33], we decided to search for the L1TD1 gene in a much larger sample of diverse mammalian genomes. In the human and mouse genomes, L1TD1 is found between the single-copy INADL and KANK4 genes (Figure 1A). We found that all mammals and even bird genomes encode INADL and KANK4 in close proximity to each other. We therefore used these flanking genes to identify the syntenic locus in other mammalian genomes, and used the sequence from this locus to identify L1TD1 coding sequences where present. We also extended our search to the rest of the genome, using repeat-masked and unmasked human and mouse L1TD1 sequences as BLAST search queries. We were never able
Two L1 ORF1-homologous regions constitute the \textit{L1TD1} gene

Human \textit{L1TD1} (RefSeq NM_001164835.1) comprises two protein-coding exons (of four total exons), which together encode an 865 amino acid protein. Each of these two exons is homologous to the first open reading frame (\textit{ORF1}) of \textit{L1} (Figure 1C), whose protein product (\textit{ORF1p}) functions as an RNA-binding protein that greatly enhances \textit{L1} retrotransposition [34]. \textit{ORF1p} appears to be important in ensuring \textit{L1} ‘cis-preference’ - the preference for the \textit{ORF2} protein (\textit{ORF2p}) to act upon the same RNA from which it was translated, so that an \textit{L1}’s machinery is less often ‘hijacked’ by other elements or other \textit{L1s} [34–36]. \textit{L1TD1} has no discernible homology to the other open reading frame of \textit{L1}, \textit{ORF2}, which encodes the enzymatic activities of \textit{L1}. The second coding exon of \textit{L1TD1} has higher conservation with \textit{ORF1p}, showing 13% amino acid identity (58% amino acid similarity) with \textit{ORF1p} of human \textit{L1}.3. In coding exon 1, only the CTD is conserved, while in coding exon 2, the CC, RRM, and CTD are all conserved. Coding exon 2 also contains a variable length glutamic acid-rich region (ER). After splicing, the human \textit{L1TD1} transcript is 3849 nucleotides in length and encodes a single 865 amino acid protein product.

do:10.1371/journal.pgen.1004531.g001

Figure 1. The mammalian \textit{L1TD1} gene was born from a tandem insertion of \textit{L1} \textit{ORF1}. A. Genomic context of \textit{L1TD1}. Single-copy genes \textit{INADL} and \textit{KANK4} flank \textit{L1TD1} in human, mouse and dog genomes; this shared synteny arrangement helped us identify \textit{L1TD1} orthologs in other mammalian genomes. B. \textit{L1TD1} evolved according to the accepted species tree. We aligned \textit{L1TD1} nucleotide sequences and generated a maximum-likelihood phylogeny (see Materials and Methods). Bootstrap values show the percentage of 1000 replicates in which descendental taxa cluster together, and the scale bar shows substitutions per site according to the GTR+I+G evolutionary model. C. \textit{L1TD1} comprises two \textit{L1-ORF1p}-like regions. \textit{L1} elements encode an approximately 6.5 kb transcript containing two open reading frames. \textit{L1 ORF1} encodes a protein, \textit{ORF1p}, with RNA-binding and chaperone activity. \textit{ORF1p} contains a coiled-coil motif (CC), a RNA-recognition motif (RRM), and a C-terminal domain (CTD). \textit{ORF2} encodes a protein with endonuclease and reverse transcription enzymatic functions. Sequence identity demonstrates that \textit{L1TD1} was formed from the domestication of two copies of \textit{ORF1} from \textit{L1}. The two copies may derive from independent insertions or from duplication after a single insertion. Human coding exon 1 and coding exon 2 share 30% and 43% amino acid identity, respectively, with \textit{ORF1p} of human \textit{L1}.3. In coding exon 1, only the CTD is conserved, while in coding exon 2, the CC, RRM, and CTD are all conserved. Coding exon 2 also contains a variable length glutamic acid-rich region (ER). After splicing, the human \textit{L1TD1} transcript is 3849 nucleotides in length and encodes a single 865 amino acid protein product.

Origin and Evolution of \textit{L1TD1} in Mammals

to identify an intact copy of \textit{L1TD1} outside the \textit{INADL-KANK4} syntenic locus, although we did find some pseudogenes with obvious inactivating mutations in other locations. Although the previously published analysis [33] suggested that dog \textit{L1TD1} lacks a portion of coding exon 1, in fact we find that \textit{L1TD1} is complete and intact in the dog genome (Table S2).

We found \textit{L1TD1} in many diverse placental mammalian genomes (Table S2, Dataset S1). However, \textit{L1TD1} is absent from both marsupial and platypus genomes. This species distribution implies that the \textit{L1TD1} gene was born in the common ancestor of placental mammals, after the split from marsupials. A phylogenetic tree using \textit{L1TD1} nucleotide sequences follows the expected species tree (Figure 1B). Together with the observed shared syntenic location, this tree demonstrates that the sequences we identified represent truly orthologous genes rather than several independent \textit{L1} domestication events. This dates the origin of \textit{L1TD1} to at least 106 million years ago, making it more ancient than the well-known domesticated retroelement genes \textit{-syncytin} and \textit{Fv1 [9,15].

Two \textit{L1} ORF1-homologous regions constitute the \textit{L1TD1} gene

Human \textit{L1TD1} (RefSeq NM_001164835.1) comprises two protein-coding exons (of four total exons), which together encode an 865 amino acid protein. Each of these two exons is homologous to the first open reading frame (\textit{ORF1}) of \textit{L1} (Figure 1C), whose protein product (\textit{ORF1p}) functions as an RNA-binding protein that greatly enhances \textit{L1} retrotransposition [34]. \textit{ORF1p} appears to be important in ensuring \textit{L1} ‘cis-preference’ - the preference for the \textit{ORF2} protein (\textit{ORF2p}) to act upon the same RNA from which it was translated, so that an \textit{L1}’s machinery is less often ‘hijacked’ by other elements or other \textit{L1s} [34–36]. \textit{L1TD1} has no discernible homology to the other open reading frame of \textit{L1}, \textit{ORF2}, which encodes the enzymatic activities of \textit{L1}. The second coding exon of \textit{L1TD1} has higher conservation with \textit{ORF1p}, showing 13% amino acid identity (58% amino acid similarity) with \textit{ORF1p} of human \textit{L1}.3. In coding exon 1, only the CTD is conserved, while in coding exon 2, the CC, RRM, and CTD are all conserved. Coding exon 2 also contains a variable length glutamic acid-rich region (ER). After splicing, the human \textit{L1TD1} transcript is 3849 nucleotides in length and encodes a single 865 amino acid protein product.
diverse mammalian genomes (Figure 2). We find that L1TD1 N-terminal regions (the protein product of coding exon 1) group together with strong bootstrap values (node A), as do L1TD1 C-terminal regions (the protein product of coding exon 2) (node B), demonstrating that the double ORF1p structure arose just once since the divergence of placental mammals and has not been subject to gene conversion between the two exons since. Furthermore, the L1TD1 N-terminal and C-terminal clades branch off from placental mammal L1-ORF1p sequences (node C) after marsupial and placental mammal L1-ORF1p sequences diverged (node D). This supports our conclusion that L1TD1 was born via L1 domestication in the ancestral eutherian mammal. Our phylogeny cannot distinguish whether the two homologous regions of L1TD1 derived from two independent L1 insertions in proximity to each other, or whether a single L1 inserted and subsequently experienced tandem duplication. Examination of the relative splice acceptor positions in coding exons 1 and 2 also does not help to distinguish these two possibilities. Nevertheless, it does appear that both coding exons were born in close temporal and physical proximity to each other, giving rise to the ancestral L1TD1 gene. Our analyses further show that this bipartite double-ORF1p organization of L1TD1 has been conserved since its birth.

Multiple, independent losses of L1TD1 in mammals

Despite its widespread conservation, our genome-wide searches nevertheless revealed that L1TD1 has been lost on at least three separate occasions during the evolution of placental mammals (Figure 3). First, we find that L1TD1 is missing in all three Afrotherian genomes we surveyed – elephant, hyrax and tenrec. These three species constitute an approximately 85 million-year-old clade. However, we find L1TD1 is intact and present in the armadillo genome, which diverged from the three sequenced members of Afrotheria nearly 100 million years ago. Using parsimony, we infer that L1TD1 was lost once in the ancestor of elephant, hyrax and tenrec between 84 and 100 million years ago.

Second, we find that L1TD1 is missing from the genomes of all Cetartiodactyla. This clade of mammals originated 64 million years ago and comprises even-toed ungulates, whales, and dolphins. All sequenced genomes of this clade (cow, sheep, dolphin, pig and alpaca, Figure 3) lack L1TD1. In contrast to its loss in Cetartiodactyla, L1TD1 is present in nearly all other members of the 80 million year old Laurasiatheria clade, which includes Cetartiodactyla. The most parsimonious explanation is that the common ancestor of Cetartiodactyla lost L1TD1 between 64 and 80 million years ago.

In both the Afrotheria and Cetartiodactyla, we find no trace of the L1TD1 gene anywhere in the genome, perhaps indicating loss of L1TD1 via genomic deletions of the whole gene. However, given that the Afrotheria and Cetartiodactyla loss events likely occurred at least 64 and 84 million years ago, it is equally likely that after pseudogenization the L1TD1 sequence has simply degenerated beyond recognition in these lineages.

The third instance of loss of L1TD1 occurred in the genome of the megabat Pteropus vampyrus (Figure 3). The microbat Myotis lucifugus also encodes an intact L1TD1, but P. vampyrus contains a pseudogenized version of L1TD1 with multiple frameshifts and stop codons. In contrast to the complete lack of a recognizable L1TD1 in the Afrotheria and Cetartiodactyla, L1TD1 is still discernible in the megabat, suggesting this loss may be more recent.

Three of these loss events in mammals strongly challenge the possibility that L1TD1 was born into a role of pluripotency maintenance in the common ancestor of placental mammals, as we would expect such a gene to be essential and not subject to subsequent loss. It is formally possible that these lineages independently acquired another L1TD1-like activity that allowed the loss of the originally acquired L1TD1. However, it is more parsimonious that L1TD1’s role in pluripotency maintenance was acquired much later than its birth.

The loss of L1TD1 in mammals parallels the multiple losses of the Fv1 restriction factor in the Mvs genus [15]. We therefore considered whether, like Fv1, L1TD1 may have rapidly evolved under positive selection due to some role in a genetic conflict.
evolutionary changes occur more often than expected given the rate of neutral mutations, with values $<1$ indicating overall conservation, and values $>1$ indicating overall positive selection.

Though we observe an overall signature of purifying selection in $L1TD1$, there could nonetheless be signatures of diversifying selection occurring on just a few domains or residues. To assess whether positive selection has acted on primate $L1TD1$, we used maximum likelihood methods (NSsites models in the PAML suite [39]) to test for positive selection at individual codons. We found that a model permitting positive selection in the alignment fit the primate data significantly better than models that disallow positive selection (MB vs. M7, M8 vs. M$\delta_8$, $p<0.01$). Both codon exons show evidence of positive selection, suggesting that both domesticated ORF1p-like regions have been recurrently selected for functional novelty. Further, the positive selection is remarkably localized. Only a few $L1TD1$ codons (~1%) show a signature of positive selection, but these codons have a high average $dN/dS (>8$, Table 1). Such highly localized signatures of recurrent positive selection often represent direct contacts with antagonistic entities (discussed below). We note there is no overlap between the positively selected sites in primate $L1TD1$ and the regions previously found in L1 ORF1p, which are largely in the coiled-coil domain [38,40]. FUBAR, carried out in the HyPhy suite of programs confirmed our finding of highly localized positive selection (Materials and Methods, Table 1) [Murrell, 2013 #].

Since we hypothesize that the ancestral function of $L1TD1$ may have been genome defense, we expect that $L1TD1$ should have evolved under positive selection within many branches of the mammalian phylogeny. To generalize our finding of pervasive positive selection in the primates, we analyzed the evolution of $L1TD1$ within the genus M. We assembled 10 complete $L1TD1$ sequences from databases, PCR-based sequencing, and RNA-seq data. Again, all species contained an intact coding sequence with a high degree of conservation (average $dN/dS = 0.46$). However, similar to the primate analysis, PAML NSsites found a small proportion of sites with a high average $dN/dS (~3\%$, $dN/dS > 12$), as well as a highly significant gene-wide signature of positive selection (Figure 4A; M8 vs. M7, M8 vs. M$\delta_8$, $p<0.001$). We also found statistical support using FUBAR for 8/12 positions identified as positively selected by PAML (Table 1; PAML M7 vs M8 BEB, $P>0.9$; FUBAR $P>0.9$). To eliminate any false positive selection signals that could arise from phylogenetic discordance, we used the HyPhy program GARD to identify potential recombination breakpoints in the Mus $L1TD1$ alignment [41]. The alternative phylogenies given by GARD attempt to correct for any recombination that may have occurred and given rise to a scenario where no single tree accurately fits the entire $L1TD1$ sequence. While GARD found no statistically significant breakpoints according to the KH test, we nonetheless tested whether Mus $L1TD1$ retained a statistically significant signature of positive selection integrating the generated alternative phylogenies (Figure S1). Using the GARD-generated trees, we still observed a strong signature of gene-wide positive selection according to PAML ($p<0.01$) and FUBAR still identified 7 positively selected positions.

In addition to phylogeny-wide selection at a few specific codons, branch-specific analyses of $dN/dS$ ratios highlighted episodic positive selection along a few specific branches of the primate phylogeny including the branch leading to the common ancestor of spider monkey and titi monkey, as well as the branch leading to the OWMs and hominoids (Figure 4A, red asterisks, Branch-site REL, $p<0.05$) [41]. Using PAML, the branch leading to the common ancestor of gorilla, human and chimpanzees (after the split from orangutan) showed a whole gene $dN/dS$ of 2.87 (1
synonymous change and 11 nonsynonymous changes); the branch preceding it shows a whole gene dN/dS of 1.671 (1 synonymous change, 5 nonsynonymous changes). While neither of these ratios is significantly greater than one, it is notable that these two branches span a time window ~9–20 Mya, overlapping the ~40–12 Mya time period in the lineage leading to humans shown to exhibit positive selection in L1-ORF1 and of particularly intense L1 activity [38].

Our finding that LITD1 has evolved under positive selection in primates and mice would be unexpected if its sole function was in pluripotency; we would expect such genes to be highly conserved. We excluded an intriguing alternative possibility that genes involved in pluripotency might not be evolving as slowly as one would intuitively assume. In a genome-wide analysis of dN/dS values calculated from trios of human-chimp-macaque orthologs [42], genes identified in a screen for determinants of ES cell identity [43] are indeed evolving more slowly than control genes (Wilcoxon p = 0.008; Figure S2). Thus, our finding of positive selection in LITD1 is indeed unexpected if its only role were in pluripotency maintenance.

We cannot evaluate the extent to which positive selection has shaped protein regions that have experienced length-changing insertion-deletion changes, because they are unsuitable for codon-based analyses of positive selection. However, we note significant divergence in the glutamic-acid rich region of LITD1 that separates the two L1 ORF1p homology regions. For instance, New World monkeys have a deletion of ~140 amino acids at the beginning of LITD1 coding exon 2 relative to the Old World monkeys and hominoids. The functional significance as well as the selective pressures that might have driven these changes is unknown; it is quite possible that these changes are completely neutral and have little impact on LITD1 function.

We also discovered a substantial restructuring of the LITD1 gene of bushbaby (Otolemur garnettii) that may have a significant impact on its function. In the bushbaby genome, we found that the latter two-thirds of the LITD1 coding region are completely typical and align well with orthologous sequences from other primates and mammals, with no stop codons or frameshifts. However, the first one-third of bushbaby LITD1 is not orthologous to any other LITD1 [Figure 4B]. Instead, it appears the bushbaby LITD1 has acquired an entirely novel 5′ end. Two processed LITD1 pseudogenes in the bushbaby genome also include this novel 5′ end. Since processed pseudogenes lack introns, they provide independent confirmation of the transcript structure of LITD1. On closer examination, the novel 1400 nucleotides at the beginning of the bushbaby LITD1 includes ~360 nucleotides of protein-coding sequence from ORF1p of a L1 element of the L1PA15-16 class, a primate-specific L1 element. The remaining ~1000-nucleotide 5′ UTR region comprises a patchwork of several other repetitive elements, and likely arose by a series of nested insertion events. Although the repetitive element portions of this 5′ UTR are found in several other loci in the bushbaby genome, the only places they are found in this particular combination are at the LITD1 syntonic locus and in the LITD1 processed pseudogenes (Figure 4B). Thus, it appears that bushbaby LITD1 has undergone a recent remodeling, replenishing its N-terminal region with a more current version of a L1 element than the version captured in the original domestication event ~100 million years ago. The newly replenished version of bushbaby LITD1 has evolved under purifying selection (Figure S3), ruling out the alternate possibility that this LITD1 rearrangement is a pseudogenization event. This novel LITD1 structure is not evident in the genomes of two lemur species, the gray mouse lemur (Microcebus murinus) and the aye-aye (Daubentonia madagascariensis), which both encode a more typical LITD1, and we have not observed evidence of bushbaby-like LITD1 restructuring in the any other mammalian genomes examined (Figure 3). This implies that the bushbaby LITD1 remodeling occurred after the split between the lemur and bushbaby lineages, nearly 60 million years ago. Closer examination of other prosimian lineages will allow a more precise dating of this event.

From these forms of variation in the primate and mouse lineages, we conclude that genetic innovation has been adaptively selected for in LITD1 through a number of mechanisms including mutation of individual amino acids and perhaps expansions and contractions of the glutamic-acid-rich region. Importantly, the fixation of a new gene structure in the bushbaby derived from a recently active L1 suggests selection for functional novelty through yet another L1-derived sequence. The signature of positive selection in both mice and primates suggests LITD1 may be coevolving with some pathogen. Intriguingly, a strong signature of positive selection, indicative of such ‘arms-races’, has been previously seen in primate L1 evolution, with an especially striking signature in ORF1p [38,40]. Combined with its origin as a domesticated L1 gene and the novelty in bushbaby (which further emphasizes the selection pressure to maintain the L1-like character of LITD1), we considered whether LITD1 could be engaged in a genetic conflict with L1 or some other retroelement. This function could parallel the known role of another domesticated retroelement gene, Fov, in defense against elements similar to its progenitor [13,14]. To more clearly understand this hypothetical relationship, we reasoned that genes that function solely in L1 restriction might be lost in species where active L1s have potentially antagonistic relationship between LITD1 and L1 by looking in bat species where the most detailed studies of L1
Table 1. Primate and mouse L1TD1 are evolving under positive selection.

<table>
<thead>
<tr>
<th>Species Group</th>
<th>Model</th>
<th>2.</th>
<th>lnL</th>
<th>df</th>
<th>p-value</th>
<th>Position</th>
<th>AA</th>
<th>PAML</th>
<th>FUBAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primates</td>
<td>M7 vs M8</td>
<td>11.93</td>
<td>2</td>
<td>0.0026</td>
<td></td>
<td>43</td>
<td>F</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>156</td>
<td>Y</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>173</td>
<td>K</td>
<td>0.56</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>229</td>
<td>K</td>
<td>0.68</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>329</td>
<td>G</td>
<td>0.71</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>887</td>
<td>S</td>
<td>0.91</td>
<td>0.42</td>
</tr>
<tr>
<td>Mice</td>
<td>M7 vs M8</td>
<td>17.23</td>
<td>2</td>
<td>3.27 x 10^-8</td>
<td>95</td>
<td>I</td>
<td>0.92</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>108</td>
<td>L</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>147</td>
<td>G</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>152</td>
<td>R</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>163</td>
<td>S</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>167</td>
<td>W</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>178</td>
<td>A</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>336</td>
<td>R</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>346</td>
<td>E</td>
<td>0.85</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>436</td>
<td>S</td>
<td>1.00</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>442</td>
<td>A</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>447</td>
<td>V</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>530</td>
<td>G</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>633</td>
<td>S</td>
<td>0.91</td>
<td>0.90</td>
</tr>
</tbody>
</table>

1 Codon positions are derived from the full-length translation-based alignments, and amino acids are those found in the spider monkey or NCBI mouse reference sequence. Sites with a high posterior probability of positive selection in PAML (Bayes Empirical Bayes, P > 0.9) or FUBAR (Empirical Bayes, P > 0.9) analyses are shown in bold.

2 The PAML p-value was calculated using twice the difference in log-likelihood between models M7 and M8 and two degrees of freedom. PAML analysis was carried out using the F61 model of codon frequencies, but similar results were obtained for the F3 x 4 model and various initial omega values.

doi:10.1371/journal.pgen.1004531.t001
species, \( L1TD1 \) shown in black, and those in which \( L1TD1 \) is discarded. Species in which \( L1TD1 \) is active, we are able to obtain part of coding exon 1 of \( L1TD1 \), but this region is intact. Presence or absence of active \( L1s \) is based upon previous data [44] and analysis of the \( M. davisii \) genome assembly (Yang and Wichman, unpublished data). doi:10.1371/journal.pgen.1004531.g005

**Figure 5. Loss of \( L1TD1 \) in megabats appears to follow the loss of \( L1 \) activity.** We obtained \( L1TD1 \) sequences from thirteen bat species (Materials and Methods). We show a species tree partly based upon a published megabat phylogeny [82], inferring placement of species (Materials and Methods). We show a species tree partly based upon a published megabat phylogeny [82], and the loss of \( L1 \) activity in rice rat genomes

six megabat species with \( Dobsonia moluccensis \) (Figure 5), we cannot formally rule out the alternate possibility that the initial loss of \( L1TD1 \) function occurred once, via a non-coding mutation. Under both scenarios, we infer that the extinction of \( L1s \) preceded the extinction of \( L1TD1 \) in the megabats [44].

Thus, it appears that loss of \( L1 \) may have led to loss of \( L1TD1 \) in megabats except in \( Dobsonia moluccensis \) (see Discussion). This evolutionary relationship between \( L1 \) elements and \( L1TD1 \) is consistent with the possibility that \( L1TD1 \) was domesticated by mammalian genomes to antagonize \( L1s \), similar to the domestication of \( Fv1 \) for defense against retroviruses. Under this model, loss of active \( L1s \) would relieve the selective pressure to maintain \( L1TD1 \) in megabat genomes.

**\( L1TD1 \) and the loss of \( L1 \) activity in rice rat genomes**

In addition to the loss of active \( L1s \) in the megabats, a group of rice rats in the subfamily Sigmodontinae are the only other group of mammals in which a loss of active \( L1s \) has been well characterized [45]. To investigate whether these species might exhibit a correlation similar to that of the megabats, in which \( L1 \) loss and \( L1TD1 \) loss co-occur, we sequenced the \( L1TD1 \) gene from ten \( L1 \)-less Sigmodontinae species (5 complete sequences, 4 coding exon 1s, 1 coding exon 2), as well as two Sigmodontinae species that retain active \( L1s \) (\( Oecomys bicolor \) and \( Reithrodontomys fulvescens \)). In all cases, we found an intact coding sequence, suggesting these species have retained a functional copy of the \( L1TD1 \) gene (Dataset S1).

We suggest two hypotheses for the apparent retention of \( L1TD1 \) in these species, despite the absence of active \( L1s \). First, it is possible that not enough time has passed for inactivating mutations to appear in the coding sequence of \( L1TD1 \) in these species; we note that the loss of active \( L1s \) in these species is much more recent than the loss of \( L1s \) in megabats (7.2–12.3 million years ago in the rodents vs 24–58 million years ago in the megabats). While we cannot definitively rule out the possibility that \( L1TD1 \) is neutrally evolving within these species, overall the tree is better fit by a single purifying dN/dS than by a dN/dS fixed at 1 \((p<0.001)\). Further, most of the Sigmodontinae sequences exhibit pairwise dN/dS ratios that suggest purifying selection within the clade (Table S1). We therefore favor the second possibility that retention of a functional \( L1TD1 \) in these species results from selection pressures independent of \( L1s \). For example, \( L1TD1 \) could play an essential role in pluripotency (despite the dispensable role of the \( Mus L1TD1 \)). Intriguingly, despite having an overall tree length similar to our \( Mus \) alignment, we do not find any evidence of positive selection in the \( L1TD1 \) genes from Sigmodontinae using PAML NSites on either full sequences (\( M7 \) vs \( MB \), \( p = 0.33 \)), coding exon 1 alone (\( p = 0.53 \)), or coding exon 2 alone (\( p = 0.84 \)). Alternatively, \( L1TD1 \) in sigmodonts could play a role in genome defense against another element active in these genomes. Indeed, the sigmodontine rodents have a highly active family of ERVs, the mysTR elements [46]. Expansion of this family preceded the death of \( L1s \), but these elements are very active, with 3500 to 7000 species-specific insertions in the L1-extinct species examined [47]. This recent ERV amplification in Sigmodontinae contrasts with the megabats (where \( L1TD1 \) has been lost in many species); there are apparently no highly active DNA or RNA elements in megabats [48]. If \( L1TD1 \) can suppress retroelements other than \( L1s \), this could explain why the gene is retained in sigmodontine rodents but not in megabats.

While these results do not provide additional support for our proposed correlation between \( L1 \) loss and \( L1TD1 \) loss, they do provide another example of the potentially stochastic co-option of \( L1TD1 \) into essential cellular processes that has driven the
variable gain and loss of this gene throughout mammalian evolution.

Discussion

Domesticated genes such as LITD1 provide a window into the opportunistic means by which host genomes can exapt new genetic functions from their resident mobile elements. L1 elements and non-autonomous SINE elements that rely on L1 for mobility comprise a substantial fraction of the human genome (at least 17 and 30 percent, respectively [49]). This abundance, together with a relatively long history of vertical transmission has provided ample opportunities for host genome domestication of L1 elements. Indeed, we see evidence of this exaptation for enhancer functions, in exonization of Alu elements and small L1 fragments, and even in long non-coding RNAs [4,6,21-23,25-28]. However, LITD1 is unique in being the only known host protein-coding gene whose protein coding sequence is derived almost entirely from the coding sequence of an L1. In addition, its exclusive expression in stem cells and the germine and the observation that LITD1 appears to be required for pluripotency maintenance in human ES cells [31], make it a powerful example of retroelement domestication for an essential function. Exaptation for an important function such as pluripotency maintenance draws immediate parallels to other retroelement domestecations such as the retroviral envelope-derived syncytin genes and the sushi-retroelement derived Peg10-related genes, both of which serve essential roles in mammalian placentation [7,11,12]. Indeed, at first glance, our estimate of the evolutionary age of LITD1 (only slightly younger than Peg10 and older than any identified syncytin gene) does suggest long-term preservation for some important function.

Nonetheless, our elucidation of LITD1's evolutionary history challenges these parallels. For instance, we found multiple instances of LITD1 loss or pseudogenization at different stages of mammalian evolution, in contrast to Peg10, which is intact in all examined mammals [11,50,51]. Furthermore, Peg10 is subject to strong purifying selection (unpublished data), presumably due to the constraint imposed by its essential functions in placentation and genomic imprinting. In contrast, we have found evidence of diversifying selection acting on LITD1 in primates and mice, along with genetic innovation in bushbaby via a partial replacement of LITD1 coding sequence with the ORF1 of a younger L1 element. This pattern of genetic innovation and sporadic loss is more reminiscent of another class of domesticated retroelement genes, exemplified by the Fo1 retroviral gag-derived gene in Mus species, which serves as a form of host defense against a broad array of incoming retroviral capsids [15,18].

It is noteworthy that germine and early embryonic tissues represent the primary battleground for the conflict between hosts and L1 elements. In order to propagate new copies to future generations, L1s must successfully retrotranspose in these cells. However, these are the very compartments where a host could incur the greatest fitness cost from L1 replication. The highly specific expression pattern of LITD1 suggests that its target could be genomic retroelements that mobilize during germine development and early embryonic stages.

Three findings lead us to consider the intriguing speculation that LITD1 may target L1 elements themselves (although other functions for LITD1 are quite possible). First, we find that extinction of L1 activity appears to have been followed by LITD1 loss in several megabat lineages. This relationship between L1 loss and LITD1 loss is far from perfect. For example, in the microbat Myotis davidi and other mammalian genomes (Afrotheria and Cetartiodactyla), LITD1 loss is not accompanied by L1 extinction. This is easily explained, because a number of different restriction factors exist to defend against L1 in mammalian genomes [52], and in particular genomes, the previously beneficial function of LITD1 may become redundant and dispensable. Also, if a rapidly evolving L1 were to escape restriction by LITD1, control over this rogue element could be regained through adaptive mutations in restriction factors besides LITD1, thereby relaxing selective pressure to maintain LITD1. Alternatively, L1 retrotransposition rates are known to have varied over evolutionary time: periods where L1 retroposition rates are low could also lead to relaxed selective pressures on LITD1 retention. A similar relaxation of selective pressure has been invoked to explain the idiosyncratic loss of Fo1 in Mus species [15]. Second, we found that bushbaby LITD1 underwent a partial replacement with a more current version of L1 ORF1. Although such ‘replenishment’ could represent a neutral event unique to the bushbaby, the fact that it has occurred suggests a model in which the ‘newer’ version of LITD1 was fixed in the bushbaby genome because it conferred a selective advantage over the pre-existing version, which itself formed from a much more ancient L1 domestication. Third, our observation that LITD1 has experienced positive selection in primates (and mice) is consistent with previous findings that L1 ORF1 has evolved under positive selection during primate evolution [38,40]. Like Fo1 action against distantly homologous retroviral capsids, the observed positive selection suggests the possibility that LITD1 may be in direct conflict with L1 ORF1p from active L1 elements. Since L1 elements also provide the machinery for non-autonomous elements like SINEs [53] and likely HAL1s [54], it is also possible that these elements represent the true targets of LITD1, rather than L1 itself. LITD1 could function as a L1 restriction factor through a number of possible mechanisms. First, analogously to Fo1, LITD1 might interfere with the homotrimerization of L1-ORF1p that is necessary for L1 retrotransposition. This trimerization normally happens through ORF1p's coiled-coil domain [55]. We note that LITD1 coding exon 2 has preserved its coiled-coil domain, which may be necessary for this interaction. Under this model, LITD1 could act as a dominant negative version of ORF1p. Its binding to ORF1p could lead to L1 restriction, which could then drive episodes of diversifying selection in L1 ORF1p to escape LITD1 restriction [40]. In response to these escape mutations in L1, we might expect LITD1 to rapidly evolve to restrict the variant L1 – a situation consistent with the positive selection we observe in LITD1. Another possibility is that LITD1 could outcompete L1 ORF1p for binding to the ORF2p protein, the likely rate-limiting component for retrotransposition. Alternatively, LITD1’s RNA-binding activity could be the source of its restrictive abilities. In this model, LITD1 could act as a competitive inhibitor of L1 ORF1p, binding the RNA of transcribed retroelements and blocking the downstream pathway that would normally create new copies. This could function through simple titration or through a more complex mechanism whereby LITD1 localizes L1 RNA to P-bodies [32] for sequestration or decapping and subsequent degradation. The previously reported association of LITD1 with the TRIM28 (KAP1) complex [56] suggests yet another possible mechanism of restriction. TRIM28 is known to function in the silencing of exogenous and endogenous retroviral integrations by recruitment of heterochromatin proteins to some retroviral elements [57]. Under this epigenetic silencing model, LITD1 may function in an analogous role to the zinc finger adaptor proteins that help target TRIM28 and consequently the silencing complex to particular genomic sequences. Thus, constant adaptation to bind either DNA or RNA at L1 integration sites could
drive diversifying selection of LITD1. It is intriguing to imagine a model wherein the silencing machinery could target any genomic parasite through a modular adapter protein that recognizes the newly integrated sequence (although it must be noted that such activities have not been ascribed to L1 ORF1p from which LITD1 derived). Further characterization will shed light into whether LITD1 encodes any or all of these biochemical activities.

When we examined the genome sequences of bats, we found that many species that have lost active L1s have also lost LITD1. However, Dobsonia moluccensis still maintains LITD1 as an intact protein-coding gene, despite the fact that this species lost active L1s many million years previously. Similarly, we found LITD1 intact in the subset of the Sigmodontinae rodents previously shown to have lost active L1s [45]. This implies that the proposed selection for restricting active L1 elements cannot be the only constraint that dictates retention of LITD1. We hypothesize that idiosyncratic, highly divergent L1 insertion patterns in different mammalian genomes could provide an explanation of both the retention of LITD1 in mammals without an active L1 and the essential role in pluripotency maintenance that LITD1 plays in humans. Under this model, LITD1’s role as either important or dispensable for pluripotency maintenance depends on the pattern of where L1s have inserted in the genome of each species. In various contexts, transposable elements have been shown to affect the expression of genes in their vicinity [58,59], and some elements are included in the untranslated regions of host mRNAs [60,61]. If LITD1 transcriptionally or post-transcriptionally silences L1 elements, this silencing could alter the expression of host genes near L1s, or with L1s in their UTRs. Under this model, in some mammalian genomes, L1 insertions would not be in proximity to genes that impact the pluripotency program (Figure S5). In these species, extinction of L1 would result in the loss of constraint on LITD1 (e.g., megalabat P. vampyrus), and experimental knockdown of LITD1 would have no effect on pluripotency maintenance (e.g., M. musculus [33]). In contrast, in other mammalian genomes where L1 insertions occurred near genes whose repression is important for pluripotency maintenance, loss of LITD1 would lead to loss of pluripotency (e.g., H. sapiens [31]).

In fact, repression could have been incorporated into any number of functionally important pathways in the cells where LITD1 is expressed, simply depending on which genes L1s landed next to during evolution. We hypothesize that in such species, extinction of L1 activity would not relax the selective pressures to maintain LITD1 (e.g., megalabat D. moluccensis, Sigmodontine rodent livers), which would now be required to recognize ‘dead’ L1 copies in order to maintain these repressive programs.

Although we present the hypothesis that LITD1 was originally domesticated as an anti-L1 restriction factor and was subsequently recruited for pluripotency regulation in humans by virtue of direct repression of L1 elements near functionally relevant genes, this idea is at present highly speculative. LITD1 has experienced a long evolutionary history functioning in the context of other factors expressed in pluripotent cells, and it is possible that LITD1 could have been coopted by these factors. For instance, LITD1 could be involved in regulating RNA involved in pluripotency via its ancestral chaperone function, or in regulating important protein complexes via protein-protein interactions. We look forward to experimental investigation of all of these hypotheses; none of them can completely explain all of our evolutionary observations (LITD1 loss and retention in various lineages, positive selection and remodeling in bushbabies), perhaps suggesting that LITD1’s function has changed multiple times over the course of mammalian evolution.

In summary, we posit that LITD1’s original function could have been in genome defense (similar to Fei1), and that it still has defense functions in many mammalian species where it is retained (including humans). We note that our evidence for LITD1 diversifying selection (and inferred genome defense) comes in part from the primate lineage, which is also where its role in pluripotency maintenance is best established. Later in evolution, by virtue of either its transcriptional or post-transcriptional silencing of L1, LITD1 may have become intricately enmeshed into the transcriptional program of L1-proximal genes in some species, meaning that LITD1 also acquired an absolutely essential function (similar to Peg10 and syncytin). An interesting feature of this model is that LITD1’s essential function (or lack thereof) is thus a consequence of stochastic L1 insertion patterns in different mammalian genomes. This stochasticity could provide an explanation for why LITD1 has been lost in some species, and why it is responsible for pluripotency maintenance in only some mammalian genomes.

Materials and Methods

Ethics statement

The animals used in this study are wild-derived laboratory animals of the species Mus musculus, Mus spretus, Mus spicilegus, Mus mattheyi and Apodemus uralensis. None of these species are protected. Permits for catching the founding members of each line were not required at the time they were caught. Some specimens were caught on the properties of private landowners, with their oral permission to enter the property and catch mice. All animal work was carried out by experienced personnel at the Max Planck Institute for Evolutionary Biology, following the legal requirements in accordance with German animal welfare law (Tierschutzgesetz) and FELASA guidelines. Permits for keeping mice were obtained from the local veterinary office “Veterinäramt Kreis Plön” (permit number: 1401-144/PLÖ-004697). The dissection of animals and organ extractions were performed according to the German Animal Welfare Act § 8a Abs. 1 Nr. 3b TierSchG; ‘Organ-/Gewebsentnahme zu wissenschaftlichen/diagnostischen Zwecken TierSchG’ V 312-72241.123-34.

Biological materials

Sigmodontine rodent liver tissue was obtained on loan from the Museum at Texas Tech University (Lubbock, TX).

For Mus samples, mice of different ages were sacrificed by CO2 asphyxiation followed by cervical dislocation. Mice were then dissected and tissues were snap-frozen within 5 minutes post-mortem. Liver (front view: front left lobe), both testes, and whole brain including brain stem were collected. For the outbred populations, Iran (AH), France (M), and Germany (CB) for Mus musculus domesticus, and Austria (WI) and Kazakhstan (KH) for Mus musculus musculus, eight individuals each were sampled. For inbred groups, Mus musculus castaneus (TA), Mus spretus (SP), Mus spicilegus (SG), Mus mattheyi (MA) and Apodemus uralensis (AP), four individuals each were sampled. All mice were obtained from the mouse collection at the Max Planck Institute for Evolutionary Biology.

Mus transcriptome sequencing, processing, and mapping/assembly

The sampled tissues of each Mus group were used for RNA extraction with the RNeasy Kit (Qiagen) and pooled at equimolar concentrations. Quality of the RNA was measured with BioAnalyzer chips (Agilent), for the individual samples and pools, and samples with RIN values below 7.5 were discarded. The pools were subsequently submitted to the Cologne Center for Genomics (CCG) for further processing and sequencing. The
sequencing of the samples was performed using a polyA tail purification step, followed by cDNA synthesis, Illumina library preparation, and sequencing with an Illumina HiSeq 2000 sequencer. Each transcriptome sample was sequenced in approximately one third of a HiSeq2000 flow-cell lane (one flow-cell lane per taxon).

All raw data files were trimmed for adaptors and quality using Trimmomatic [62]. The quality trimming was performed base-wise, removing bases below quality score of 20 (Q20), and keeping reads whose average quality was of at least Q30. Reads whose trimmed length was shorter than 40 bases were excluded from further analyses, and pairs missing one member because of poor quality were also removed from any further analyses.

Quality-filtered transcriptome reads were aligned against the mm10 version of the mouse reference genome from UCSC [63] using NextGenMap [64]. Reads which were ambiguously or poorly mapped (MAPQ < 20) were removed from the analyses. Quality-filtered transcriptome reads for each taxon were merged into a single input file, discarding tissue information, and assembled de novo with the Trinity platform [65] using default parameters.

Sequence collection

\textit{L1TD1} sequences were obtained from publicly available primate genome databases using PSI-BLAST [66] against the NR database or TBLASTN [66] against the HTGS database, with human \textit{L1TD1} as a search seed (Table S2). For sequencing of other primate, Sigmodontinae, and \textit{Mus} species (Table S2), exon 1 and exon 2 were PCR amplified individually from genomic DNA using oligonucleotides designed against intronic regions. PCR primers were designed against intronic regions neighboring each exon based upon an alignment of mouse, rat, and Chinese hamster for rodents or human, rhesus, and squirrel monkey for primates (Primates Exon1 Sense: CAGAATCCAGTCTTGAACACATA-TCC; Primates Exon1 Antisense: CAGGAGAACCTTTAGAC-CTGGG; Primates Exon2 Sense: GTCAAGATGGAAAGCCA-TATTTAAAAATGTG; Primates Exon2 Antisense: GCTATTA-GCCTGCCCATTCTTCCTG; Rodent Exon1 Sense: GYAAG-WAMAYTTTCATTGTTAKTCT; Rodent Exon1 Antisense: CCGATCTTCTCCTGGACCTCCYCCCRTCA; Rodent Exon2 Sense: GGMAAGYATACAAATTYGAGGCTRA-AATAG; Rodent Exon2 Antisense: AASTCAAAACACMYK-CAGKARGT). PCR products were sequenced using standard Sanger sequencing.

For \textit{Mus} sequences, informative reads were obtained with samtools [67] from alignments overlapping with the \textit{L1TD1} annotation. Known sequences from the coding sequences of \textit{L1TD1} were used to identify the most similar assembled contigs in each taxon using nucleotide-nucleotide blast [68] (e-value < 1e-10).

\textit{L1TD1} sequences from primate, Sigmodontinae, and \textit{Mus} species have been deposited in Genbank under accession numbers KJ994281-KJ994329.

Alignments and positive selection analysis

Primate and \textit{Mus} nucleotide sequences were aligned using the ClustalW translation align’ function in Geneious Pro (Biomatters Ltd.). Alignments were refined manually, including truncation of the poorly aligned glutamic-acid rich region, and this alignment and an established primate phylogeny [69] or \textit{Mus L1TD1} gene tree were input into the CODEML sites model of PAML [70] to detect positive selection at individual sites. Positively selected sites were classified as those sites with a M8 Bayes empirical Bayes posterior probability > 90%. FUBAR was performed using the web-based implementation of HyPhy (www.datamonkey.org) [71,72]. To test for signatures of positive selection along individual branches of the primate phylogeny, we used Branch-site REL in HyPhy [71] or the branch model of PAML. For PAML, the statistical significance of any branch that showed dN/dS > 1 in the free ratio model (model = 2) was tested using a two-ratio tree (model = 2) by specifying the branch of interest as a foreground branch and all other branches as background branches. We then compared the likelihood of a model where the foreground branch had a freely estimated dN/dS with the likelihood of a model where that branch had dN/dS fixed at the neutral value of 1 [39].

For the analysis shown in Figure S2, we obtained dN/dS values from a genome-wide analysis of trios of human-chimpanzee-macaque orthology [42], made available via Adam Siepel’s website (http://compgen.bscb.cornell.edu/orthology). To be conservative, we filtered out dN/dS values that might be artificially high for technical reasons: we eliminated values calculated from alignments of fewer than 100 codons, eliminated alignments that contained fewer than 5 evolutionary changes of any type, and included only RefSeq genes and not other genes derived from less confident annotation sets. In this dataset, alignments with no synonymous changes had apparent dN/dS values of 999; again, to be conservative, we replaced dN/dS values for these genes with a more conservative estimate of dN/dS = 2. We converted sequence identifiers to gene symbols using Bioconductor [73], allowing us to cross-reference dN/dS values with genes identified in a screen for ES cell determinants [43]. For the plot shown in Figure S2, we selected a list of 127 ES cell determinants validated by secondary screening using three markers of stem cell identity [43]; other overlapping lists of validated hits from this screen show similar evolutionary patterns.

Mammalian sequence collection and alignment

To identify previously unannotated copies of \textit{L1TD1} in sequenced genomes, the genomic sequence between flanking genes \textit{INADL} and \textit{KANK4} was extracted and compared with the corresponding human genomic region (\textit{INADL-L1TD1-KANK4}) using Dotter [74]. Visual inspection of Dotter output enabled extraction of \textit{L1TD1} protein-coding sequences. Multispecies \textit{L1TD1} alignments were generated using CLUSTALW [75] with manual adjustment. For all mammals mentioned in the text, genome sequences were searched for non-syntenic and/or pseudogene copies of \textit{L1TD1} using TBLASTN with a RepeatMasked version [48] of human and mouse \textit{L1TD1} sequences, and in some cases \textit{L1TD1} sequence from a more closely related species.

To identify representative \textit{L1} ORF1p sequences (Dataset S2) used to build the tree shown in Figure 2, we first selected a small number of \textit{L1} ORF1p consensus sequences from RepBase [76] and obtained the sequence of an active human L1 from Genbank (LRE2, accession AA860344.1). We then used TBLASTN [66] against whole genome assemblies (Table S2) to identify a single intact \textit{L1}-ORF1p sequence from the genome assemblies of each species shown in Figure 2, selecting a copy arbitrarily from among the blast hits that did not show inactivating mutations in ORF1p.

Phylogenetic inference

To construct the tree shown in Figure 1B, we used our alignment of \textit{L1TD1} nucleotide sequences and the jModelTest 2 program [77] to determine that the best-fitting evolutionary model for this alignment is the GTR+I+G model with invariant sites and gamma distributed rates (GTR+I+G). We then generated a maximum-likelihood phylogeny using PhyML [78] and the GTR+I+G evolutionary model with four site categories for the
gamma distribution. 1000 replicate trees were constructed, and the tree with the highest log-likelihood was chosen and displayed using MEGAS [79].

To construct the tree shown in Figure 2, predicted protein sequences were aligned by hand (Dataset S2). We used ProtTest 3 [80] to determine that the best-fitting evolutionary model for this alignment is the JTT model with gamma-distributed rates (JTT+G). We then used PHYML to generate maximum-likelihood trees with the JTT+G model (four site categories for the gamma distribution). 1000 replicate trees were constructed, and the tree with the highest log-likelihood was chosen and displayed using MEGAS [79]. Bootstrap values represent the percentage of trees in which the descendent taxa cluster together.

Dates of divergence mentioned in the text were generated using the TimeTree web service [81].

Bat sequencing and database sequence collection

We obtained L1TD1 sequences for Pteropus alecto, Pteropus vampyrus, Rhinolophus ferrumequinum, Myotis lucifugus, Myotis davidi, Myotis brandti and Eppticus fuscus from publicly available genome sequences. For other bat species, genomic DNA samples were obtained from tissues from The Museum, Texas Tech University. Sample accession numbers were previously published [44]. Degenerate PCR primers were designed against intronic regions neighboring each exon based upon an alignment of P. vampyrus and M. lucifugus sequences (Exon1 Sense: TTTCAAGATGATTITKAAWGAAGATTTGGG; Exon1 Antisense: YYTMYWYWATWAMASTGTTAAS-YYWTTSTTCC; Exon2 Sense: TGGGGWTCCMAGCCTTYYAGAAMAAATC; Exon2 Antisense: CATCMCCAGATATA-CGTGTTACGTTCCTCCATC). For Ronsettus amplexicaudatus and Rhinolophus elorus we designed a second set of primers based upon the new sequences we generated to amplify Exon1 (Sense: AAATATCACCCACATGGAAAGAATTAG; Antisense: TTTCAGATTITKAAWGAAGATTTGGG; Exon1 Antisense: YYTMYWYWATWAMASTGTTAAS-YYWTTSTTCC; Exon2 Sense: TGGGGWTCCMAGCCTTYYAGAAMAAATC; Exon2 Antisense: CATCMCCAGATATA-CGTGTTACGTTCCTCCATC). PCR products were sequenced by Sanger sequencing. For Nycitine albonodon, Megaderos niphanae and Cynopterus sphinx, we were unable to amplify or sequence any product from either exon with multiple PCR optimizations of annealing temperature, amount of genomic DNA template and concentration of magnesium ion. For two bat species, it appeared that more than one closely related L1TD1 sequence was present. Upon cloning and sequencing multiple PCR products, we found that M. woernanni had 4 sequence variants and R. amplexicaudatus had 2 variants. For each species, at least one inactivating mutation was shared by all variants.

Supporting Information

Dataset S1 In-frame nucleotide alignment of L1TD1 sequences. The fasta-formatted alignment includes all L1TD1 sequences described in the manuscript. The inactivating mutations in L1TD1 that are shared by subsets of megabat species (Figure 5) are at the following nucleotide positions of the alignment: P. hypomelanus, P. alecto and P. vampyrus share three mutations: a stop codon at bp 34–36; a stop codon at bp 106–108; a 1 bp insertion at bp ~774. H. monstrosus and M. woernanni share a different set of three mutations: a stop codon at bp 385–387; a 1 bp deletion at bp 720; a 1 bp deletion at bp 939. R. amplexicaudatus does not share any of the above mutations, but has numerous lineage-specific inactivating mutations.

Dataset S2 Amino acid alignment of L1-ORF1p and the two ORF1p-homologous regions of L1TD1. The fasta-formatted alignment was used to generate the phylogenetic tree shown in Figure 2.

(PDF)

Figure S1 Positive selection in Mus L1TD1 is robust to alternative phylogenies from potential recombination breakpoints. We used HyPhy’s GARD program to test for recombination breakpoints in the Mus L1TD1 alignment that could give rise to false signatures of positive selection. We found three potential breakpoints, though none was statistically significant (KH test, p> 0.1). To ensure that our detection of positive selection was robust to the use of these alternative phylogenies, we performed PAML NSsites on slices of the alignment corresponding to each of the GARD trees, and we used the built-in functionality of DataMonkey to use these GARD-generated trees to identify positively-selected positions using FUBAR. Top, the whole gene tree is shown in black. The three breakpoints are shown on the schematic of the L1TD1 gene with vertical lines, and the slices they delimit are shown in different colors. Below each gene segment is shown the GARD-generated tree that best describes that region. Both the PAML NSsites signature of selection and FUBAR-identified selected sites are robust to the use of these alternative phylogenies. (PDF)

Figure S2 Determinants of embryonic stem cell identity evolve more slowly than control genes. The boxplots show the distribution of dN/dS values for genes identified in a screen for determinants of human ES cell identity [43] and for all other genes in a genome-wide dataset of dN/dS values calculated from trios of orthologs from human, chimpanzee and macaque genomes [42]. Outliers with dN/dS>1 are omitted; gray dots represent other outliers - the large number of data points for "other genes" precludes visualization of individual data points. The stem cell determinants are evolving more slowly than other genes (Wilcoxon p = 0.008). L1TD1’s dN/dS value in this genome-wide dataset is shown using a red horizontal line; it is evolving faster than most other pluripotency genes. Although L1TD1 did not meet the arbitrary threshold (Fay score <−2) used to identify stem cell determinants in the published RNAi screen, its score in the screen (Fay = −0.90) is well below the genome-wide average, consistent with previous results that human L1TD1 is a pluripotency factor [31,32].

(PDF)

Figure S3 Intact L1TD1 from bushbaby Otolemur garnetti has evolved under purifying selection. We used codeml’s free-ratio model to estimate selective pressures on L1TD1 on each branch of the two trees shown. Above each branch we show estimated dN/dS ratios, and in parentheses below each branch we show the estimated number of non-synonymous and synonymous changes, respectively. L1TD1 pseudogenes are shown with their labels in gray. For three selected branches we performed likelihood tests of whether the estimated dN/dS ratio is significantly better fit to the data than dN/dS = 1 for that branch, by assuming only two dN/dS ratios for the entire tree (one ratio for the branch in question, and one for all other branches), and comparing a model where dN/dS for the branch in question was fixed at 1 with a model where dN/dS was freely estimated. Results of these tests are shown with red superscripts. A. We examined evolution of the full-length ORF of the novel bushbaby L1TD1 structure, comparing it to two processed pseudogenes in the bushbaby genome that arose after this novel L1TD1 structure formed. The intact bushbaby L1TD1 is more likely evolving under purifying than neutral selection (p = 0.0003; **). B. We examined the portion of bushbaby L1TD1 that aligns to the ancestral L1TD1 gene, including human and aye-aye L1TD1 genes as outgroups. The
intact bushbaby *LITD1* is more likely evolving under purifying than neutral selection (*p* < 10^-5; **); there is also weaker support (*p* = 0.07; *) for purifying selection on the shared ancestor of the intact bushbaby *LITD1* and bushbaby processed pseudogene B. (PDF)

**Figure S4** *LITD1* from megabat *Dobsonia mulocensis* has evolved under purifying selection. We used codeml’s free-rate model to estimate selective pressures on *LITD1* on each branch of the species tree shown. Above each branch we show estimated dN/dS ratios, and in parentheses below each branch we show the estimated number of non-synonymous and synonymous changes, respectively. Species in which *LITD1* is a pseudogene are shown with their labels in gray. For three branches, we performed likelihood tests of whether the estimated dN/dS ratio is a significantly better fit to the data than dN/dS = 1 for that branch, by assuming only two dN/dS ratios for the entire tree (one ratio for the branch in question, and one for all other branches), and comparing a model where dN/dS for the branch in question was fixed at 1 with a model where dN/dS was freely estimated. Results of those tests are shown with red superscripts: the *Dobsonia mulocensis* branch is more likely evolving under purifying than neutral selection (*p* = 0.003), as is the branch ancestral to *D. moluccensis* and the three *Pteropus* species (*p* = 0.002). In contrast, for the *Pteropus alecto* branch, a neutral model is as good a fit to the data as a model invoking purifying selection (denoted by n.s., for non-significant). (PDF)

**Figure S5** *LITD1* could become essential based upon the pattern of L1 insertions in a specific genome. We present a scheme under which *LITD1* could idiosyncratically adopt a pluripotency role in a species-specific fashion. A. We schematize the genomes of three exemplar species with an identical stretch of seven genes (rounded boxes), one of which must be silenced for pluripotency maintenance (orange boxes). B. L1s (black rounded boxes) insert randomly into each genome. One insertion (species 3) happens to be near the silenced gene. C. According to our hypothesis, *LITD1* silences the newly inserted L1s (grayed areas), as well as genes near *L1TD1* insertions. In the case of species 3, the gene that must be silenced is now silenced redundantly by the ancestral mechanism and *LITD1*. Because of this redundancy, either *LITD1* or the ancestral silencing could be lost. D. Loss of the ancestral silencing mechanism would render *LITD1* essential. A loss of *LITD1* in this case (species 3) would result in the expression of the orange gene, which must remain silenced for survival. In this way, *LITD1* could be co-opted as an essential regulator of pluripotency. If *LITD1* instead targets L1s near arbitrary genes with no influence on the pluripotent state of a cell (species 1 and 2), loss of *LITD1* would not affect the maintenance of pluripotency; in this case, *LITD1* would be retained only if its restriction or other functions were beneficial. (PDF)

**Table S1** Maximum likelihood estimates of pairwise dN/dS for *LITD1* from Sigmodontinae rodents. (XLSX)

**Table S2** *LITD1* Genbank accessions or coordinates in mammalian genome assemblies searched. (XLSX)

**Acknowledgments**
We thank Michael Emerman, Lucie Etienne, Patrick Mitchell, Jacob Gable, Maulik Patel, and members of the Malik lab for critical comments on the manuscript. We thank the Museum of Texas Tech University for generously providing rodent tissue samples. We thank Christine Pfeifle and the mouse team at the Max Planck Institute for Evolutionary Biology for their help with mouse samples, and Nicole Thomsen for her assistance in sample preparation. We also thank colleagues and genome sequencing centers for public release of genome sequences.

**Author Contributions**
Conceived and designed the experiments: RNM JMY HAW HSM. Performed the experiments: RNM JMY LY RN HSM. Analyzed the data: RNM JMY L1TD1 LY RN HSM. Contributed reagents/materials/analysis tools: LY RN HAW. Wrote the paper: RNM JMY HAW HSM.

**References**